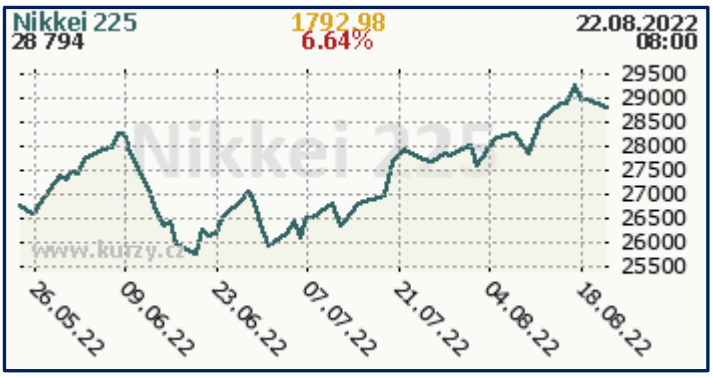
IT a anatomie firmy

(Podniková analytika)

(pracovní dokument)

MBI tým

VŠE Praha, 2024



[B)] Obsah a principy podnikové analytiky ***(úlohy, procesy, metriky, dimenze, datové zdroje, role, faktory, metody)***	
[C)] Nástroje a řešení pro deskriptivní analytiku ***(business intelligence, self service business intelligence, competitive intelligece, mobil BI)***	
[D)] Komponenty podnikové analytiky ***(datové báze analytiky, transformace dat, analytika, real time DWH)***	**[E)] Reporting a vizualizace dat** ***(reporting, dashboardy, vizualizace)***
[F)] Pokročilá podniková analytika – nástroje, řešení ***(Data Science, data mining, text mining, machine learning, prediktivní analytika)***	
[G)] Data pro podnikovou analytiku ***(správa dat, master data management, data governance, zajištění datové kvality)***	**[H] Podniková analytika na velkých datech** ***(big data, big data analytics, data driven company)***
[I)] Analytika a cloud computing ***(BI cloud computing, využití cloud BI, možnosti cloud BI)***	**[J)] Řízení podnikové výkonnosti** ***(CPM, řízení výkonnosti IT, řízení výkonnosti eGovernmentu, marketingu, prodeje)***

Text tohoto dokumentu by měl obsahovat následující oblasti informací:

- analýza ***obsahové náplně řízení*** firmy, který z pohledu analytiky tvoří především tyto komponenty:
  - ***úlohy***, resp. procesy podnikové analytiky, tedy co a jak se řeší (***kapitola 1.2***),

  - ***systém podnikových metrik***, resp. ukazatelů a jejich dimenzí, tedy, co je předmětem řešení (***kapitola 1.3***),
  - ***data a datové zdroje***, které jsou vstupem do řešení a které jsou i nositeli již zmíněných metrik,
- na řešení úloh podnikové analytiky se podílejí ***analytici, IT specialisté i uživatelé*** se specifickou funkční náplní a specifickými kvalifikačními nároky (***kapitola 2***),
- obdobně jako je tomu i u jiných projektů, jejich charakter a přístupy k řešení jsou ovlivněny celou ***škálou faktorů***, které je nezbytné pochopit a brát v úvahu (***kapitola 3***),
- pro řešení podnikové analytiky je k dispozici řada ***metod a metodik***, v tomto textu jsme se zaměřili výlučně na dimenzionální a datové modelování (***kapitola 4***),
- v rámci podnikové analytiky se využívá velmi široká ***škála nástrojů, konceptů a technologií***.

Z pohledu řešení a odpovídajících nástrojů a přístupů je možné podnikovou analytiku ***pracovně členit*** do následujících ***oblastí***:

- ***základní*** podniková analytika zahrnující business intelligence self service business intelligence, mobile business intelligence a competitive intelligence (***oddíl C)***,
- jednotlivé ***dílčí komponenty*** podnikové analytiky, které tvoří dílčí nebo komplexní systémy, a to i na základě různých architektur (***oddíl D),***
- ***firemní reporting***, tj. obsahové, formální i organizační přístupy při definování, vytváření a poskytování reportů manažerům a specialistům firmy pro jejich řídící aktivity ***(oddíl E),***
- ***pokročilá*** podniková analytika zahrnující celou škálu nástrojů a technologií, jako např. data science, data mining, text mining, predictive analytics a další (***oddíl D)***,
- ***data, datové báze*** a principy jejich řízení v rámci podnikové analytiky (***oddíl G),***
- v souvislosti s uplatňováním nových technologií (IoT a dalších) nabývá na významu a rozšíření ***řešení podnikové analytiky na velkých objemech dat*** – „*Big Data*“ (***oddíl H),***
- velmi významným trendem je využívání technologií a služeb ***cloud computingu v podnikové analytice*** (***oddíl D),***
- s podnikovou analytikou úzce souvisí koncepty a aplikace ***řízení podnikové výkonnosti*** a její různé formy (***oddíl J).***

Následující části obsahují doporučené postupy řešení vybraných problémů s využitím relevantních dokumentů MBI-AF, a to:

- řízení a řešení podnikové analytiky, celkový pohled,
- řešení analytické úlohy

Řízení a řešení podnikové analytiky s využitím dokumentů MBI-AF

Účel:

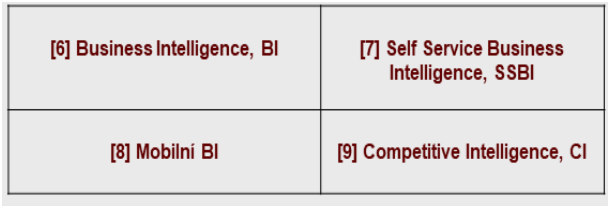
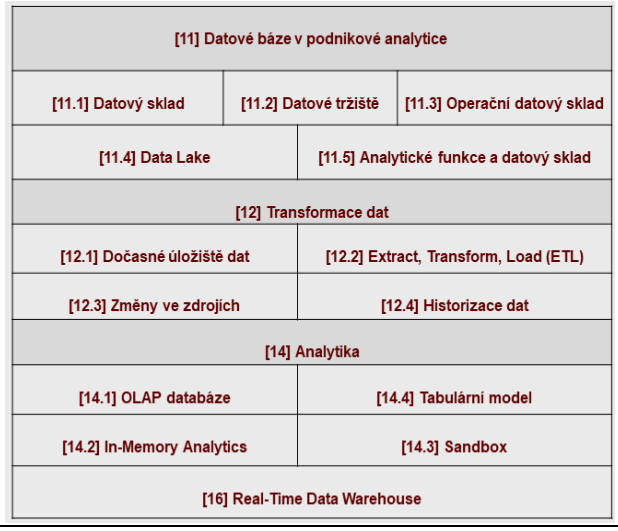
Podniková analytika představuje přípravu a realizaci ***podnikového reportingu, analytických úloh, plánovacích úloh a úloh pokročilé analytiky,*** a to s adekvátním zajištěním příslušných komponent a datových zdrojů a jejich kvality. ***Účelem*** v tomto případě je:

- připravit ***pro řešení projektu*** podnikové analytiky ***potřebné podklady*** s možným využitím dokumentů MBI_AF, případně z portálu MBI,
- je nutné ***vybrat všechny ty oblasti, komponenty a další aspekty*** řízení, které jsou pro další řešení projektu relevantní,
- vyhodnotit (podle potřeby) ***podstatné faktory firemního a byznys prostředí***, které další řešení podnikové analytiky ve firmě budou výrazně ovlivňovat (např. odvětvová působnost, vlastnictví firmy, byznys model apod.),
- pro vlastní průběh projektu připravit a ***formulovat specifické analytické otázky*** ve vazbě na charakter analytických úloh a zvažované metodiky, metody a faktory, a to ***jako základnu pro jednání se zákazníkem***, zaměřené na obsah analytických úloh, jejich funkcionalitu a realizační možnosti,

- definovat analytické úlohy v jednotlivých oblastech řízení (např. řízení prodeje) s respektováním ***kontextu podnikové analytiky celé firmy***, tedy určit podstatné vazby na analytiku v ostatních oblastech řízení,
- vymezení ***základních schémat analytických úloh***, zejména jejich funkcionalitu, a to přes všechny vybrané oblasti řízení,
- formulace navrhovaného ***zajištění oblasti řízení IT aplikacemi*** (business intelligence, self service business intelligence, případně další) s příslušnou argumentací,
- ***zapracování příslušné dokumentace*** a specifikace jejího ***využití v řízení a řešení*** projektů.

Doporučený postup:

	Krok	Podklad v dokumentech MBI-AF
1	***Posouzení všech relevantních faktorů ovlivňujících uplatnění podnikové analytiky:*** ▪ dokument: AF II.05: Podniková analytika, kapitola [3], viz schéma, ▪ rozhodnutí, zda podnikovou analytiku a uplatnění příslušných aplikací řešit, či odmítnout ▪ ***v případě přijetí – viz další kroky***	[3.1] Byznys prostředí; [3.2] Řízení a organizace firmy; [3.3] Řízení podnikové analytiky; [3.4] Metody a metodiky řešení podnikové analytiky; [3.5] Kvalita informačního systému
2	***Nastavení rolí pro podnikovou analytiku:*** ▪ specifické role pro PA: dokument: AF II.05: Podniková analytika, kapitola [2], viz schéma ▪ kompletní přehled rolí: dokument: *„AF II.02: Komponenty“*, kapitola 5,	[2.1] Sponzor projektu; [2.2] Manažer projektu; [2.3] Byznys analytik; [2.4] Datový analytik; [2.5] Architekt; [2.6] Vývojář; [2.7] Uživatel podnikové analytiky; [2.8] Klíčový uživatel
3	***Výběr relevantních metod a metodik pro podnikovou analytiku:*** ▪ specifické metody: dokument: AF II.05: Podniková analytika, kapitola [4], viz schéma ▪ ostatní využitelné metody: dokument: *„AF II.02: Komponenty“*, kapitoly 8, 9, 10	[10.1] PMBOK; [10.2] PRINCE2; [10.3] SCRUM; [10.4] Azure DevOps; [10.5] RUP; [10.6] UML; [10.7] CRISP-DM; [10.8] Datové modelování; [10.9] Procesní modelování; [10.10] Dimenzionální modelování
4	***Konkretizace postupů pro řešení úloh BI a SSBI:***	
5	***Určení, v kterých oblastech řízení budou úlohy podnikové analytiky řešeny,*** jaké typy úloh (reporting, analytické, plánovací, pokročilá analytika), jejich funkcionaloita a s jakými prioritami: ▪ dokument: *„AF II.01: Oblasti řízení“*, viz schéma ▪ dokument: *„AF II.05: Podniková analytika“*, kapitola [1.2]	[1] Strategické řízení firmy; [2] Finanční řízení; [3] Závazky; [4] Pohledávky; [5] PAM; [6] Controlling; [7] Prodej; [8] Nákup; [9] Sklady; [10] Personál; [11] Majetek; [12] Marketing; [13] Doprava; [14] Energie; [15] Řízení IT

	Krok	Podklad v dokumentech MBI-AF
6	***Určení obsahu řešení podnikové analytiky (metriky, dimenze)***: ▪ dokument *„AF II.05 Podnikové analytika“,* kapitola [1.3] ▪ detailní vymezení: *„AF II.02 Komponenty“:* ○ kapitoly 2 (metriky), viz první schéma ○ kapitola 3 (dimenze), viz druhé schéma	
7	***Určení typů aplikací pro základní podnikovou analytiku***: ▪ určení typů aplikací: dokument *„AF II.05 Podnikové analytika“,* oddíl [C)], viz schéma	
8	***Specifikace zaměření na pokročilou podnikovou analytiku:*** ▪ určení typů aplikací: dokument *„AF II.05 Podnikové analytika“,* oddíl [C)], viz schéma	
9	***Určení typů odpovídajících technologií a komponent pro podnikovou analytiku:*** ▪ určení jednotlivých komponent, jejich charakteristik, efektů a omezení: dokument *„AF II.05 Podnikové analytika“,* oddíl [D)], viz schéma	

	Krok	Podklad v dokumentech MBI-AF
10	***Návrh pravidel pro řešení obsahu reportingu a principů vizualizace dat***: ▪ dokument *„AF II.05 Podnikové analytika“,* oddíl [E)], viz schéma	[18] Reporting [19] Dashboardy
11	***Specifikace a řízení datových zdrojů pro podnikovou analytiku:*** ▪ specifikace obsahu datových zdrojů: dokument *„AF II.02 Komponenty“, kapitola 4, viz první schéma,* ▪ řízení datových zdrojů pro PA: dokument *„AF II.05 Podnikové analytika“,* oddíl [G)], *viz druhé schéma,*	[4.1] Data a dokumenty strategického řízení [4.2] Finanční řízení | [4.3] Řízení závazků | [4.4] Řízení pohledávek | [4.5] Práce a mzdy [4.6] Controlling | [4.7] Řízení prodeje | [4.8] Řízení nákupu | [4.9] Řízení skladů [4.10] Personální řízení | [4.11] Řízení majetku | [4.12] Marketing [4.13] Řízení dopravy | [4.14] Řízení energií | [4.15] Řízení IT [25] Správa dat [25.1] Metadata | [25.2] Master data | [25.3] Data profiling [25.4] Data Governance | [26] Zajištění kvality a čištění dat
12	***Specifikace a rozhodnutí o zaměření na Big Data***: ▪ návrh a řízení uplatnění konceptu Big Data: dokument *„AF II.05 Podnikové analytika“,* oddíl [H)], *viz schéma*,	[27] Big Data – principy [28] Big Data Analytics | [29] Data Driven Company
13	***Specifikace a rozhodnutí o zaměření na využití služeb cloud computingu pro PA***: ▪ posouzení a řízení uplatnění služeb cloud computingu v PA: dokument *„AF II.05 Podnikové analytika“,* oddíl [I)], *viz schéma*,	[31] Cloud computing pro BI [32] Využití Cloud řešení a Cloud BI [33] Možnosti Cloud BI současnosti
14	***Posouzení a rozhodnutí o uplatnění konceptu řízení podnikové výkonnosti (CPM) ve vazbě na podnikovou analytiku***: ▪ řízení uplatnění CPM v souvislisti s PA: dokument *„AF II.05 Podnikové analytika“,* oddíl [J)], *viz schéma*,	[35] Řízení podnikové výkonnosti [36] Řízení výkonnosti podle uplatnění
15	***Kompletace dokumentu*** pro řízení rozvoje podnikové analytiky.	

Využití:

Výsledná dokumentace, nebo její části, může být zejména ***podkladem:***

- pro přípravu manažerské dokumentace firmy k podnikové analytice a jejímu užití v řízení,
- pro řešení projektů business intelligence, self service business intelligence nebo projektů pokročilé analytiky

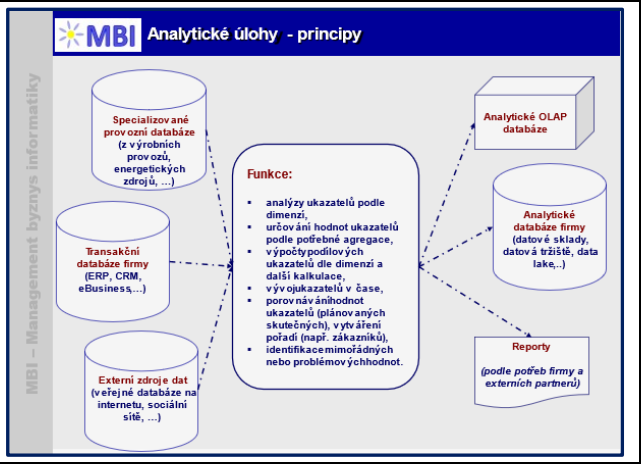
Řešení analytické úlohy s využitím dokumentů MBI-AF

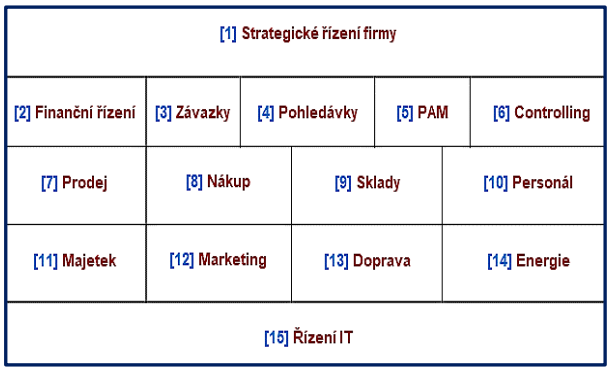
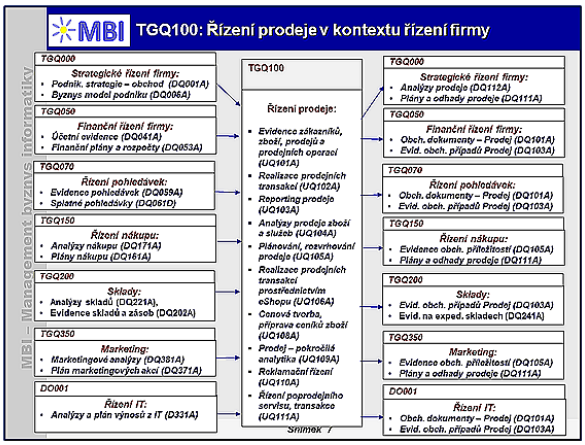
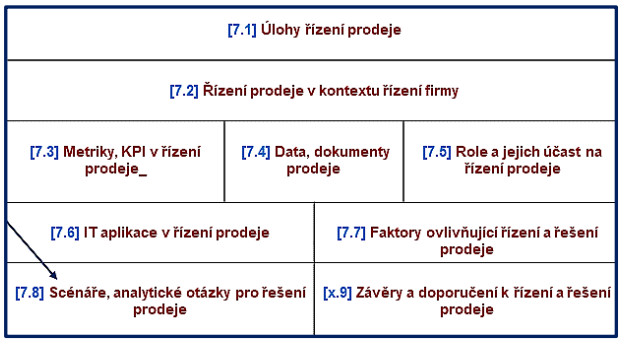
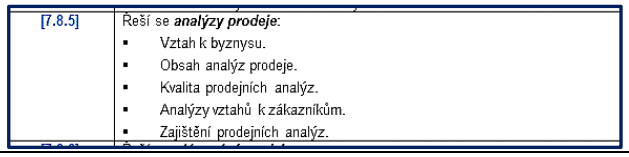
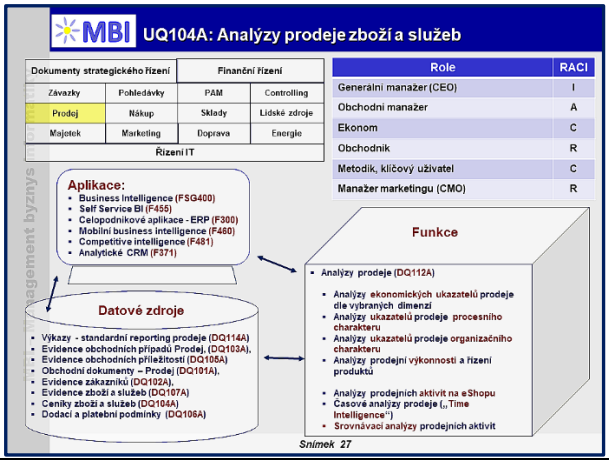
Účel:

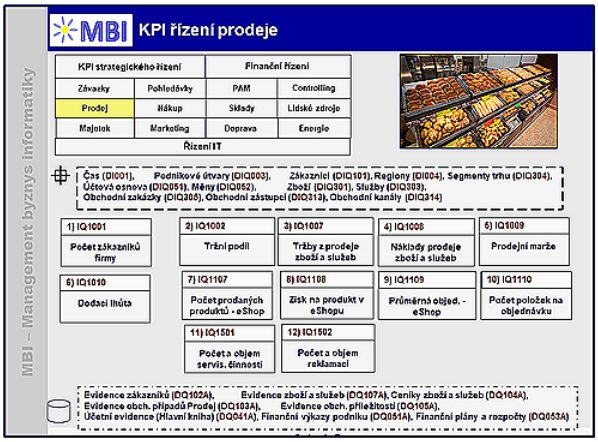
Řešení ***analytických úloh a plánovacích úloh*** mají specifický charakter***. Účelem*** je:

- připravit ***pro řešení*** analytické úlohy ***potřebné podklady*** s možným využitím dokumentů MBI_AF, případně z portálu MBI,
- je nutné ***vybrat všechny ty oblasti, komponenty a další aspekty*** řízení, které jsou pro další řešení projektu relevantní,
- vyhodnotit (podle potřeby) ***podstatné faktory firemního a byznys prostředí***, které další řešení podnikové analytiky ve firmě budou výrazně ovlivňovat (např. odvětvová působnost, vlastnictví firmy, byznys model apod.),
- pro vlastní průběh projektu připravit a ***formulovat specifické analytické otázky*** ve vazbě na charakter analytických úloh a zvažované metodiky, metody a faktory, a to ***jako základnu pro jednání se zákazníkem***, zaměřené na obsah analytických úloh, jejich funkcionalitu a realizační možnosti,
- formulace navrhovaného ***zajištění oblasti řízení IT aplikacemi*** (business intelligence, self service business intelligence, případně další) s příslušnou argumentací,
- ***zapracování příslušné dokumentace*** a specifikace jejího ***využití v řízení a řešení*** projektů.

Doporučený postup:

	Krok	Podklad v dokumentech MBI-AF
1	***Vymezení zadání, principů a nároků*** řešení reportingu, analytických úloh, plánovacích úloh a pokročilé analytiky: ▪ dokument: AF II.02: Komponenty, ▪ kapitoly 1.3, 1.4, 1.5 a 1.6 ▪ příklad: Principy analytických úloh.	

	Krok	Podklad v dokumentech MBI-AF
2	***Určení, v kterých oblastech řízení budou analytické úlohy řešeny*** a s jakými prioritami: ▪ dokument: „*AF II.01: Oblasti řízení*“,	
3	***Specifikace relevantních vazeb analytiky dané oblasti řízení (např. prodej) k ostatním oblastem řízení.***: ▪ dokument: „*AF II.01: Oblasti řízení*“, kapitola x.2	
4	***Příprava analytických otázek*** k řešení úloh podnikové analytiky pro jednání se zákazníkem, výběr „scénáře“, např. v řízení prodeje: ▪ dokument: „*AF II.01: Oblasti řízení*“, kapitola x.8. Scénáře,	
5	***Úprava a konkretizace analytických otázek*** speciálně pro podnikovou analytiku: ▪ dokument: „*AF II.01: Oblasti řízení*“, kapitola x.8. Scénáře, ▪	
6	***Vymezení základní funkcionality*** analytické úlohy: ▪ dokument: *AF II.01: Oblasti řízení*“, kapitola x.1	

	Krok	Podklad v dokumentech MBI-AF
7	***Výběr metrik a KPI pro podnikovou analytiku*** v dané oblasti (např. prodeje): ▪ dokument: „*AF II.01: Oblasti řízení*“, kapitola x.3	
8	***Detailnější vymezení jednotlivých charakteristik metrik*** (podle potřeby), včetně souvisejících metrik, analytických dimenzí a zdrojů dat: ▪ dokument: „*AF II.02: Komponenty*“, kapitola 2, ▪ zahrnuje: ○ vymezení metriky, ○ související metriky, ○ adekvátní dimenze, ○ datové zdroje pro metriku.	[2.1] Metriky strategického řízení [2.2] Finanční řízení | [2.3] Řízení závazků | [2.4] Řízení pohledávek | [2.5] PAM [2.6] Controlling | [2.7] Řízení prodeje | [2.8] Řízení nákupu | [2.9] Řízení skladů [2.10] Personální řízení | [2.11] Řízení majetku | [2.12] Marketing [2.13] Řízení dopravy | [2.14] Řízení energií | [2.15] Řízení IT
9	***Specifikace a úpravy obsahu analytických dimenzí*** využívanými v metrikách: ▪ dokument: „*AF II.02: Komponenty*“, kapitola 3, ▪ zahrnuje: ○ vymezení dimenze, ○ strukturu prvků dimenze	[3.1] Základní dimenze [3.2] Podniková organizace | [3.3] Ekonomické dimenze | [3.4] Externí partneři [3.5] Lidské zdroje a mzdy | [3.6] Obchodní dimenze | [3.7] Skladové hospodářství [3.8] Dimenze majetku | [3.9] Interní doprava | [3.10] Hospodaření s energiemi [3.11] IT služby a zdroje | [3.12] IT ekonomika
10	***Specifikace obsahu datových zdrojů*** pro řešené metriky: ▪ dokument: „*AF II.02: Komponenty*“, kapitola 4, ▪ zahrnuje obvykle: ○ vymezení datového zdroje, ○ strukturu obsahu datového zdroje	[4.1] Data a dokumenty strategického řízení [4.2] Finanční řízení | [4.3] Řízení závazků | [4.4] Řízení pohledávek | [4.5] Práce a mzdy [4.6] Controlling | [4.7] Řízení prodeje | [4.8] Řízení nákupu | [4.9] Řízení skladů [4.10] Personální řízení | [4.11] Řízení majetku | [4.12] Marketing [4.13] Řízení dopravy | [4.14] Řízení energií | [4.15] Řízení IT
11	***Návrh IT aplikací a nástrojů*** s určením jejich efektů a případných problémů: ▪ dokument: „*AF II.01: Oblasti řízení*“, kapitola 6	
12	***Kompletace dokumentu*** pro řešení úloh podnikové analytiky.	

Využití:

Výsledná dokumentace, nebo její části, může být zejména ***podkladem:***

- pro řešení projektů business intelligence, resp. self service business intelligence

Obsah

A) Úvod

Předkládaný text představuje součást souboru postupně připravovaných textů s názvem „***IT a anatomie firmy***“ jejímž společným záměrem je prezentovat ***otázky analýzy a návrhu informačních systémů zejména vzhledem k obsahu řízení*** firem a organizací. Tyto texty jsou s ohledem na rychlost využití realizovány na bázi ***stejných formálních i strukturálních pravidel***. Každý z jednotlivých textů se chápe primárně jako ***podklad pro řešení*** problémů a projektů, s kterými se analytik, nebo manažer v praxi obvykle střetává.

Cílem tohoto textu je poskytnout, pokud možno ***komplexní přehled*** o podnikové analytice, a to jak z hlediska jejího ***obsahu***, tak použitých ***technologií, metod i postupů***. Protože jde o poměrně rozsáhlou oblast odkazujeme se v některých částech na další detailní informace uvedené v ostatních dokumentech této řady.

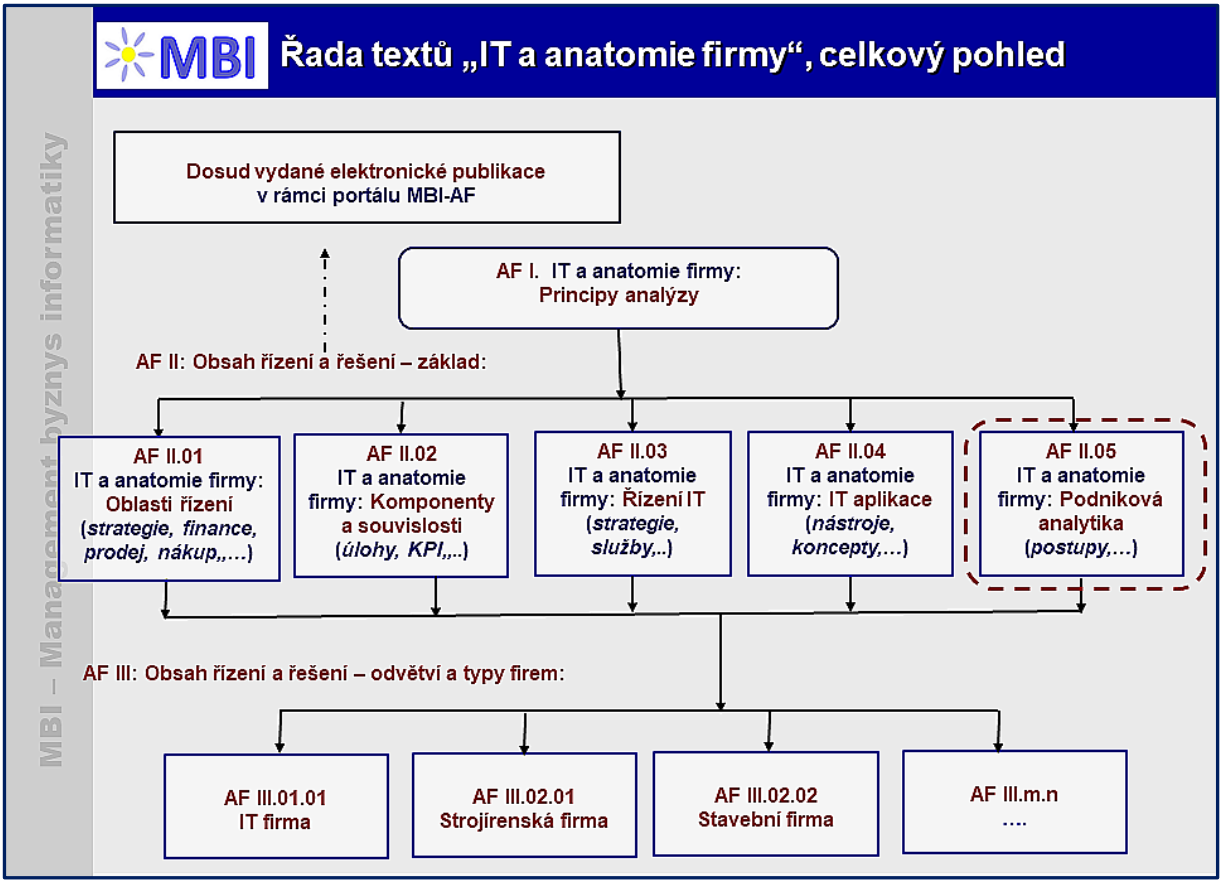
S rychlostí využití textů souvisí i začlenění tzv. „***map***“ na úvod celých dokumentů a jednotlivých kapitol, obsahujících jejich obsah s číselnými odkazy na příslušné části. ***Odkazy*** na jiné podkapitoly v rámci tohoto textu jsou standardně uvedeny ***v hranatých závorkách***, např. [2.3.4]. Odkazy na kapitoly v jiných textech nebo dokumentech nejsou automatické a jsou označeny symbolem „→“.

Obsahem jsou vzájemně na sebe navazující texty vycházející většinou ze stránek obdobného portálu MBI, který je na adrese https://mbi.vse.cz, ale ten je však ***momentálně z bezpečnostních důvodů přístupný*** uživatelům využívajících ***pouze doménu vse.cz***. V obrázcích a dokumentech se proto můžeme setkat s logem portálu MBI a s kódy objektů, které se na MBI používají.

Vzhledem k otázkám rychlého vývoje v byznysu i IT prostředí, jsou jednotlivé texty dostupné ***pouze elektronicky***. Předpokládáme tak, že se jednotlivé části budou ***průběžně aktualizovat*** a umisťovat na daný portál.

Ještě terminologická poznámka: V textu jsou použity termíny „***podnik***“ i „***firma***“ ve stejném nebo obdobném smyslu. Termín „***firma***“ pokládáme za základní, ale v mnohém kontextu je využití termínu „***podnik***“, „***podnikový***“ apod. přirozenější. Využíváme je tak podle obvyklých použití v praxi.

Celkový pohled na řadu dokumentů dokumentuje Obrázek A-1.



Obrázek A-1: Řada textů "IT a anatomie firmy"

Ke schématu a ***uspořádání*** řady doplníme ***několik poznámek*** s rozdělením na výše uvedené ***úrovně***. Na nejvyšší úrovni je speciální ***souhrnný dokument nabízející doporučené postupy*** při řešení obvyklých analytických nebo manažerských úkolů, a to s využitím dokumentů nebo jejich částí na dalších úrovních, tedy AF I – AF III.

A.1) Úroveň 1: AFI

Nejvyšší „vrstvu“ publikací představuje „***AF I. IT a anatomie firmy: Principy analýzy***“, kde jsou vymezeny ***hlavní přístupy a principy*** celého konceptu.

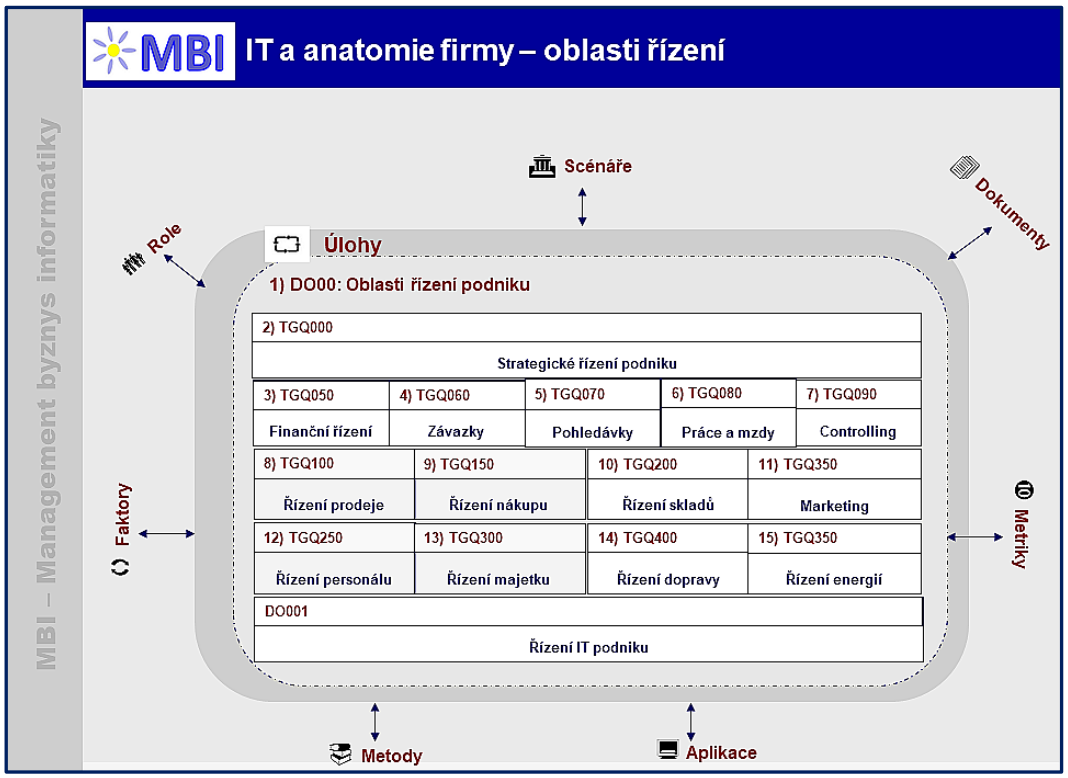
A.2) Úroveň 2: AF II

Texty na úrovni II navazují na „*Principy analýzy*“ a ***charakterizují firmu a její obsah řízení*** z analytických pohledů, ale ***bez odvětvových specifik***. Zahrnují dva základní texty, a to:

- „*AF II.01: IT a anatomie firmy: Oblasti řízení*“.
- „*AF II.02: IT a anatomie firmy: Komponenty a souvislosti*“.
- „*AF II.03: IT a anatomie firmy: Řízení IT*“:
- „*AF II.04: IT a anatomie firmy: IT aplikace, nástroje, koncepty*“.
- „*AF II.05: IT a anatomie firmy: Podniková analytika*“.

A.2.1) AF II.01. IT a anatomie firmy: Oblasti řízení

První publikací ***na dané úrovni*** je „*AF II.1. IT a anatomie firmy: Oblasti řízení*“. Kapitoly zde představují podstatné aspekty obsahu řízení ***podle jednotlivých oblastí řízení*** firmy. Oblasti řízení, na které se text orientuje dokumentuje Obrázek A-2.



Obrázek A-2: Oblasti řízení v anatomii firmy

Dokumentace každé z oblastí řízení je založena ***na těchto standardních pravidlech***:

- Vlastní obsah každé oblasti řízení je postaven na ***charakteristikách komponent***, které jsou v rámci jednotlivých kapitol ***prezentovány ve stejném pořadí***, tedy i ***se stejnými symboly i čísly*** podkapitol (např. x.3).
- Autoři předpokládají, že uživatel může využít ***samostatně pouze informace pro jednu oblast řízení***, a tedy pouze jednu z kapitol. Proto se využívá jejich standardní struktury, a navíc se řada částí mezi kapitolami ***může ve své podstatě opakovat***, i když s modifikacemi odpovídajícími dané oblasti řízení.
- Oblast ***řízení IT*** má speciální postavení, a následně i ***speciální text*** „*AF II.03. IT a anatomie firmy: Řízení IT*“.

Pohled na strukturu jednotlivých kapitol podle oblastí řízení dokumentuje následující tabulka:

Tabulka A-1: Přehled komponent řízení a jejich symbolické značení

	Přehled a obsah ***úloh*** tvořících oblast řízení, jsou náplní každé oblasti řízení a ***vymezují její hlavní obsah***. ▪ U základních oblastí řízení (strategie, finance atd.) jsou uvedeny odkazy na dokument „***AF.II.01: Oblasti řízení***“, kde je obsah vymezen detailněji.
	Řízení oblasti v kontextu řízení celé firmy, tj. ***vazby k ostatním oblastem řízení***, včetně řízení IT.
	Metriky a KPI využívané především v reportingu, analytických a plánovacích úlohách. ▪ Jsou součástí obsahu ***každé oblasti*** řízení. ▪ Přehled a obsah ***nejpodstatnějších metrik*** pro řízení daného typu firmy obsahuje ***příloha 1***. ▪ Celkový ***přehled metrik*** řízení firmy a jejich vymezení nabízí dokument „***AF.II.02: Komponenty a souvislosti***“ v kapitole ***2***. ▪ Celkový ***přehled analytických dimenzí*** a jejich vymezení nabízí dokument „***AF.II.02: Komponenty a souvislosti***“ v kapitole ***3***.

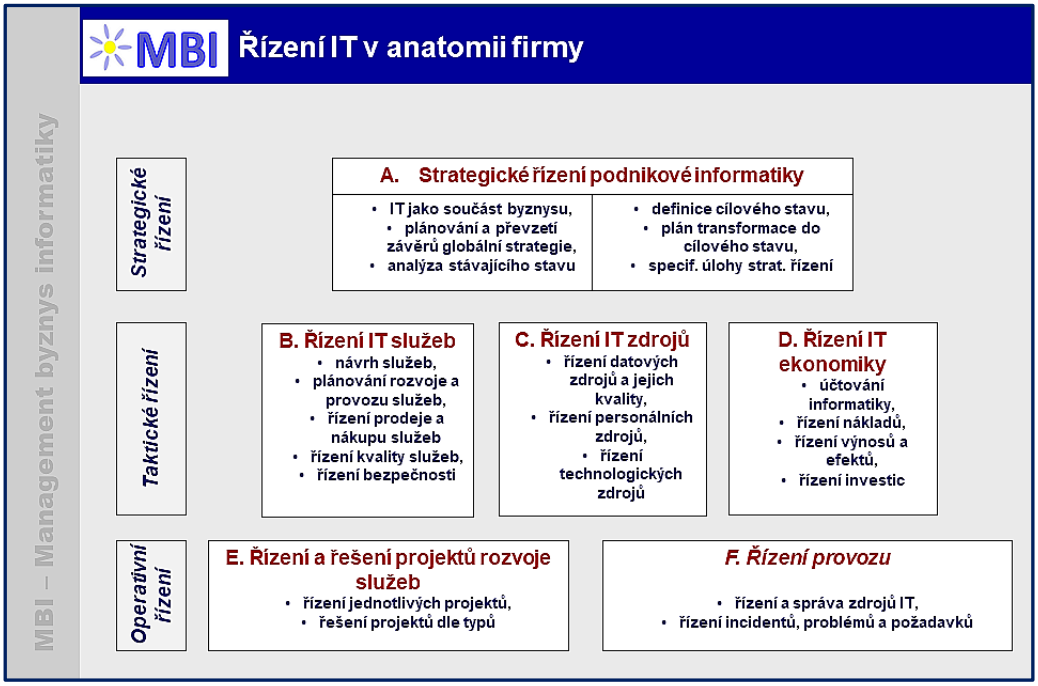
	Přehled hlavních ***dokumentů a dat*** využívaných v jednotlivých úlohách dané oblasti řízení. ▪ Přehled a obsah ***nejpodstatnějších dokumentů*** pro řízení daného typu firmy obsahuje ***příloha 2***. ▪ Celkový ***přehled dat a dokumentů*** řízení firmy a jejich vymezení nabízí dokument „***AF.II.02: Komponenty a souvislosti***“ v kapitole ***4***.
	***Role*** zajišťující úlohy řízení daného typu firmy. ▪ Celkový ***přehled rolí*** řízení firmy a jejich náplň nabízí dokument „***AF.II.02: Komponenty a souvislosti***“ v kapitole ***5***.
	Možnosti ***využití IT aplikací***, zejména aplikací, jejich výhody a nevýhody. ▪ Celkový ***přehled aplikací*** v řízení firmy a jejich charakteristiky, výhody a nevýhody nabízí dokument „***AF.II.04: IT aplikace***“.
	Podstatné ***faktory*** ovlivňující řízení, její rozvoj, celkový charakter i úspěšnost daného typu firmy. ▪ Celkový ***přehled faktorů*** řízení firmy a jejich vymezení nabízí dokument „***AF.II.02: Komponenty a souvislosti***“ v kapitolách ***6 a 7***.
	Scénáře představující ***analytické otázky*** k řešení problémů a projektů v oblasti řízení, orientované na hlavní nebo potenciální problémy dané oblasti. ▪ S ohledem na ***specifický charakter*** pro každou oblast řízení jsou uvedeny ***v rámci kapitol*** každé jednotlivé oblasti řízení.
	Hlavní ***doporučení*** k analýze a návrhu řešení informačního systému pro danou oblast řízení.

A.2.2) AF II.02. IT a anatomie firmy: Komponenty a souvislosti

V daném případě je text „***AF II.2. IT a anatomie firmy: Komponenty a souvislosti***“ založen ***na specifikaci jednotlivých komponent*** anatomie firmy a jejich vzájemných vazbách (viz předchozí tabulka)***.*** V tomto dokumentu je obsahem jejich ***detailnější specifikace a přehledy*** napříč firmou.

A.2.3) AF II.03. IT a anatomie firmy: Řízení IT

Dokument obsahuje vyčleněné řízení IT rozdělené do celé řady specifických domén a oblastí, kde každé z nich odpovídá jedna kapitola, a to ve stejné struktuře a se stejnými symboly jako v případě dokumentu „*AF II.01 Oblasti řízení*“.



Obrázek A-3: Oblasti řízení IT v anatomii firmy

A.2.4) AF II.04. IT a anatomie firmy: IT aplikace, nástroje, koncepty

Dokument pokrývá přehled a podstatné charakteristiky zejména jednotlivých typů aplikací, nástrojů a konceptů. Nezahrnuje konkrétní produkty, pouze odkazy na analytické portály, kde jsou tyto informace k dispozici. Charakteristiky IT produktů jsou pojaty z analytického pohledu, tedy kromě základní specifikace i potenciální efekty a na druhé straně problémy, které uživateli přinášejí. Součástí jsou i doporučené postupy implementací u vybraných typů aplikací.

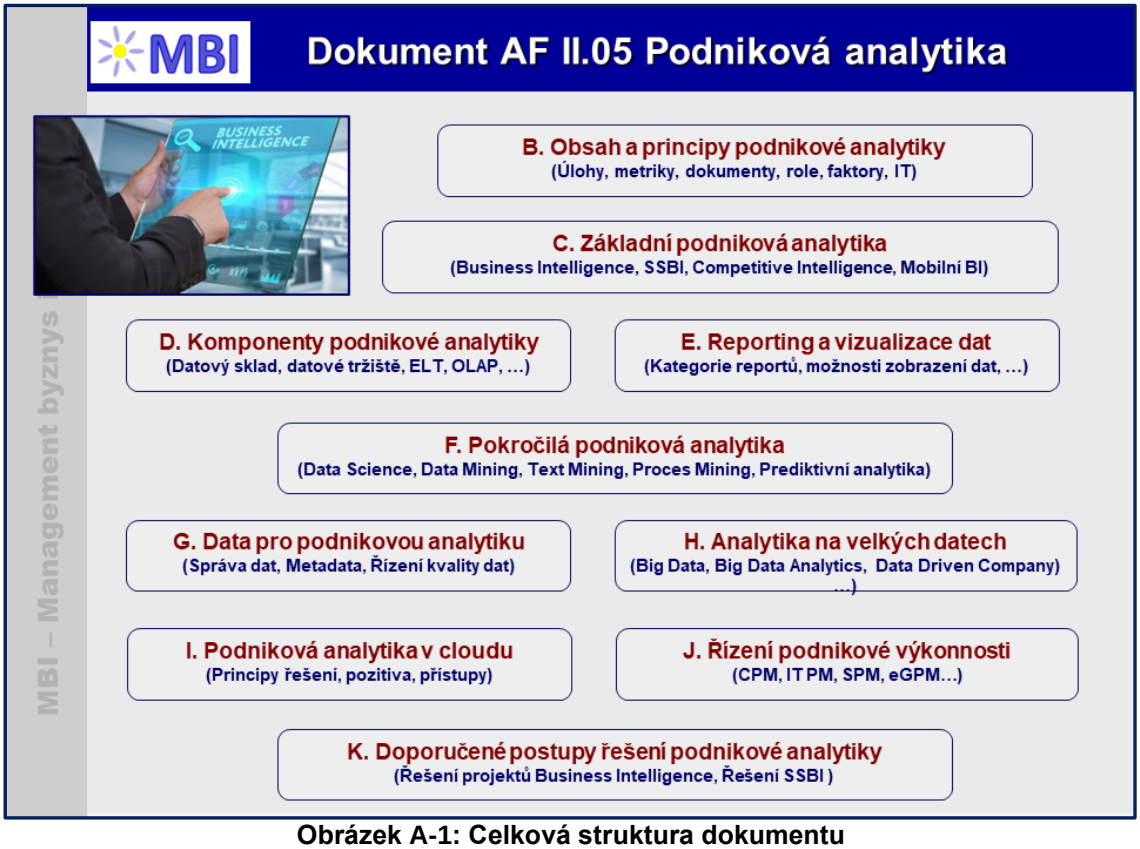
A.2.5) AF II.05. IT a anatomie firmy: Podniková analytika

Dokument podává analýzu nástrojů, konceptů a doporučených postupů v oblasti podnikové analytiky Charakteristiky IT produktů jsou pojaty z analytického pohledu, tedy kromě základní specifikace i potenciální efekty a na druhé straně problémy, které uživateli přinášejí. Dokument je rozdělen na „*základní podnikovou analytiku*“ (business intelligence, self service business intelligence, competitive intelligence) a „*pokročilou podnikovou analytiku*“. Součástí jsou i doporučené postupy implementací u vybraných typů aplikací.

A.3) Úroveň 3: AF III.01 – AF III.0n

Texty na úrovni 3 vycházejí a navazují na dokumenty na prvních dvou úrovních ovšem se zaměřením na ***určité typy firem a vybraná odvětví ekonomiky*** (výroba, maloobchod atd.). To znamená, že staví ***na společném základu*** oblastí řízení firmy v dokumentu „*AF II.1 IT a anatomie firmy: Oblasti řízení*“ a ***využívají komponenty*** definované v textu „*AF II.2. IT a anatomie firmy: Komponenty a souvislosti*“. Předmětem jsou zde zejména obsahová specifika u ***standardních oblastí řízení*** (jako např. finance, personalistika, majetek apod.) doplněné o ***specifické oblasti řízení*** (jako např. u výroby operativní řízení výroby apod.). Specifické oblasti řízení jsou dokumentovány ve stejné struktuře a na základě stejných pravidel, jako u společného základu.

Tento dokument se komplexně orientuje na podnikovou analytiku a její nejrůznější formy, nástroje, technologie i zdroje. ***Celkový pohled*** na obsah dokumentu ***podle jednotlivých oddílů textu*** nabízí následující schéma:



Obrázek A-1: Celková struktura dokumentu

B) Obsah a principy podnikové analytiky

[1] Obsah podnikové analytiky	**[2] Role v podnikové analytice**
[3] Faktory podnikové analytiky	**[4] Metody a metodiky podnikové analytiky**

Vstupní oddíl do podnikové analytiky (po úvodu) pokrývá zejména její ***obsahovou stránku***, tedy co je nebo by mělo být prvotním zájmem analytika při řešení projektů. Patří sem tak všechny ***aspekty tvořící logiku podnikové analytiky a současně základnu*** pro využití nejrůznějších ***technologií a nástrojů***, které jsou předmětem dalších navazujících oddílů.

Oddíl zahrnuje 4 kapitoly, jak ukazuje schéma, jejichž náplň stručně objasní ***následující přehled***:

- Východiskem je celkové ***vymezení a kategorizace*** podnikové analytiky (***kapitola 1.1***), tj specifikace hlavních témat řešení i s odvolávkou na příslušené části, resp. kapitoly tohoto a dalších dokumentů.
- Řešení podnikové analytiky se vztahuje k jednotlivým ***úlohám nebo procesům řízení firmy*** (viz Obrázek 1-1), přičemž evidenční a transakční úlohy představují z pohledu analytiky pouze přípravu dat a na tomto místě se jimi přímo nezabýváme. Úlohy s bezprostředním vztahem k podnikové analytice jsou (Obrázek 1-1):
  - úlohy reportingu,
  - analytické úlohy,
  - plánovací úlohy,
  - úlohy pokročilé analytiky.
- ***Kapitola 1.2*** vymezuje ***obsah a obecně funkcionalitu*** uvedených úloh a současně scénáře, resp. ***hlavní otázky*** pro jejich řešení a diskuze se zákazníky a uživateli v rámci projektů. Předpokládá se, že mohou tvořit ***základ jejich konkretizace*** ve vztahu k jednotlivým oblastem řízení firmy uvedeným v dokumentu „***AF.II:01: Oblasti řízení***“.

- Další klíčovou komponentou řešení podnikové analytiky je celý ***systém metrik*** a příslušných ***analytických dimenzí***. Rozsah jejich dokumentace je mimořádně velký, a proto na tomto místě (***kapitola 1.3***) se pouze rekapitulují jejich základní ***principy*** a specifikuje se ***umístění jejich dokumentace***, především v dokumentu „***AF.II.02: Komponenty a souvislosti***“ a následně i v dokumentech odvětvových řešení.
- V souvislosti s metrikami i dimenzemi je podstatnou otázkou, kde se pro ně získají data, tedy ***datové zdroje***. I v tomto případě jde o poměrně rozsáhlou dokumentaci, a proto jako v předchozím případě pouze odkazujeme na jejich umístění v dokumentu „***AF.II.02: Komponenty a souvislosti***“ a odvětvových řešení.
- Pro řešení projektů musí být definovány příslušné ***role pracovníků***, a to jak na řešitelské, tak uživatelské straně. ***Kapitola 2*** obsahuje vymezení vybraných rolí specifických pro potřeby podnikové analytiky s určením jejich ***funkční náplně*** i potřebných ***znalostí nebo dovedností***.
- Obsahem ***kapitoly 3*** je charakteristika ***faktorů*** majících obvykle ***podstatný vliv*** na průběh i úspěšnost řešení analytických projektů. Představují zde poměrně široké spektrum těchto faktorů rozdělených do skupin podle povahy faktorů.
- Poslední kapitola oddílu, ***kapitola 4*** se zaměřuje na poněkud detailnější prezentaci ***dvou vybraných metod***, které se k řešení podnikové analytiky obvykle vztahují, a to ***multidimenzionálního a datového modelování***. Metodik a metod je ve spojení s podnikovou analytikou ale celá řada, a proto i zde je v těchto případech uvedena pouze navigace do dokumentu „***AF.II.02: Komponenty a souvislosti***“.
- V závěru oddílu B je dobré projít a posoudit ***souhrnná doporučení*** k řešení projektů podnikové analytiky z pohledu jejího obsahu.

1. Obsah podnikové analytiky

Obsah podnikové analytiky, tedy ekonomickou nebo obchodní náplň řízení firmy lze ***vyjádřit úlohami***, které jsou předmětem řešení, ***metrikami, adekvátními dimenzemi a daty*** a datovými zdroji pro podnikovou analytiku.

Účelem kapitoly je:

- ***vymezit*** podnikovou analytiku, ***kategorizovat součásti*** jejího obsahu ***a úrovně*** řešení,
- ***určit charakteristiky vybraných typů úloh*** ve vztahu k podnikové analytice,
- ***určit*** podstatné ***charakteristiky metrik***, které jsou předmětem řešení úloh podnikové analytiky a souvisejících ***analytických dimenzí***,
- ***určit*** podstatné ***charakteristiky datových zdrojů*** pro podnikovou analytiku a jejich využití.

1.1 Vymezení a kategorizace podnikové analytiky

Podniková analytika v současnosti tvoří integrální součást podnikového řízení a s tím i součást IT i anatomie firmy. Zatímco termínem ***analýza*** („*analysis*“) se obvykle označují principy a aktivity spojené s řešením informačních systémů firem, pak termín ***analytika*** („*analytics*“) se využívá pro vyhodnocování a prezentování firemních aktivit a jejich výsledků v podobě celé škály jejich ukazatelů a souvisejících analytických dimenzí. Zahrnují převážně analytické nebo plánovací činnosti. ***Podniková analytika*** pak představuje svým způsobem zastřešující pojem, případně i synonymum pro různé specifické, či dílčí disciplíny a pojetí jako je business analytika, datová analytika a další a s nimi spojené dnes již obvyklé nástroje a technologie, jako je business intelligence, competitive intelligence a další.

Pokud bychom měli rozdělit, resp. ***kategorizovat*** podnikovou analytiku z pohledu jejích aplikací a využívaných metod, pak je pracovně můžeme ***členit na dvě skupiny***:

- ***základní***, ***deskriptivní,*** která má v praxi zřejmě největší rozšíření s aplikacemi a nástroji, které mají již většinou delší historii a kam patří zejména business intelligence (BI), self service business intelligence (SSBI), competitive intelligence (CI) a technologická varianta předchozích možností, tedy mobilní business intelligence apod. další),
- ***pokročilá,*** která zejména v poslední době nabírá rychle na významu a praktickém rozšíření a která je většinou založená na velmi sofistikovaných metodách, modelech i algoritmech vycházejících z matematiky a statistiky. Sem zřejmě patří zejména data mining, process mining, text mining, prediktivní analytika, data science, umělá inteligence, strojové učení a další. Je nutné současně přiznat, že tyto disciplíny se často vzájemně překrývají nebo doplňují a je obtížné mezi nimi hledat ostré hranice.

Z pohledu řešení a uplatnění podnikové analytiky je dobré upozornit na následující její ***výrazné aspekty***:

- Aplikace a řešení podnikové analytiky jsou převážně ***orientovány na potřeby pracovníků vyšších úrovní řízení*** firmy, tedy manažery, podnikové specialisty, vlastníky, členy správních rad apod. Tomu musí ***odpovídat i způsob a forma poskytovaných informací***. Na druhé straně s rozvojem i nových technologií uplatnění analytiky postupně dostává i na nižší úrovně firemního řízení.
- ***Na rozdíl od transakčních úloh a aplikací*** je při řešení úloh analytiky nutné ***respektovat tyto momenty***:
    - I když i v analytické oblasti jsou v nabídce trhu určité standardní (typové) aplikace, přesto převažující smysl poskytovaných analytických aplikací je ***v jejich individualizaci*** podle potřeb konkrétních manažerů, specialistů a dalších.
    - Právě individualizace řešení je i základem obvykle skloňovaného jejich efektu, tedy hlavní ***přínos pro konkurenceschopnost firmy*** a získávání konkurenčních výhod. Standardní, typové aplikace příliš konkurenčních výhod nepřinášejí, naopak čím

„chytřejší“ budou individuální aplikace podnikové analytiky (analytické i plánovací), tím vyšší je potenciál budoucích výhod.

- Podstatný rozdíl transakčních aplikací a aplikací podnikové analytiky spočívá v tom, že zatímco ***transakční aplikace obvykle datové zdroje*** vytvářejí (např. na základě vstupních dokladů), pak ***podniková analytika tyto zdroje využívá*** pro realizaci analýz, plánů, prognóz. To současně vytváří na tyto zdroje i specifické nároky, např. v oblasti jejich dostupnosti, kvality dat apod.

- ***Při praktickém užití aplikací podnikové analytiky*** manažeři a specialisté firmy pracují a vyhodnocují podnikové ukazatele podle různých dimenzí, v delším časovém vývoji, s predikcemi na delší časové horizonty, s identifikací problémů, kterou jsou jinak obtížně identifikovatelné. To vše znamená, že uplatnění podnikové analytiky vede přirozenou cestou ke ***zvyšování manažerské kvalifikace***.

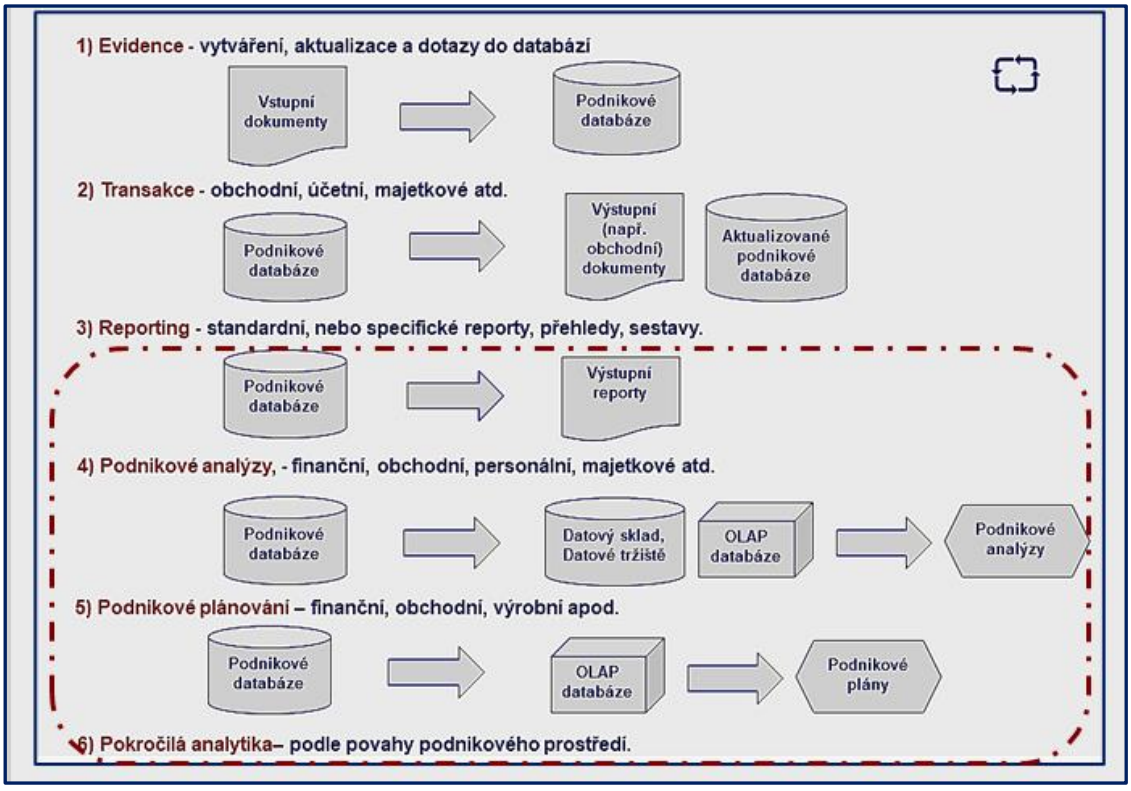
1.2 Úlohy, procesy

Úloha je v anatomii firmy ***základní komponentou,*** a tedy i v podnikové analytice. Úlohy mají zachytit a dokumentovat řízení firmy, tedy co se realizuje, řídí a řeší. ***Termín „úloha“*** se jeví jako ***obecnější***, a to proto, že její součástí ***může nebo nemusí být „proces“*** *nebo* naopak úloha může být součástí nějakého komplexního procesu. „*Úlohy*“ mohou v podnikové analytice primárně zahrnovat přehled tzv. „***klíčových aktivit***“, tj. činností, které se nejvýraznějším způsobem podílejí na finálním výsledku úlohy. Z pohledu analytických nebo plánovacích úloh to mohou být aktivity spojené s analýzou nebo plánováním vybraných skupin metrik, se sledování časového vývoje jejich hodnot, porovnáním plánovaných a skutečných hodnot, plánování hodnot podle určitých časových horizontů apod.

Počet úloh je v řízení firmy obvykle značný, a proto je dobré je nějakým způsobem ***členit, resp. kategorizovat***, v našem případě využíváme hlediska dvě:

- úlohy rozdělené ***podle oblastí řízení,*** např. finance, prodej atd.,
- úlohy rozdělené podle ***standardních operací s daty,*** tj. evidence, transakce, reporting, analýzy, plánování, pokročilá analytika, případě specifické úlohy.

Kategorizace úloh z obou pohledů je využita v publikaci „***IT a anatomie firmy (Oblasti a komponenty řízení)***“. Na tomto místě pouze pro rekapitulaci uvádíme přehled typů úloh podle operací s daty se zvýrazněním těch, které se přímo váží k podnikové analytice (Obrázek 1-1).



Obrázek 1-1: Typy úloh v podnikové analytice

Typy úloh v rámci této pracovní klasifikace zahrnují:

- ***Evidenční*** úlohy zahrnují především pořizování dat a jejich kontroly.
- ***Transakční*** úlohy zajišťují provádění obchodních, finančních a dalších transakcí, tj. např. realizace obchodního případu „Prodej“, “Nákup“ apod.
- ***Reportingové*** úlohy obsahují návrh, zpracování a distribuci reportů.
- ***Analytické*** úlohy představují analýzy finančních, obchodních a dalších aktivit firmy.
- ***Plánovací*** úlohy zajišťují plánování na bázi podstatných plánovacích ukazatelů a jejich dimenzí.
- ***Úlohy pokročilé analytiky*** znamenají realizaci pokročilých analytických funkcí, jako je např. prediktivní analytika atd.

Evidenční a transakční úlohy představují z pohledu podnikové analytiky zejména vytváření a aktualizace datových zdrojů jako vstupů do jejích úloh. Další typy úloh, tj. ***reportingové, analytické, plánovací analytiky*** se k podnikové analytice již bezprostředně váží. Jejich obsah byl rovněž vymezen v publikaci „***IT a anatomie firmy (Oblasti a komponenty řízení)***“ a proto se zde na ni pouze odvoláváme. Úlohy pokročilé analytiky budou obsahem jedné z následujících publikací.

→

Vymezení ***obsahu všech úloh*** zařazených do anatomie firmy obsahuje dokument „*AF II.01: Oblasti řízení*“.

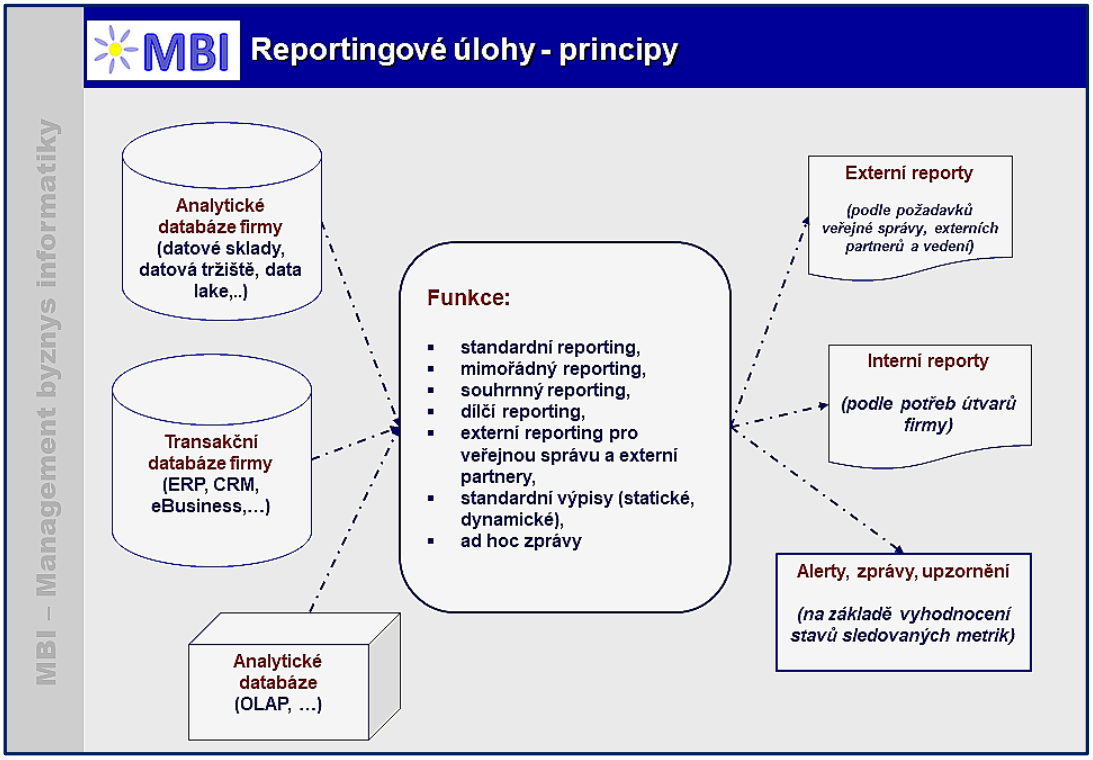
Charakteristiky obsahu ***úloh základních i specifických oblastí řízení*** v rámci odvětvových řešení obsahují dokumenty úrovně III, např. „*AF III 01: IT firma*“.

Každá z dalších podkapitol věnovaná jednotlivým typům úloh zahrnuje kromě vymezení jejich podstaty a funkcionality i „***Scénáře, analytické otázky***“, které by měly být předmětem diskusí analytika se zákazníkem, resp. uživatelem a „***Některá doporučení***“, k řešení těchto úloh.

Tyto otázky i doporučení se promítají a ***konkretizují podle jednotlivých oblastí řízení*** firmy a jsou obsaženy v příslušných podkapitolách dokumentu „*AF.II.01: Oblasti řízení*“.

1.2.1 Reportingové úlohy

Podstatu reportingových úloh dokumentuje další obrázek. ***Cílem reportingu*** je nejen prezentovat informace vzhledem k uživatelským požadavkům, ale i zajištění takové jejich distribuce, která pracovníkům v podniku poskytne přístup pouze k jim relevantním údajům a bude chránit důvěrná data před nežádoucím šířením (Fibírová a Šoljaková, 2010).



Obrázek 1-2: Principy reportingových úloh

1.2.1.1 Kategorizace reportů

Kategorizace reportů zahrnuje:

Interní reporting:

- ***Standardní reporting*** se vyznačuje zprávami dodávanými v pravidelných časových úsecích, obvykle jednou za měsíc, za čtvrtletí nebo za celý rok. Struktura zprávy z hlediska obsahu informací, výpočtů nebo analýz je předem stanovená.
- ***Mimořádný reporting*** představuje zprávy generované na požadavek nebo reporty, které se běžně negenerují, například analýza obchodního rizika, analýza sortimentních skupin apod.
- ***Souhrnný reporting*** podává přehled o činnostech firmy za určité období, např. základní finanční ukazatelé včetně srovnání těchto hodnot s hodnotami plánovanými nebo s hodnotami minulé časové periody.
- ***Dílčí reporting*** se věnuje konkrétnějším informacím, struktura více odpovídá daným požadavkům pracovníků firmy.

Externí reporting:

- ***Povinnost ze zákona*** předkládat výsledky hospodaření formou auditovaných finančních výkazů obsahuje velkou část externího reportingu. Jde o standardní periodické výkazy o hospodaření, které vycházejí jednou ročně, tedy o rozvahy, výsledovky a výkazy peněžních toků.
- ***Další kategorie reportů*** zahrnují:
  - ***Standardní* výpisy** představují předem definované dotazy, jež se zpravidla spouští v jim určený a nastavený čas.
  - Standardní reporty ještě lze dále členit na statické a dynamické. Uživatel si ***statický výpis*** již nemůže upravovat, kdežto s daty ***v dynamickém reportu*** může manipulovat podle svých potřeb. Je však vždy omezen nástroji, které mu byly vyhrazeny.
  - ***Ad hoc zprávy*** si může vytvořit sám uživatel vytvořením konkrétního jednorázového dotazu nad databázemi.

1.2.1.2 Scénáře, analytické otázky

Vztah k byznysu:

- Jaký ***obsah*** mají mít reporty a dashboardy vzhledem k potřebám uživatelů? Jak zajistit jejich ***flexibilitu*** vzhledem ke změnám těchto potřeb?
- Jak ***minimalizovat duplicity*** reportů a dashboardů, resp. jejich překrývání v obsahu i funkcionalitě?
- Jak zajistit efektivní, přehledné a ***souhrnné informování*** uživatelů o disponibilních reportech a dashboardech a jejich dostupnosti?
- Jaké ***legislativní aspekty*** je nezbytné při přípravě reportů a dashboardů brát v úvahu?
- Jaké ***standardy*** (mezinárodní, národní i firemní) je nezbytné při navrhování reportů?
- Jak specifikovat ***potenciální efekty*** kvalitního reportingu, a to ***ekonomické i mimoekonomické***?
- Které ***faktory*** budou pro řešení a uplatnění reportingu ***rozhodující***?
- Jak zajistit kvalitní a systematické ***kooperace*** jednotlivých útvarů firmy při řešení reportingu?
- Jaké ***vlivy*** má úroveň a využití reportingu ***na řízení a organizaci firmy***, jaké jsou hlavní problémy nebo omezení?

Vztah k datovým zdrojům:

- Jaký je aktuální ***obsah datových zdrojů*** vzhledem k potřebám reportingu?
- Jaká je ***kvalita datových zdrojů*** vzhledem k nárokům reportingu?

Vztah k personálu:

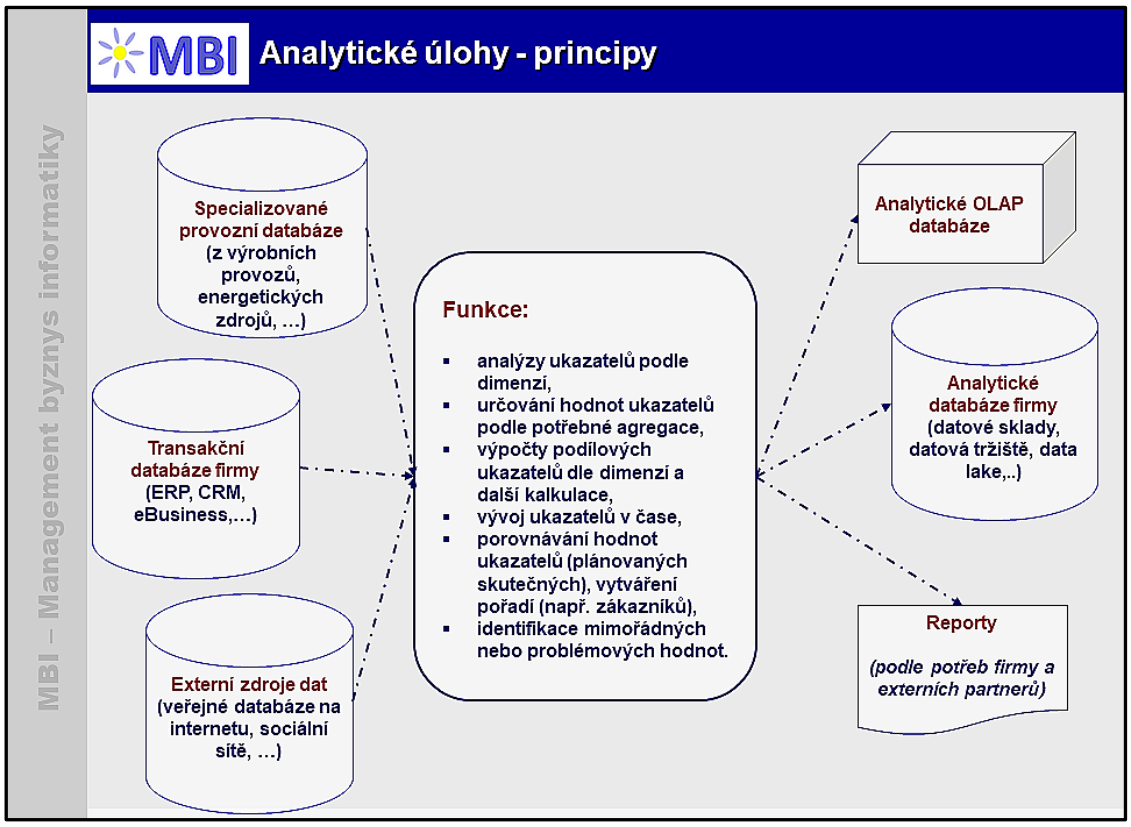
- Jak dosáhnout potřebné ***úrovně kooperace*** manažerů, specialistů a ostatních uživatelů v oblasti rozvoje podnikového reportingu?
- Jak ***připravovat pracovníky*** na uplatnění nových technologií při práci s reporty a jejich přípravě?

Vztah k IT:

- Jak zajistit vysokou ***prezentační úroveň*** výsledků řešení reportů a dashboardů?
- Jaké ***IT produkty,*** např. standardní kancelářské, speciální reportovací, business intelligence a další bude účelné v reportingu vzhledem k potřebám i možnostem firmy využít?
- Jak dosahovat ***zkracování doby a časové náročnosti*** na přípravu reportů?
- Jak využít dostupná ***cloudová řešení*** na druhé straně ***on premis***, případně jejich kombinace v realizaci reportů?

1.2.2 Analytické úlohy

Podstatu a principy analytických úloh dokumentuje další obrázek:



Obrázek 1-3: Principy analytických úloh

Analytické úlohy a jejich softwarové zajištění tvoří běžnou ***součást řízení firmy*** (finanční analýzy, analýzy závazků, pohledávek, analýzy mezd, analýzy prodeje a řada dalších). Jsou ***realizovatelné různými prostředky*** od standardních kancelářských produktů (Excel, Access), přes analyticky zaměřenou funkcionalitu transakčních systémů (ERP, CRM apod.) až po produkty Business Intelligence a Self Service Business Intelligence, které nabízejí v této oblasti největší možnosti.

1.2.2.1 Podstata a cíle analytických úloh

Analytické úlohy v prostředí BI a SSBI jsou z***aložené na vyhodnocování a presentaci ukazatelů podle vybraných dimenzí,*** a to na různé úrovni detailu odpovídající úrovním hierarchických struktur jednotlivých dimenzí (jak to vyplývá z předchozích částí kapitoly). Analytické úlohy nabízejí širokou škálu funkcí (viz další paragraf) a mají z pohledu podnikového řízení plnit následující ***cíle a požadavky***:

- zajistit vysokou ***kvalitu a komplexnost*** analytických operací v řízení podniku, tj. poskytovat údaje ke všem potřebným ukazatelům a podle všech relevantních dimenzí, a to nejen v dané oblasti řízení (řízení financí, prodeje atd.), ale i souvisejících napříč celým podnikovým řízením,
- nabízet adekvátní ***flexibilitu*** zpracovávaných analýz vzhledem k momentálním potřebám podnikových manažerů, analytiků a specialistů, tj. možnost vybrat pro analýzy pouze požadované ukazatele, pouze vybrané dimenze a jejich kombinace, případně pouze vybrané prvky těchto dimenzí, možnost upravovat, nebo doplňovat kalkulace nad ukazateli a dimenzemi apod.,
- poskytovat analýzy na aktuálně požadované úrovni ***granularity dat***, tj. na takové úrovni podrobnosti výstupů, která přesně odpovídá požadavkům i časovým možnostem pracovníků podniku,
- zajistit potřebnou ***dostupnost*** analytických operací a výstupů v místě a čase, tedy na základě, pokud možno nejaktuálnějších údajů, zpřístupnit analytické výstupy v místě pracoviště i mimo, tj. u zákazníků, dodavatelů, partnerů apod.,
- umožnit i v rámci analytických operací realizovat relevantní ***predikce hodnot***, tj. např. odhady objemu prodejů v následujících obdobích, kvalifikace případných obchodních rizik atd.,
- umožnit realizovat analýzy i na základě ***heterogenních a externích datových zdrojů***, tj. mimo

vlastních standardních strukturovaných dat i analýzy textových dat, obrazů, databází externích, z internetu, zdrojů ze sociálních sítí.

Cíle a požadavky kladené na analytické úlohy se samozřejmě ***liší mezi odvětvími, podniky i jednotlivými uživateli***. Jejich analýza a návrh jsou jednou z klíčových činností v rámci BI a SSBI projektů, v jejichž rámci se přesněji formulují, konkretizují a jsou základem dalších řešení. Výše uvedený přehled proto považujme pouze za východisko ukazující, s čím je dobré při koncipování těchto úloh a aplikací počítat.

1.2.2.2 Základní funkcionalita analytických úloh

Obdobně, jako v předchozím paragrafu, se zde zaměříme na vymezení pouze obecně formulované funkcionality analytických úloh, jako základu pro její další konkretizace v rámci jednotlivých projektů. Funkcionalitu těchto úloh můžeme pracovně rozčlenit následujícím způsobem:

- ***základní přehledy*** hodnot vybraných ukazatelů podle specifikovaných dimenzí a jejich vzájemných kombinací s možností nastavení filtrů na prvky dimenzí a jejich skupiny (*slice and dice*),
- vytváření ***filtrů a průřezů*** nad uloženými daty podle stanovených kritérií,
- operativní určování aktuálně ***požadované úrovně agregace***, resp. úrovně detailu pro vybrané ukazatele, resp. pohyb po různých úrovních detailu hodnot odpovídajících hierarchickým strukturám dimenzí (*drill down, drill up*),
- ***zjišťování detailních informací*** z primárních datových zdrojů odpovídajících vybranému ukazateli (faktu) a jeho hodnotě (*drill through*),
- ***výpočty a sledování podílových hodnot*** ukazatelů, např. podíl tržeb skupin zákazníků, resp. jednotlivých zákazníků na celkovém objemu tržeb, odpovídajících obvykle podílům podle úrovní jednotlivých dimenzí,
- zjišťování a analýzy hodnot ukazatelů ***statistického charakteru,*** jako např. maxima, minima, průměru, mediánu, percentilu, směrodatné odchylky, rozptylu,
- ***výpočty dalších odvozených ukazatelů*** ze základních podle okamžité potřeby pracovníků,
- ***časové charakteristiky a vývoj hodnot*** vybraných ukazatelů („*time intelligence*“),
- ***porovnávání rozpočtových, plánovaných a skutečně dosahovaných hodnot*** ukazatelů, tj. výstupy typu *scorecard*,
- ***vizuální identifikace problémových hodnot ukazatelů*** nebo naopak vysoce pozitivních hodnot podle aktuálně stanovených pravidel (např. Top 10), nebo referenčních hodnot.

1.2.2.3 Analytické úlohy v kontextu řízení firmy

Analytické úlohy jsou ***součástí prakticky každé oblasti podnikového řízení***, případně řízení celého podniku. Na základě výše uvedených základních funkcí analytických úloh lze formulovat jejich ***uplatnění v oblastech řízení firmy***, na tomto místě na obecné úrovni s tím, že je pak konkretizováno dle oblastí řízení (v dokumentu „*AF.II.01: Oblasti řízení*“). ***Pracovně*** lze přijmout následující ***klasifikaci***:

- Analýzy ***ekonomických a obchodních ukazatelů*** podle zvolených dimenzí, jako např. objem tržeb za zboží a služby, objem nakupovaného zboží, objem odpisů z majetku apod.
- Analýzy ***výkonových ukazatelů*** podle zvolených dimenzí, jako např. počet přijatých objednávek, počet vydaných faktur, počet servisních zásahů apod.
- Analýzy ukazatelů ***organizačního charakteru*** podle zvolených dimenzí, jako např. počet zákazníků, počet nových zákazníků, počet dodavatelů apod.
- Analýzy ***trendů, resp. časového vývoje*** vybraných ukazatelů podle dimenzí (tzv. „*time intelligence*“), tzn. podle jednotlivých let, čtvrtletí, měsíců, sledování hodnot ukazatelů k počátečnímu datu, např. začátku roku, meziroční srovnání nebo srovnání mezi odpovídajícími obdobími, výpočty a sledování různých typů indexů, např. řetězových nebo bazických (předpokladem je zde však dostupnost dat za delší časová období).
- ***Srovnávací analýzy*** ukazatelů podle dimenzí, např. porovnávání plánovaných hodnot se skutečnými, případně podle variant plánů, porovnání dosažených obchodních výsledků podle vybraných regionů, typů produktů, skupin zákazníků apod.
- Zpracování ***pořadí určitých objektů*** podle zvolených hodnot ukazatelů („*ranking*“), např. zákazníků podle objemu tržeb, dodavatelů podle objemu dodávek, reklamací podle jejich počtu nebo objemů apod.

- Zpracování výpočtů a ***analýz statistického charakteru***, např. zjišťování rozptylu hodnot apod.

1.2.2.4 Scénáře, analytické otázky

Vztah k byznysu:

- Jak zvýšit ***úspěšnost a výkonnost*** byznysu díky využití analytických úloh?
- Jak specifikovat ***potenciální efekty*** užití analytických úloh, a to ***ekonomické i mimoekonomické***?
- Které ***oblasti řízení*** jsou a budou z hlediska uplatnění analytických úloh ***prioritní***?
- Které ***metriky*** budou pro analytické úlohy významné, které budou mít charakter KPI, které ***dimenze*** ve vztahu k metrikám budou relevantní?
- Jak správně nastavit ***analytická pravidla*** ve vztahu k metrikám pro generování varovných zpráv (alertů)?
- Které ***faktory*** budou pro řešení a uplatnění analytických úloh ***rozhodující***?
- Jak zajistit požadovanou ***komplexnost a kvalitu*** analytických úloh?
- Jak dosáhnout požadované ***flexibility*** analytických úloh vzhledem k momentálním potřebám manažerů a specialistů firmy?
- Jaké ***vlivy*** má úroveň a využití analytických úloh ***na řízení a organizaci firmy***, jaké jsou hlavní problémy nebo omezení?
- Jak navrhnout a řešit ***analytické úlohy zaměřené na externí subjekty*** nebo partnery (zákazníci, dodavatelé apod.), jakou mají poskytovat funkcionalitu, na základě jakých zdrojů?

Vztah k datovým zdrojům:

- Jaký je aktuální ***obsah datových zdrojů*** vzhledem k potřebám analytických úloh?
- Jaká je ***kvalita datových zdrojů*** vzhledem k analytickým úlohám, jak řešit problémy s nekvalitou dat?
- Jak správně stanovit ***perspektivu datových zdrojů*** a příslušných transakčních aplikací vzhledem k plánovanému rozvoji analytických úloh?
- Jak dosáhnout požadovanou ***granularitu dat*** v rámci jednotlivých operací analytických úloh? Jak k tomu zajistit ***potřebné datové zdroje*** na požadované úrovni granularity?
- Jak se daří ***minimalizovat časová zpoždění aktualizace dat*** v datových zdrojích?

Vztah k personálu:

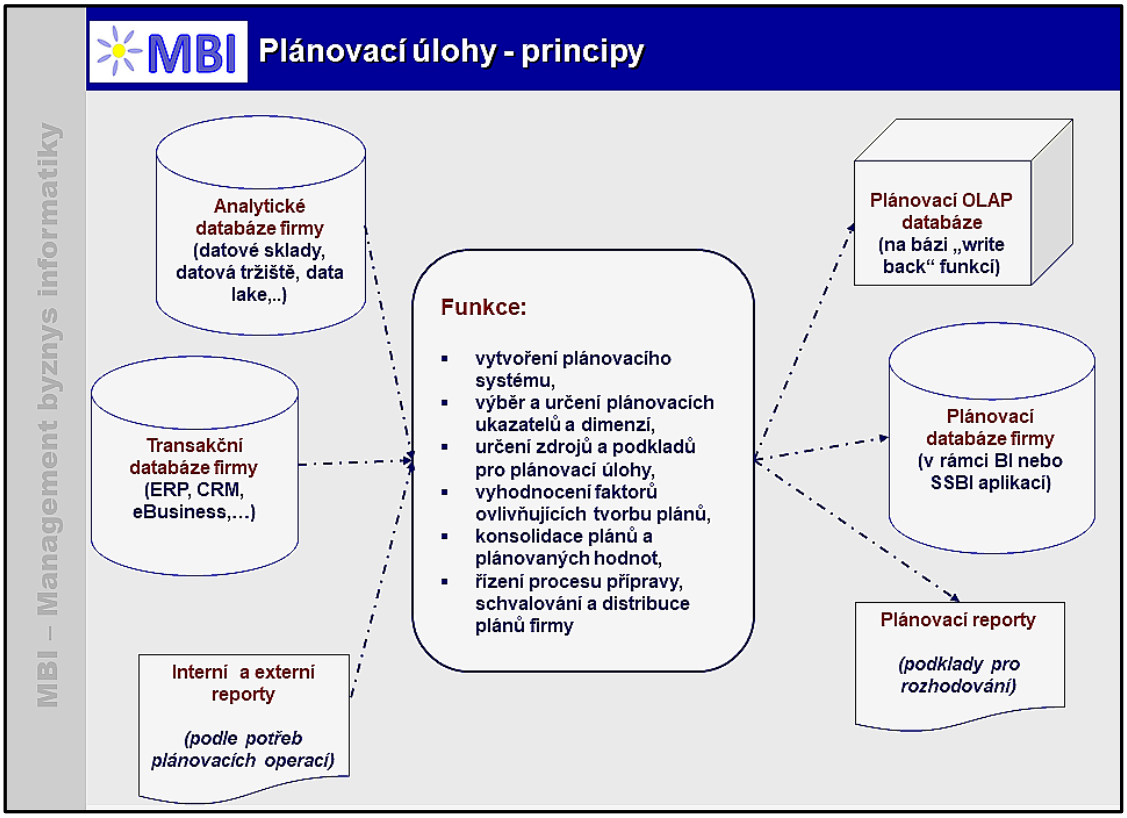
- Jak dosáhnout potřebné ***kvalifikace a motivace*** manažerů a specialistů v oblasti využití a případně i řešení analytických úloh?
- Jak dosáhnout ***konsensu mezi pracovníky*** firmy na navrženém obsahu a strukturách analytických úloh?
- Jak posilovat ***samostatnost pracovníků firmy*** při řešení analytických úloh a využívání analytických nástrojů?

Vztah k IT:

- Jaké ***IT produkty*** (BI, SSBI,…) bude účelné vzhledem k potřebám i možnostem firmy využít?
- Je k dispozici ***podpora*** analytických úloh ***v provozovaných transakčních aplikacích*** (ERP, CRM,…), jaké jsou možnosti a omezení jejich využití?
- Jak dosahovat ***zkracování doby a časové náročnosti*** na přípravu analytických úloh a aplikací?
- Jak řešit ***sourcing*** v oblasti řešení a implementací analytických úloh?
- Jak využít dostupná ***cloudová řešení*** na druhé straně ***on premis***, případně jejich kombinace v realizaci analytických úloh?

1.2.3 Plánovací úlohy

Podstatu a principy plánovacích úloh dokumentuje další obrázek.



Obrázek 1-4: Principy plánovacích úloh

Hlavním smyslem plánovacích úloh je ***snížení nejistoty budoucího vývoje,*** tj. plány a rozpočty poskytují kontrolní mechanismus, zdali jsou naplánované cíle plněny s určitými povolenými odchylkami (Žůrková 2007, s. 9). Klíčovým aspektem systému plánů a rozpočtů.je ***časový předstih***, s nímž jsou možná rizika a úzká místa plnění cílů a řízení firmy identifikována.

Úrovně plánování zahrnují 3 základní úrovně plánů:

- strategické, cca na 10 let,
- taktické (manažerské), 1 – 3 roky,
- operativní, denní, týdenní, měsíční či kvartální.

Klouzavé plánování a na ně navázané klouzavé rozpočty představuje v praxi tzv. ***prognózování (forecasting)***. Funguje na principu neustálé aktualizace plánů a rozpočtů, a to na základě skutečně sledovaného vývoje. Umožňuje průběžně porovnávat plány a rozpočty se skutečností a vyhodnocovat jejich dosavadní přesnost a příslušně je upravovat pro další období.

1.2.3.1 Podstata plánovacích úloh

Využití BI a SSBI technologií je velmi účelné i pro řešení plánovacích úloh (plánů tržeb, nákladů, investic atd.), kdy je ***třeba nejen data zpřístupňovat podle nejrůznějších definovaných dimenzí*** a analyzovat je podle nich, ale i nová data, tedy ***plány tvořit, rovněž s respektováním těchto dimenzí*** a jejich hierarchických struktur. Plánovací úlohy sledují ***několik základních cílů***, zejména:

- vytvořit a ***využít plánovací systém*** respektující v podniku uplatňované plánovací a rozvrhové metody,
- ***zajistit konsolidace vytvářených plánů*** vznikajících na různých organizačních jednotkách, tj. závodech, divizích, odděleních, nebo naopak rozpouštění centrálně stanovených plánů na tyto jednotky,
- ***zajistit konsolidace hodnot z různých druhů plánů***, např. plánu investičního, výrobního nákupního, prodejního, personálního apod. do výsledného, obvykle finančního plánu,

- ***zajistit konsolidace plánů z pohledu různých měn*** a přepočet na výslednou požadovanou měnu,
- automatizovat ***řízení pracovního toku (workflow)*** při přípravě plánu, resp. plánů, na kterém se podílejí různí manažeři, plánovači a další pracovníci podniku,
- efektivně ***zpřístupňovat sestavené plány*** zainteresovaným pracovníkům podniku,
- zajistit ***potřebnou bezpečnost a nastavení přístupových práv*** pro zpracování plánů i pro jejich presentaci v podniku, případně mimo podnik, kde jde o možnosti jejich čtení, zápisu a schvalování.

Základem plánovacích úloh v BI jsou ***plánovací a rozvrhovací algoritmy*** a z technologického hlediska i funkce zpětného zápisu (*write back*) do analytických databází.

1.2.3.2 Principy plánovacích úloh v BI

Data do analytických databází se obvykle ***nahrávají z primárních zdrojů***. Jinou možností je však tzv. ***zpětný zápis*** (*write back*), což je funkce, kdy data do analytické databáze může zadávat přímo uživatel, nebo určitá, většinou plánovací aplikace. To znamená, že uživatel, resp. uživatelská aplikace zde může měnit hodnoty ukazatelů a dimenzí. Tyto změny jsou zde uloženy a je možné je i zpětně promítat do primárních zdrojů. Takové možnosti vytváření plánů a realizace změn prováděných uživatelem zpětným zápisem jsou tak základem řešení efektivních a výkonných plánovacích aplikací, které přinášejí ***několik podstatných efektů***:

- ***využití již existujících hodnot*** dříve vytvořených plánů, nebo hodnot o skutečnosti z minulých období, které se tak stávají základem pro tvorbu nových plánů,
- ***pružné zapracování změn plánů***, kdy je třeba promítnout změny v celé plánovací struktuře, např. pokud se změní plán v jednom oddělení, je nutné měnit i konsolidované plány na vyšších úrovních podnikového řízení, a naopak kdy je nutné promítat změny z vyšších organizačních jednotek na nižší,
- ***promítání změn zpět do zdrojových databází*** transakčních aplikací lze efektivně využívat k distribuci aktuálních plánů do aplikací nebo databází podřízených útvarů, divizí, jednotlivých poboček apod.,
- při využití efektivních ***rozvrhových algoritmů aplikovaných na struktury jednotlivých dimenzí*** a jejich kombinace (např. procentuálních nebo obdobných rozvrhových schémat) lze podnikové plány sestavovat podstatně efektivněji a zajistit jejich konsistenci,
- je možné definovat a využívat tzv. ***podniková pravidla*** (*business rules*) nebo omezení (*business constraints*), tj. pravidla pro kontroly plánů, např. objem investic nepřesáhne 20 % plánovaných nákladů celého podniku apod.,
- podnikové plány jsou ***centrálně uložené v analytické databázi***, a to znamená jejich rychlé a konzistentní vytváření s rozpadem, či agregací plánovacích ukazatelů dle definovaných dimenzí a možnost rychlého porovnání skutečnosti s plánem, pokud je použita stejná struktura ukazatelů a dimenzí.

1.2.3.3 Plánovací úlohy v kontextu řízení firmy

Plánovací úlohy jsou rovněž ***součástí většiny oblastí podnikového řízení*** i celého podniku. Na základě definovaných principů lze určit jejich ***uplatnění v oblastech řízení firmy***, na tomto místě na obecné úrovni s tím, že je pak konkretizováno dle oblastí řízení (v dokumentu „*AF.II.01: Oblasti řízení*“). ***Pracovně*** lze přijmout následující ***klasifikaci***:

- ***Sestavení plánu*** za vybranou oblast ***podle vybraných ukazatelů*** a s využitím relevantních dimenzí, např. plán tržeb v daném období, plán obejmu nákupů vybraných materiálů, plán počtu servisních zásahů na dodaných zboží apod.
- ***Zpracování výhledu podle vybraných ukazatelů a s využitím relevantních dimenzí***, např. na 12 měsíců klouzavě v průběhu roku.
- ***Sestavení rozpočtu,*** tedy finančních hodnot ukazatelů, které silně závisí na prostředí konkrétní organizace, charakteristice trhu, druhu nabízeného produktu a dalších

1.2.3.4 Scénáře, analytické otázky

Vztah k byznysu:

- Jak zvýšit ***úspěšnost a výkonnost*** byznysu díky kvalitě a využití plánovacích úloh?
- Jak specifikovat ***potenciální efekty*** užití plánovacích úloh, a to ***ekonomické i mimoekonomické***?
- Které ***oblasti řízení*** jsou a budou z hlediska uplatnění plánovacích úloh ***prioritní***?
- Které ***metriky*** budou pro plánovací úlohy významné, které budou mít charakter KPI, které ***dimenze*** ve vztahu k metrikám budou relevantní?
- Které ***faktory*** budou pro řešení a uplatnění plánovacích úloh ***rozhodující***?
- Jak zajistit požadovanou ***komplexnost a kvalitu*** plánovacích úloh?
- Jak dosáhnout požadované ***flexibility*** plánovacích úloh vzhledem k momentálním potřebám manažerů a specialistů firmy?
- Jak zajistit kvalitní a systematické ***shromažďování podkladů*** z jednotlivých útvarů firmy pro přípravu plánů?
- Jak nastavit ***efektivní postupy projednávání*** a schvalování připravovaných plánů?
- Jaké ***vlivy*** má úroveň a využití plánovacích úloh ***na řízení a organizaci firmy***, jaké jsou hlavní problémy nebo omezení?

Vztah k datovým zdrojům:

- Jaký je aktuální ***obsah datových zdrojů*** vzhledem k potřebám plánovacích úloh?
- Jaká je ***kvalita datových zdrojů*** vzhledem k plánovacím úlohám?
- Jak dosáhnout požadovanou ***granularitu dat*** v rámci jednotlivých operací plánovacích úloh? Jak k tomu zajistit ***potřebné datové zdroje*** na požadované úrovni granularity?

Vztah k personálu:

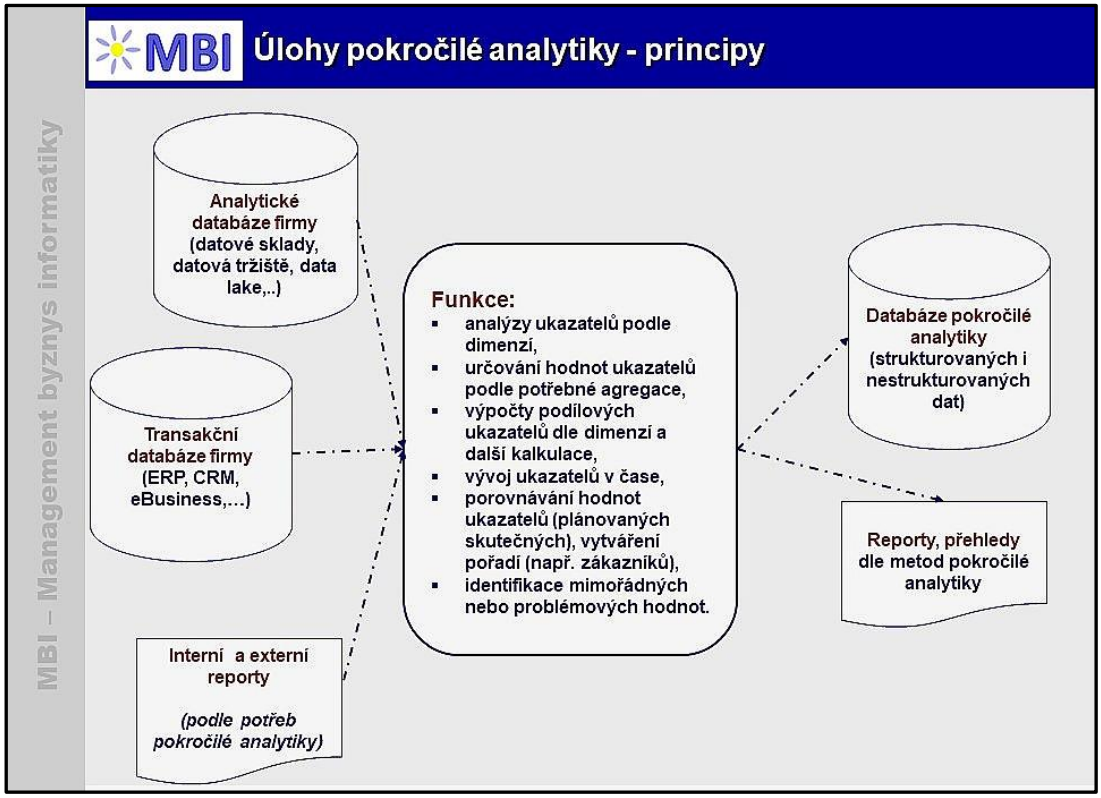
- Jak dosáhnout potřebné ***kvalifikace a motivace*** manažerů a specialistů v oblasti využití a případně i řešení plánovacích úloh s využitím nových technologií?
- Jak ***připravovat pracovníky*** na uplatnění nových technologií a aplikací v plánování firmy?

Vztah k IT:

- Jaké ***IT produkty,*** např. standardní kancelářské, business intelligence a další bude účelné v plánování vzhledem k potřebám i možnostem firmy využít?
- Je k dispozici ***podpora*** plánovacích úloh ***v provozovaných transakčních aplikacích*** (ERP, CRM,…), jaké jsou možnosti a omezení jejich využití?
- Jak dosahovat ***zkracování doby a časové náročnosti*** na přípravu plánů?
- Jak využít dostupná ***cloudová řešení*** na druhé straně ***on premis***, případně jejich kombinace v realizaci plánovacích úloh?

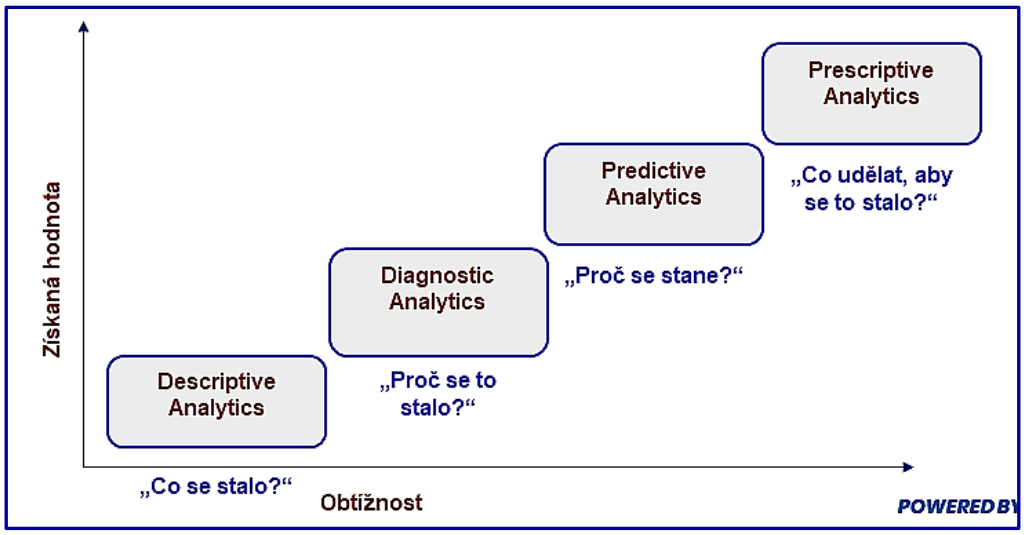
1.2.4 Pokročilá analytika

Principy úloh pokročilé analytiky dokumentuje další obrázek.



Obrázek 1-5: Principy úloh pokročilé analytiky

Pokročilá analytika představuje ***soustavu konceptů, přístupů, metod a produktů***, jejichž společnou charakteristikou je sofistikovanost analytických a plánovacích funkcí. Přesto je účelné ji zasadit do celého širšího komplexu metod a nástrojů podnikové byznys analytiky, jejíž rozmanitost je v současné době mimořádná. Jako východisko charakteristik pokročilé analytiky je účelné využít schéma společnosti Gartner (Obrázek 1-6):



Obrázek 1-6: Úrovně řešení pokročilé analytiky (Zdroj: Gartner Analytic Ascendency Model)

1.2.4.1 Scénáře, analytické otázky

Vztah k byznysu:

- Jak zvýšit ***úspěšnost a výkonnost*** byznysu díky využití pokročilé analytiky?

- Jak specifikovat ***potenciální efekty*** užití pokročilé analytiky, a to ***ekonomické i mimoekonomické***?
- Které ***oblasti řízení*** jsou a budou z hlediska uplatnění pokročilé analytiky ***prioritní***?
- Které ***faktory*** budou pro řešení a uplatnění analytických úloh ***rozhodující***?
- Jak zajistit požadovanou ***komplexnost a kvalitu*** pokročilé analytiky?
- Jaké ***vlivy*** má úroveň a využití pokročilé analytiky ***na řízení a organizaci firmy***, jaké jsou hlavní problémy nebo omezení?

Vztah k datovým zdrojům:

- Jaký je aktuální ***obsah datových zdrojů*** vzhledem k potřebám pokročilé analytiky?
- Jaká je ***kvalita datových zdrojů*** vzhledem k pokročilé analytice, jak řešit problémy s nekvalitou dat?
- Jak správně stanovit ***perspektivu datových zdrojů*** a příslušných transakčních aplikací vzhledem k plánovanému rozvoji pokročilé analytiky?

Vztah k personálu:

- Jak dosáhnout potřebné ***kvalifikace a motivace*** manažerů a specialistů v oblasti využití a případně i řešení úloh pokročilé analytiky?
- Jak dosáhnout ***konsensu mezi pracovníky*** firmy na navrženém obsahu a strukturách úloh pokročilé analytiky?

Vztah k IT:

- Jaké ***IT produkty, metody, modely*** bude účelné vzhledem k potřebám i možnostem firmy využít?
- Jak dosahovat ***zkracování doby a časové náročnosti*** na přípravu úloh pokročilé analytiky a aplikací?

1.3 Metriky, dimenze

Předmětem řešení v úlohách podnikové analytiky jsou zejména metriky reprezentované ***ukazateli, odpovídajícími dimenzemi a datovými zdroji***. Každá metrika je ***dokumentována*** standardně v těchto bodech:

- obsahové vymezení metriky, resp. ukazatele, její využití i s případným kalkulačním předpisem,
- související metriky, odvozené od základního vymezení, kde jde o pracovní vytváření jejich skupin s ohledem na jejich velký počet, a tedy i zajištění větší přehlednosti celého textu,
- přehled analytických dimenzí, které se v souvislosti s metrikou obvykle využívají,
- datové zdroje pro metriku, obvykle databáze nebo jejich části nebo speciální dokumenty a výkazy.

Z hlediska obsahu metrik je účelné v jednotlivých oblastech řízení sledovat:

- metriky ***finančního, resp. ekonomického*** charakteru (objem nákladů, výnosů, komplexní finanční metriky, jako cash flow, úroveň zadluženosti firmy apod., objem majetku, objem odpisů, počet a objem reklamací apod.),
- metriky ***výkonového, nebo procesního*** charakteru (počet zpracovávaných daňových dokladů, počet prodejních nebo nákupních transakcí, počet vyřízených reklamací, počet realizovaných školení apod.),
- metriky ***organizačního charakteru*** (vytížení kapacit, počet pracovníků, útvarů, počet zákazníků, dodavatelů apod.),
- metriky ***kvalitativního charakteru***, jako sledovaná úroveň kvality produktů a služeb, dodací lhůty, kvality dodávaných materiálů atd.

Z hlediska dopadů metrik do systému řízení existují jejich 3 typy – ***KPI, KRI a KGI***.

Pro správné určení ***KPI ukazatele*** je vhodné vypsat ***seznam kritických faktorů úspěchů*** podniku, který úzce souvisí s KPI. ***KPI (Key Performance Indicator)***, klíčový ukazatel výkonnosti splňuje zpravidla tyto body:

- je měřitelný velice často, zpravidla denně,
- musí mu rozumět všichni zaměstnanci,
- má významný vliv na organizaci,
- není finanční povahy,
- má zásadní vliv na výkonnost podniku,
- ukazuje, jak dobře fungují klíčové procesy podniku,
- v podniku existuje maximálně 10 KPI ukazatelů.

U každého ***KPI*** ukazatele je především podstatný ***potřebný zdroj dat, granularita, jeho výpočet a odchylka***:

- ***Zdroje dat*** identifikují, kde se nachází potřebné ***informace pro daný typ KPI*** ukazatele. Samotný způsob výpočtu KPI ukazatele se může počítat z více datových zdrojů, proto je důležité hodnotit analytické nástroje podle nezávislosti a existence velkého množství konektorů na různé typy datových zdrojů.
- ***Granularita***, která ***určuje detail výpočtu pro KPI***. Jednoduše si to lze představit u zjišťování KPI za celý stát, nebo i města, obce nebo ve vazbě na čas, kdy KPI může mít souhrnnou hodnotu za celý rok, nebo měsíc, týden či den.
- ***Výpočet KPI*** je většinou vyjádřen ***v procentech***, nebo jako součet či průměr, mluvíme-li o ***jednom datovém zdroji***. Je-li zdrojem pro výpočet KPI potřeba ***více datových zdrojů***, je nutné počítat ***s podmíněnými výpočty*** a v některých případech i s různou kvalitou dat.
- ***Výpočet změny oproti určitému časovému období***, například porovnání tržeb ve stejné pobočce oproti minulým obdobím, porovnání zisků za různá období, plnění plánu výroby oproti minulému čtvrtletí atd.

KRI (Key Result Indicator) je ***výstup aktivit*** podniku. Sděluje managementu, jak je podnik úspěšný, ale už nelze zjistit směr, jakým způsobem dále zlepšit výsledek.

KGI (Key Goal Indicator) ukazuje, jakým způsobem se firmě daří dosahovat výsledků a cílů a je provázán se strategií podniku.

Dokument „*AF II.01. IT a anatomie firmy – Oblasti řízení*“ určuje ***využití metrik pro jednotlivé oblasti*** řízení a jsou v rámci tohoto dokumentu ***standardně umístěny v podkapitole x.2***, kde x je číslo kapitoly, např. „*7.2 Metriky řízení prodeje*“.

→ Vymezení ***obsahu všech metrik*** zařazených do anatomie firmy obsahuje dokument „*AF II.02: Komponenty*“, ***kapitola 2***. Jsou uspořádány podle oblastí řízení, včetně řízení IT, ***kapitola 2.15***.

Charakteristiky obsahu základních ***metrik odvětvových řešení*** (např. tržby, zisk, počet zákazníků apod.) obsahuje rovněž dokument „*AF II.02: Komponenty*“, ***specifické metriky pro dané odvětví*** obsahuje příslušný dokument úrovně III, např. „*AF III 01: IT firma*“.

Celkový přehled a ***vymezení obsahu*** analytických dimenzí napříč oblastmi řízení poskytuje rovněž dokument „*AF II.02: Komponenty*“. Základní struktura dokumentace dimenze je zde následující:

- vymezení obsahu a účelu dimenze, např. dimenze „*Zboží*“,
- vnitřní struktura dimenze, např. „*Kategorie zboží – Skupiny zboží – Zboží*“,
- datové zdroje pro dimenzi.

→ ***Přehled a obsah všech dimenzí*** zařazených do anatomie firmy obsahuje dokument „*AF II.02: Komponenty*“, ***kapitola 3***. Jsou uspořádány ***do logicky souvisejících skupin***, jako např. „*Obchodní dimenze*“, *Ekonomické dimenze*“ apod.

1.4 Data a datové zdroje

Pro analýzy a užití metrik jsou podstatné charakteristiky odpovídajících datových zdrojů. Každý datový zdroj je ***dokumentován*** jejich obsahem, účelem a alespoň vybranými nebo hlavními atributy. Ve vztahu k řešení podnikové analytiky je ale s data a datovými zdroji spojena řada témat a metod, z nichž se zde ve stručnosti zaměříme pouze na následující:

- řízení a správa metadat,
- zajištění kvality a čištění dat,
- MDM, Master Data Management,
- Data Governance.

1.4.1 Řízení a správa metadat

1.4.2 Správa metadat

Metadata jsou ve své podstatě ***strukturovaná data o datech***. Metadata představují údaje ***nejen o samotných datech***, ale také o technických prostředcích, softwaru, nebo sítích, kde se data nacházejí. Specifikují jejich ***kontext, obsah, předpokládanou interpretaci a dostupnost dat***. Hlavní ***účelem*** metadat je ***poskytování informací k analýze***, návrhu, vývoji, implementaci a užití jednotlivých aplikací i celé podnikové analytiky. V souvislosti s řešeními projektů je ***obsah metadat*** např. následující:

- celkový ***popis zdrojových systémů***,
- u data staging area to jsou popisy dat ***ve slovníku datového skladu***,
- u datového skladu ***popisy transformačních pravidel*** pro každou tabulku a každý datový element a popisy business názvů a transformačních pravidel pro každou tabulku a každý datový element,
- ***pro reporting*** vysvětlení každého pole na reportu.

Podstata metadat je tedy zřejmá, jejich uplatnění jako faktor v řešeních podnikové analytiky je dáno několika důvody:

- řešení podnikové analytiky ***se vztahují převážně na celý podnik***, jsou proto velmi komplexní, rozsáhlá a komplikovaná. Uspořádané, jasně strukturované a dostupné informace o tom, co tato řešení obsahují, jaké datové struktury, v jakých vazbách apod. jsou ***při této složitosti nutnou podmínkou realizace*** těchto projektů;
- s rozsahem řešení analytiky ***roste i rozsah jejich metadat***. Pro efektivní zajištění projektů i provozu aplikací se využívají ***celé systémy pro správu metadat***, tedy databáze metadat s příslušnou aplikační nadstavbou pro práci s nimi;
- jako jeden z efektů podnikové analytiky, vedle své analytické a plánovací funkcionality, se běžně zdůrazňuje i jejich ***úloha ve zvyšování pořádku*** (např. čistoty dat) v celém informačním systému. K uskutečnění této úlohy je nutné ***disponovat dokonalým přehledem a evidencí*** o stávajících datových a dalších zdrojích podnikové informatiky a takovou evidenci nabízejí metadata.

Proto se otázkám metadat věnuje ve velkých společnostech určení specialisté a na jejich využití je postavena naprostá většina projektů.

1.4.3 Zajištění kvality a čištění dat

Řešení kvality data představuje celý ***komplex činností, procesů a metod***, které se v rámci systému ***řízení kvality*** dat realizují. Systém řízení kvality dat (***DQMS, Data Management Quality System***) je

obvykle vymezen jako „celkový úhrn aktivit organizace, sloužících k zajištění kvality dat“ a ***zahrnuje*** tyto hlavní funkce:

- opravu chyb a jiných nedostatků v datech hlášených organizaci zákazníkem a jinými subjekty,
- periodické čištění datových zdrojů a databází,
- čištění aktuálně používaných dat,
- prevenci chyb.

Kvalitní nastavení řízení kvality dat přináší ***nejrůznější ekonomické efekty***, např. v úsporách za ztrát zákazníků vzhledem k nekvalitně poskytovaných informací, ve ***zvýšení důvěryhodnosti firmy*** vzhledem k zákazníkům. Podstatným momentem ve využití řízení datové kvality je ***snížení časových ztrát při dohledávání chyb*** v rámci podnikového reportingu.

Na druhé straně realizace kontrolních a opravných operací na datech představují i často ***značnou časovou náročnost (až 80 % celkového času řešení) i zvýšené náklady***. S tím je nutné počítat, ale pro úspěch podnikové analytiky jsou právě tyto aktivity klíčovým předpokladem.

1.4.4 Master Data Management

Master data jsou vysoce ceněná a klíčová data používaná ***k podpoře kritických podnikových procesů*** napříč celým podnikem. Tato data jsou důležitá v každé obchodní transakci, aplikaci, reportu a rozhodovacím procesu,

Master Data Management představuje ***sadu procesů, přístupů, metodologií, nástrojů, technologií*** a řešení, které pomáhají organizacím jednoznačně identifikovat a integrovat celopodnikově sdílená klíčová data. Jako příklad mohou sloužit data zákaznická, data o produktech, službách či lokalitách a další. Obdobné řešení a přístup představuje ***Reference Data Management (RDM)*** - vztahující se k určitým organizačním jednotkám.

MDM zajišťuje i ***správu významu – sémantiky*** – klíčových dat a také správu vztahů mezi daty klíčovými a ostatními.

V rámci MDM se rozlišuje ***provozní a analytické MDM***:

- ***provozní MDM*** se soustřeďuje na zajištění „jednotného pohledu“ na master data v hlavních transakčních systémech, tj. na místě, kde jsou master data poprvé vytvářena,
- ***analytické MDM*** se orientuje na zajištění „jednotného pohledu“ na master data v datovém skladu.

Podstatnými přínosy MDM jsou např.:

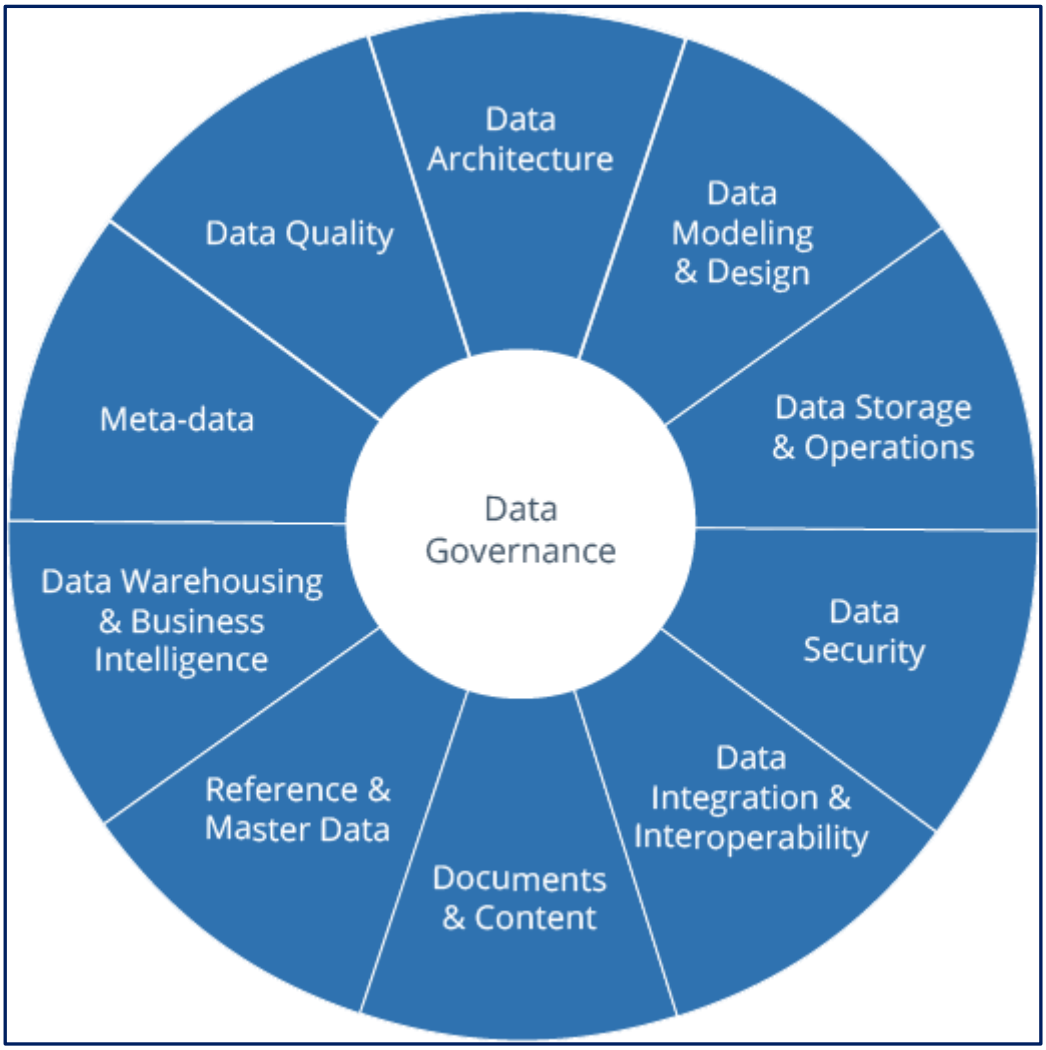
- výrazné ***zvýšení kvality klíčových dat, jejich konzistence***, dostupnosti a důvěryhodnosti. Konzistence dat podporovaná MDM snižuje nedůvěru k datům a zefektivňuje rozhodovací procesy,
- ***poskytování komplexních informací o zákazníkovi***, tj. úložiště master dat pro zákaznická data poskytuje jediný zdroj pro konsolidaci všech zákaznických aktivit v konzistentní formě a využívá se v transakčních i analytických aplikacích,
- ***zkvalitnění zákaznických služeb*** – díky dostupnosti informací a schopnosti reagovat na poptávky, objednávky, dotazy atd.,
- ***snížení nákladů na podnikovou analytiku*** – snižuje se objem replikací stejných dat, a tedy činností a nákladů s nimi spojených (licenční náklady, náklady na infrastrukturu apod.).

Předpokladem úspěchu je ale efektivní ***systém motivace*** mezi pracovníky uživatelské sféry i IT, jasně deklarovaná ***pravidla*** pro hodnocení kvality dat a zejména ***efektivní začlenění MDM*** do řízení datových zdrojů v rámci řízení IT, a to se všemi podstatnými souvislostmi, včetně ekonomických.

1.4.5 Data Governance

Společnost ***Profisee*** uvádí k Data governance následující ***definici***: „*Data governance je sadou principů a pokynů zaručujících vysokou kvalitu dat po celou dobu jejich životního cyklu.*“. Data governance je tedy ***programem zaručující patřičnou kvalitu dat*** v celé organizaci. Aby došlo ke skutečnému posunu při práci s daty a jejich ***kvalitou či bezpečností***, tak je důležité, aby jednotliví zaměstnanci dodržovali jasně nastavená pravidla při jakékoliv práci s daty. (Profisee, 2020).

Kromě kvality dat postihuje data governance ještě ***další oblasti spojené s daty***. Základní oblastí je ***bezpečnost***. Další oblasti jsou zaměřené na ***architekturu, integraci*** nebo například na ***datové sklady***.



Obrázek 1-7: Oblasti postihované data governance (Zdroj: JANOSCHEK, 2019)

Na obrázku (Obrázek 1.2) je uvedeno, kolik oblastí data governance pokrývá. Několik alespoň ***základních bodů spojených a data governance*** je v následujícím přehledu:

- určení ***jednoznačných pravidel*** týkající se metadat, bezpečnosti dat, ale také požadované kvality a přístupu k datům,
- určení ***majitelů a osob zodpovědných za data***. Data jsou využívána na různých úrovních a každé oddělení užívající nebo produkující data by se mělo podílet na jejich správnosti a nést určitou zodpovědnost za tato data. Nemělo by se tedy stát, že majitelem dat bude pouze jedno oddělení,

- ***specifikace kontroly dat***, určení časového úseku, kdy se budou data testovat, zda data odpovídají požadované kvalitě, případně jak budou probíhat změny v datech a zápisy o těchto změnách,
- ***realizace datových auditů*** a zjišťování, na jaké úrovni se data nachází a jak je dodržována nastavená politika.

1.5 Závěry k obsahu podnikové analytiky

- Podniková analytika představuje ***komplex zejména analytických a plánovacích úloh*** v řízení firmy a lze ji pracovně ***rozlišit na základní*** (business intelligence a další) ***a pokročilou*** (data mining, prediktivní analytiku apod.).
- ***Řešení úloh a aplikací*** podnikové analytiky ***se podstatně liší od transakčních aplikací***, způsobem ***práce s daty***, ***použitými metodami***, ***účastí uživatelské sféry*** na řešení.
- Kvalitně připravené aplikace podnikové analytiky přispívají nejen ke ***zvyšování kvality řízení a konkurenceschopnosti firmy***, ale svojí povahou často znamenají i ***přínos v pozitivních posunech manažerské kvalifikace.***
- ***Znalost obsahu řízení*** firmy je pro řešení úloh podnikové analytiky ***klíčová***, a to ve všech jeho uvedených součástech (úlohy, metriky, dimenze, data).
- ***Základem řešení*** úloh podnikové analytiky je ***návrh a realizace analytických a plánovacích funkcí nad systémem metrik*** firmy, které vyjadřují jednotlivé sledované ukazatele a jim odpovídající analytické dimenze.
- ***Nejvýznamnější metriky*** pro řízení firmy se charakterizují jako ***KPI (Key Performance Indicator)*** a jsou hlavní náplní manažerských reportů a dashboardů.
- S metrikami souvisí i ***komplex otázek a problémů datových zdrojů***, z nichž jsou tyto metriky čerpány.
- Klíčovou otázkou datových zdrojů je ***zajištění jejich kvality,*** pro které je k dispozici celá řada metod a nástrojů.
- ***Obsah řízení a podnikové analytiky*** je samozřejmě ***velmi variantní*** podle odvětví, typu a zaměření firem a jeho zvládnutí je pro analytiky hodně náročné. Přesto je účelné disponovat potřebnou orientací v tomto směru a tu konkretizovat podle konkrétních podmínek dané firmy.

2. Role v podnikové analytice

Účelem kapitoly je:

- vymezit ***funkční náplň*** a hlavní požadované znalosti pracovníků v oblasti podnikové analytiky, zejména manažerů, analytiků, vývojářů.,
- nabídnout čtenáři alespoň ***základ pro jejich další konkretizaci*** podle podmínek vlastní praxe.

Účelem kapitoly je:

- vymezit ***funkční náplň*** pracovníků v oblasti podnikové analytiky, zejména manažerů, analytiků, vývojářů.,
- nabídnout čtenáři alespoň ***základ rolí v podnikové analytice pro jejich další konkretizaci*** podle podmínek vlastní praxe.

Role určují, jaké ***funkce a kompetence*** bude mít pracovník v řešení problémů v určité specifické oblasti podnikové analytiky. ***Smyslem rolí*** je zde přesněji vyjádřit, jak jsou, nebo mají být, úlohy v řízení a řešeni podnikové analytiky personálně zajištěny, resp. jaká je personální náročnost na realizaci jednotlivých úloh. Vedle toho je dobře dokumentovaný systém rolí solidním ***základem pro přípravu a realizaci kvalifikačních programů a školení*** ve vazbě na IT projekty a v širším kontextu i ve vazbě na řízení celé firmy.

Do kapitoly jsou vzhledem k řešení podnikové analytiky zařazeny tyto ***role***:

- ***Uživatelské role***:
  - ***Vlastník***, resp. člen správní rady nebo dozorčí rady, většinou posuzuje a rozhoduje o investicích do IT včetně podnikové analytiky.
  - ***Manažer*** představuje v našem případě společnou roli pro různé typy manažerů podle jejich působnosti.
  - ***Klíčový uživatel*** podnikové analytiky („*Power User*“) definuje požadavky na obsah a funkcionalitu úloh podnikové analytiky a současně s tím se podílí na analytické přípravě řešených úloh v rámci projektů.
  - ***Uživatel služeb a aplikací*** podnikové analytiky využívá disponibilní analytické a plánovací aplikace, kvalifikovaně interpretuje dodávané výsledky v rámci řízení dané oblasti.
- ***IT role:***
  - ***Informační manažer*** zodpovídá za začlenění aplikací podnikové analytiky do celé aplikačního portfolia a za jejich efektivní provoz.
  - ***Sponzor projektu***, pracovník odpovídající za celý projekt a jeho výsledky. Jde většinou o osobnost, která je schopná řešení a uplatnění analytiky ve firmě prosadit.
  - ***Manažer projektu*** řídí projekt a projektový tým v celém průběhu jeho řešení, zajišťuje komunikaci a kooperaci jak s vedením firmy, tak s externími partnery a dodavateli.
  - ***Architekt podnikové analytiky*** řeší návrhy IT infrastruktury a technologickou realizaci aplikací, vytváří potřebnou aplikační a technologickou architekturu podnikové analytiky.
  - ***Byznys analytik*** zajišťuje řešení všech aspektů obsahové náplně úloh v rámci projektu, komunikuje a kooperuje s vedením projektu a ostatními členy analytických týmů.
  - ***Datový analytik*** zajišťuje mapování dat ve zdrojových databázích a jejich transformace do analytických databází, podílí se na řešení potřebné kvality dat a jejich konsolidaci.
  - ***Vývojář softwaru*** pro podnikovou analytiku vyhodnocuje požadavky na analytické a plánovací aplikace, na jejich základě je navrhuje a zajišťuje jejich technologickou realizaci a dokumentaci.

- ***Dodavatel*** v případě řešení externími kapacitami zajišťuje veškeré činnosti řešení projektu a následnou podporu podle uzavřené smlouvy.

Vedle již zmíněného účelu rolí vzhledem k řízení firmy a k řešení podnikové analytiky je dobré uvést i ***podstatný předpoklad*** jejich úspěšného užití, a tím je jejich ***průběžná aktualizace***. Celková funkcionalita, architektura i použité technologie podnikové analytiky ***se s vývojem relativně často mění***. A s tím je účelné upravovat i náplň jednotlivých rolí, jejich kompetence a nároky na kvalifikační přípravu. Úspěšnost aplikací podnikové analytiky je podstatně více, než je tomu u jiných typů aplikací, ***závislá na kvalitě, motivaci a znalostech lidí,*** které je realizují a následně využívají. Proto je jejich příprava jedním z klíčových faktorů úspěchu a systém dobře dokumentovaných rolí může v tomto směru dobře napomáhat. V dalších podkapitolách jsou uvedeny ***podstatné funkční charakteristiky vybraných rolí*** podnikové analytiky:

2.1 Uživatelské role

Rolí, které se podílejí na přípravě a užití podnikové analytiky, je celá řada a s ohledem na již výše zmíněnou individualizaci řešení je jejich vymezení nezbytné věnovat intenzivní pozornost. Bylo by samozřejmě možné podrobněji rozlišovat i role na uživatelské straně, ale s ohledem na rozsah textu se omezujeme na ***charakteristiky vlastníka, manažera, uživatele a klíčového uživatele***, které je už možné promítat a upravovat podle jednotlivých konkrétních uživatelských rolí.

2.1.1 Vlastník

Vlastník firmy, resp. člen některého z řídících orgánů, jehož ***vztah k řešení podnikové analytiky*** se ***výrazně liší*** podle toho, zda jde o velkou nebo malou firmu, zda jde o českou firmu nebo součást některé nadnárodní korporace, kultura firmy, kterou vlastní a ovlivňuje je nakloněná k řešení a užívání podnikové analytiky apod. Obvykle zajišťuje tyto ***funkce***:

- posuzování, případně ***formulace celkové koncepce rozvoje*** firmy včetně uplatňování možností podnikové analytiky,
- monitorování a ***hodnocení výkonnosti podniku a jeho managementu*** i s využitím aplikací podnikové analytiky,
- posuzování a ***schvalování investic*** do podnikové analytiky a posuzování efektivnosti těchto investic,
- výběr nebo ***schvalování řídících pracovníků*** ve vazbě na aktivity spojené s rozvojem podnikové analytiky,
- využívání specifických ***aplikací pro prognózování a plánování*** rozvoje firmy.

2.1.2 Manažer

Rolí manažera jsou zde chápany ***zobecněné charakteristiky pro různé typy a úrovně manažerů***, jako např. generálního manažera (CEO), finančního manažera (CFO), marketingového manažera (CMO) a dalších. ***Jejich funkce*** vzhledem k podnikové analytice jsou obvykle tyto:

- určování ***cílů, strategie a politiky v rozvoji podnikové analytiky*** vzhledem k aktuálním i budoucím potřebám firmy,
- podíl na ***plánování, řízení a koordinaci jednotlivých aktivit*** spojených s podnikovou analytikou,
- monitorování a ***hodnocení výkonu pracovníků*** podle jednotlivých oblastí řízení v souvislosti s účastí na projektech podnikové analytiky,
- ***schvalování rozpočtů, kontrolování výdajů*** a zajišťování účelného využívání finančních zdrojů určených pro podnikovou analytiku,
- ***řízení spolupráce mezi útvary*** firmy i s externími partnery na rozvoji a užití podnikové analytiky,
- ***návrh případných organizačních změn*** ve firmě v souvislosti s podnikovou analytikou, např. vznik kompetenčních center apod.,

- řešení ***transformace byznys modelů a byznys procesů*** v souvislosti s výsledky aplikací podnikové analytiky.

2.1.3 Klíčový uživatel, Power User

Klíčový uživatel určuje ***základní nároky a požadavky na realizaci*** aplikací podnikové analytiky ve firmě. Podílí se na analytických činnostech v oblasti úloh podnikové analytiky. Podílí se podle jednotlivých fází a úloh řešení projektů na těchto ***činnostech****:*

- zajišťuje ***hodnocení funkcionality zdrojových aplikací*** a IT služeb vzhledem k vývoji analytických a plánovacích potřeb firmy, požadavků partnerů i vývoji v legislativě,
- ***konzultuje požadavky*** na řešení podnikové analytiky v rámci své oblasti řízení, a to z pohledu potřeb firmy i aktuální legislativy,
- definuje ***specifikace sledovaných ukazatelů***, struktury a další charakteristiky dimenzí a navrhuje podstatné ***vazby ukazatelů a dimenzí***,
- definuje nároky na ***obsah a strukturu hlavních reportů***, na jednotlivé sledované ukazatele a jim odpovídající dimenze, na hlavní analytické a plánovací funkce,
- určuje ***charakter složitějších analytických operací***, data-miningu, analýzy nestrukturovaných dat a dalších pokročilých technik datové analýzy,
- konzultuje ***požadavky na funkcionalitu*** analytických a plánovacích aplikací,
- ***analyzuje definované požadavky uživatelů***, resp. celé uživatelské sféry vzhledem k podnikovým metodikám, směrnicím a standardům,
- poskytuje informace a zkušenosti o ***stavu a kvalitě relevantních zdrojových databází***,
- formuluje ***zadání nových analytický a plánovacích aplikací*** a reportů, specifikuje nároky na jejich kvalitu, tj. funkcionalitu, dostupnost, uživatelské rozhraní apod.,
- konzultuje analytické otázky ***transformací dat***,
- podílí se na řešení ***problémů spojených s kvalitou dat***, definuje potřebné kontroly dat a podílí se na úpravách dat pro podnikovou analytiku,
- spolupracuje na přípravě a ***vytvoření databází*** datového skladu a datových tržišť z existujících datových zdrojů,
- podílí se se ***kvalifikační přípravě*** všech ostatních uživatelů.

2.1.4 Uživatel služeb a aplikací podnikové analytiky

Uživatelem služeb podnikové analytiky se rozumí ***nejen manažeři***, jimž byly věnována předchozí část, ale i další ***pracovníci převážně na střední úrovni řízení*** nebo ***specialisté firmy***, např. finanční analytici, obchodní a marketingoví specialisté a další. V tomto kontextu se jimi rozumí pracovníci, kteří na rozdíl od klíčových uživatelů jsou pasivními uživateli podnikové analytiky. Realizují tyto základní ***funkce***:

- kvalifikovaně ***využívá realizované analytické a plánovací aplikace*** a výstupy reportingu,
- ***konzultuje specifické problémy a otázky*** analytických a plánovacích aplikací v rámci své působnosti,
- formuluje dílčí ***nové požadavky*** na aplikace a technologie podnikové analytiky, vzhledem k potřebám podniku i ke změnám v legislativě,
- ***vyhodnocuje vlastní způsob využití*** provozovaných aplikací, jejich dostupnost, úroveň komunikace, kvalitu vizualizace a další.

2.2 Role IT

Skupina rolí IT ve vazbě na podnikovou analytiku zahrnuje hlavní ***manažerské role, role analytiků a vývoje.***

2.2.1 Informační manažer (CIO)

Informační manažer podílí se na plánování úloh a aplikací podnikové analytiky, zajišťuje ***průběžnou kontrolu*** nad průběhem projektů podnikové analytiky a podílí se na ***řešení zásadních problémů*** v jejich rámci. Monitoruje ***využití jednotlivých analytických a plánovacích aplikací*** a navrhuje případné provozní nebo koncepční změny. Uskutečňuje tyto hlavní ***činnosti***:

- ***formulace a realizace strategie*** v oblasti podnikové analytiky, zejména definování jejích cílů v souladu s celopodnikovými cíli, určení způsobu a termínů splnění navržených cílů, metrik splnění cílů, zodpovědnosti za splnění cílů,
- určení ***strategických úloh řízení firmy*** ve vazbě na aplikace podnikové analytiky,
- ***zajištění souladu*** řešení podnikové analytiky s podnikovou architekturou, s IT architekturou podniku, vývojovými trendy na trhu i s aktuálními potřebami firmy,
- uplatňování ***strategie sourcingu*** při přípravě i realizaci projektů podnikové analytiky,
- ***vyhodnocování nákladů a přínosů*** podnikové analytiky, jejich možná kvantifikace a prezentace managementu firmy,
- formulace ***pravidel řízení rozvoje a provozu*** podnikové analytiky, změny organizačních struktur v relaci na rozvoj podnikové analytiky,
- ***řešení vztahů k externím dodavatelům*** produktů a služeb podnikové analytiky, schvalování smluv s dodavateli, koordinace smluv s interními i externími zákazníky a dodavateli IT,
- ***řešení vazeb útvaru informatiky na ostatní útvary*** firmy v relaci k podnikové analytice,
- ***plánování a vyhodnocování investic*** do podnikové analytiky,
- dohled nad řešením ***operativních úkolů*** souvisejících s řízením rozvoje a provozu podnikové analytiky.

2.2.2 Sponzor projektu podnikové analytiky

Sponzor projektu ***zodpovídá za celý projekt*** a dosažení jeho očekávaných ***efektů. Je jmenován vlastníky*** společnosti, nebo jejím nejvyšším vedením (CEO). Obvykle představuje ***osobnost*****,** která nejen že je schopná zajistit finanční prostředky, ale i ***přesvědčit a získat pracovníky firmy*** na úrovni manažerů a podnikových specialistů pro řešení, a především i následné využití projektu. Je tak schopen projekt prosadit a postarat se o jeho pozitivní přijetí. Je obvykle *členem vrcholového vedení*, se silnou pozicí ve společnosti. Zajišťuje zejména tyto ***činnosti***:

- zajišťuje ***komunikaci a kooperaci s vedením firmy*** v kontextu daného projektu,
- s vedením podniku spoluurčuje celkovou ***strategii v oblasti podnikové analytiky***, které typy aplikací vzhledem k podnikovým potřebám budou prioritní,
- je obvykle ***představitelem řídící komise projektu***, má hlavní rozhodovací pravomoc,
- ***jmenuje*** členy řídící komise, resp. řídícího týmu, včetně manažera projektu,
- zodpovídá za to, že ***projekt je přijatelný z pohledu relace cena / výkon*** a za naplnění cílů projektu a očekávaných ekonomických i mimoekonomických efektů,
- získává potřebné ***finanční a lidské zdroje*** a případně se účastní na získání potřebných technických prostředků,
- při řešení koncepčních a organizačních úkolů projektu ***kooperuje převážně*** s informačním manažerem, manažerem projektu podnikové analytiky a dodavateli.

2.2.3 Manažer projektu

Manažer projektu podnikové analytiky ***je jmenován*** sponzorem projektu, resp. vedením firmy. Jeho ***hlavním úkolem*** je řídit projekt a projektový tým, tak aby dosáhl ***potřebné kvality a očekávaných efektů*** z pohledu vlastníků, vedení firmy, pracovníků firmy, případně i externích partnerů. Role manažera projektu v podnikové analytice představuje zejména tyto ***činnosti***:

- kooperuje a ***komunikuje se sponzorem projektu*** na řešení klíčových koncepčních a organizačních otázek projektu a jeho prostřednictvím komunikuje s vedením firmy,
- zajišťuje ***komunikaci i s externími partnery***, resp. dodavateli podílejícími se na řešení projektu,
- je obvykle ***členem řídící komise projektu***,
- zodpovídá sponzorovi za ***naplnění cílů projektu*** a očekávaných ekonomických i mimoekonomických efektů,
- v kooperaci ***s architektem*** podnikové analytiky určuje základní ***přístup k řešení*** projektu na základě posouzení efektů a omezení možných ***variant***, případně jejich modifikací a kombinací,
- s informačním manažerem a sponzorem projektu se podílí ***na rozvoji celkové koncepce podnikové analytiky***, tj. na určení dalších aplikací včetně pokročilé analytiky, a na specifikaci jejich priorit,
- připravuje ***zadání projektu***, specifikaci cílů projektu, metriky projektu, určuje strategii a výběr metod řízení a postupu projektu.

2.2.4 Architekt podnikové analytiky

Architekt podnikové analytiky je specialistou v oblasti celkového návrhu aplikací a návrhu IT infrastruktury pro podnikovou analytiku. Analyzuje a navrhuje vhodnou aplikační a technologickou architekturu celé podnikové analytiky firmy. Připravuje podklady pro konfigurování hardware a software. Monitoruje a řeší problémy s optimalizací výkonu aplikací. Realizuje např. ***tyto činnosti***:

- řeší ***datovou a technologickou architekturu*** řešení s respektováním zvoleného přístupu,
- určuje ***umístění datového skladu na servery***, umístění jednotlivých datových tržišť, technické řešení OLAP databází,
- zajišťuje výběr a realizaci ***produktů pro klientské aplikace***, využití portálů pro reportování z aplikací,
- analyzuje současný ***stav infrastruktury*** ve vztahu k dosavadnímu řešení podnikové analytiky,
- vyhodnocuje ***vytížení kapacit*** zdrojů,
- analyzuje současný ***stav infrastruktury vzhledem k aktuálním uživatelským požadavkům***,
- analyzuje ***stav datových zdrojů*** a jejich technickou dostupnost,
- zajišťuje analýzy, hodnocení a monitorování ***technologické infrastruktury*** s cílem zajistit, aby byla optimálně konfigurována a škálovatelná,
- zpracovává návrh ***technologické infrastruktury s respektováním požadavků*** na její výkon, včetně špičkových zatížení, bezpečnost provozu, flexibilitu, možnosti škálování i pracovní náročnost obsluhy,
- podílí se na ***výběru vhodných technologických prostředků*** pro projekty,
- řeší ***technologické problémy*** spojené s návrhy a předpokládaným provozem transformačních procedur.

2.2.5 Byznys analytik / Konzultant

Byznys analytik řeší obsahovou a logickou stránku jednotlivých úloh a přípravu analytických a plánovacích aplikací v rámci projektů podnikové analytiky. Podle povahy projektu zajišťuje podle jednotlivých úloh a fází řešení projektů např. tyto ***činnosti***:

- ***konzultuje s uživateli*** jejich problémy a požadavky na analytické a plánovací aplikace,
- posuzuje ***kvalitu zdrojových databází***,
- specifikuje rámcový ***obsah a strukturu hlavních reportů***, ukazatelů a jim odpovídajících dimenzí,

- definuje základní ***funkcionalitu*** analytických a plánovacích aplikací,
- řeší ***celkovou koncepci transformací dat*** (ETL / ELT apod.),
- zajišťuje ***analýzu současného stavu řešení*** a disponibilních zdrojových databází a aplikací,
- kooperuje na ***analýzách požadavků*** uživatelů a jejich konsolidaci,
- v kooperaci s dodavatelem zajišťuje ***analytické a projektové činnosti*** na návrhu ***dočasného úložiště dat,***
- navrhuje a projednává s uživateli ***hrubý dimenzionální model*** řešení,
- navrhuje ***datový model datového skladu i datových tržišť*** v rámci daného přírůstku,
- navrhuje strukturu a způsob ***využití OLAP databází***,
- řeší ***analytické otázky transformací dat***, tj. vlastní transformace dat (např. formáty dat), granularitu transformovaných dat,
- ***navrhuje kontrolní a opravné procedury*** v souvislosti s čištěním, resp. zajištěním kvality dat,
- navrhuje způsob ***aktualizací při změnách dimenzí***,
- kooperuje na ***přípravě datové základny*** pro zahájení provozu aplikací,
- podílí se na ***řešení analytických problémů nebo chyb*** vyplývajících s přípravou provozu.

2.2.6 Datový analytik

Datový analytik zajišťuje zejména správné ***mapování ze zdrojových do cílového systému***, konzistenci dat apod. Zajišťuje podle jednotlivých úloh řešení projektů tyto ***činnosti***:

- zajišťuje ***analýzu datových struktur zdrojových systémů***, jejich kvalitu a dostupnost,
- spolupracuje s architektem podnikové analytiky na návrhu ***datové architektury***,
- spolupracuje na ***vyhodnocení stavu a kvality dat*** na formulaci principů řízení jejich kvality,
- konzultuje a posuzuje možné ***varianty přístupů k řešení*** projektů z pohledu datové architektury,
- definuje ***datové struktury cílového systému***, tj. datového skladu, datových tržišť a dalších,
- spolupracuje na ***specifikaci nástrojů pro transformace dat***,
- spolupracuje s byznys analytiky na ***určení oblastí***, jež se budou v rámci datových transformací ***prioritní,***
- definuje ***tabulky nutných pro plnění cílové databáze,*** např. datového skladu a tržišť, které se nenacházejí ve zdrojových systémech,
- definuje ***detailní pravidla procedury transformací dat***, popis transformací polí mezi zdrojovými a cílovým systémem (datové typy, délka polí, plnění konstantami),
- spolupracuje na ***implementaci databázových schémat*** a transformačních procedur,
- spolupracuje na ***vytváření prvotních databází*** datového skladu a datových tržišť.

2.2.7 Vývojář softwaru pro podnikovou analytiku

Vývojář software pro podnikovou analytiku hodnotí požadavky na stávající nebo nové softwarové aplikace a navrhuje, vyvíjí, dokumentuje a provádí údržbu aplikací. Realizuje ***následující činnosti***:

- zajišťuje analýzu a ***hodnocení požadavků na softwarové aplikace*** a související základní software,
- podílí se na ***návrhu vývojové a provozní platformy*** aplikačního softwaru,
- konzultuje s architektem datové analytiky ***specifikace potřebných technických parametrů infrastruktury***,

- realizuje návrh a ***vývoj zejména transformačních procedur***, analytických, plánovacích aplikací a reportů,
- kooperuje na ***řízení testování aplikací*** a validačních postupů,
- zajišťuje ***zpracování dokumentace*** transformačních procedur a aplikací,
- kooperuje na posouzení, vývoji, zdokonalování a dokumentaci ***postupů údržby*** operačních systémů, komunikačních prostředí a aplikačního softwaru.

2.2.8 Dodavatel

Externí dodavatel představuje ***větší či menší firmy*** poskytující jednak produkty, ale zejména ***komplexní či specializované služby*** pro řešení projektů nebo zajištění provozu podnikové analytiky. Rozsah a způsob dodávek ***se odvíjí od sourcingové strategie*** přijaté na úrovni vedení firmy. Dodavatel ***zajišťuje tyto funkce***:

- spolupráce na celkové strategii a ***přístupu k řešení*** podnikové analytiky***,***
- ***formulace jednotlivých typů architektur*** relevantních vzhledem k projektům,
- ***analýza a návrh řešení***, tj. funkcionality, datového zajištění, technologických zdrojů, jejich testování a dokumentace,
- ***implementace a zavedení do provozu*** realizovaných aplikací a IT služeb, zajištění instalačních, integračních, školicích a dalších služeb,
- ***zajištění kompletního postimplementačního servisu*** a dalšího rozvoje řešení.

2.3 Závěry k rolím v podnikové analytice

- Řešení podnikové analytiky je charakteristické ***vysokými nároky na kooperaci analytiků a uživatelů***. S rostoucí komplexností a složitostí analytických systémů tyto nároky dále rostou. Proto ***vymezení struktury a náplně jednotlivých rolí*** je z tohoto pohledu velmi významná.
- Výše uvedené role tvoří pouze ***podmnožinu***, které se úzce váže na jednotlivé součásti řešení podnikové analytiky. Je nezbytné pak ***brát v úvahu i další role, zejména manažerské a role firemních specialistů***
- Náplň jednotlivých rolí je kromě zmíněné kooperace, účelná i pro systematické plánování a ***řízení kvalifikačních programů*** pro podnikovou analytiku, a nejen pro ni, de facto pro vytváření celého komplexu školicích programů a aktivit.
- Náplň jednotlivých rolí je nezbytné ***modifikovat podle potřeb*** a možností firmy a zejména struktury jejích personálních zdrojů.
- Ekonomické i technologické prostředí se postupně a často i rychle mění. Proto je nezbytné ***nastavit i systém průběžné aktualizace*** definovaných rolí.
- Výše uvedené vymezení rolí se s ohledem na rozsah textu omezuje na jejich funkční náplň, resp. klíčové činnosti. Pro praktické využití je ***účelné je doplnit i požadavky na kvalifikaci, resp. znalosti,*** a to podle konkrétních potřeb firmy.
- Uplatnění rolí, jejich náplně a požadavků na znalosti je velmi dobré rovněž při ***specifikaci náborových akcí*** a následně i v personálním řízení při určování obsahu jednotlivých pracovních pozic.

3. Faktory ovlivňující podnikovou analytiku

Každý projekt i aplikace podnikové analytiky jsou ovlivněny celou škálou organizačních, technických a dalších faktorů, které je nutné při jeho řešení, nasazení a užití jejích aplikací respektovat.

Účelem kapitoly je:

- určit ty ***faktory řízení firmy i byznys prostředí,*** které významně ovlivňují řešení a užití úloh podnikové analytiky,
- vymezit ***jejich podstatné charakteristiky a vlivy*** na podnikovou analytiku,
- vytvořit tak předpoklady, aby na základě ***vyhodnocení podmínek, resp. faktorů*** bylo možné posoudit, jaké řešení podnikové analytiky bude účelné, případně, zda má taková implementace vůbec smysl.

Faktorem se rozumí takové uplatnění osvědčených přístupů, postupů a zkušeností, které povedou ke splnění stanovených cílů a požadavků kladených na projekt a k dosažení očekávaných efektů. Faktory jsou ***založeny na různých zdrojích***. Jednak vycházejí ***z nejlepších praktických zkušeností*** získaných při realizaci obdobných řešení a projektů, jednak z poznatků a znalostí získaných na bázi výzkumných aktivit, průzkumů, hodnocení vývojových trendů, nebo teoretických prací, které jsou předmětem činnosti analytických společností i akademické sféry. Jinak řečeno, faktory v sobě zahrnují ve velmi komprimované formě všechny ***osvědčené poznatky a zkušenosti z dané oblasti***, v našem případě z implementací a provozu aplikací podnikové analytiky.

Faktory podnikové analytiky jsou svým rozsahem, strukturou i obsahem ***velmi rozmanité*** a je proto účelné je členit a klasifikovat. Při každé klasifikaci v informatice je však zřejmé, že hranice mezi jednotlivými skupinami faktorů nemohou být vždy ostré, a navíc každá taková ***klasifikace má subjektivní charakter*** a představuje pouze určitý pohled na danou věc. Proto považujme i další klasifikaci jako za pouze pracovní.

Kapitola ***obsahuje tyto faktory a jejich skupiny***:

- ***Byznys prostředí*** vyjadřuje ***vnější podmínky*** a vlivy pro uplatnění podnikové analytiky v daném podniku. Sem např. patří:
  - ***velikost firmy*** daná zejména počtem zaměstnanců a velikostí ročního obratu,
  - ***původ a vlastnictví firmy***, zda je firma v českém vlastnictví, nebo je pobočkou nadnárodní společnosti, nebo výlučně zahraniční firmou,
  - ***konkurenční prostředí***, představující zejména sílu konkurenčních tlaků, a tedy i tlaku na uplatnění podnikové analytiky,
  - ***odvětví působnosti firmy***, z něhož vyplývají zejména požadavky na obsahovou stránku řešených analytických a dalších úloh.

- ***Řízení a organizace firmy*** představují ***vnitřní podmínky řízení*** ovlivňují obsah, rozsah i užití úloh podnikové analytiky. Patří sem:
  - ***firemní kultura***, tj. systém hodnot, které podnik vyznává, zaběhnutá schémata jednání a rozhodování atd., má vliv na styl řízení a na úroveň detailu, na jakém se otázky podnikové analytiky řeší a další aspekty,
  - ***organizace firmy*** je prostředí pro racionální kooperaci pracovníků a pracovních týmů, výrazem efektivní dělby práce,
  - ***dislokace firmy*** vyjadřuje regionální rozmístění centrály firmy a jejích jednotlivých obchodních poboček, detašovaných skladů, výrobních provozů apod.,
  - ***existence business modelu*** je forma pro kvalitní pochopení základního fungování firmy a pro řešení zejména obsahu podnikové analytiky má klíčový význam,
  - ***agilní organizace*** používá „***agile***“ jako základ svého fungování, kde jednotlivé týmy jsou organizovány do vyšších celků, které se operativně řídí a orientují se na průběžné poskytování výsledků řešení,

- ***regineering podnikových procesů*** je zaměřen zejména na snižování časové a finanční náročnosti procesů firmy při zachování jejich vysoké kvality,
- ***vztah podniku ke svým obchodním a dalším partnerům*** je založen v tomto kontextu na poskytování výstupů, resp. reportů podnikové analytiky v určitých oblastech řízení (stav zakázek apod.) pro obchodní partnery firmy,
- ***personální zdroje*** jsou rovněž klíčovým faktorem ovlivňujícím provoz a rozvoj podnikové analytiky, zejména profesní, kvalifikační i věková struktura pracovníků a úroveň jejich znalostí (viz kapitola věnovaná rolím).

- ***Řízení podnikové analytiky*** zahrnuje podmínky řízení spojené s rozvojem a provozem IT ve firmě. Patří sem:
  - ***řízení IT ve firmě*** musí respektovat specifické nároky IT aplikací, v případě podnikové analytiky je to např. vysoká míra účasti uživatelů na řešení aplikací, specifické nároky na jejich provoz apod.
  - ***existence silného sponzora***, tedy osobnost se značnou mírou vlivu a s nezbytnými rozhodovacími pravomocemi, která je také schopna vidět firmu a její aktivity ve všech podstatných souvislostech,
  - ***strategie řešení podnikové analytiky*** je velmi podstatným faktorem řešení s ohledem na značný rozsah těchto projektů, na jejich finanční a časovou náročnost a rychlý rozvoj technologií,
  - ***identifikace potenciálních efektů podnikové analytiky*** je podstatná s ohledem na to, že analytické přípravě a využití těchto aplikací musí věnovat svůj čas i manažeři a specialisté firmy, pak musí vědět, proč ho do podnikové analytiky mají investovat,
  - ***kompetenční centra*** organizačně a systematicky sdružují pracovníky uživatelských a IT útvarů a vytvářejí tak kvalifikovanou základnu pro kvalitní rozvoj podnikové analytiky ve firmě,
  - ***využití cloud computingu*** založeného na sdílení hardwarových a softwarových prostředků v infrastruktuře internetu umožňující rozvoj podnikové analytiky při nižších nákladech a často i s nižšími časovými nároky,
  - ***metody řešení nárůstu dat***, kdy při enormních nárůstech objemu dat je nutné hledat a aplikovat efektivní metody řešení tohoto problému.

- ***Kvalita informačního systému*** představuje zejména hodnocení kvality jeho zdrojů, zejména úrovně řešení kvality dat a kvality podnikových aplikací. Patří sem:
  - ***úroveň dokumentace a optimalizace podnikových procesů*** vytváří nezbytné prostředí pro racionální využívání zdrojů pro podnikovou analytiku a současně její efektivní využívání,
  - ***kvalita datových zdrojů***, včetně dostupnosti dokumentace datových zdrojů – tento faktor je zde zahrnut (i když byl již zmíněn) jako zdůraznění jeho zásadního významu pro úspěšnost podnikové analytiky,
  - ***kvalita podnikových aplikací,*** zejména transakčních, je významná právě jako předpoklad kvality datových zdrojů i s ohledem na předpokládaný rozvoj zdrojových aplikací (jejich náhrady, změny, upgrade atd.).

Analýza, návrh, provoz i užití jakékoli aplikace podnikové analytiky, jakéhokoli projektu jsou tak vždy ovlivňovány celou řadou ***faktorů, které určují, nebo spoluurčují jejich výslednou kvalitu a konečný úspěch***. Přitom platí, že tyto faktory se mohou u jednotlivých typů projektů a aplikací výrazně lišit. Je proto vždy účelné tyto faktory identifikovat a analyzovat vzhledem k typu řešené aplikace, resp. projektu. Jejich poznání se pak nutně promítá do projektových a provozních postupů, tj. do využití jejich positivního potenciálu a do aktivit směřujících k minimalizaci jejich negativních dopadů.

Detailněji jsou jednotlivé uvedené faktory analyzovány ***v dalších podkapitolách***.

3.1 Byznys prostředí

Skupina faktorů označených jako „*byznys prostředí*“ vyjadřuje převážně ***vnější podmínky*** pro uplatnění podnikové analytiky v daném podniku. Do této skupiny spadají především tyto faktory:

- velikost firmy,
- původ a vlastnictví firmy,
- konkurenční prostředí,
- odvětví působnosti firmy.

3.1.1 Velikost firmy

Velikost firmy výrazně ***ovlivňuje způsob a potřeby řešení podnikové analytiky***, a nakonec i očekávané a následně reálné výsledky. Je obvykle dána počtem zaměstnanců a výší ročního obratu, obvykle se dělí na 3 kategorie:

- ***Malé*** firmy: s počtem 1–100 zaměstnanců, s ročním obratem, který nepřesahuje 30 milionů Kč,
- ***Střední*** firmy: s počtem od 101 do 500 zaměstnanců a s ročním obratem od 31 do 100 milionů Kč,
- ***Velké*** firmy: s počtem nad 500 zaměstnanců a s obratem vyšším než 100 milionů Kč.

Zatímco v minulosti byla řešení podnikové analytiky záležitostí převážně velkých, případně středních firem, v současnosti se uplatňuje prakticky na všech úrovních jejich velikosti. Patří sem i následující charakteristiky:

- Pro ***velké a střední*** firmy je typické ***využití komplexních řešení*** na bázi rozsáhlých datových skladů, datových tržišť a dalších technologií. Zejména se zesilují potřeby na ***standardní reporting***, na integraci a konsolidaci datových zdrojů a samozřejmě na podstatně rozsáhlejší a komplexnější analytické aplikace.
- ***Malé***, resp. menší firmy se zaměřují spíše na menší řešení na bázi self service business intelligence (***SSBI***).
- ***Velké a střední*** firmy disponují obvykle ***potřebnými finančními a personálním zdroji*** nutnými pro trvalý rozvoj analytických řešení.
- Na druhé straně ***u větších firem*** jsou často podstatně ***složitější zdrojové systémy***, a tedy i podstatně náročnější transformace dat do analytických databází, stejně tak je podstatně náročnější řízení a zajištění potřebné kvality dat.
- U menších firem jsou ***obvykle vlastníci současně manažery***, a tedy jsou často do rozvoje podnikové analytiky přímo zapojeni, u velkých firem je tento problém s vysokým počtem vlastníků složitější.

Všechny uvedené dílčí faktory je nezbytné pečlivě zvažovat zejména již při plánování a zadávání těchto projektů.

3.1.2 Původ a vlastnictví firmy

Vlastnictví firmy je jeden z faktorů, který ***vyjadřuje formy a složitost vlastnických vztahů*** a v tomto případě zejména také to, zda je firma v českém vlastnictví, nebo je pobočkou nějaké nadnárodní společnosti, nebo je výlučně zahraniční firmou působící v českém prostředí a na českých trzích. S tím jsou spojeny tyto dopady:

- Podniková analytika a její nástroje jsou vedle managementu ***směrovány na potřeby vlastníků***, pokud mají zájem se v dané oblasti angažovat. Podstatné je to, že ***schvalují investice do jejího rozvoje***. Jejich souhlas je pak často výrazně ovlivněn tím, do jaké míry jsou aktivními uživateli produktů podnikové analytiky a nejlépe i jejich spolutvůrci, resp. konzultanty.
- Specifické problémy jsou v situaci, kdy firma představuje ***pobočku zahraniční nadnárodní firmy.*** V tomto případě nejde pouze o schvalování investic, ale i souhlas s navrhovaným řešením, kdy ***mateřská společnost často vyžaduje dodržování celofiremních standardů*** – funkčních i technologických. To je na jedné straně pochopitelné, na druhé straně to omezuje invenci a iniciativu tuzemských pracovníků v rozvoji podnikové analytiky.

3.1.3 Konkurenční prostředí

Potřeby řešení analytických projektů jsou výrazně ovlivněny i ***tržním, zejména konkurenčním prostředím***, kde podnik působí. V segmentech trhu s relativně nízkým konkurenčním tlakem, např. v hutnictví nebo těžkém strojírenství a případně ve veřejné správě je potřeba projektů podnikové analytiky, a především zájem uživatelů na využívání těchto aplikací omezen a tím i výsledky a efekty z nich jsou často nepříliš výrazné. Stejně tak je omezen i tlak konkurence na efektivitu a kvalitu obchodních aktivit podniku a tím na existenci a úroveň analytických a plánovacích aplikací.

Konkurenční prostředí firem je popsáno v řadě publikací a modelů, jde o ***faktor, který v řešení rozvoje podnikové analytiky tak často hraje klíčovou roli***:

- Konkurence, zákazníci, dodavatelé a další partneři, jejich hodnocení, jejich očekávaný vývoj, jejich ***nároky a očekávání se stávají velmi významnou součástí podnikové analytiky*** a obvykle tvoří jádro aplikací zaměřených na strategické řízení, marketing, nebo prodej.
- Síla konkurence je faktor, který, zejména u velkých firem, ***posiluje potřeby investic zejména do pokročilé analytiky,*** jako např. prediktivní analytiky.
- Faktor konkurenčního prostředí znamená ***výraznější potřebu a uplatnění externích datových zdrojů*** a s tím spojené nezbytné aktivity, jako např. jejich kvalifikované hodnocení kvality, dostupnosti, finanční náročnosti.

3.1.4 Odvětví působnosti firmy

Společné charakteristiky ve vztahu k odvětví, kde firma působí jsou následující:

- Odvětví působnosti firmy je ***významné zejména pro obsahové zaměření analytických a plánovacích úloh*** a náročnost a složitost jejich funkcionality.
- Odvětví ekonomiky ***se liší i dosavadním historickým vývojem v uplatňováním nástrojů podnikové analytiky*** a tím i rozsahem zkušeností, které týmy těchto firem v průběhu vývoje získaly. Odvětví s nejdelší historií v této oblasti je evidentně bankovnictví a finanční služby vůbec, nebo telekomunikace.
- Firmy v různých odvětvích ekonomiky svojí složitostí řízení, tlakem na efektivitu, rozsahem různorodých datových zdrojů, již existující IT infrastrukturou ***vytvářejí jak různé předpoklady pro rozvoj a uplatňování analytiky, tak i vyšší či nižší objektivní náročnost na funkcionalitu*** a technologickou úroveň aplikací.

3.2 Řízení a organizace firmy

Skupina faktorů označených jako „*řízení a organizace firmy*“ vyjadřuje převážně ***vnitřní podmínky řízení*** pro uplatnění podnikové analytiky v daném podniku. Do této skupiny spadají především tyto faktory:

- firemní kultura,
- organizace firmy,
- dislokace firmy,
- existence business modelu,
- agilní organizace,
- regineering podnikových procesů,
- vztah podniku ke svým obchodním a dalším partnerům,
- personální zdroje.

3.2.1 Firemní kultura

Zřejmě nejpodstatnějším faktorem je ***potřeba*** a samotný ***zájem vedení společnosti*** o aplikace tohoto typu. Ty jsou určeny primárně ***pro manažersky nebo analyticky orientované uživatele***, i když s ohledem na jejich stále vyšší dostupnost se předpokládá jejich užití i na středních a nižších úrovních řízení. Z toho vyplývá, že kvalita jejich ***řešení a zejména užití není primárně dána předpisy,***

metodikami, resp. disciplínou pracovníků, ale zájmem, motivací a invencí na uživatelské i dodavatelské straně.

Kultura firmy, tj. ***systém hodnot, které podnik vyznává***, zaběhnutá schémata jednání a rozhodování atd., má vliv na styl řízení a na úroveň detailu, na jakém se otázky v podniku řeší – např. jak detailně jsou předepsány činnosti jednotlivých podnikových procesů:

- Úroveň kultury firmy má také ***vliv na iniciativu a kreativitu pracovníků*****,** která je pro účast na řešení projektů podnikové analytiky často rozhodující.
- Vysoká podniková kultura umožňuje obvykle ***méně problémů s rychlejším zavádění nových aplikací*** a technologií, včetně analytických.
- Firemní kultura se ***promítá do efektivnosti řízení projektů*** i do efektivnosti řízení celé podnikové analytiky.
- Kultura firmy určuje i priority **v *orientaci informatiky na určité typy aplikací,*** zejména v případě podnikové analytiky.
- Z pohledu kultury firmy má ***klíčový vliv*** na pozici, rozvoj a užití podnikové analytiky ***přístup managementu.***

3.2.2 Organizace firmy

Organizace firmy je prostředí pro racionální kooperaci pracovníků a pracovních týmů, výrazem efektivní dělby práce. ***Organizační struktura*** řeší problém ***přijatelného rozpětí řízení***, tj. počtu pracovníků, který je schopen daný řídící pracovník efektivně řídit. Na základě toho vznikají organizační úrovně:

- Jasně definovaná a dokumentovaná organizace frimy ***přispívá k efektivnosti řízení*** a k uplatnění podnikové analytiky.
- Kvalitní organizace přispívá ke ***zvyšování výkonnosti celé firmy***, včetně využití metody Corporate Performance Management, CPM.
- Řešení organizace ***navazuje na firemní strategii***, resp. je její součástí. K efektivnímu promítnutí strategie do firemní organizace ***se využívá metoda Balance Scorecard, BSC***.
- Jasně definovaná podniková organizace ***přispívá k rychlé identifikaci problémů***, jejich zdrojů a příčin, a nakonec k jejich řešení i na bázi podnikové analytiky.
- Organizace firmy má být natolik ***flexibilní, aby byla schopna rychle reagovat na vývoj*** podnikatelského prostředí a vztahů k externím partnerům.
- Organizace firmy ***má efektivně respektovat i možnosti podnikové analytiky***.
- Organizace firmy musí, i ***s předstihem, reagovat na předpokládané potřeby*** v transformacích byznysu a byznys modelu dané firmy.

3.2.3 Dislokace firmy

Dislokace firmy vyjadřuje ***regionální rozmístění centrály firmy a jejích jednotlivých obchodních nebo výrobních*** poboček, detašovaných skladů apod.:

- Vysoká ***decentralizace umožňuje často lepší využití dostupných kapacit*** (personálních, technických nebo materiálových) pro řešení projektů analytiky.
- S vysokou nebo novou dislokací ***vznikají nároky na změny v řízení firmy*** a obvykle i na transformaci celého byznysu a současně i nároky na změny analytických a plánovacích aplikací.
- U vysoké dislokace ***do zahraničních regionů dochází i k řešení kulturních rozdílů***, pracovních návyků, disciplíny s dopady na řešení a využití analytických úloh.
- U dislokací ***do zahraničí*** je nutné řešit i tuzemská ***legislativní omezení a rozdíly***.

3.2.4 Business model

Business model je ***forma a přístup pro kvalitní pochopení základního fungování firmy***.

- Podporuje ***uvědomění si souvislostí jednotlivých částí a aspektů firmy*** a souvislostí v analytických i plánovacích úlohách.
- Umožňuje ***kvalifikovanou aplikaci tohoto přístupu při řešení strategických úloh*** v řízení firmy.
- Vyžaduje ***motivaci a zájem ze strany vedení firmy*** pro uplatnění business modelu v analytice a plánování.
- Zahrnuje rovněž postupné ***naplnění tzv. „Lean Canvas“*** (viz další obrázek):

<table>
<tr><td rowspan="2">Problem

1. Nejdříve identifikujte problém, který chcete řešit.</td><td>Solution

4. Popište základní prvky Vaší firmy.</td><td rowspan="2">Unique Value Proposition

3. Pojmenujte klíčové hodnoty Vašeho produktu, kvůli kterým bude mít zákazník zájem koupit.</td><td>Unfair Advantage

5. Identifikujte výhodu, kterou konkurence nemá/nezíská.</td><td rowspan="2">Customer Segments

2. Určete typické zákazníky, kteří tento problém mají a budou mít zájem ho řešit.</td></tr>
<tr><td>Key Metrics

8. Jak budete měřit úspěch v jednotlivých fázích podnikání?</td><td>Channels

9. Jaké kanály zvolíte k obsluze zákazníků?</td></tr>
<tr><td colspan="2">Cost Structure

7. Specifikujte strukturu nákladů potřebných pro rozjezd a fungování firmy.</td><td colspan="3">Revenue Streams

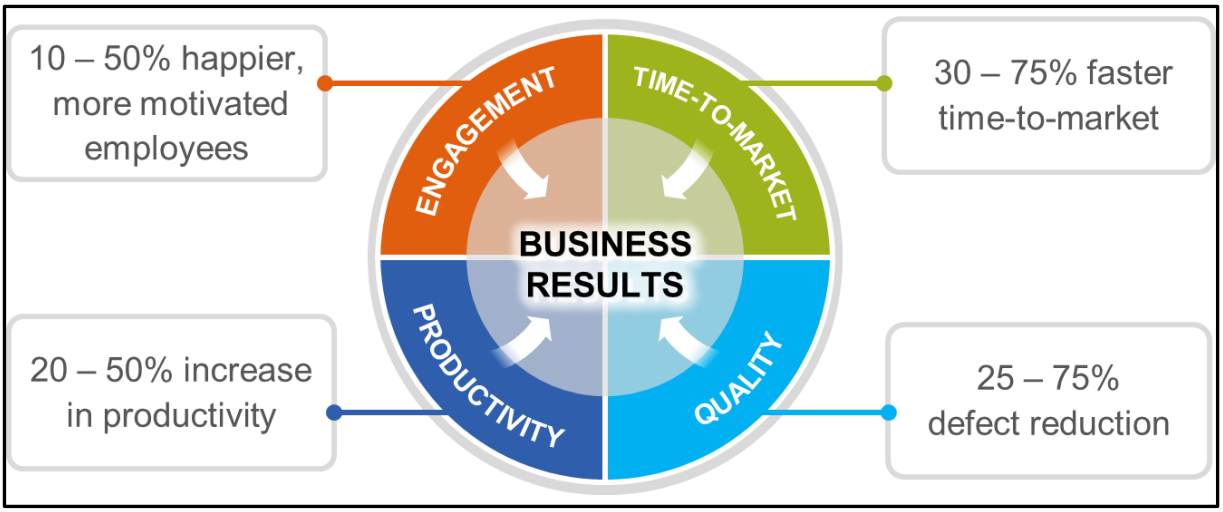
6. Stanovte, z čeho budou plynout příjmy.</td></tr>
</table>

Obrázek 3-1: Lean canvas

3.2.5 Agilní organizace

Agilní organizace je podnik nebo jeho část, která ***používá agile jako základ svého fungování***. Jednotlivé týmy (využívající např. SCRUM) jsou organizovány do vyšších celků. Existuje několik přístupů/frameworků (např. Spotify, SAFe, LeSS), které podporují orchestraci agilní organizace.

- Některé společnosti ***zvolily agilní transformaci v plném rozsahu***, tedy v řešení projektů podnikové analytiky.
- Agilní transformace velkého rozsahu ***musí být vedena vrcholovými manažery***. Efektivní vedení však není možné bez osvojení agilních metod a zapojení do transformace.
- Efekty agilní organizace dokumentuje následující obrázek:



Obrázek 3-2: Efekty agilní organizace (Zdroj: scaledagileframework.com/case-studies)

3.2.6 Reengineering podnikových procesů

Hlavní přístupy představují úlohy řízení podnikových procesů (BPM, Business Process Management), modelování podnikových procesů a reengineeringu podnikových procesů (BPR, Business Process Reengineering) spojené s možnostmi snižování jejich časové a finanční náročnosti. S tím jsou spojené následující aspekty:

- Vytvoření podkladů pro ***realizaci organizačních změn***, kdy organizační změny budou založeny na optimalizovaných procesech, tedy optimalizovaném fungování firmy.
- ***Změny organizační příslušnosti a kvalifikace pracovníků*** se budou provádět v kontextu zlepšení průběhů všech navržených procesů.
- Procesní rengineering je především záležitostí zvyšování kvality řízení firmy. Proto by ***hlavní zájem na jeho řešení a úspěchu měli mít podnikoví manažeři*****,** zejména nejvyšší podnikový management.
- Projekty procesního reenginneringu vyžadují ***aktivní účast většiny pracovníků*** firmy, musí se však zajistit jejich kvalitní příprava v oblasti procesního modelování, což může narážet na časové nebo finanční bariéry.
- Pro projekty BPR je nutné volit i ***adekvátní nástroje pro specialisty*****,** včetně nástrojů pro podporu analytiky, tedy specializované modelovací nástroje, ale následně je nutné zajistit dostupnost výsledků projektu široké uživatelské sféře.

3.2.7 Vztah firmy ke svým obchodním a dalším partnerům

Výstupy, resp. reporty z datových skladů a tržišť jsou umisťovány na firemní portály, a to nejen pro využití interními pracovníky podniku, ale v určitých oblastech řízení (stav zakázek apod.) ***pro obchodní partnery*** firmy. Nabídka této funkcionality s pochopením a respektováním potřeb i externích subjektů tvoří jeden z podstatných efektů podnikové analytiky a ovlivňuje tak i její výsledný úspěch.

3.2.8 Personální zdroje, úroveň jejich znalostí

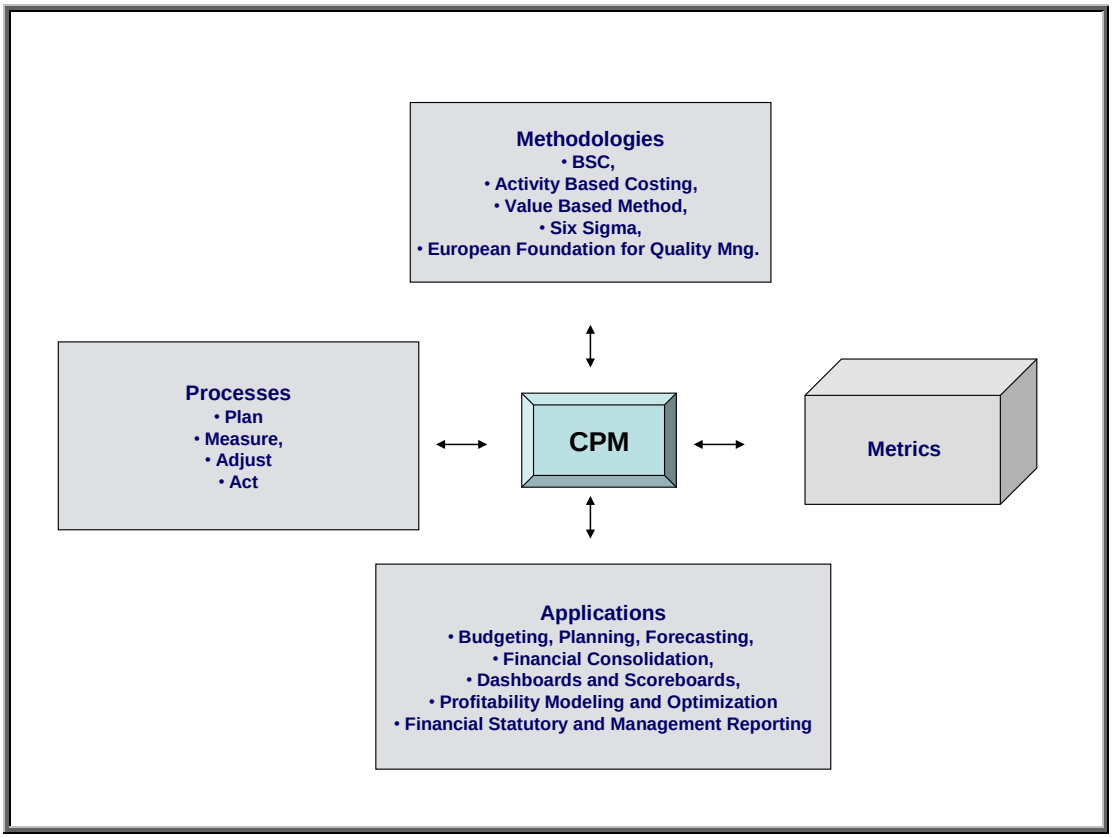
Ve fázi analýzy jsou projekty podnikové analytiky charakteristické tím, že ***předpokládají velmi intenzivní a kvalifikovanou kooperaci*** dodavatelů a řešitelů. To je dáno tím, že vesměs jde o aplikace značně specializované, šité na potřeby konkrétních uživatelů (tedy ne standardní) a jde o aplikace, jejichž ***finální přínos*** pro firmu bude více než jinde závislý na tom, ***jak budou tyto aplikace „chytré“***, v jakém rozsahu v nich budou uplatněny znalosti a zkušenosti jejich dodavatelů i budoucích uživatelů. Čím kvalifikovanější a motivovanější budou tito pracovníci na obou stranách, tím lze očekávat kvalitnější výsledky. A to platí u analytických aplikací v podstatně větší míře než u jiných typů aplikací.

Klíčovým faktorem ovlivňujícím provoz a rozvoj podnikové analytiky je tak ***profesní, kvalifikační i věková struktura pracovníků*** a úroveň jejich znalostí, a to jak znalosti pracovníků IT útvaru, tak znalosti jejích uživatelů (uživateli mohou být i lidé mimo podnik). S tím jsou spojené potenciální efekty a případné problémy:

- ***Kvalifikační struktura pracovníků*** vyšší úrovně je předpokladem a zdrojem pro rozvoj podnikové analytiky a její efektivní využívání.
- ***Znalosti pracovníků*** orientované jak na byznys a manažerské metody, tak na analytické metody v IT jsou dobrým základem pro kvalitní řešení projektů v podnikové analytice.
- Je účelné orientovat ***kvalifikační rozvoj pracovníků IT*** na ty oblasti, které nelze efektivně získat v rámci outsourcingu.
- Problém často spočívá v nedostatečné, nebo ***špatně orientované motivaci uživatelů a někdy i informatiků*** na dalším rozvoji podnikové analytiky.
- Není vytvářen ***časový prostor*** pro práci klíčových uživatelů na projektech podnikové analytiky.
- ***Top management*** firmy často ***nemá dostatečný nadhled*** nad rozvojem a perspektivními možnostmi podnikové analytiky.
- ***Kvalifikace CIO*** je především technologická, nikoli ekonomická a manažerská.

3.2.9 Uplatnění konceptu řízení podnikové výkonnosti

Řízení podnikové výkonnosti, Corporate Performance Management, CPM je kombinace managementu, metodik a metrik podporovaná aplikacemi, nástroji a infrastrukturou, která umožňuje uživatelům definovat, monitorovat a optimalizovat výsledky a výstupy tak, aby bylo dosaženo cílů osobních či cílů organizační jednotky v souladu se strategickými cíli podniku. Corporate Performance Management (CPM) je hlavním představitelem systémů řízení výkonnosti. Základní vymezení CPM vytvořila společnost Gartner, a to je v souladu s obecnou definicí systémů řízení výkonnosti. CPM tvoří čtyři základní segmenty, resp. komponenty řešení, které jsou vzájemně provázané (viz Obrázek 35-1).



Obrázek 3-3: Vymezení CPM (Zdroj: Chandler, 2008)

Základní ***segmenty CPM*** tvoří:

- Komplex ***manažerských metod***, které tvoří metodologický logický základ podnikového řízení a jejichž principy se respektují v ostatních segmentech CPM. Do těchto metod obvykle patří Balanced Scorecard (BSC), Activity Based Costing (ABC), Value Based Management (VBM), Six Sigma a další.
- ***Podnikové procesy***, tj. plánovací, analytické, monitorovací vytvářející ve svém komplexu procedurální logiku podnikového řízení a navazující na uvedené manažerské metody.
- ***Metriky*** pro podnikové řízení postavené ***na principech podnikové analytiky***, tj. představované klíčovými ukazateli ve vztahu k odpovídajícím dimenzím (zákaznické, komoditní atd.).
- Metody, procesy a metriky jsou základem pro ***plánovací a analytické aplikace*** postavené na technologiích a přístupech business intelligence.
- CPM tak představuje ***komplex metod, procesů, dat a aplikací***, který je nezbytné pro konkrétní podnikové potřeby vždy přiměřeně přizpůsobit. Na druhé straně tento koncept je aplikovatelný jak pro firmy různé odvětvové orientace, tak velikosti.

CPM je založeno na principech a technologiích podnikové analytiky, zejména business intelligence, ale není jeho synonymem. CPM je ***rozšiřuje o koncept „řízení“,*** který zahrnuje procesy jako plánování, prognózování a základní východiska podnikové strategie a úzce se váže na metody a metodiky pro řízení výkonnosti (např. BSC, ABC a další). Uplatnění aplikací podnikové analytiky podporují uvedené metodiky a je proto jádrem současného konceptu CPM).

K ***hlavní funkcionalitě*** úloh a nástrojů CPM tak patří komplexní podnikové analýzy na všech úrovních řízení, plánování, rozpočtování a prognózování na bázi klíčových firemních metrik. Lze pracovně vymezit ***tři hlavní pojetí řízení výkonnosti***:

- ***analytické*** pojetí řízení výkonnosti – zaměřené na vymezení a provázanost podstatných komponent řízení podniku,
- ***ekonomické*** pojetí řízení výkonnosti – obvykle chápané jako součást manažerského účetnictví,

- ***manažerské*** pojetí řízení výkonnosti – definující zejména manažerské postupy pro nastavení a hodnocení výkonnosti dílčích útvarů a pracovníků podniku.

Řízení podnikové výkonnosti v analytickém pojetí směřuje k ***určení a vymezení hlavních podnikových procesů***, resp. úloh, ***metrik, metod a aplikací***, tj. komponent, které významně ***ovlivňují úspěšnost podniku***. Definování vazeb mezi uvedenými komponentami ***podporuje systematický charakter*** řízení. Vymezení postupů a ***zodpovědností pracovníků za plnění cílů*** ve výkonnosti podniku (manažerské pojetí) zvyšuje úspěšnosti jejich dosažení,

Ekonomické pojetí vede k výběru nejdůležitějších ekonomických, převážně finančních ukazatelů a soustřeďuje na ně hlavní pozornost. Hlavním efektem je ***provázanost všech tří uvedených pojetí*** a vytvoření předpokladů pro kvalitní a systematické řízení podnikové výkonnosti. Principy CPM ***posilují i vazby a komunikaci v řízení*** na strategické, taktické i operativní úrovni. To znamená i stejné pochopení strategických cílů firmy u pracovníků na všech úrovních řízení a jejich podporování.

3.3 Řízení podnikové analytiky

Skupina faktorů označených jako „*řízení podnikové analytiky*“ vyjadřuje převážně ***podmínky řízení*** vyvolané charakterem a nároky IT ve firmě. Do této skupiny ***spadají především tyto faktory***:

- řízení IT ve firmě,
- existence silného sponzora,
- strategie řešení podnikové analytiky,
- identifikace potenciálních efektů podnikové analytiky,
- kompetenční centra,
- využití cloud computingu.

3.3.1 Řízení IT ve firmě

Aplikace podnikové analytiky jsou dnes již obvyklou, přesto specifickou součástí IT ve firmě. Řízení IT ***musí proto respektovat i některé specifické nároky*** těchto aplikací, zejména:

- pro analytické aplikace jsou charakteristické ***velmi těsné vazby na ostatní aplikace*** podnikové informatiky. Jejich úspěch závisí proto i ***na kvalitě podnikové a aplikační architektury***, úrovni zajištění integrace podnikové informatiky, kvalitním plánování a zadávání nových projektů s respektováním možností využití podnikové analytiky (např. využití její analytické funkcionality a nezatěžováním tím transakční aplikace apod.),
- podniková analytika je postavena na ***využití datových zdrojů*** vznikajících převážně ***v transakčních aplikacích***. Je nutné řešit celý ***komplex otázek zajištění kvality dat***, což není záležitostí pouze projektů analytiky, ale procesů řízení celé podnikové informatiky, resp. IT. S jejich nízkou kvalitou klesá kvalita nebo úplně zaniká řešení podnikové analytiky,
- pro kvalitní a systematické řízení rozvoje podnikové analytiky je předpokladem i ***systematické a kvalitní řízení datových zdrojů***, v současné době obvykle založené ***na principech Data Governance***,
- s předchozí poznámkou souvisí ***řízení změn*** v IT aplikacích. Pokud není zajištěno ***efektivní a včasné předávání informací*** o změnách v primárních aplikacích správcům analytických aplikací řešení dochází ***k chybám ve vstupních datových strukturách*** v ETL, resp. transformačních procedurách, načítání chybných dat a následně k chybám ve výstupních reportech a a dashboardech,
- úspěch podnikové analytiky ovlivňuje i ***úroveň řízení provozu*** celé informatiky, zejména správa databázových serverů, plánování a kontrola průběhu ETL procesů, zařazení problematiky analytiky do služeb help-desku atd.

3.3.2 Existence silného sponzora

Pravidlo ***silného sponzora*** je v oblasti podnikové analytiky již všeobecně známé. S ohledem na konečný úspěch by tyto projekty (BI a další) měly být uvnitř firmy vždy podporovány ***osobností se značnou mírou vlivu a s nezbytnými rozhodovacími pravomocemi*** (viz kapitola 2.2.2). Jde o osobnost, která navíc vedle svého vlivu a presentovaného zájmu o řešení, je schopna vidět podnik a jeho aktivity ve všech podstatných souvislostech, ve vztahu k podnikovému okolí, je schopna formulovat a rozhodovat o klíčových prioritách řešení a samozřejmě je schopna řešit finanční zajištění projektu a dalšího provozu.

3.3.3 Strategie řešení podnikové analytiky

Strategické otázky řešení podnikové analytiky se promítají na nejvyšší úrovni ***do informační strategie*** firmy vztahující se k IT firmy jako celku a na nižší úrovni do strategie analytických projektů. Při převládajícím ***značném rozsahu*** těchto projektů, jejich ***finanční a časové náročnosti***, mimořádně rychlém ***rozvoji technologií***, na nichž jsou založeny, je určení správné strategie odpovídající potřebám a možnostem firmy ***velmi podstatným faktorem*** řešení. Stanovení strategie by tak mělo zohlednit i vyhodnocení všech dílčích faktorů, a to zejména:

- určení ***rozsahu*** projektu a jeho ***priorit*** ve vztahu k podnikovým procesům, resp. oblastem řízení firmy,
- určení ***přístupu*** k řešení projektu,
- zaměření řešení na určitý ***typ produktu a produkt***, kde se nabízejí následující hlavní varianty:
    - využití ***specializovaných analytických aplikací*** vytvářených přesně dle zadání uživatele s využitím nástrojů pro jejich tvorbu, a to ***integrovaných do databázových systémů***, jako např. v MS SQL Serveru, nebo ***specializovaných nástrojů***, jako např. produkty Informatica apod.,
    - využití ***standardních analytických aplikací***, např. pro finanční analýzy, marketing apod.,
    - využití analytických funkcí integrovaných do jiných typů aplikací, např. do aplikací ERP, CRM, APS/SCM apod.,
- určení rozsahu ***outsourcingu*** v implementaci a provozu analytických řešení, určení podílu vlastních řešitelských kapacit a ***způsobu výběru*** dodavatelské společnosti,
- možnost využití ***specifických modelů*** zajištění implementačních a provozních kapacit, např. využití cloud computingu a s ním spojených služeb.

3.3.4 Identifikace potenciálních efektů podnikové analytiky

K tomu, aby ***potřeba aplikací*** podnikové analytiky byla kvalifikovaně posuzována, je nezbytné ***formulovat jejich potenciální efekty*** s ohledem na danou situaci firmy. Určování a ***posuzování efektů*** analytiky je v porovnání s ostatními typy aplikací poněkud ***specifické***. V každém případě je vymezení očekávaných efektů a sledování jejich naplnění podstatné s ohledem na to, že analytické přípravě a využití těchto aplikací musí ***věnovat čas na uživatelské straně manažeři*** a podnikoví specialisté, jejichž časové možnosti jsou vesměs omezené. Musí proto, pokud možno, přesně vědět, co jim takto vynaložená časová i finanční investice přinese.

Na druhé straně však efekty podnikové analytiky ***nemusí*** být vždy zcela ***přesně kvantifikovatelné***, resp. ve finančním vyjádření. Často se v těchto případech jedná o ***kvalitativní efekty*** znamenající dosažení vyšší konkurenceschopnosti firmy, získání lepší pozice na trhu, poskytování kvalitnějších informačních služeb obchodním partnerům atd. Je dobré si v tomto kontextu i položit ***otázku „jaký bude mít dopad na firmu situace, kdy nebude investovat do podnikové analytiky, zatímco konkurence ano?“***.

Navíc má dosažení těchto efektů ***delší časový horizont*** způsobený potřebnou dobou na vytvoření a naplnění datového skladu, vytvoření časových řad sledovaných ukazatelů, osvojení si náročnějších analytických aplikací uživateli apod.

3.3.5 Kompetenční centra

Úspěch řešení podnikové analytiky je silně ***závislý na efektivní kooperaci dodavatelů a uživatelů*** především v analytické fázi řešení. Průzkumy v české i zahraniční praxi však mnohokrát ukázaly, že právě ***nedostatečná kooperace*** a komunikace mezi byznysem a IT specialisty je zde překážkou kvalitnějších výsledků.

Jednou z cest je ***vytvoření kompetenčních center*** pro podnikovou analytiku, které byly i v praxi mnohokrát ověřeny. Jejich ***podstatou*** je to, že formálně organizačně a systematicky ***sdružují pracovníky uživatelských a IT útvarů***, kde společně řeší klíčové problémy a úlohy projektů podnikové analytiky. Jsou postaveny na jasně definované organizaci, pracovních procedurách, dokumentačních a dalších standardech a znamenají tak obvykle významný posun v racionalizaci řešení i provozu aplikací.

3.3.6 Využití cloud computingu

Koncept cloud computingu je charakterizován sdílením hardwarových a softwarových prostředků v infrastruktuře internetu. Služby cloud computingu jsou významné tím, že uživatelé platí pouze za užívání těchto kapacit a vstupní náklady jsou tak minimální.

Cloud podnikové analytiky, resp. business intelligence nabízí možnost ***soustředit se více na předmět podnikání*** a analýzy z byznysového hlediska než na správu IT infrastruktury. ***Cloud business intelligence*** lze vymezit jako ***model, ve kterém je alespoň některá komponenta BI řešení umístěna v cloudovém prostředí*** (tj. mimo infrastrukturu firmy). Komponenty, které se obvykle umisťují do cloudového prostředí jsou ETL/ELT pumpy, datový sklad a reportingové, analytické a další vizualizační nástroje (Borovec, 2020).

Rovněž lze na Cloud Business Intelligence pohlížet ***z hlediska modelů dodávky služeb***. V daném případě poskytování cloudových služeb v modelu software jako služba (***SaaS***) znamená nabízení ***předpřipravených BI software***, přičemž z firemního prostředí zákazníka je do cloudového řešení ***potřeba migrovat pouze data***. Při dodávání cloudového BI formou platformy jako služby (***PaaS***) se jedná o BI systém, jehož ***funkcionalitu si může zákazník upravit podle potřeby***, tak aby ukazatele z BI analýz mohly přesně měřit výkonost byznysu (tj. sledovat mimo jiné i naplnění takzvaných klíčových ukazatelů výkonnosti). V daném případě pak jde dodavateli o ***nabídnutí co nejširší funkcionality na svých platformách zákazníkům***. Poslední možností je ***poskytnutí pouhé infrastruktury*** jako služby (tj. hardware a software s využitím virtualizace, přičemž veškeré dané infrastrukturní součásti jsou spravované dodavatelem), na kterém ***si zákazník samostatně buduje své BI*** řešení.

Veškerá cloudová řešení v podnikové analytice, a tedy i v BI by měla být podložena smlouvou o dodávce služeb (***SLA***) uzavřenou mezi dodavatelem a odběratelem. Součástí by měla být ***definice způsobu platby za služby, pojištění*** při výpadku služeb včetně ***vymezené hierarchie priorit a kritičnosti*** pro zákazníka, tj. které systémy musí být provozuschopné a jaká je jejich kritičnost pro byznys včetně definic, co přesně pro zákaznickou firmu znamená provozuschopnost systémů a co má učinit dodavatel, když nebudou provozuschopné a jaké budou ***sankce***, pokud tak neučiní a jak se budou v čase sankce zvyšovat (Borovec, 2020).

3.4 Kvalita informačního systému

Skupina faktorů spojených s informačním systémem podniku představuje zejména ***hodnocení kvality jeho zdrojů***, resp. z čeho může řešení podnikové analytiky vycházet, tj. zejména úrovně řešení podnikových procesů, kvality dat a kvality podnikových aplikací. ***Spadají sem tyto faktory***:

- úroveň dokumentace a optimalizace podnikových ***procesů***,
- kvalita ***datových zdrojů***, včetně dostupnosti dokumentace datových zdrojů,
- kvalita ***podnikových aplikací***.

3.4.1 Úroveň dokumentace a optimalizace podnikových procesů

Procesní modely a jejich dokumentace má obvykle primární užití v transakčních systémech. V podnikové analytice je jejich význam ***dán těmito aspekty***:

- díky procesním modelům a reegneeringu podnikových procesů jsou ***přesněji definovány problémy a požadavky*** i na analytické aplikace, resp. tyto aplikace pak mohou přesněji odpovídat i potřebám podnikových procesů,
- celkové procesní modely podniků jsou ***základem pro přesnější a objektivnější specifikaci priorit*** v řešení projektů podnikové analytiky, resp. přesnější definici obsahu a pořadí jednotlivých případných přírůstků při přírůstkovém přístupu k řešení,
- současná řešení podnikové analytiky představují nejen jednotlivé analytické a plánovací aplikace, ale zahrnují i ***definování analytických a plánovacích procesů*** včetně jejich realizací v aplikacích workflow integrovaných do systémů podnikové analytiky a procesní podnikové modely pro ně vytvářejí potřebný základ,
- procesní modely jsou rovněž důležitým předpokladem pro ***uplatňování konceptu řízení podnikové výkonnosti*** – *Corporate Performance Management*, v němž podnikové procesy jsou jednou ze čtyř součástí.

3.4.2 Kvalita datových zdrojů

Význam kvality datových zdrojů pro řešení podnikové analytiky byl již několikrát zdůrazněn. Do této kapitoly je tento faktor zahrnut pouze jako konstatování jeho klíčového významu pro úspěšnost podnikové analytiky. ***Dílčí faktory*** ovlivňující datovou kvalitu lze vymezit v následujících ***třech skupinách***:

- ***technické prostředí*** zahrnující celopodnikový slovník dat, datový katalog, centralizaci aplikací a jejich datových zdrojů, např. jednotná identifikace zákazníků, kontroly definovaných business pravidel,
- úroveň použité ***metodiky***, tj. podnikové metodiky a směrnice, kvalita číselníků a kódových tabulek, systém řízení změn,
- ***přístupy k řešení*** informačního systému, tj. způsob přípravy dat, přípravy uživatelů, systém motivačních kritérií.

S tím souvisí ještě další podstatné aspekty, a to ***dostupnost dokumentace produkčních datových zdrojů***, případně možnost poskytování potřebných dat poskytovateli nebo provozovateli. V některých případech ***není zcela jednoduché tyto dokumentace, resp. data získat*** s ohledem na autorská práva nebo smlouvy mezi zákazníkem a poskytovateli těchto primárních systémů. V každém případě je dobré si tyto podmínky a možnosti ještě před zahájením projektu ověřit a podle možností je začít řešit.

3.4.3 Kvalita podnikových aplikací

Otázka kvality podnikových aplikací, zejména transakčních, je v souvislosti s projekty podnikové analytiky ***posuzována v několika úhlech pohledu***:

- do jaké míry jsou tyto aplikace ***schopné poskytovat úplná, konsistentní a přesná data***, tj. jak je navržena jejich ***datová základna***, jaký ***systém kontrol*** zahrnuje jejich ***funkcionalita***, jak odpovídají potřebám podniku z pohledu poskytovaných funkcí i vytvářených a zpracovávaných dat,
- zda zahrnují ***vlastní analytickou a plánovací funkcionalitu***, do jaké míry je využívána, zda je účelné ji nahradit funkcionalitou aplikací podnikové analytiky,
- zda v sobě integrují technologie a aplikace podnikové analytiky, jako je tomu např. u systémů ERP, u CRM s analytickými komponentami, tedy *customer intelligence*, nebo u aplikací APS / SCM pro analýzy a plánování provozu a rozvoje dodavatelských řetězců.

3.5 Závěry k faktorům ovlivňujícím podnikovou analytiku

- „***Faktor***“ představuje významné podmínky firemního prostředí ovlivňující způsob řešení rozvoje firmy, jednotlivých projektů, a nakonec i jejich i úspěšnost.
- Jedním z ***klíčových předpokladů*** pro úspěšné řešení a využití analytických aplikací je tak ***existence jejich potřeby*** z pohledu cílové skupiny uživatelů, tj. manažerů,

podnikových analytiků a specialistů. Tato potřeba je buď ***dána čistě odborným zájmem jednotlivců, nebo zájmem vedení společnosti*** na jejím celkovém úspěchu. Pokud vedení firmy nepovažuje analytické aplikace za účelné, nebo systém řízení firmy je založen spíše na citu a zkušenostech vedoucích pracovníků, pak je lepší projekty podnikové analytiky nezahajovat, nebo je přesunout na pozdější období podle vývoje situace.

- Jednotlivé faktory, zejména ***velikost a původ vlastnictví*** obvykle výrazně ***určují přístupy vlastníků a managementu k projektům podnikové analytiky***, které jsou obvykle pro konečnou úspěšnost projektů zásadní. Pokud není podnikové prostředí ze všech výše uvedených pohledů připravené ***analytické aplikace přijmout*** a reálně je využívat, pak je jejich řešení problematické a investice do nich nepřinášejí očekávané nebo vůbec žádné efekty. Prakticky všechny uvedené ***faktory je proto nezbytné důsledně analyzovat*** a posoudit především ve fázi plánování projektu a rozhodnutí o jeho zahájení, případně o rozsahu jeho řešení.
- Je rovněž účelné velmi dobře posoudit dopady ***rozdílů vyplývající z odvětvové orientace*** firmy na obsah řešení podnikové analytiky, právě pochopení a uplatnění rozdílů v obsahové náplni přináší často ***konkurenční výhody***, a tedy zásadní efekty vyplývající z těchto projektů.
- ***Klíčovým faktorem*** je zde „***Firemní kultura***“ určující prostředí a zájem a motivaci pracovníků firmy na jejím rozvoji, jejich přístup a iniciativu při řešení podnikové analytiky, postavení a zájem managementu na řešení takových projektů a další.
- Je účelné ještě před zahájením projektu ***dobře pochopit a vyhodnotit jednotlivé faktory*** a na základě takového vyhodnocení modifikovat celkovou strategii a přístup k řešení projektu.
- Na základě vyhodnocení identifikovaných faktorů je na začátku i účelné ***posoudit, zda*** řešení některých součástí podnikové analytiky ***má vůbec smysl***.

4. Metody a metodiky podnikové analytiky

Účelem kapitoly je:

- poskytnout ***vymezení vybraných metod*** uplatňovaných v řešení podnikové analytiky,
- vymezit alespoň hlavní charakteristiky především metod ***dimenzionálního a datového modelování***.

→ ***Celkový přehled a obsah všech metod a metodik*** zařazených do anatomie firmy obsahuje dokument „*AF II.02: Komponenty*“, a to

- Metodiky a metody v řízení firmy: ***kapitola 8***.
- Metodiky, metody a rámce řízení IT: ***kapitola 9.***
- Metodik a metody řešení IT: ***kapitola 10.***

..

Na úvod je třeba zdůraznit, že ***text této kapitoly vychází a navazuje na odpovídající část publikace*** „MARYŠKA, M., POUR, J., STANOVSKÁ, I., ŠEDIVÁ, Z.: Self Service Business Intelligence, Praha. Grada, 2018. ISBN 978-80-271-0616-5.“ S ohledem na již dřívější rok vydání této publikace a současně na skutečnost, že právě tyto metody představují určité ***jádro řešení*** podnikové analytiky pokládáme za účelné uvedenou část i do této publikace zařadit.

V oblasti podnikové analytiky existuje ***celá řada metodik obecného i čistě firemního charakteru***. Jako příklady je možné uvést metodiky společností Adastra, IBM, Oracle, Teradata a další. Tyto metodiky jsou velmi detailní a obvykle vázané na produkty a služby poskytované danými společnostmi. Nemělo by zřejmě smysl je zde opakovat, je k nim dostatek zdrojů a omezíme se pouze na zobecněný a stručný doporučený postup řešení projektu business intelligence

Bez ohledu na jejich povahu je nutné zmínit jejich existenci a zejména ***racionální způsob využití*** jako jednoho z faktorů kvality výsledného řešení. Důraz na uplatnění metodiky je ***dán následujícími aspekty***:

- na analýze i provozu analytických aplikací se většinou podílí ***řada pracovníků různých profesí*** a je proto účelné, aby v jednotlivých částech řešení uplatňovali, pokud možno ***společné, standardní postupy***,
- projekty podnikové analytiky jsou často ***velmi rozsáhlé***, pokrývají celé nebo velkou část řízení firmy, probíhají ***v delším časovém období***, a proto je nezbytné uplatňovat procesní, dokumentační a komunikační ***standardy definované metodikou*** a omezit tak běžné chyby v kooperaci, nebo nedorozumění mezi řešiteli a uživateli,
- ***uplatnění*** standardních metodik by však mělo být ***v praxi efektivní***, tj. mělo by podporovat a racionalizovat standardní operace a postupy, ale na druhé straně neomazovat potřebnou iniciativu a invenci jednotlivých řešitelů (v tom jsou tyto projekty rovněž specifické).

Další podkapitoly se váží pouze ke stručnému vymezení ***metody dimenzionálního a metody datového modelování*** v prostředí podnikové analytiky.

4.1 Dimenzionální modelování

Dimenzionální modely, zejména ***hrubý dimenzionální model*** by měly vytvořit potřebný základ pro návrh a implementaci analytických aplikací, a to s těmito vlastnostmi:

- jejich výsledná ***podoba je relativně jednoduchá***, založená na určité symetrii a standardech (fakta – dimenze), což umožňuje poměrně dobrou a rychlou orientaci v logice řešení jak informatikům, tak uživatelům,
- jednoduchost řešení nabízí i lepší možnosti ***zvyšování provozního výkonu*** aplikací,
- jednoduchost a přehlednost je rovněž základem i pro ***snadnější úpravy a rozšiřování obsahu*** řešení, tedy doplňování nových dimenzí, atributů apod.

Podstatou ***hrubého dimenzionálního modelu*** je ***vymezit obsah*** řešených BI a SSBI aplikací bez ohledu na jejich technickou realizaci v konkrétní databázi. Součástí řešení je i ***stanovení granularity***, tj. úrovně detailu sledovaných hodnot v analytické databázi. Ta by měla být co nejvyšší, aby bylo možné realizovat co nejpodrobnější analytické operace. Na druhé straně je nutné ***při dané granularitě odhadnout objem databáze a její očekávaný růst***. Znamená to ***určit formou běžného textu, nebo s použitím tabulek, matic, případně specifických schémat:***

1. ***dimenze***, s jejichž pomocí se budou ukazatelé analyzovat,
2. ***ukazatele*** (fakta, metriky), která budou v aplikaci sledována a analyzována,
3. ***vazby ukazatelů a dimenzí***, tj. který ukazatel se bude analyzovat podle kterých dimenzí.

V následujícím textu jsou uvedeny ***charakteristiky dimenzí i ukazatelů***, které má obsahovat hrubý dimenzionální model. I v tomto případě je nutné zdůraznit ***možnosti modifikace nebo redukce*** jeho obsahu. Řada charakteristik totiž vychází z potřeby sladění představ o výsledných řešení mezi širokou škálou uživatelů, což je typické zejména pro komplexní systémy Business Intelligence. Na druhé straně aplikace zaměřené pouze pro individuální potřeby jednotlivců rozsáhlou dokumentaci dimenzionálních modelů někdy nepotřebují. Je proto na úvaze konkrétních uživatelů i IT analytiků, v jakém rozsahu a podrobnosti budou hrubé dimenzionální modely tvořit.

V případě návrhu ***ukazatelů se určují tyto charakteristiky***:

- jednoznačná ***identifikace*** ukazatele – např. *Trzby_Objem* - pro účely definování vazeb ukazatele na jednotlivé dimenze a dalšího navrhování tabulek faktů a datových modelů,
- plný název ukazatele, např. *Tržby z prodeje v Kč*,
- obsahové ***vymezení a účel*** ukazatele – je dobré vymezit ukazatel co nejpřesněji a nejkonkrétněji tak, aby byl vytvořen podklad pro shodu mezi uživateli na jeho interpretaci,
- ***zdroje dat*** pro ukazatel - z jakých databází, nebo souborů lze hodnoty ukazatele získávat, případně je pořizovat manuálním způsobem, expertním odhadem apod.,
- ***kalkulace*** ukazatele – výpočty pro základní i související ukazatele,
- ***typ***, formát dat, např. numerické, textové,
- ***měrná jednotka*** – např. Kč, % apod.
- možnost ***agregace*** ukazatele, tj. aditivní (A), neaditivní (N), semiaditivní (S),
- ***KPI***, tj. zda ukazatel představuje klíčový indikátor výkonnosti (*Key Performance Indicator*), nebo klíčový cílový indikátor byznysu (*KGI, Key Goal Indicator*),
- ***vazby*** ukazatele na definované dimenze formou jednoduché matice.

V případě návrhu ***dimenzí se určují a dokumentují následující charakteristiky***:

- ***identifikace*** dimenze – podle stanovených dohodnutých standardů, např. *DI_Zbozi*,
- plný ***název*** dimenze, např. *Zboží prodávané firmou*,
- ***obsah***, tj. detailnější textové vyjádření obsahu dimenze tak, aby umožňovalo sjednocené její chápání mezi různými uživateli a analytiky,
- ***typ*** dimenze, tedy zda jde o dimenzi časovou, STAR, SNOWFLAKE, degenerovanou (viz dále),
- ***zdroj dat*** pro dimenzi, resp. její prvky, např. databáze *Zbozi*, tabulka číselníku v Excelu, textový soubor apod.,
- struktura ***prvků*** dimenze, resp. jejich hierarchie tak, aby bylo zřejmé, v jaké celkové struktuře budou prvky dimenze uspořádány, např.:

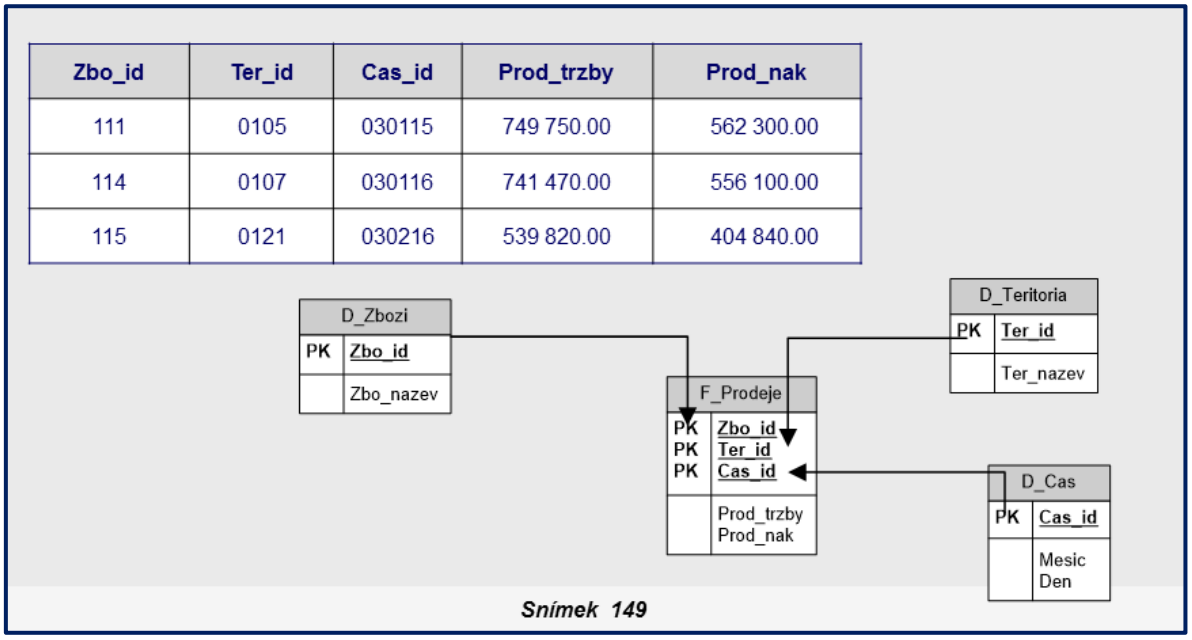
  Kategorie zboží – Skupiny zboží – Jednotlivá zboží, zbožové položky
- ***atributy*** dimenze, tj. struktura záznamu dimenzionálních tabulek,
- doplňující ***poznámky*** k vytvoření a využití dimenze, např. požadavky legislativy apod.

Speciální místo v modelu má ***časová dimenze***, tj. jaká bude struktura časových intervalů (roky, kvartály, měsíce), zda těchto časových struktur bude současně více, zda se bude k aktuálnímu datu nějakým způsobem měnit (např. na dekády, dny), zda se budou nějaké starší časové úseky přesouvat z provozního řešení do archivu (tzv. *aging*) apod.

4.1.1 Principy návrhu tabulek faktů

Základní principy návrhu tabulek faktů dokumentuje Obrázek 4-1. ***Sloupce*** tabulky faktů jsou vesměs buď ***klíčové atributy, resp. primární a cizí klíče*** (Zbo_id, Ter_id, Cas_id), ***a hodnoty*** ukazatelů (Prod_trzby, Prod_nak). ***Klíčové atributy, cizí klíče reprezentují jednotlivé dimenze*** a jejich hodnoty jsou prvky těchto dimenzí, např. 111 je prvkem dimenze Zboží.

Řádky v tabulce představují ve své podstatě ***jednotlivá měření*** (v obchodě, výrobě apod.) a většinou jsou přiřazovány ***co nejdateilněji***, tj. pouze na úrovni listů ve strukturách použitých dimenzí. Klíčové položky (listy ve struktuře dimenzí) jsou, s ohledem na nároky na paměťový prostor, vesměs celočíselného typu.



Obrázek 4-1: Principy návrhu tabulek faktů

Příklad dokumentující vazby tabulky faktů na tabulky dimenzí dokumentuje Obrázek 4-1, kardinalita vazeb tabulek dimenzí na tabulku faktů je vždy 1:N.

4.1.2 Granularita v tabulce faktů

Granularita určuje ***úroveň podrobnosti údajů, resp.*** faktů ***uložených v tabulce*** faktů. Granularita údajů v tabulce faktů je přímo ***závislá na počtu a úrovni podrobnosti dimenzí*** odpovídajících příslušné tabulce faktů. Pro řešení úrovně granularity existují některé ***obecná doporučení***:

- pokud to technické kapacity dovolují, měla by být data uložena ***s nejvyšší možnou granularitou,***
- data vstupující do datového skladu z různých zdrojů je účelné ***transformovat na stejnou nebo srovnatelnou granularitu.***

V případě druhého uvedeného doporučení existují v praxi určité problémy a s nimi spojené analytické úlohy. Typickým příkladem mohou být data získávaná z obchodních objednávek, faktur apod. Některé údaje (fakta) se mohou vázat k objednávce jako celku (tak zvaně k její hlavičce), např. náklady na dopravu, a některé k jednotlivým objednávaným zbožovým položkám (např. hodnota dodávky příslušného zboží). Jde evidentně o různou granularitu uvedených dat.

Úlohou analytika je ***převést tyto údaje na stejnou, vyšší granularitu***, tedy v našem případě rozpočítat náklady na dopravu na jednotlivé zbožové položky. Tato operace se v dimenzionálním modelování označuje jako ***alokace*** (nikoli pouze nákladů, jak je známá z ekonomických disciplín). Při řešení této úlohy je však nutné nejprve rozhodnout, zda taková alokace je možná a má smysl a současně, kdo z uživatelské (byznys) sféry se na formulaci pravidel pro alokaci bude podílet. Pokud se alokace ukáže jako neúčelná, pak souhrnné hodnoty (např. dopravné) musí vytvořit samostatnou agregovanou faktovou tabulku. ***Nedoporučuje se kombinovat fakta s různou granularitou do jedné faktové tabulky.***

4.1.3 Typy tabulek faktů

V datových skladech existují ***tři hlavní typy tabulek faktů*** vzhledem k jejich granularitě dat, viz další obrázek.

Charakteristika	Transakční	Periodická snímkovaná	Akumulovaná snímkovaná
Časová perioda	Časový okamžik (čas transakce)	Pravidelné, předem určené intervaly	Nedefinovaný časový rozsah
Granularita	1 záznam = 1 transakce	1 záznam = 1 časový interval	1 záznam (postupně aktualizovaný)
Plnění (load) tabulky faktů	Přidávání záznamů (Insert)	Přidávání záznamů (Insert)	Přidávání záznamů (Insert) a aktualizace (Update)
Aktualizace záznamů tabulky faktů	Nerealizuje se	Nerealizuje se	Realizuje se vždy při změně
Časová dimenze	Datum transakce	Datum konce časového intervalu	Více datumů pro standardní provádění změn
Fakta (ukazatele)	Obsah transakční aktivity	Obsah odpovídající definovanému časovému intervalu	Obsah odpovídající celému životnímu cyklu dat

Obrázek 4-2: Typy tabulek faktů

Transakční tabulky faktů jsou založeny na tom, že detailní informace vstupující do datového skladu jsou ***vázány na jednotlivé transakce*** a pohybují se na nejvyšší možné granularitě dat. Z toho vyplývá, že časový úsek nebude stejný, ale bude záviset na době výskytu jednotlivých transakcí. Transakční tabulky faktů patří v praxi k těm velmi často využívaným.

Periodické snímkované tabulky faktů jsou ***v praxi nejpoužívanější***. Data vstupují do datového skladu ***v pravidelných, předem definovaných časových úsecích*** (snímcích, např. dnech) a vyjadřují souhrnné hodnoty ukazatelů za celý časový snímek (např. celkový objem transakcí za daný časový úsek). Tento typ tabulek je nejvíce užívaný i pro odhadování, resp. predikci trendů vybraných ukazatelů.

Akumulované (též někdy stavové) snímkované tabulky faktů jsou rovněž ***závislé na výskytu transakcí,*** ale jejich hodnoty ***se v čase postupně aktualizují***. Například při postupném objednávání zboží na sklad se tak udržuje přehled o aktuálním stavu a vývoji dané objednávky. Tyto tabulky se v praxi využívají méně často než předchozí dva typy.

4.1.4 Měrné jednotky, rozsah, zdroje a kalkulace ukazatelů,

Tabulky faktů obsahují ukazatele, které ***potřebují různí uživatelé sledovat v různých měrných jednotkách***, např. počty vyrobených produktů v kusech, v tisících, krabicích, v paletách apod. Nabízejí se dvě možnosti, buď umístit jednotky a přepočítací koeficienty např. do produktové dimenze, nebo je umístit přímo do jednotlivých záznamů tabulky faktů. S ohledem na riziko chyb a možné změny v koeficientech se doporučuje využívat spíše druhou variantu, tedy ***umisťovat je do záznamů tabulky faktů.***

Tabulky faktů zabírají v datovém skladu obvykle kolem 90 % jeho celkové kapacity (oproti cca 10 % tabulek dimenzí). Je pro ně charakteristické, že tento rozsah je dán obrovským počtem jejich řádků, záznamů (např. každý prodej, každý telefonní hovor apod.). Na druhé straně je proto ***snaha omezit jejich rozměr co do počtu sloupců a rozsahu jednotlivých sloupců***. Dalším způsobem řešení je určení granularity dat podle období, např. pro posledních aktuálních 60 dnů se využije denní granularita tabulky faktů, pro starší období, pak granularita nižší.

Tabulka faktů obsahuje ***základní, elementární hodnoty ukazatelů vstupující ze zdrojových databází i hodnoty kalkulované***, tedy v tomto případě kalkulace v rámci jednoho záznamu (např. Prod_zisk = Prod_trzby – Prod_nak). Kalkulace se mohou provádět na úrovni ETL, datového skladu, resp. tržišť, nebo na úrovni analytických aplikací. Obvykle je ***užitečné u aditivních faktů ukládat kalkulované hodnoty přímo do datového skladu, r***esp. tržišť, neboť se tak zajistí dostupnost těchto dat všem uživatelům bez nutnosti kalkulace opakovat v různých aplikacích. Oproti tomu stojí někdy fakt, že výpočty v analytických nástrojích (Excel atd.) mohou být jednodušší.

Dimenzionální modely obsahují i tzv. ***tabulky faktů bez ukazatelů***, faktů (***factless fact table***), které nemají žádné ukazatele a využívají se např. pro ***zjišťování počtu určitých událostí***. To znamená, že každý výskyt záznamu ve fakt tabulce s daným klíčem indikuje vznik události, např. daná činnost je součástí procesu, zboží bylo zařazeno do marketingové akce apod. U nich lze pak sledovat souhrnné hodnoty pouhou sumarizací počtu záznamů v členění podle klíče.

4.1.5 Tabulky dimenzí – principy návrhu

Princip návrhu tabulek dimenzí dokumentuje Obrázek 4-3.

Produkt_Id	Skupina_Id	Produkt_Skupina	Kateg_Id	Produkt_Kateg	Nazev
111	11	notebook	1	počítač	Acer Travelmate 292
114	12	PDA	1	počítač	Pocket LOOX 410
116	21	monitor LCD	2	monitor	LG L1730S

Obrázek 4-3: Principy návrhu tabulek dimenzí

Tabulky dimenzí jsou de facto ***podnikové číselníky***, např. dimenze zboží, zákazníků, dodavatelů atd., a to se všemi možnostmi a problémy, které jsou s nimi spojeny. Dimenzionální tabulka obsahuje ***vedle klíčových (obvykle numerických) atributů další řadu většinou textových atributů*** popisujících podstatné charakteristiky jednotlivých produktů, zákazníků apod. Počet atributů je dán nároky na zpracování nejrůznějších podnikových reportů vycházejících z dat v datovém skladu. Rozdíl mezi dimenzionálními tabulkami a tabulkami faktů je vedle jejich samotného účelu v tom, že tabulky faktů mají relativně malý počet sloupců, ale enormní počet řádek, zatímco u tabulek dimenzí je to právě naopak.

Kvalita návrhu celého datového skladu je tak často dána zejména ***kvalitou návrhu dimenzí***.

Jedna řádka tabulky je vymezena pouze pro ***jeden prvek dimenze***. Každý řádek dimenzionální tabulky musí být identifikován svým primárním klíčem (v případě obrázku - *Produkt_Id*), který pak také reprezentuje ***vztah k tabulce faktů (ve vazbě 1:N),*** kde je cizím klíčem a zajišťuje podmínku spojení (*join*) mezi tabulkou faktů a tabulkou dimenzí. Musí tedy zachovávat pravidla referenční integrity. Jednotlivé další popisné atributy slouží jako výběrová kritéria v dotazech, jako obsah hlaviček v reportech a pro další operace v uživatelských aplikacích. Hodnoty dalších ***atributů*** by měly být ***převážně textové a diskrétní*** a měly by spíše obsahovat plné a jasné vyjádření dané charakteristiky ***s co nejmenším používáním různých kódů a zkratek***.

Do atributů dimenzionální tabulky se řadí i některé numericky vyjádřené charakteristiky, např. velikost zboží, ale v kontextu tabulky mají čistě popisný charakter. Při návrhu datového skladu je často problémem, ***zda příslušný numerický atribut má být součástí určité tabulky faktů nebo dimenzí***. Řešení spočívá v zodpovězení otázky, jestli takový atribut nabývá velkého množství hodnot, a tedy se i často mění, jestli se podílí na běžných kalkulacích, v tom případě patří do tabulky faktů. Pokud jde o atribut, který nabývá diskrétních hodnot, relativně konstantních a používá se převážně pro výběrové operace v dotazech, pak patří do dimenzionální tabulky. Příkladem může být již zmíněná cena zboží, pokud je relativně stálá, často se nemění, pak je účelné ji zařadit do dimenzionální tabulky, pokud se mění téměř s každým prodejem, patří do jednotlivých řádků prodejů v tabulce faktů.

Další běžnou otázkou při řešení dimenzionálního modelu je to, ***zda využívat více jednodušších dimenzí, nebo méně dimenzí, ale složitějších a komplexnějších***. Příkladem mohou být jedna dimenze geografická, resp. teritorií a vedle toho jedna dimenze jednotlivých prodejen. Oproti tomu stojí varianta jedné společné dimenze prodejen, kde na vyšší úrovni budou teritoria, tedy státy, regiony apod. a na nejnižší úrovni jednotlivé prodejny podle jejich umístění v regionech.

Varianta více jednodušších dimenzí má tyto výhody:

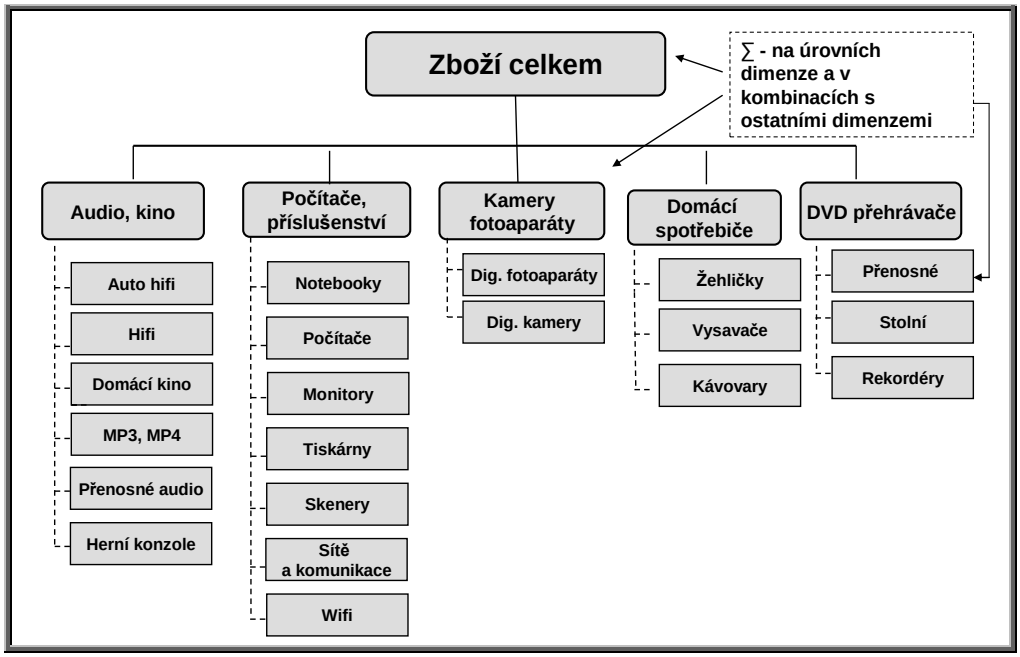
- je jednodušší pro pochopení uživateli,
- je jednodušší jejich správa a provádění úprav,
- nabízí více možných kombinací mezi různými dimenzemi v analytických aplikacích.

Oproti tomu ***varianta méně dimenzí, ale komplexních je výhodná v tom, že***:

- zjednodušuje a zpřehledňuje celý dimenzionální model, dimenzí je méně,

- je efektivnější při prohlížení celé její struktury a ukazatelů, k nimž se tato dimenze váže.

Podstatnou charakteristikou dimenzionálních tabulek je také především to, že ***data v dimenzích jsou hierarchicky strukturovaná*** tak, aby bylo možné na základě těchto struktur získávat agregované hodnoty ukazatelů v připojených tabulkách faktů. Ve své podstatě struktury dimenzí vyjadřují i struktury řízení v podniku.



Obrázek 4-4: Struktura dimenze *Zboží*

Pro vyjádření hierarchie existují ***dvě možnosti její realizace – STAR a SNOWFLAKE***.

4.1.6 Dimenze - STAR

U STAR schématu jsou v základní dimenzionální tabulce ***zahrnuty i všechny další sloupce pro nadřízené (parent) úrovně v hierarchii*** (kategorie zboží - *Zbo_Kategorie* a skupina zboží - *Zbo_Skupina*, *resp. identifikátory prvků nadřízených úrovní*). To představuje poměrně ***vysokou redundanci dat,*** kdy se hodnoty atributů nadřízených úrovní hierarchie vícenásobně opakují (Obrázek 4-5).

Zbo_id	Zbo_Kategorie	Zbo_Kat_Nazev	Zbo_Skupina	Zbo_Sku_Nazev	Zbo_Nazev
501	1	Audio, kino	101	Auto hifi	Autoradio Logik
502	1	Audio, kino	101	Auto hifi	Autoradio LG LAC3800
503	1	Audio, kino	101	Auto hifi	Autoradio Pioneer
504	1	Audio, kino	101	Auto hifi	Autoradio Logik
505	1	Audio, kino	102	Hifi	Mikro systém Hitachi AXM717
506	1	Audio, kino	102	Hifi	Mikro systém Panasonic SCPM45
507	1	Audio, kino	102	Hifi	Mikro systém Sony CMTEH25
508	2	Počítače	201	Notebooky	Acer Aspire One A150
509	2	Počítače	201	Notebooky	HP Compaq 2133
510	2	Počítače	201	Notebooky	Asus X51L
511	2	Počítače	203	Monitory	LG W1934S BN
512	2	Počítače	203	Monitory	ASUS VW193B
513	4	Domácí spotřebiče	401	Žehličky	Philips GC2528
514	4	Domácí spotřebiče	401	Žehličky	Bosch 250
515	4	Domácí spotřebiče	402	Vysavače	Zelmer 3500
516	4	Domácí spotřebiče	402	Vysavače	Zelmer 5000
517	4	Domácí spotřebiče	403	Kávovary	Espresso Delonghi ESAM2600
518	4	Domácí spotřebiče	403	Kávovary	Espresso Delonghi E2100

Obrázek 4-5: Dimenze STAR

Pro STAR schéma jsou z hlediska provozu a uživatelských aplikací ***podstatné tyto charakteristiky***:

- je evidentně ***rychlejší v době odezvy*** pro poskytování výstupů, neboť odpadají operace spojování (join) mezi tabulkami jednotlivých úrovní a stačí zpravidla jedno spojení mezi tabulkou

faktů a dimenzí, proto se využívá u těch komponent analytiky bezprostředně vázaných k vlastnímu využití (datová tržiště, datové sety v Power BI apod., viz dále),

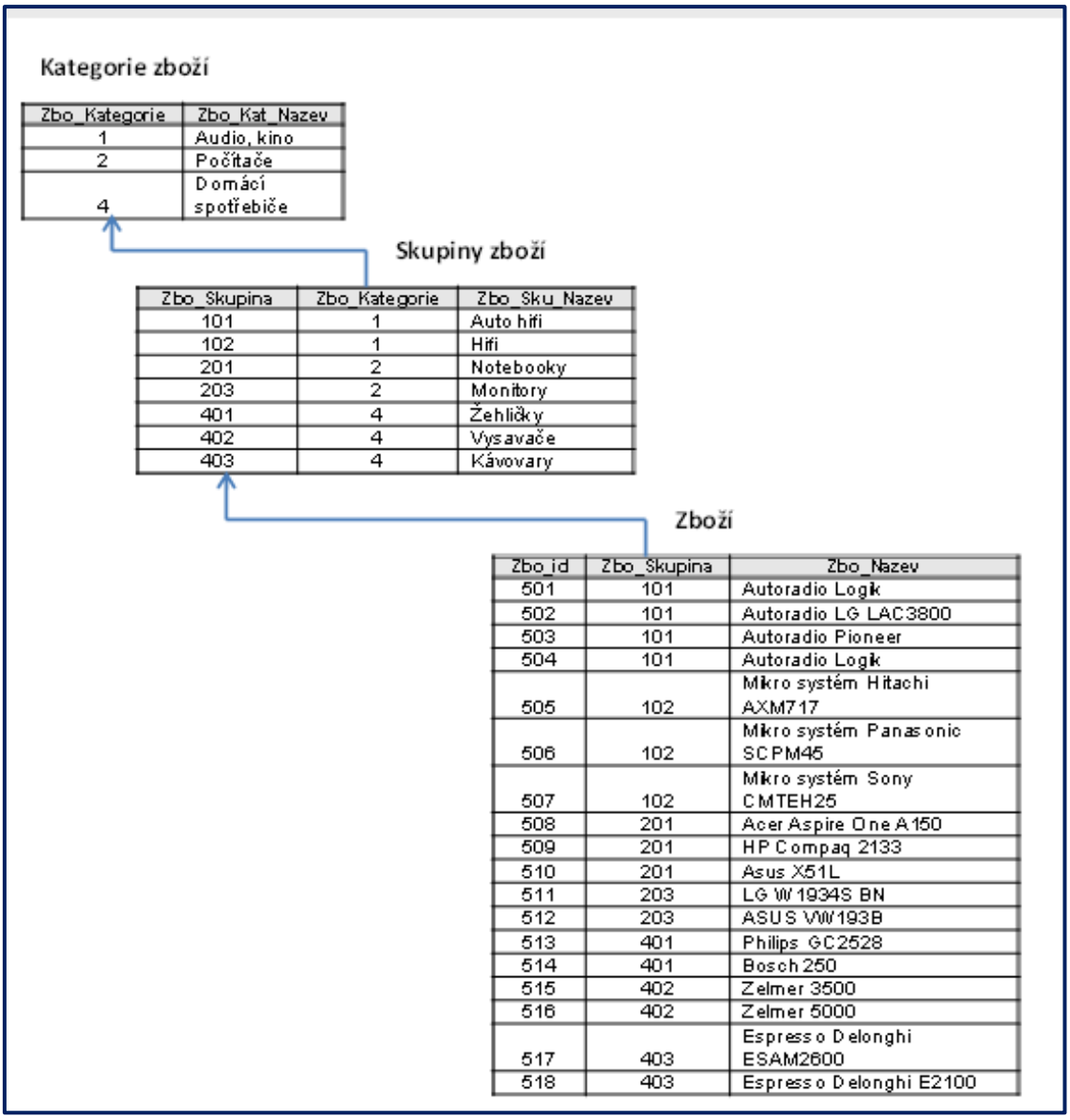
- umožňuje ***jednodušší prohlížení (browsing) dimenzí***, a zadávání filtrů pro všechny hierarchické úrovně dimenze,
- je však ***neefektivní při častých změnách*** v hierarchiích prvků dimenze, neboť jedna a tatáž změna se musí promítnout do mnoha řádek tabulky.

4.1.7 Dimenze - SNOWFLAKE

Hierarchie ve SNOWFLAKE schématu je ***založena na řetězci provázaných tabulek*** vždy s kardinalitou ***1:N*** pro dvě související úrovně hierarchie v dimenzi (v našem případě *Zbo_Zbozi – Zbo_Skupina* a *Zbo_Skupina – Zbo_Kategorie*). To znamená, že ***došlo k normalizaci dat*** v tabulkách, redundance dat je minimální, viz Obrázek 4-6.

Pro SNOWFLAKE schéma jsou ***podstatné tyto charakteristiky***:

- díky normalizaci dat je toto řešení ***výhodné při častých změnách*** v dimenzích a v hierarchické struktuře jejich prvků,
- vede ***k úspoře místa*** v databázi datového skladu, což je ale v důsledku nízkého objemu dat v dimenzionálních tabulkách v relaci k objemu dat v tabulkách faktů často minimální a z hlediska celkového řešení skladu nevýznamný faktor,
- umožňuje využívat ***prostředky pro vynucení referenční integrity*** mezi jednotlivými úrovněmi tabulek v hierarchii dimenze,
- poskytuje výhody pro ***efektivní tvorbu agregačních tabulek***,
- je ***méně přehledné*** než schéma STAR,
- realizace ***spojení tabulek (joinů) je složitá a komplexní,*** a i v současných databázových systémech i časově náročná.



Obrázek 4-6: Dimenze SNOWFLAKE

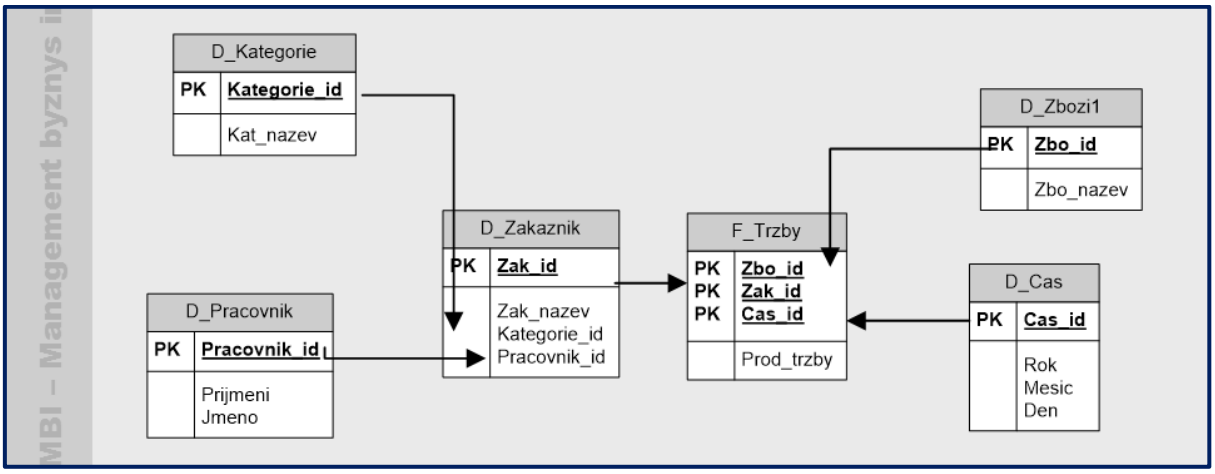
V rámci jednoho datového skladu a jednoho schématu mohou být ***současně definovány některé dimenze ve schématu STAR, další tabulky dimenzí ve schématu SNOWFLAKE*** a záleží pouze na efektivnosti řešení jednotlivých dimenzí. V dalším textu se budeme zabývat různými variantami řešení dimenzí.

4.1.8 Referenční dimenze a vazby M : N

V terminologii dimenzionálního modelování se pro ***dimenzi, která se na tabulku faktů odvolává prostřednictvím jiné dimenze,*** používá termín ***referenční dimenze***, nebo referenční dimenzionální tabulka. V našem případě je referenční dimenzí *D_Zakaznik*, viz Obrázek 4-7.

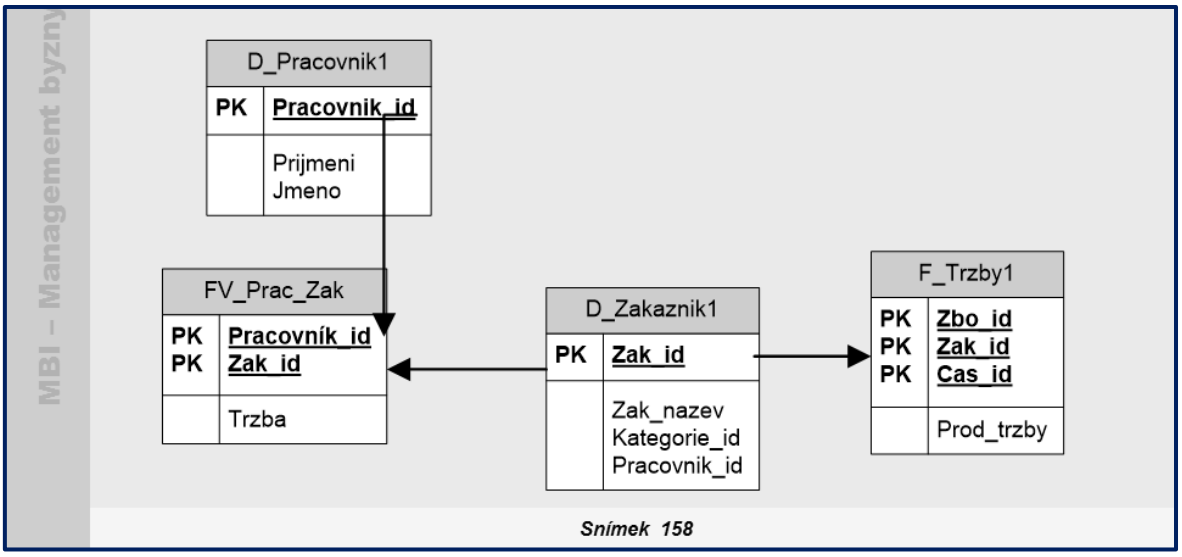
Tyto vztahy nemusí být vždy tak jednoduché, vyjadřující hierarchii určité dimenze, ale mohou ***řešit relevantní vazby i na jiné objekty*** a mohou být i ***více rozvětvené***. Příklad na obrázku ukazuje, kdy na tabulku faktů sledující *Tržby od zákazníků* je vázána dimenze Zákazník a na ni další dvě dimenze *Kategorie zákazníků* a *Pracovníci* (mající daného zákazníka na starosti).

Z obrázku vyplývá, že jsme schopni sledovat a agregovat tržby podle kategorií zákazníků a v rámci toho podle jednotlivých zákazníků, nebo pracovníků zodpovědných za zákazníky a v rámci toho podle jednotlivých zákazníků. Toto řešení je postaveno na předpokladu kardinality vazeb mezi nadřazenou a podřízenou dimenzí ***1:N (One-to-Many).*** To je zcela obvyklé např. u vztahu *Kategorie zákazníka – Zákazník*.



Obrázek 4-7: Referenční dimenze

U vztahu *Pracovník – Zákazník* to ale tak nemusí být. Se zákazníkem může pracovat více pracovníků a pak se kardinalita vztahu změní na **M*:N (Many-to-Many).*** Řešením je využití vazební (pracovní, *bridge*) tabulky faktů řešící dílčí tržby zákazníků s vazbou na jednotlivé pracovníky (pokud je samozřejmě věcně možné tržby na jednotlivé pracovníky alokovat), viz Obrázek 4-8.



Obrázek 4-8: Vazby dimenzí M : N

4.1.9 Degenerované dimenze

V reálných řešeních existují i tzv. ***degenerované dimenze*****,** tzn., že existuje ***dimenze pouze na základě příslušného atributu v tabulce faktů a nemusí pro ni existovat dimenzionální tabulka***. Příklady jsou dimenze *Objednávky, Faktury,* jejichž prvky jsou určeny pouze číslem objednávky nebo faktury bez slovního vyjádření. To znamená, že číslo objednávky, faktury apod. je součástí struktury tabulky faktů bez vazby na dimenzionální tabulku. ***Analýzy lze tak realizovat i podle těchto identifikací dokumentů,*** a tedy i dopovídajících transakcí. Navíc je možné, v případě potřeby, následně získávat na jejich základě i další potřebné informace z OLTP systémů, tj. k jednotlivým objednávkám atd. Prvky degenerovaných dimenzí se mohou účelně využívat v kombinaci se standardními dimenzemi. Např. číslo nákupu (nákupní transakce) v maloobchodě představuje degenerovanou dimenzi, která se často váže na identifikaci pokladny (*POS, Point of Sale*), která je standardní dimenzí.

V případě ale, že ***atribut a současně primární klíč takové dimenze je příliš rozsáhlý*** (např. u čísla faktury), nebo je třeba s dokumentem (fakturou, objednávkou apod.) vést další atributy, pak se

původní klíč nahrazuje umělým klíčem a příslušnou dimenzionální tabulkou a pak se charakter degenerované dimenze logicky ztrácí.

4.1.10 Dimenze parent-child

Existují a využívají se i tzv. ***parent-child*** dimenze, což je např. ***dimenze zaměstnanců***, kde:

- na úrovni listu je vždy pouze jeden *zaměstnanec*,
- na vyšší úrovni je *manažer*, který je většinou nadřízený pro několik zaměstnanců, ale ten se odkazuje zpět na jednoho konkrétního zaměstnance – manažera.

4.1.11 Klíče, umělé klíče

Pro identifikaci jednotlivých prvků i vyšších úrovní v dimenzi se doporučuje využívat především ***umělých***, systémem ***automaticky generovaných klíčů*** (*surrogate key*) oproti ***operačním,*** ve zdrojových ***transakčních systémech*** využívaným klíčům. Tyto umělé klíče tedy slouží jako primární klíče v tabulkách dimenzí a jako cizí klíče v tabulkách faktů, a tak současně k řešení vazeb (*join*) mezi tabulkami faktů a tabulkami dimenzí. V tabulkách faktů tyto cizí klíče vytvářejí složený primární klíč tabulky faktů.

Důvody pro umělé primární klíče v dimenzích jsou:

- umělé klíče ***odstiňují datové sklady a tržiště od změn klíčů*** ve zdrojových databázích, jsou na těchto změnách nezávislé,
- umožňují ***efektivnější a kvalitnější konsolidaci dat*** v situacích, kdy operační klíče více transakčních systémů se překrývají, nebo jsou vzájemně nekonsistentní,
- jejich řešení je ***jednodušší, menší rozsahem*** (obvykle typu integer, 4 byty) a přispívají tak i k vyššímu výkonu datového skladu,
- i při malém rozsahu (4 byty) nabízejí ***identifikaci pro cca 2 miliardy záznamů,***
- ***úspory v rozsahu*** jsou zejména patrné u složených klíčů ve faktových tabulkách (s miliardami záznamů), kde celý rozsah klíče je dán n x 4 byty, kde n je počet cizích klíčů (tj. dimenzí), oproti podstatně rozsáhlejším operačním klíčům transakčních systémů,
- mohou ***vyjádřit i specifické stavy prvků*** dimenze (např. „není známo“ apod.),
- pro časové dimenze a při dodržení standardního pořadí, očíslování časových úseků nabízejí pak ***lepší uspořádání dat ve fakt tabulce*** a lepší možnosti pro rozdělení faktové tabulky na části, resp. partitioning,
- protože datový sklad nebo tržiště obsahují historická data, ***u operačních klíčů*** se může stát, že budou po určitém období v transakčních systémech ***restartovány,*** a to může evidentně způsobit problémy,
- jsou výhodné ***pro řešení změn v dimenzích, SCD (Slowly Changing Dimension)***.

Umělé klíče ***se obvykle vytvářejí v pracovní oblasti DSA*** a současně se zde řeší i vazby dimenzionálních a faktových tabulek. V případě, že složené klíče ve faktových tabulkách nezajistí jednoznačnost záznamů (např. při řešení obchodních operací v maloobchodě), pak se tato ***jednoznačnost zajišťuje umělým klíčem*** faktové tabulky, nebo kombinací složeného klíče a časového razítka, doplněním čísla obchodní transakce, faktury apod., tedy degenerovanou dimenzí.

Přirozené klíče (nikoli umělé) se mohou využít pouze v případech klíče času pro datum (např. 20120910 vyjadřující strukturu data ve tvaru *rrrrmmdd*), pro klíče identifikující intervaly hodnot umožňující lépe vyjadřovat jednotlivé intervaly a pro klíče sběrných (*junk*) dimenzí.

4.1.12 Chyby a NULL hodnoty v klíčích

V některých případech se lze setkat s výskytem NULL hodnot v cizích klíčích. Tuto možnost nabízí následně ***řešení OLAP kostek, které se tak mohou s NULL hodnotami v cizích klíčích vyrovnat***. Na druhé straně se velmi silně doporučuje se NULL hodnotám v cizích klíčích vyhnout. Je v tomto případě vždy efektivnější příslušnou hodnotu do dimenze doplnit a zajistit tak korektní vazbu mezi

dimenzionální a faktovou tabulkou. Souhrnně je tak nutné řešit veškeré možné chyby v konsistenci dat a vazbách mezi primárními a cizími klíči.

4.1.13 Alternativní struktury dimenzí

Pohyb v datech tabulky faktů podle hierarchie prvků v příslušné dimenzi znamená pohybovat se po různé úrovni detailu dat. Jde tedy o operaci drill down / drill up. V rámci jedné dimenze se mohou tvořit tzv. ***alternativní struktury***, tj. např. v časové dimenzi (*rok – měsíc – den*, *rok – kvartál – den* apod.), pro různé organizační struktury v rámci jednoho podniku apod. V tomto případě se uživateli nabízí více možností, jak se pohybovat po datech v různých strukturách hierarchie. ***Realizace řešení spočívá v tom, že:***

- nejnižší úrovně hierarchie – listy obsahují sloupce pro identifikátory, resp. klíče pro více nadřízených struktur,
- agregační operace se pak realizují rozdílně pro různé hierarchické struktury dimenze.

4.1.14 Dimenze času

Dimenze času je v řešeních BI jejich ***standardní součástí***. Prakticky vždy se vývoj ukazatelů sleduje v čase, a proto musí být tato dimenze prakticky vždy definována. Obvykle se ***generuje z podnikového kalendáře,*** nebo se vytváří manuálně, a to tak aby pokryla existující data v potřebném rozsahu zpět a s potřebným počtem let dopředu (např. 10 let).

Příkladem je vyjádření časové dimenze *20120610*, která jednoznačně vyjadřuje strukturu *rok – měsíc – den*).

Z praktických důvodů ***se obvykle dělí na dvě dimenze***:

- dimenzi data (*date dimension*) - rok, kvartál, měsíc, den apod.,
- dimenzi času dne (*time-of-day dimension*) – hodina, minuta, příp. sekunda.

Není účelné tyto dvě dimenze spojovat dohromady, neboť dimenze času dne má naprosto standardní strukturu a pro každý den by se tedy její řádky stále opakovaly. To znamená, že každý den v této dimenzi by při rozpadu jen na minuty představoval 1440 řádků, pro jeden rok by to pak bylo 525 600 řádků navíc.

V dimenzi data se jako atributy, kromě celočíselného primárního klíče, ***definují jednak běžné části data podle stanovené struktury***, např. číslo roku, měsíce, dne. Využívají se i ***další atributy*** jinak označující stanovené časové jednotky, a to ***pro reporty a různé výpočty***, např.:

- název měsíce, dne (např. březen, středa apod.),
- pořadí dne v týdnu (1 – 7), v měsíci, v kvartálu, v roce pro účely porovnávání stejných dnů,
- číslo kvartálu,
- datum fiskálního roku (kvartál, měsíc, den),
- identifikátor všedního dne, soboty, neděle, svátků,
- časová pracovní kapacita dne,
- určení sezónního dne, např. prodejní sezóny,
- případně další.

4.1.15 Dimenze různých rolí

Různé role jedné dimenze (*Role-Playing Dimension*) se využívají v těch situacích, kdy ***potřebujeme jednu a tutéž dimenzi využívat ve vztahu k tabulce faktů s různým kontextem,*** resp. různým obsahovým zaměřením. Příkladem může být uplatnění dimenze času (viz kapitola 4.3.11) ve vztahu k tabulce faktů skladových zásob. V tomto případě můžeme v této tabulce disponovat různými časy nebo daty, např. datem objednání na sklad, datem přijetí na sklad, datem výdeje ze skladu apod. (viz historizace dat,). Pracujeme zde ***s několika atributy data a v různém kontextu***. Jedním z možných řešení je vytvořit fyzicky různé časové dimenze, což je ale neefektivní a náchylné k chybám, vzhledem k zachování konsistence jednotlivých tabulek. Druhou možností je ***na základě jedné fyzické***

dimenzionální tabulky vytvořit různé databázové pohledy, s respektováním různých názvů položek (pro uplatnění v různých reportech) a ty pak připojovat k faktové tabulce podle aktuálního kontextu. Např. pro výše zmíněnou tabulku zásob může vytvoření pohledu nad časovou tabulkou *DATUM* vypadat takto:

CREATE VIEW DATUM_PRIJMU (DATUM_PRIJMU_KLIC, DATUM_PRIJMU_ROK, DATUM_PRIJMU_MESIC, DATUM_PRIJMU_DEN)

AS SELECT DATUM_KLIC, DATUM_ROK, DATUM_MESIC, DATUM_DEN.

K tabulce faktů bude pak jako dimenze připojen tento pohled a do hlavičky příslušných reportů se vloží titulky odpovídající datu příjmu zboží na sklad. Obdobně lze definovat pohledy i pro další data spojená s průběhem zboží na skladě. Takto využitá dimenze v různých rolích se označuje jako ***Role-Playing Dimension.***

4.1.16 Sběrná dimenze

Ve zdrojových systémech se často pro rozlišení ukazatelů využívají ***nejrůznější kódy a příznaky*** obsahující vždy pouze několik znaků. Příkladem může být způsob platby (hotově, kartou, šekem), plánovaná, nebo skutečná hodnota ukazatele apod. Jednou z možností je ***pro každý z těchto příznaků vytvořit samostatnou dimenzi*** (pak narůstá jejich počet), ponechat je ve faktové tabulce (narůstá rozsah záznamů), ***nebo z nich vytvořit společnou, sběrnou dimenzi (junk dimension),*** která bude v jednotlivých záznamech obsahovat kombinace příznaků, kde každý záznam s takovou kombinací bude mít svůj umělý klíč. Příklad dokumentuje Obrázek 4-9.

Klíč	Plán_Skut	Způsob platby
1	Skutečná	Hotově
2	Skutečná	Kartou
3	Skutečná	Šekem
4	Plánovaná	Hotově
5	Plánovaná	Kartou
6	Plánovaná	Šekem

Obrázek 4-9: Sběrná dimenze

Na celou kombinaci se pak záznamy ve faktové tabulce odkazují ***příslušným klíčem***. To znamená, že v našem případě nahradí ve faktové tabulce dva cizí klíče pouze jeden, v případě 5 různých příznaků nahradí 5 různých cizích klíčů rovněž 1 apod.

Při větším počtu různých kódů a příznaků a jejich hodnot samozřejmě narůstá počet kombinací a záznamů ve sběrné dimenzi. Při jejím vytváření je proto účelné ***do ní zařazovat pouze smysluplné kombinace*** a případně je pak doplňovat podle skutečných výskytů. Již vytvoření takové dimenze je užitečnou analytickou úlohou, která určuje, co s čím (jaké kódy) má smysl kombinovat, a navíc výsledné řešení datového skladu zjednodušuje. Tím se tak zvyšuje i orientace uživatele v uspořádání datového skladu a jeho využívání.

Dalším účelným způsobem využití sběrné dimenze je v případě ***doplňování poznámek a komentářů k jednotlivým ukazatelům ve faktové tabulce.*** Kombinace těchto komentářů mohou být rovněž založeny na principu sběrné dimenze s tím, že jednou z hodnot musí být vždy *Bez komentáře*.

Na druhé straně je se sběrnými dimenzemi spojeno i určité ***riziko, kdy při narůstajícím počtu výskytů jednotlivých znaků*** a tím i jejich kombinací lze dojít k závěru, že sběrnou dimenzi bude třeba ***rozdělit do více základních***. Pak to obvykle znamená i přepracování struktury datového skladu a poměrně vysokou pracnost takového řešení.

4.1.17 Změny v dimenzích - SCD (Slowly Changing Dimensions)

Jedním ze závažných problémů, které se musí v BI, resp. SSBI řešit, jsou ***změny v dimenzionálních tabulkách.*** Tyto změny se musí ***rozlišovat dle významu*** zejména z pohledu analýz dle dimenze času. Některé změny v dimenzích mají ***pro hodnocení vývoje a srovnatelnost sledovaných ukazatelů zásadní význam,*** některé minimální nebo žádný. Principy řešení změn v dimenzích se označují termínem ***SCD - Slowly Changing Dimensions***.

SCD – Slowly Changing Dimensions představuje ***změny ve struktuře a prvcích dimenzí, tedy číselníků v čase,*** přičemž při řešení úloh BI a SSBI existuje reálná potřeba zachovat konzistenci dat z časového hlediska. V tomto kontextu existují následující ***možné změny v dimenzích***, tj. jejich prvků, příp. vyšších úrovní v rámci hierarchie dimenze:

- ***přidání*** nových prvků (např. nová zboží, zákazníci apod.),
- ***zrušení*** prvků (např. zrušení zboží),
- ***změny hodnot atributů*** (názvy produktů, jména pracovnic, adresy zákazníků nebo dodavatelů apod.),
- ***změny ve struktuře*** – zatřídění (zboží, pracovníků) do jiných skupin v rámci hierarchické struktury prvků dimenze.

Při řešení problémů spojených se změnami obsahu dimenzí je třeba ***počítat se složitějším návrhem analytické databáze,*** s většími nároky na paměťový prostor dimenzí i s prodloužením běhu datových transformací.

Historii v datech lze uchovávat dvojím způsobem, a to ***v rámci tabulky faktů***, nebo tzv. ***historickými tabulkami dimenzí.*** Řešení historie prostřednictvím tabulky faktů je založeno na principu, kdy se měnící se atribut přesune z tabulky dimenze (číselníku) do tabulky faktů. To je výhodné v těch případech, když je takový atribut součástí dalších kalkulovaných ukazatelů, jako je např. prodejní plocha, cena apod.

V dalším textu se budeme zabývat řešením změn přímo v dimenzionálních tabulkách. To je založeno ***na různých přístupech k řešení***, označovaných v dimenzionálním modelování jako ***typ*** *(Type)* a ***číslo typu***, např. *Typ 1*. V každém případě ***řešení změn*** musí být proaktivní, a tedy již ***na úrovni hrubého dimenzionálního modelu*** musí být pro každý atribut dimenzionální tabulky uveden typ řešení jeho změn.

Typ 0 určuje, že změny jsou ***ignorovány*** a do datového skladu se nepromítají. Skutečné promítání změn dimenzí se rozlišuje několika dalšími typy řešení.

4.1.17.1 Typ 1 – Přepis hodnoty atributu

Řešení *Typ 1* je nejjednodušší, neboť ***pouze nahrazuje původní hodnotu atributu hodnotou novou***, změněnou ve stejné řádce dimenzionální tabulky. Předpokládejme, že máme dimenzionální tabulku *Zboží*, kde každý záznam má mimo jiné atributy *Primární klíč* (umělý klíč), *Číslo zboží* (operační klíč z transakčního systému), *Zodpovědný pracovník* (pracovník, který kompletně zodpovídá za prodej daného produktu), viz Tabulka 4-1.

Tabulka 4-1: Původní záznamy dimenzionální tabulky zboží

Primární klíč	Název zboží	Číslo zboží	Číslo skupiny zboží	Název skupiny	Zodpovědný pracovník
111	Acer Travelmate 292	NT-0015	01	Notebook	J. Malá
112	IBM ThinkPad R40	NT-0023	01	Notebook	P. Novák

Pokud dojde např. ke změně zodpovědného pracovníka a místo J. Malé bude produkt NT-0015 zajišťovat F. Klas, pak se tabulka změní pouze přepisem původního jména novým, viz Tabulka 4-2.

Tabulka 4-2: Dimenzionální tabulky zboží s aktualizovaným jménem zodpovědného pracovníka

Primární klíč	Název zboží	Číslo zboží	Číslo skupiny zboží	Název skupiny	Zodpovědný pracovník
111	Acer Travelmate 292	NT-0015	01	Notebook	F. Klas
112	IBM ThinkPad R40	NT-0023	01	Notebook	P. Novák

Pro změny v dimenzi **typu 1** ***je charakteristické***:

- ***nebyl změněn žádný klíč*** (primární, ani operační), a to ani v dimenzionální, ani ve faktové tabulce,
- jde o ***nejjednodušší*** způsob řešení změn dimenzí,
- ***faktová tabulka se nijak nemění***,
- na základě uplatnění typu jedna se však ***ztrácí historie sledovaných ukazatelů*** podle změn v atributech (nevíme, za jaké prodeje byla ještě zodpovědná J. Malá),
- typ 1 je ***vhodný*** především tam, kdy ***jde o opravu chybných hodnot atributů***, nebo původní hodnoty nemají dle požadavků uživatelů na výstupy nebo aplikace žádný význam,
- všechny ***předchozí výstupy***, přehledy, agregace ***se musí znovu vytvořit*** s novou hodnotou atributu (tam, kde je daný atribut ve výstupech uplatněn).

4.1.17.2 Typ 2 – Přidání nové řádky do dimenzionální tabulky

Typ 2 je ***nejčastěji používaný způsob*** řešení změn dimenzí. Znamená ***přidání nového řádku do dimenzionální tabulky***, který obsahuje změněnou hodnotu atributu, případně atributů. Tím se zachovají historické údaje, ale narůstá objem dimenze. V tomto případě ***záznam dimenze obvykle obsahuje***:

- ***nový primární klíč*** pro nový, resp. doplněný záznam,
- ***operační klíč*** (původní z transakčního systému), který se se změnami nemění, tj. v novém záznamu obsahuje stejnou hodnotu,
- ***datum počátku*** platnosti záznamu (DATUM_OD),
- ***datum konce*** platnosti záznamu (DATUM_DO) – u aktuálních záznamů se vyplňuje dostatečně vysokou konstantou (např. 1.1.2200).

Způsob řešení předchozí změny u typu 1, dokumentuje pro typ 2 Tabulka 4-3.

Tabulka 4-3: Změna dimenzionální tabulky zboží pro typ 2

Primární klíč	Název zboží	Číslo zboží	Číslo skupiny zboží	Název skupiny	Zodpovědný pracovník	Datum od	Datum do
111	Acer Travelmate 292	NT-0015	01	Notebook	J. Malá	1.1.2011	5.5.2011
321	Acer Travelmate 292	NT-0015	01	Notebook	F. Klas	6.5.2011	1.1.2200
112	IBM ThinkPad R40	NT-0023	01	Notebook	P. Novák	1.1.2011	1.1.2200

Pro změny v dimenzi **typu 2** je, kromě již uvedeného způsobu realizace zněny, ***charakteristické***:

- ***data ve faktové tabulce nebudou*** v důsledku změny dimenzionální tabulky ***změněny***, tj. záznamy ve faktové tabulce se budou po datu změny (5.5.2011) odkazovat na záznam v tabulce dimenze zboží s již novým primárním klíčem, tedy nikoli 111, ale na záznam 321,
- ***výběry dat*** podle *zodpovědného pracovníka* ***budou probíhat správně*** podle různých hodnot daného atributu, tedy dostaneme např. správné objemy prodejů pro J. Malou i pro F. Klase,
- tím, že zůstává stejný operační klíč (NT-0015) dostáváme ***správné hodnoty i pro celkové objemy prodeje*** tohoto produktu, bez ohledu na změnu odpovědného pracovníka,
- ***hodnoty atributů Datum_od, Datum_do*** slouží ***pro pokročilé analýzy*** a přesné sledování historie prodeje podle změn atributů, pro základní přehledy (viz předchozí body) se ale nemusí využívat,

- výhodou typu 2 je zejména ***možnost sledování vývoje ukazatelů podle změn atributů*** a současně možnost sledovat tolik změn a u tolika atributů v dimenzionální tabulce, kolik je třeba,
- rovněž platí, že ***nadřízené (master) tabulky*** (ve SNOWFLAKE schématu) ***využívající změny typu 2 musí mít i podřízené tabulky (detail) se změnami typu 2,*** naopak to ale není nutné,
- vzhledem k tomu, že je přesně dáno, kdy došlo ke změně, pak ***není třeba zpětně vytvářet původní výstupy a agregace.***

4.1.17.3 Typ 3 – Přidání nového sloupce do dimenzionální tabulky

Typ 3 představuje ***přidání nového atributu do dimenze*** (nového sloupce) a ***přesunutí staré hodnoty do tohoto nového atributu.*** Do původního (aktuálního) atributu je pak možné vložit novou hodnotu. Příklad řešení změny typu 3 ukazuje Tabulka 4-4.

Tabulka 4-4: Změna dimenzionální tabulky pro typ 3

Primární klíč	Název zboží	Číslo zboží	Číslo skupiny zboží	Název skupiny	Zodpovědný pracovník	Původní zodpovědný pracovník
111	Acer Travelmate 292	NT-0015	01	Notebook	F. Klas	J. Malá
112	IBM ThinkPad R40	NT-0023	01	Notebook	P. Novák	P. Novák

Pro změny v dimenzi **typu 3** je charakteristické:

- lze efektivně ***kombinovat výběry*** podle původní i nové hodnoty atributu, např. pro srovnání výkonnosti obou pracovníků apod.,
- na rozdíl od typu 2 ***může být předchozí i současná hodnota atributu využívána jako platná*** současně,
- typ 3 se však ***prakticky nepoužívá často***,
- při větším očekávaném počtu změn se může ***využít i větší počet nových atributů***, ale to je možné pouze v případech, kdy počet změn lze u dimenze přesněji odhadovat.

4.1.17.4 Typ 4 – Vytvoření nové historické dimenze

Typ 4 je založen na vytvoření nové ***historické dimenze, do níž se přesouvají historické údaje***. Tento přístup se používá obvykle tehdy, kdy v dimenzích dochází k rychlým, respektive častým změnám. Jestliže dojde ke změně, je daná ***existující dimenze nejdříve zkopírována do historické dimenze*** a následně jsou ***v původní tabulce dimenze provedeny potřebné změny***.

Při dimenzionálním modelování ve vztahu k SCD je třeba brát v úvahu, že ***různé atributy***, které jsou součástí záznamu dimenzionální tabulky, ***budou mít různý typ SCD***. Např. datum narození zákazníka se nebude měnit (kromě chyby v datech), bude mít tedy typ 1, zatímco adresa zákazníka se měnit bude a bude účelné ji zpětně sledovat, tedy typ 2, nebo vyšší. V každém případě je nutné typy SCD u jednotlivých atributů definovat a realizovat v transformačních procedurách (ETL), což samozřejmě zvyšuje jejich složitost.

Nejčastěji užívanými typy SCD jsou 1 a 2. Při návrhu analytické databáze se v této souvislosti můžeme rozhodovat ***mezi dvěma variantami***:

- ***jednu dimenzi lze rozdělit na dvě dimenzionální tabulky***, přičemž v jedné budou atributy typu 1 a v druhé typu 2. Toto řešení je relativně jednoduché a často populární, ale nese s sebou dvě rizika – zvyšuje se počet dimenzí, v dotazech je třeba řešit navíc vazby mezi těmito dvěma dimenzemi, což je tvoří složitějšími a časově náročnějšími;
- ***zachovat jednu společnou dimenzi s atributy typu 1 i 2***, což řeší problémy předchozí varianty, ale naopak vede ke složitějším transformacím dat.

Provádění změn v dimenzích ***při přípravě dat*** je poměrně ***časově náročné***. Používá se proto kontrolních součtů v každém databázovém záznamu (jako provozní položka). Teprve v případě nesouhlasu kontrolních součtů se kontrolují jednotlivé položky záznamů a zjišťují se změny, které se pak u jednotlivých atributů řeší podle stanoveného typu změny.

4.2 Datové modelování

I když datové modelování je svébytnou disciplínou, v souvislosti s analytikou je dobré rekapitulovat některé základní principy.

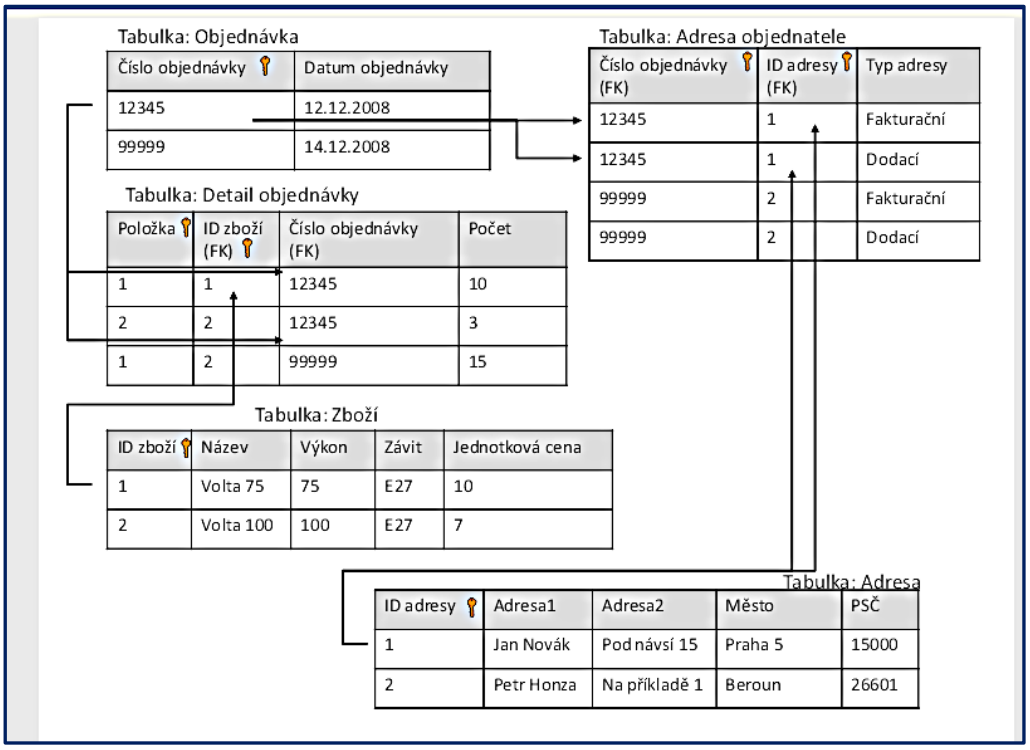
4.2.1 Účel datového modelování

Datové modelování je ***metoda***, která slouží ***k navrhování struktury dat v databázích*** a k jejich dokumentaci. ***Účelem*** datového modelování je přispívat ***ke kvalitnímu návrhu databází, jak transakčních, zdrojových, tak analytických***. To znamená, že:

- databáze jsou ***konstruovány efektivně s minimalizací redundancí*** dat (tj. jejich duplicit, či multiplicit), se zajištěním požadovaného výkonu a doby odezvy při nejčetnějších dotazech a dalších operacích,
- databáze se navrhují tak, ***aby poskytovaly data uživatelům, která odpovídají jejich*** aktuálním ***požadavkům***, ale i požadavkům, které lze v budoucnu očekávat. To znamená tak, aby databáze bylo možné relativně bezpečně rozšiřovat, udržovat a modifikovat. Úpravy databáze nesmí znamenat nepřiměřená rizika pro kvalitu dat,
- v relační databázi je ***řešena integrita na úrovni položek (atributů), tabulek i vazeb mezi tabulkami*** zajišťující přesnost a konsistenci uložených dat po celou dobu provozu,
- ***aplikace postavené na databázi lze pak řešit efektivně*** a návrh databáze neznamená významné omezení pro jejich další údržbu a rozvoj.

4.2.2 Relační databáze

Datové modelování je spojené obvykle s ***relačními databázemi***. Je účelné v tomto kontextu ***zrekapitulovat*** několik základních termínů.



Obrázek 4-10: Tabulky a jejich vazby v relační databázi

4.2.2.1 Relace

Relace – dvourozměrná datová struktura představovaná ***tabulkou***, kde jeden řádek tabulky odpovídá jednomu záznamu (např. jednomu záznamu zboží) a sloupec odpovídá položce (např. *ID zboží*, *Název*, *Jednotková cena* atd.).

V databázovém pojetí se ***sloupec tabulky*** označuje jako ***atribut,*** jehož hodnoty spadají do tzv. ***domény.***

4.2.2.2 Integrita

Integrita**,** což je pravidlo omezující možné hodnoty atributů, případně možné manipulace se záznamy, které existují ve vazbě k záznamům jiné tabulky.

K významným integritním omezením patří ***existence primárního klíče*** (tzv. ***entitní integrita***). Tím se zajišťuje ***jednoznačná identifikace záznamu*** v tabulce (žádné dva záznamy nemohou mít stejnou hodnotu primárního klíče). S termínem primární klíč souvisí i ***umělý primární klíč (surrogate key),*** nemající nějaký specifický podnikový význam (např. *ID zboží*), ale vzniká postupným načítáním stanovené hodnoty (většinou 1) při přidávání dalších záznamů do tabulky databáze.

Další typ integritního omezení je ***doménová integrita***, která zajišťuje, aby údaj uvedený jako hodnota atributu byl vybrán z množiny definovaných přípustných hodnot. Například pro sloupec (doménu) *Typ adresy* v tabulce *Adresa* objednatele může být takovým doménovým omezením možnost uvedení buďto hodnoty *Fakturační* nebo hodnoty *Dodací*.

Posledním typem integritního omezení je tzv. ***referenční integrita***, která je ***reprezentována cizím klíčem (foreign key)***. Cizí klíč zajišťuje provázanost tabulek, např. mezi tabulkou *Zboží* a *Detail objednávky* je cizím klíčem *ID zboží* umístěný v tabulce *Detail objednávky.* Na obrázku (Obrázek 4-10) je položka, která je cizím klíčem označena zkratkou FK. ***Princip referenční integrity zajišťuje***:

- ***aby nemohla být vložena jako hodnota atributu***, který je označen jako **cizí klíč**, hodnota, jež nemá odpovídající hodnotu atributu v tabulce s primárním klíčem. Tak je např. zajištěn vztah mezi tabulkou *Objednávka* a *Detail objednávky*, kdy platí, že v tabulce *Detail objednávky* se jako hodnota atributu *Číslo objednávky* nesmí objevit jiné číslo než to, které je uvedeno u některého ze záznamů v tabulce *Objednávka*,
- aby z tabulky ***nebylo možno odstranit záznam***, jehož hodnota je uvedena jako cizí klíč v tabulce jiné. To např. znamená, že nelze odstranit záznam z tabulky *Zboží*, jehož ID zboží je rovno 1, protože se tato hodnota vyskytuje jako hodnota cizího klíče v tabulce *Detail objednávky*.

4.2.3 Řešení datových modelů

Uvedené základní ***principy relačních databází*** se promítají do řešení datových modelů. Návrh datových modelů se realizuje na několika úrovních, resp. v několika základních krocích. Tyto kroky následují obvykle od konceptuálního návrhu, přes logický k fyzickému:

- ***Konceptuální návrh*** databáze – prvotní návrh databáze její ***struktury, vazeb a všech potřebných charakteristik:***
  - je zaměřen na ***věcnou podstatu*** dat a nezohledňuje fyzické, resp. technologické charakteristiky databáze, ale pouze její požadovanou strukturu a logiku uspořádání dat, které jsou ***nezávislé na konkrétním softwaru*** pro řízení databáze,
  - ***účelem*** konceptuálního návrhu databáze je specifikovat typy datových objektů (***typy entit***), jejich vlastnosti (***atributy***), přirozenou ***identifikaci, vazby*** mezi entitami a jejich vlastnosti,
  - - v entitách ***se nemodelují cizí klíče***, vazby mezi entitami jsou vyjádřeny pouze graficky.
- ***Logický návrh*** databáze na základě konceptuálního návrhu databáze se řeší logický návrh databáze:
  - ***cílem*** je *detailně vymezit strukturu* databáze založené již na určitém logickém schématu dat, tj. v našem případě relačním,
  - především definovat ***struktury databázových tabulek***, jejich atributy, primární klíče a relace mezi tabulkami prostřednictvím cizích klíčů,

- Vytvořením, případně generováním ***fyzického návrhu*** databáze na základě logického návrhu databáze se definuje fyzický datový model:
  - specifikuje již potřebné ***implementační charakteristiky*** ve vztahu k příslušnému ***databázovému prostředí***, v němž bude databáze realizována,
  - např. ***formáty jednotlivých datových položek***, jejich rozsah, resp. délka, definování cizích klíčů a vztahů mezi tabulkami a další charakteristiky.

4.2.4 Podstatné charakteristiky datových modelů

Entita představuje z pohledu konceptuálního datového modelu dále ***nedělitelnou jednotku reálného světa***, o které se budou v databázi vést příslušná data:

- v databázi entitě obvykle odpovídá ***jeden řádek*** v databázové tabulce,
- ***typ entity*** pak představuje zobecnění pro všechny entity se stejnými vlastnostmi (např. typem entity je *Zákazník* a entitou jeden každý zákazník),
- typu entity odpovídá v databázi jedna ***databázová tabulka***,
- typ entity má ***vlastní název, identifikátor*** a její ***obsah*** je dále definován skupinou atributů, které jsou následně, v rámci logického návrhu databáze, základem pro definování struktury databázové tabulky,
- v rámci logického návrhu databáze je základem pro ***definování struktury*** databázové tabulky,
- ***Poznámka*****:** Entita a Typ entity se oba běžně používají pro termín „typ entity“, Výskyt pak pro jednu konkrétní entitu, tj. data o konkrétním sledovaném objektu.

V*azby* se definují mezi dvěma entitami použitím symbolu vazby s doplněním názvu vazby a dalších jejich charakteristik, včetně kardinality.

Kardinalita vazby – násobnost znamená vyjádření ***možnosti výskytu počtu entit*** na obou stranách vazby:

- existují základní vztahy ***1:1***, ***1:M, M : N,***
- pro vazbu ***1 : M, (One to Many),*** to znamená, že na příklad jeden obchodník odpovídá za více objednávek zboží a jedna objednávka zboží je vyřizována pouze jedním obchodníkem,
- vazba ***M : N (Many to Many)*** to bude znamenat, že jeden obchodník vyřizuje více objednávek, ale na druhé straně na jedné objednávce se může podílet více obchodníků,

Povinnost vazby vyjadřuje ***nutnost existence*** dané vazby ke každému výskytu entity, například obchodník nemusí být odpovědný za žádnou objednávku, ale objednávka musí mít obchodníka, který je za ní odpovědný.

4.2.5 Normalizace databáze

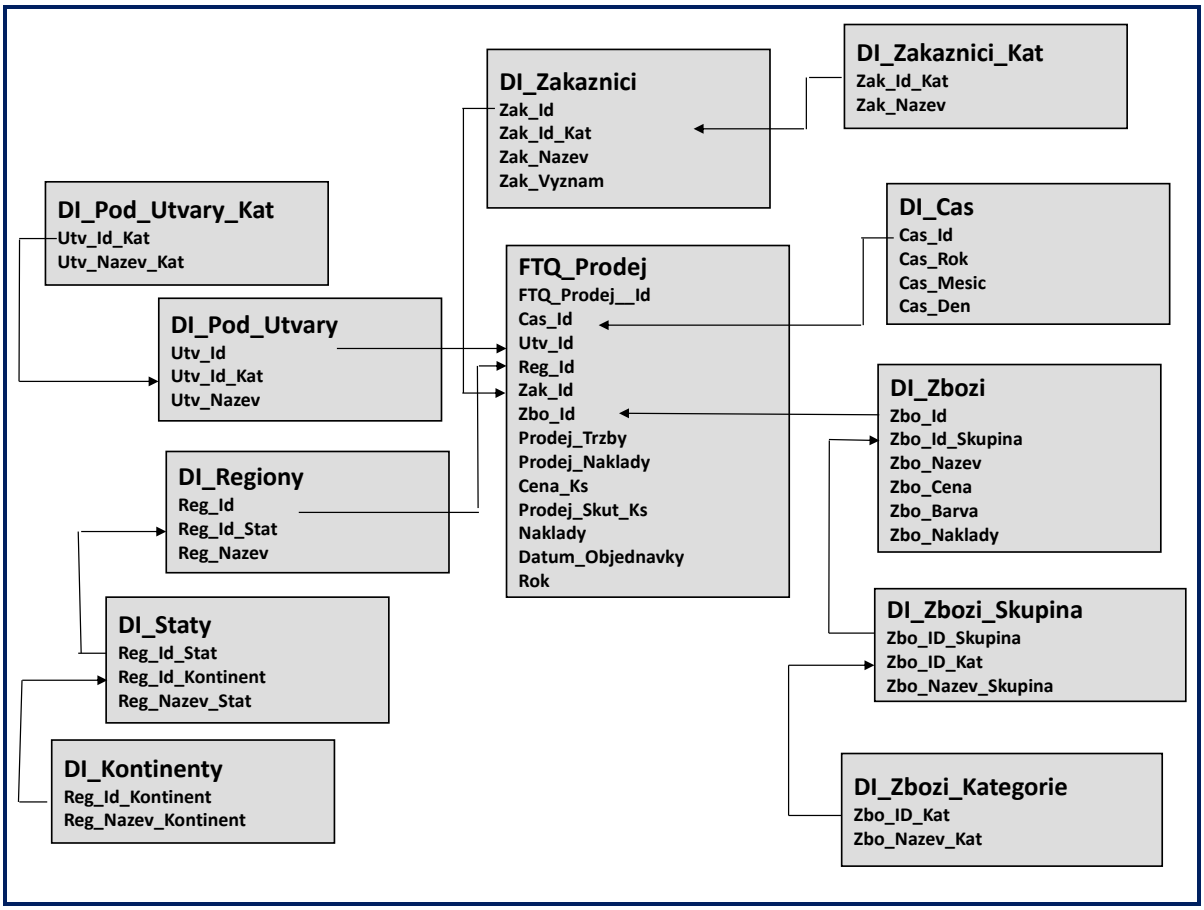
Normalizace databáze se chápe jako postup, kdy je ***struktura dat*** v relační databázi ***přeorganizována*** tak, aby byly ***minimalizovány duplicity nebo multiplicity*** uložení dat v databázi tak, aby ideálně jedna změna v realitě vyžadovala pouze jednu změnu v datech v databázi.

4.2.6 Návrh datového modelu pro BI a SSBI

Další aktivitou dimenzionálního modelování je analýza a návrh dimenzionálních modelů na bázi principů datového modelování. V návaznosti na hrubý dimenzionální model ***se upřesňuje obsah a vazby databázových tabulek dimenzí a faktů s určením běžných charakteristik*** (formátu, rozsahu) jejich jednotlivých atributů.

V rámci této fáze je třeba ještě znovu verifikovat, zda data pro navrhované dimenzionální modely jsou ve zdrojových databázích k dispozici a jsou kvalitní. Podstatou této aktivity je transformace výsledků

hrubého dimenzionálního modelu do relačních databázových struktur. Příklad datového modelu pro analytickou aplikaci řízení prodeje dokumentuje Obrázek 4-11.



Obrázek 4-11: Příklad: datový model pro BI a SSBI aplikace¨

4.3 Závěry k metodám řešení podnikové analytiky

- Před zahájením řešení je účelné rozhodnout a použitých ***analytických metodách, metodikách a pracovních postupech.***
  - Tento text obsahuje hlavní principy 2 vybraných metod, a to ***dimenzionálního modelování a datového modelování***.
  - S metodami by se měli seznámit a ovládat je nejen informatici (analytici, vývojáři), ale i uživatelé, především ***klíčoví uživatelé***. To je významné především u aplikací SSBI, kde se vyšší aktivita uživatelů při řešení předpokládá.
- V souvislosti s přípravou analýzy v rámci řešení jednotlivých projektů je dobré si připravit ***sady analytických otázek*** a pak by měly tvořit ***náplň pracovních schůzek*** analytiků s jednotlivými manažery a specialisty firmy. Uvedené a další otázky je nutné ***konkretizovat podle jednotlivých oblastí řízení*** a typu firmy.
- Pro naprostou většinu úloh podnikové analytiky je nutné specifikovat odpovídající ***sady metrik***, resp. podnikových ukazatelů a analyzovat jejich hlavní charakteristiky, zejména ***analytické dimenze a datové zdroje*** pro ukazatele i dimenze.
- Po nezbytném vyjasnění obsahu úloh podnikové analytiky je třeba řešit ***technologickou realizaci,*** výběr příslušných nástrojů a aplikací a jejich implementace.

- ***Charakteristiky a hodnocení*** potenciálních efektů a problémů nástrojů ***základní analytiky*** je náplní ***oddílu B***.
- Řešení aplikací podnikové analytiky je obvykle složeno z celé škály ***různých technologických komponent***.
- Součástí realizace aplikací jsou specifické nároky na ***reporting a vizualizace dat***.

C) Základní podniková analytika – nástroje, řešení

[5] Business Intelligence, BI	**[6] Self Service Business Intelligence, SSBI**
[7] Mobilní BI	**[8] Competitive Intelligence, CI**

Podniková analytika je v tomto dokumentu pracovně ***rozdělena na „základní“ a „pokročilou***“. Tento oddíl se vztahuje k řešením, nástrojům a aplikacím, které mají ***charakter základní podnikové analytiky***, tj. obvykle tvoří její jádro a je dnes již prakticky běžnou součástí informačních systémů.

Do této skupiny ***patří***:

- ***Business Intelligence (BI)***, tj. řešení založená na celém spektru ***velmi výkonných komponent*** (zejména databázových systémů, transformačních nástrojů, analytických a dalších nástrojů), které jsou vzájemně propojeny a tvoří často velmi mohutný ***komplex*** datových skladů, datových tržišť a dalších komponent. Jsou určeny obvykle pro ***rozsáhlé týmy pracovníků*** různých útvarů a často i dislokovaných obchodních jednotek.
- ***Self Service Business Intelligence (SSBI)*** jsou řešení navazující na původní Business Intelligence, ale založené na ***technologiích nižších kapacit***. Jsou určeny pro relativně menší týmy, resp. menší podniky, případně jako doplňky komplexních BI systémů. Pro řešení i užití je charakteristický jejich ***operativní charakter***, kratší doba řešení a podstatně nižší náklady, než je tomu u systémů Business Intelligence.
- ***Mobilní Business Intelligence*** se vztahuje jak ke komplexním BI systémům, tak aplikacím SSBI a jak název napovídá, tak jako koncová (a prezentační) zařízení se užívají tablety nebo smartphony.

- ***Competitive Intelligence*** se zabývá sběrem, zpracováním a ochranou informací ***s cílem získat konkurenční výhodu pro firmu.*** Je postaveno primárně na externích informačních zdrojích, často nestrukturovaných s cílem provádět na nich analýzy převážně prostředí trhu.

U každého uvedeného typu řešení nebo aplikace je v jednotlivých kapitolách obsaženo jejich ***funkční vymezení*** a hlavní převážně ***technologické charakteristiky*** a následně (podle principů MBI) i jejich ***hlavní efekty*** pro firmu na druhé straně ***omezení*** nebo potenciální problémy. Závěr kapitol tvoří podstatná ***doporučení*** pro jejich řešení a užití.

5. Business intelligence, BI

Účelem kapitoly je:

- poskytnout ***podklady pro hodnocení IT aplikací***, nástrojů a technologií zaměřených na byznys intelligence a určit doporučení pro přístupy k jejich analýze, návrhu a implementaci,
- definovat ***podstatné charakteristiky*** těchto aplikací a nástrojů,
- vymezit významné ***efekty a případná omezení*** aplikací a nástrojů byznys analytiky jako podstatné vstupy pro jejich výběr a užití v implementaci.

5.1 Základní principy Business Intelligence

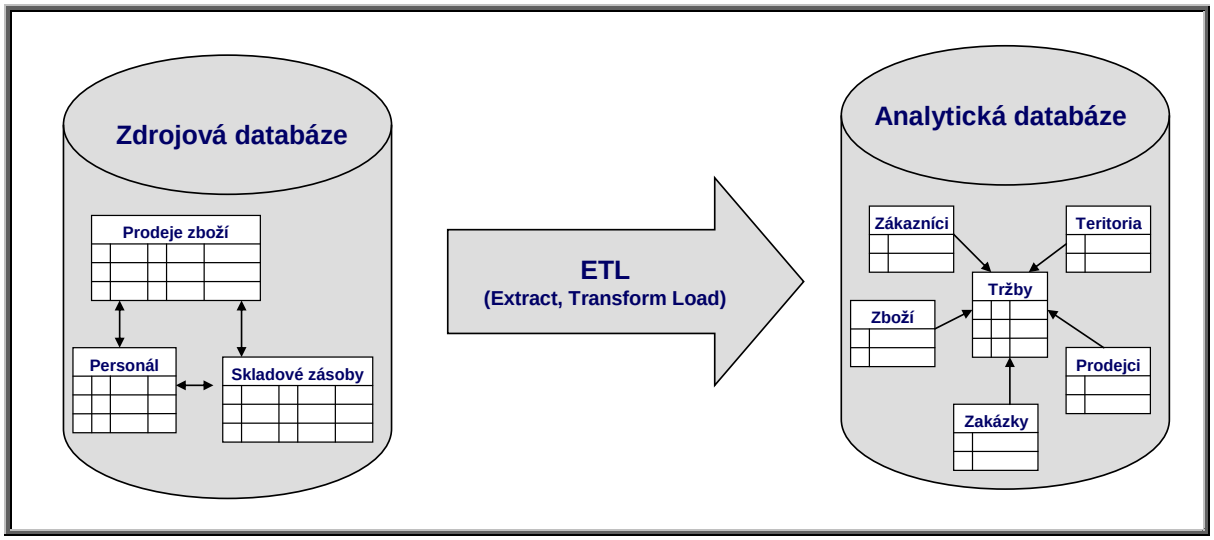
Pokud zatím odhlédneme od realizačních charakteristik a jednotlivých produktů, pak ***základní principy*** řešení Business Intelligence můžeme shrnout do následujících bodů:

- zatímco transakční systémy jsou primárně určeny pro pořizování a aktualizace dat a tomu odpovídá i organizace dat v databázi, pak řešení Business Intelligence jsou určená pro analytické aplikace a tomu musí odpovídat i ***výběr dat*** ze zdrojových databází a jejich ***organizace*** v analytických databázích,
- zatímco transakční systémy udržují data na maximální úrovni detailu (většinou na úrovni jedné transakce se všemi jejími detailními atributy), BI řešení ukládají pouze data relevantní pro analýzy, tedy na potřebné ***úrovni detailu (granularity)***, tedy ***detailní i agregovaná*** podle požadovaných hledisek řízení firmy,
- analytické aplikace pracují primárně s daty podnikových ukazatelů a ty vyhodnocují podle nejrůznějších hledisek, dimenzí a jejich kombinací. To znamená, že BI řešení jsou vesměs založena na ***multidimenzionalitě*** uložení a zpracování dat,
- zatímco transakční systémy obvykle udržují databáze většinou s aktuálními daty, BI řešení jsou založeny na využití ***časové dimenze***, to znamená, že ukládají data do analytických databází většinou postupně v jednotlivých časových snímcích,
- uvedené předchozí principy (agregace, multidimenzionalita, časová dimenze) vedou ke specifickým, podstatně vyšším nárokům Business Intelligence na ***kvalitu dat***.

Business Intelligence je postavena na celé řadě dalších dílčích pravidel a principů a současně i jejich ***různých modifikací podle konkrétního technologického prostředí***. Výše uvedená pravidla však pokládejme pouze za výchozí.

5.1.1 Výběr a organizace dat

Aplikace Business Intelligence (kromě plánovacích aplikací) nevytvářejí nebo nepořizují nová data, ale využívají data vytvořená transakčními aplikacemi (ERP, CRM atd.). Databáze těchto transakčních aplikací se proto z pohledu BI označují jako ***zdrojové.*** Podstatnou vlastností těchto databází je organizace jejich dat podporující přístupy k detailním datům, ukládání a aktualizace dat, tj. že jsou pro tyto operace optimalizované (např. na základě normalizace datových struktur). Oproti tomu analytické BI aplikace jsou optimalizované na efektivní poskytování analytických informací, tj. data zde musí být organizována ve shodě s potřebami analytických úloh, tj. musí obsahovat ***hodnoty ukazatelů ve vazbě na analytická hlediska, tedy dimenze***. Z toho vyplývá, že mezi zdrojovými databázemi a analytickými databázemi musí proběhnout transformace dat. Detailněji tento princip dokumentuje Obrázek 5-1.



Obrázek 5-1: Transformace dat

ETL, resp. ***ELT*** (Extract, Transform, Load – viz dále), nebo také ***datová pumpa*** je program, resp. soustava programů, který zajišťuje výběr dat (*Extract*) ze zdrojových databází, jejich transformace (*Transform*) do jiných datových struktur (jiných tabulek a struktur záznamů) a fyzické uložení dat (*Load*) do analytických databází. Z toho dále vyplývají následující charakteristiky ETL:

- ze zdrojových databází musí být vybrána pouze taková ***data***, která jsou určena ***pro analytické, plánovací a rozhodovací aktivity*** podniku, tedy nikoli všechna. To je jedním z prvních úkolů analytiků Business Intelligence takový kvalifikovaný výběr dat určit, tj. rozhodnout, co se bude z datových zdrojů vybírat;
- data jsou transformována do nových datových struktur analytických databází, které musí být předem navrženy tak, aby nejlépe odpovídaly potřebám řízení podniku. To souvisí s možnostmi řešení ***multidimenzionality a granularity dat***, k nimž se vrátíme v dalších paragrafech;
- data do Business Intelligence vstupují z různých zdrojových databází (ERP, e-Business, CRM atd.), přičemž v těchto různých zdrojích, mohou být jedna a tatáž data uložena vícekrát (a k tomu ještě různě), např. různé databáze prodejců, zákazníků apod. Ale do analytických databází musí vstoupit pouze jednou. V této souvislosti se o BI často mluví jako o „jedné verzi pravdy“. V transformační vrstvě musí dojít ke ***konsolidaci dat,*** tj. určení vstupujících dat s vyloučením duplicit, či multiplicit. I toto je důležitý úkol pro analytiky a následně provozovatele aplikací Business Intelligence;
- se zajištěním konsolidace dat úzce souvisí i dosažení potřebné ***kvality dat, tj. vyloučení chyb, nepřesností atd.***

Právě tato část projektů BI, tedy řešení transformační vrstvy je pracovně, časově i finančně nejnáročnější a obvykle představuje cca 60 - 80 % vynaložených pracovních kapacit. Pro úspěšné BI řešení však znamená zcela nezbytný předpoklad.

5.1.2 Multidimenzionalita uložení a práce s daty

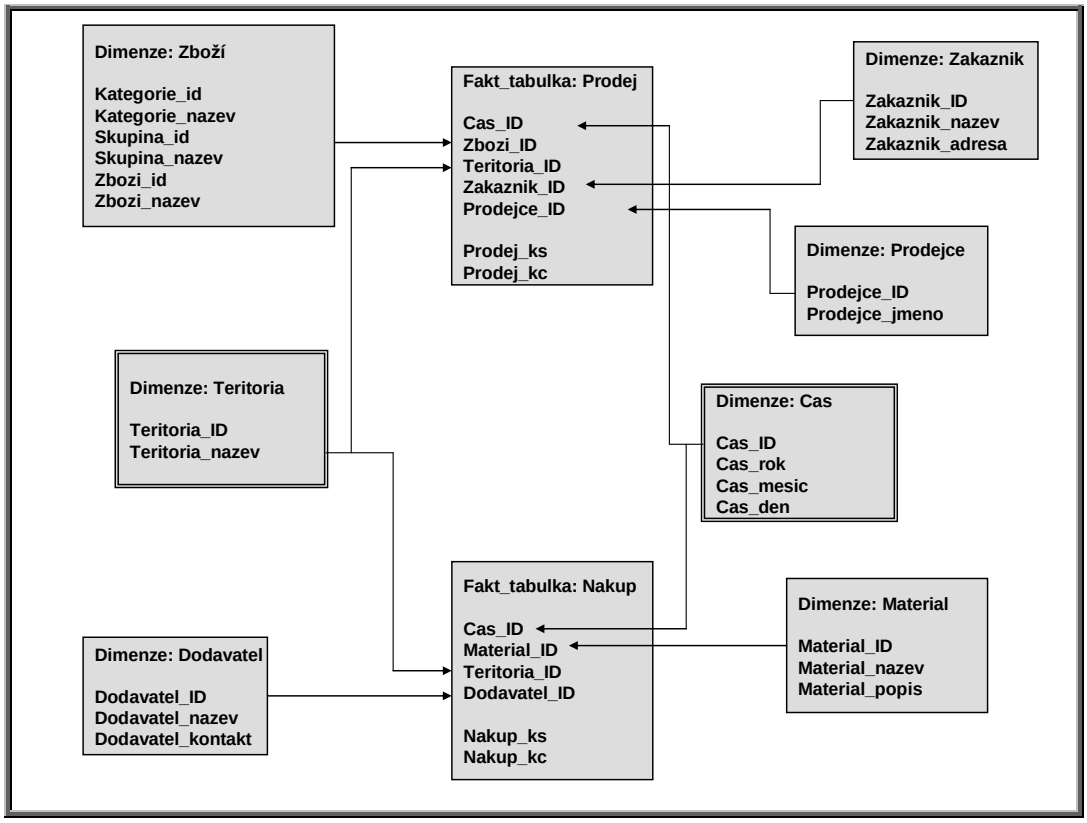
Požadavek pohledů uživatele na sledované ukazatele z více hledisek (dimenzí) a jejich kombinací je základem pro užití principu ***multidimenzionality*** uložení a práce s daty. To s sebou současně přináší i požadavek na ***specifickou organizaci dat*** v databázi. Způsob realizace multidimenzionality v datech poskytuje klasicky základní možnosti a podle nich rozdělíme i další objasnění tohoto principu, a to:

- multidimenzionalitu vyjádřenou v relačních databázích,
- multidimenzionalitu dat realizovanou pomocí tzv. OLAP technologie (On Line Analytical Processing), tj. technologie vyvinuté speciálně pro tyto účely,
- další možnosti, jako je např. tabulární forma organizace dat.

Multidimenzionalita dat v prostředí relační databáze

Datové modely produkčních systémů jsou komplexní, obsahují mnoho tabulek a jejich vazeb. Takto organizovaná data jsou z hlediska jejich vytváření a aktualizace velmi efektivní, ale pro běžného uživatele se stávají dosti nepřehledná. Pro výše uvedený nedostatek se objevily snahy o zjednodušení takového uložení dat a jeho přizpůsobení pro tvorbu BI řešení. Vznikl tak relační **„*dimenzionální model*“,** kterému se také běžně říká **„Schéma hvězdy**“ (*STAR scheme*), resp. "***Schéma sněhové vločky***" (*SNOWFLAKE scheme*) – viz kapitoly 4.1.6 a 4.1.7.

Je zřejmé, že databáze datového skladu, či datových tržišť (viz dále) nemůže být řešena jedním schématem typu STAR či SNOWFLAKE, ale těchto schémat je v jejich databázích obvykle obrovské množství. Základní princip je v tom, že se sledované ukazatele dělí do tabulek faktů podle logické příbuznosti a tedy i podle stejných přiřazených dimenzí a stejných složených klíčů. Takto se vytvářejí jednotlivá schémata, která se pak provazují pomocí společných, resp. **sdílených dimenzí**. Obrázek 5-2 tento princip dokumentuje na příkladu, kde vidíme, že těmito sdílenými dimenzemi mezi tabulkami faktů *Prodej* a *Nakup* jsou dimenze *Teritoria* a *Cas*.



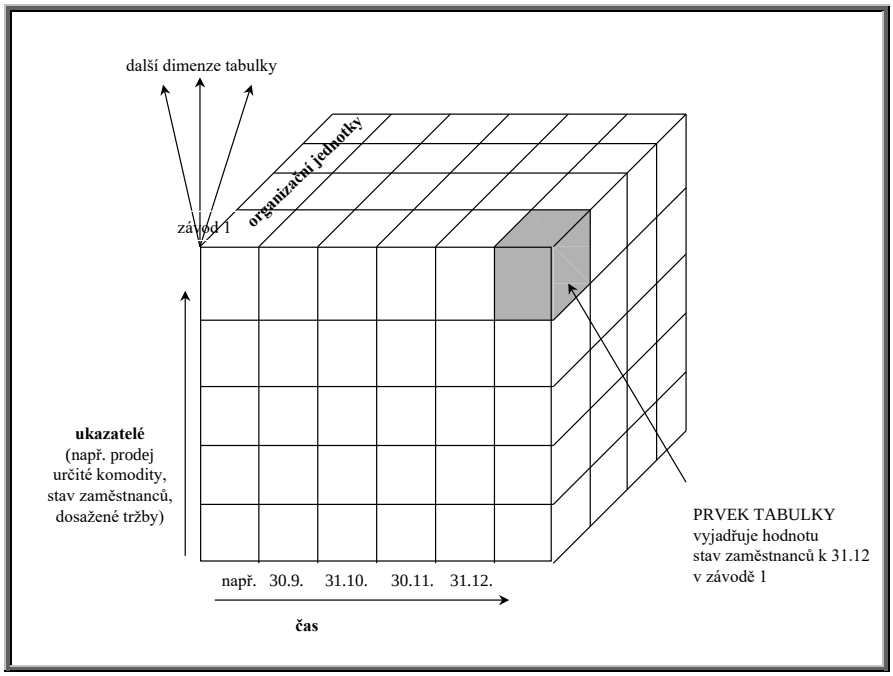
Obrázek 5-2: Propojení jednotlivých databázových schémat pomocí sdílených dimenzí

Multidimenzionalita dat v prostředí OLAP technologie

Multidimenzionální databáze jsou optimalizované pro uložení a interaktivní využívání multidimenzionálních dat. Výhodou multidimenzionality, resp. nasazení OLAP (On Line Analytical Processing) technologií je rychlost zpracování a efektivní analýzy multidimenzionálních dat (drilling, slice and dice apod.). OLAP technologie tak nabízí efektivní přístup k datům v jednoduché struktuře vhodné pro analytické činnosti podnikových manažerů a podporu jejich rozhodování. Se zavedením pojmu BI a současně s rozvojem nástrojů a technologií pro širokou podporu analytických činností v organizaci se však výraz OLAP poněkud zúžil.

Užší význam definuje ***OLAP*** čistě technologicky, tedy jako „informační technologii založenou především na koncepci multidimenzionálních databází. Jejím hlavním principem je několikadimenzionální tabulka umožňující rychle a pružně měnit jednotlivé dimenze a měnit tak pohledy uživatele na modelovanou ekonomickou realitu.“ Tento text bude nadále pracovat s užším – technologickým významem výrazu OLAP.

Základním ***principem*** technologie OLAP je ***několikadimenzionální tabulka*** umožňující velmi rychle a pružně měnit ***jednotlivé dimenze*** a nabízet tak uživateli různé pohledy na ekonomickou realitu a která nabízí možnosti drill-down, drill-up nebo slice and dice pro výběr („vykrojení“) dat z OLAP kostky, jak dokumentuje Obrázek 5-3. OLAP kostky tak zahrnují předzpracované agregace dat podle výše uvedených hierarchických struktur dimenzí a jejich kombinací. Jde tak v podstatě o princip "n-dimenzionální Rubikovy kostky" naplněné nejdůležitějšími podnikovými daty.



Obrázek 5-3: Princip multidimenzionální databáze na bázi OLAP

Z obrázku vyplývá, že standardními dvěma dimenzemi jsou tu ukazatelé (ekonomické proměnné) a čas**.** Ostatní dimenze se pro jednotlivé modely definují podle potřeby, např. organizační jednotka, zboží, zákazník, dodavatel, teritorium, konkurent apod. Obsah dimenzí je tvořen ***prvky dimenzí*** (viz předchozí text). Promítnutí všech dimenzí do jednoho bodu tvoří ***prvek OLAP kostky*****.** Na základě technologie OLAP jsou vytvářeny ***OLAP databáze***, které představují jednu nebo několik souvisejících OLAP kostek.

Technologie OLAP se prakticky realizuje ***v řadě variant***, z nichž uvedeme alespoň základní:

- MOLAP (*Multidimensional OLAP*) je charakteristická specielním uložením dat v multidimenzionálních – binárních OLAP kostkách,
- ROLAP (*Relational OLAP*) řeší multidimenzionalitu s využitím technologie relačních databází,
- HOLAP (*Hybrid OLAP*) je kombinací předchozích přístupů, kdy detailní data jsou uložena v relační databázi a agregované hodnoty jsou uloženy v binárních OLAP kostkách.

Multidimenzionální databáze v prostředí OLAP jsou ***optimalizované pro uložení a interaktivní využívání multidimenzionálních dat*** a jejich výhodou je zejména rychlost zpracování a krátká doba odezvy.

Tabulární model

Tabulární model lze definovat jako „sadu metadat, zahrnující tabulky, vztahy, metriky, KPI, skupiny kalkulací (calculation groups), hierarchie, překlady, bezpečnostní role a mnoho dalších prvků, které tvoří tzv. sémantický model, který zajišťuje systém navigace v klientských nástrojích, jako je realizace reportů v Power BI nebo v Excelu.

Model odpovídá datové sadě ve službě Power BI anebo databázi ve službě Analysis Services (Azure Analysis Sevices anebo SQL Server Analysis Service). Datové sady a databáze jsou modely naplněné daty. Definice modelu bez jakýchkoli dat je „tabulární model“. (Russo 2022)

Datové sady anebo databáze založené na tabulárních modelech jsou speciální ***analytické databáze***, které pro ukládání dat mohou používat 2 druhy řešení:

- ***In-memory***, což znamená, že vykonávají dotazy na data v paměti aplikace nad celou databází,

- ***DirectQuery***, což znamená, že data nejsou kopírována do aplikace, přes kterou k nim přistupujeme (například Power BI Desktop, Microsoft SQL Server Analysis Services, Azure Analysis Services), ale dotazy jsou předávány přímo do zdrojových databází, kde probíhá zpracování a do aplikace je předáván až výsledek dotazu. V tomto případě je tedy tabulární model založen v příslušném nástroji, ale data nejsou jeho součástí.

Tabulární modely ***kombinují technologii MOLAP a relačních databází***. (Rabeler et al. 2018)

Zatímco ***in-memory*** modely jsou ***výchozí***, ***DirectQuery režim se používá pro*** příliš velké objemy dat na to, aby se vešly do paměti, nebo pokud volatilita dat vylučuje rozumnou strategii zpracování. ***DirectQuery*** dosahuje shody s in-memory modely prostřednictvím ***podpory široké škály datových zdrojů***, schopností zpracovávat kalkulované tabulky a sloupce v DirectQuery modelu, zabezpečení na úrovni řádků prostřednictvím výrazů DAX, které se dostanou do databáze typu back-end, a optimalizace dotazů, jejichž výsledkem je rychlejší průchodnost (propustnost). (Rabeler et al. 2018)

Tabulární modely jsou definovány strukturami a protokoly dokumentovanými Microsoftem (“. (Russo 2022):

- MS-SSAS-T: SQL Server Analysis Services Tabular Protocol
- Tabular Object Model (TOM)
- Tabular Model Scripting Language (TMSL)
- XML for Analysis (XMLA) Reference

Sama definice tabulárního modelu má díky TOM strukturu JSON souboru, známá jako model.bim. Díky tomu lze model vytvářet prostřednictvím:

- Power BI
- Microsoft Visual Studio (Analysis Services Porject) ,
- Tabular Editoru,

Tabulární modely mohou být ***nasazeny do služby Azure Analysis Services*** nebo na instance služby ***SQL Server Analysis Services (SSAS)*** nakonfigurované na režim Tabular server anebo do služby Power BI Premium. Nasazené tabulární modely je možné spravovat přes SQL Server Management Studio. (Rabeler et al. 2018)

Nástroje tabulárního původně podporovaly mezi tabulkami ***pouze vazby 1:1 nebo 1:N s jednosměrnou i obousměrnou filtrací***. Od roku 2020 jsou implementovány i vazby M:N, i když je doporučeno tento typ vazeb používat pouze ve specifických scénářích užití, stejně jako obousměrné vazby 1:N.

5.1.3 Nároky na kvalitu dat

Jedním z častých problémů současné podnikové informatiky je nízká kvalita dat. To se nejvíce promítá právě do řešení aplikací Business Intelligence, které jsou na kvalitu dat zvláště citlivé, neboť každá dílčí chyba se může v souhrnných reportech nebo analytických aplikacích mnohonásobně projevit nebo zvětšit. Kvalita dat se posuzuje podle čtyř základních hledisek:

- ***dostupnost*** (*availability*) – vyjadřuje možnosti, jak může uživatel přistupovat k informacím v okamžiku aktuální potřeby a dělí se na:
  - dostupnost v čase,
  - dostupnost v místě, lokalitě, kde mají být informace využity,
  - dostupnost v požadované struktuře,
  - dostupnost v požadovaném formátu,
- ***přesnost*** (*accuracy*) – sleduje, zda všechna data jsou ve správném kontextu (např. zda odpovídá dané PSČ zadané adrese),
- ***úplnost*** (*completeness*) – určuje, do jaké míry jsou k dispozici všechny informace pro daný kontext, např. existuje správné PSČ, ale celá adresa je neúplná. Dalším příkladem je situace, kdy nemusí platit, že jsou všechny požadované transakce faktů neimportovány, tj. celkový jejich souhrn má pravdivý smysl jenom tehdy, když ho tvoří všechny detailní transakce,
- ***konzistence*** (*consistency*) – specifikuje možné problémy v porušení standardů nebo vazeb mezi daty. Např. databáze zákazníků v různých lokalitách mají pro stejného zákazníka různá identifikační čísla.

Významným aspektem kvality dat je určení a zjišťování možných zdrojů chyb, nepřesností, neúplností a porušených konsistencí. Při analýze zdrojových datových bází je proto nutné identifikovat nejen

syntaxi dat (tabulky, atributy, klíče), ale i jejich sémantiku (význam jednotlivých atributů, kombinace různých příznaků apod.).

Největším zdrojem chyb a různých poruch v datech je obvykle provoz transakčních aplikací, resp. jejich využití koncovými uživateli. V souvislosti s tím existují následující případy způsobující ***obvyklé chyby v datech:***

- chyby, které vznikají při manuálních vstupech dat, např. prohození číslic, pravopisné chyby, špatně zadané kódy, hodnoty zapsané do nesprávného pole,
- data pocházející z různých zdrojů (různých zdrojových aplikací), s tím jsou spojeny problémy sjednocení číselníků, dodržení formálních standardů v datech apod.,
- neoprávněné zkoušení a testování aplikací uživateli na „ostrých" datech,
- rozdíly v interpretaci dat, kdy různí uživatelé chápou sémantiku datové položky různě, což je dáno většinou nepřesným vymezením datových objektů, nebo nedostatky v dokumentaci,
- zásahy do zdrojových aplikací, kdy např. administrátor přidává do struktury záznamu novou datovou položku a zapomene to oznámit správcům ostatních aplikací, včetně BI,
- chyby přicházející z existujících externích datových zdrojů v důsledku špatně nastavených vazeb k externím datům,
- chyby vznikající v důsledku přímých vstupů do databází externími partnery (zákazníky, dodavateli) v rámci sdílených nebo otevřených aplikací pro vnější subjekty,
- chyby v návrhu databází, zejména nedodržovanou doménovou nebo referenční integritou,
- chyby v nedodržení relačních zásad – neexistují primární klíče, není vynucována povinnost atributů apod.,
- problémy s konverzí datových typů (zejména typu datum).

Přesné ***specifikace a analýzy zdrojů chyb*** jsou tak velmi důležitým předpokladem úspěšnosti řešení BI aplikací

5.2 Efekty BI

- Umožňuje ***lépe pochopit a analyzovat*** podstatu vlastní obchodní a ***manažerské činnosti***, proniknout do jejich hlubších a složitějších souvislostí, umožňují posun k multidimenzionálnímu řízení v reálném čase.
- Uplatnění časové dimenze nabízí ***sledování vývojových trendů*** z nejrůznějších pohledů.
- Analytická pravidla podle stanovených limitních hodnot jednotlivých ukazatelů a jejich dimenzí ***umožňují upozorňovat na kritické nebo mimořádné stavy***.
- Možnosti rychle se pohybovat ***na různé úrovni detailu informací*** (agregačních úrovních) odpovídajících právě řešenému problému v řízení.
- Podpora ***řešení skrytých problémů*** na základě identifikace složitých závislostí mezi daty, odhalováním podobností mezi např. zákazníky, obchodními případy apod.
- ***Kvalifikační efekty*** – BI přirozenou cestou posiluje schopnosti manažerů a specialistů při řešení svých úloh „multidimenzionálně uvažovat".
- Aplikace BI pro práci s ukazateli ve vazbě na podnikové procesy přinášejí i do manažerské sféry své specifické efekty a ***racionalizují jak řízení*** jednotlivých ***dílčích úloh*** a procesů, tak ***celého podniku***.
- ***Integrační efekty*** podporou integrace dezintegrovaných informačních zdrojů (např. z věcného či geografického hlediska).

5.3 Omezení, problémy, předpoklady BI

- Úspěšnost BI řešení ***silně závisí na zájmu uživatelů*** a zejména vedení podniku BI implementovat.
- BI aplikace jsou velmi silně ***závislé na kvalitě dat transakčních systémů***, ze kterých data čerpají.
- BI řešení jsou velmi ***náročné na konsolidaci*** vstupních dat z různých datových zdrojů.

- Efekty BI jsou ***obtížně měřitelné*** a mají spíše měkký charakter.
- Pro úspěšnost BI je podstatné ***zvolit správně přístup*** k řešení a navrhnout správně architekturu.

5.4 Závěry k řešením Business Intelligence

- Nástroje ***Business Intelligence***, BI představují již dlouhodobě ***základ řešení*** analytických a plánovacích aplikací a souvisejícího reportingu.
- Řešení aplikací Business Intelligence je charakteristické relativně vysokou individualizací, tj. aplikace jsou obvykle vytvářeny ***podle konkrétních specifických potřeb jednotlivých manažerů*** a specialistů firmy.
- Na trhu je mimořádně ***silná nabídka produktů BI, včetně souvisejících služeb***. To vytváří i podstatně vyšší nároky na kvalifikovaný výběr těchto produktů a služeb na základě vyhodnocení reálných potřeb a možností firmy.
- Systémy BI jsou založené ***na architekturách zahrnujících celou škálu komponent*** různých typů a určení (datový sklad, datová tržiště, ETL a další). Návrh takové architektury musí vycházet z rozsáhlé ***analýzy stavu a potřeb řízení*** jak celé firmy, tak řízení a architektury IT.
- Přehled a ***základní charakteristiky komponent*** obsahuje následující text, kde pro každou komponentu jsou uvedeny obvyklé ***efekty a případné problémy***. Ty je třeba při návrhu architektury BI kvalitně konkretizovat a analyzovat pro další návrhy.
- Systémy IT ve firmě, jak zdrojové, tak BI procházejí trvale ***vývojem a řadou změn***. To při provázanosti BI na zdrojové databáze a aplikace vytváří vysokou ***potřebu kvalitního řízení, právě na rozhraní zdrojů a BI***.
- BI produkty poskytují, kromě vlastního uložení dat a analytiky, ***množství nástrojů*** pro zajištění kvality dat, správu metadat i pro pokročilou analytiku.
- BI systémy prošly vývojem směřujícím ***ke konvergenci nástrojů***, tj. současné databázové systémy zahrnují celou funkcionalitu pro BI. Uživatelům tak odpadají časté starosti se zajištěním integrace mezi různými nástroji. I tak na druhé straně existuje nabídka specializovaných nástrojů, např. pro ETL.
- Užití BI aplikací vytváří u manažerů dobré předpoklady pro ***zvyšování manažerské kvalifikace***, nutí je přirozenou cestou uvažovat o problémech „*multidimenzionálně*“.
- Pro řešení úloh BI je účelné definovat ***sadu analytických otázek***, které by měly být předmětem řešení a diskusí mezi manažery a analytiky.

6. Self Service Business intelligence, SSBI

Účelem kapitoly je:

- vymezit hlavní ***charakteristiky aplikací a produktů SSBI***, jejich pozitiva a problémy,
- vytvořit základní podklad pro ***posouzení jejich možností*** při jejich návrzích samotnými uživateli.

6.1 Podstatné charakteristiky SSBI

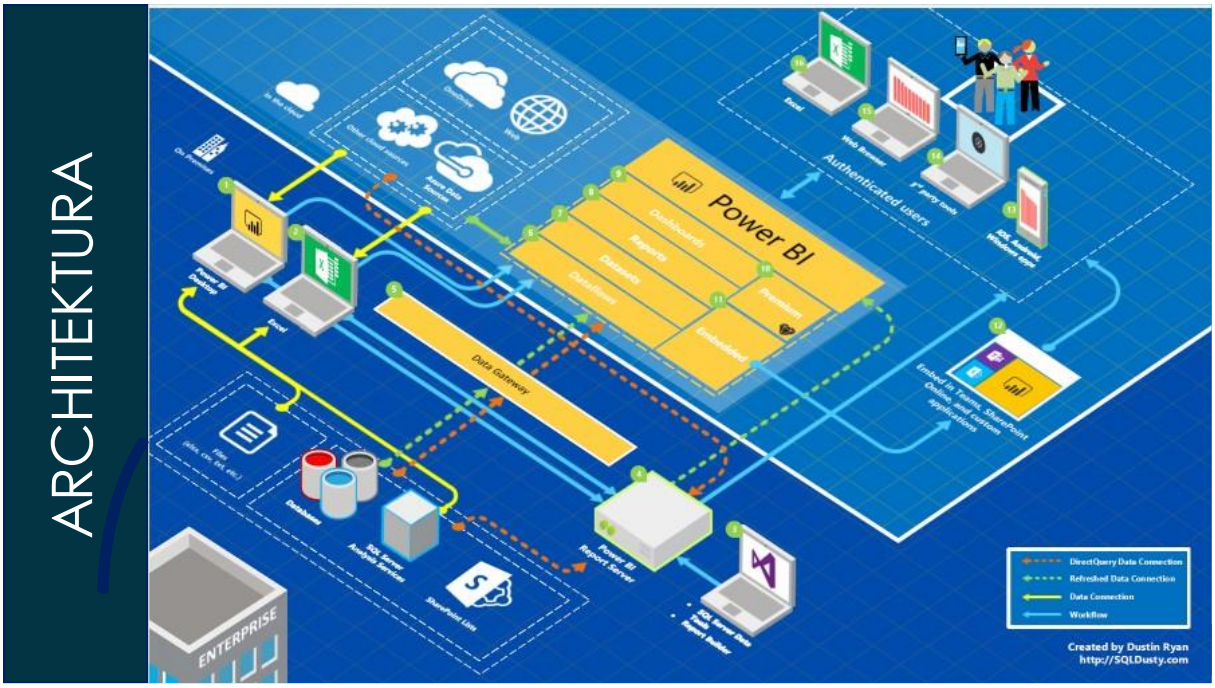
Smyslem SSBI je na základě nových technologií poskytnout uživatelům prostředí pro realizaci svých analytických úloh bez nutnosti využívání komplexních a obvykle velmi složitých systémů BI.

Samoobslužné BI umožňují např. realizovat ***multidimenzionální uložení a zpracování dat*****,** nabízejí efektivní a ***jednoduché přístupy k datům***, poskytují prostředky, např. ***jazyk DAX***, pro výpočty a další operace v dimenzionálním prostředí apod. Kromě toho jsou tyto aplikace i ***vhodným prostředkem pro pochopení podstaty a způsobu využití větších*** Business Intelligence ***systémů***.

Self-Service BI ***se snaží vyřešit rozpor mezi dvěma protichůdnými podnikovými silami*** – potřebou ***flexibility a svobody*** při analýze a zkoumání dat koncovými „spotřebiteli“ BI výstupů, které ovšem bohužel nelze docílit s pomocí „dodavatele“ ve formě podnikového IT oddělení. Proti pak stojí ***potřeba IT oddělení mít neustálou kontrolu nad daty*** a vytvářením a distribucí informací uvnitř firmy.

SSBI rozšiřuje klasické tradiční BI prostředí o ***možnosti provádění vlastních analýz*** nad zpřístupněnými daty a jejich reporting bez nutnosti potřeby zásahu IT oddělení. Dává ne-IT pracovníkům ***více možností, větší flexibilitu a větší samostatnost***, kdy je možné získat odpověď na danou otázku ve výrazně kratší době. Přístup SSBI tedy obchází ***problém s neustále se měnícími a novými požadavky*** a potřebami uživatelů, kdy každý může získat potřebnou informaci, kdykoliv si zamane. Podnikové IT se tak staví do zcela nové role.

Typickým ***představitelem*** produktů SSBI je ***Power BI*** společnosti Microsoft, který v současné době patří i nejrozšířenějším. Schéma jeho architektury představuje Obrázek 6-1:



Obrázek 6-1: Architektura Power BI (Zdroj: Seyfor)

6.2 Efekty SSBI

- Podstatně se zkracuje ***doba potřebná na implementaci*** analytických aplikací.
- Analýza dat ***samotnými byznys uživateli*** **a** dodání požadovaných výstupů/informací je rychlejší. Je možné využít i jiná data než z podnikového datového skladu.
- Self-Service BI aplikace nabízejí obdobnou ***flexibilitu a výkonnost***, jako je tomu většinou u aplikací založených na OLAP databázích.
- Základní ***příprava aplikací*****,** včetně transformací zdrojových dat je výrazně zjednodušená, takže je dostupná i ne IT pracovníkům, samozřejmě po nezbytném zaškolení.
- Součástí technologií jsou i ***programovací prostředky***, jako např. jazyk DAX (Data Analysis Expressions) pro Microsoft Power BI.
- ***Finanční, zdrojová i provozní náročnost*** Self-Service BI je oproti standardním projektům a provozovaným aplikacím výrazně nižší.
- Self-Service poskytuje i velmi dobrou podporu pro tvorbu ***analyticky i vizuálně náročnějších aplikací***, jako např. dashboardů, klikovacích map apod.
- ***Uživatelé dostávají do rukou data a přístup k nim***, která si mohou kdykoliv dle svých aktuálních potřeb zanalyzovat a vyhodnotit za pomoci poskytnutých nástrojů. Nemusí tak dlouho čekat na výstupy, které jim dodá pomalé IT na základě jejich požadavku.
- Namísto toho, aby ***IT oddělení*** bylo zahlceno stále novými a měnícími se požadavky na reporty a analýzy, předává „moc“ práce s daty do rukou samotných uživatelů. Zároveň ***se může soustředit na řízení a aktivity, které přinášejí společnosti větší hodnotu***.
- Změna vede ***k celkové větší spokojenosti uživatelů s IT oddělením*** v organizaci, které se v jejich očích přetváří z podoby úzkého hrdla v rychlosti dodávání BI výstupů, do role flexibilního pomocníka, který lépe dokáže uspokojovat uživatelské byznys požadavky.

6.3 Omezení, problémy, předpoklady SSBI

- Self-Service BI řešení se mohou realizovat ***pouze pro úlohy určité kategorie***, tedy relativně jednodušší a izolovanější úlohy.

- Self-Service BI má ***omezené možnosti čištění a konsolidace dat***, které u standardních BI aplikací představují jednu z pracovně nejnáročnějších, ale i finálně nejefektivnějších jejich částí.
- Problémem je dosažení ***integrace dat*** v rámci podniku, které nabízejí celopodnikové datové sklady.
- S předchozí poznámkou souvisejí i podstatně omezené ***možnosti celopodnikového reportingu***, např. centrálního reportingu v rámci nadnárodních společností apod.
- I když je příprava analytických aplikací na bázi Self-Service BI uživatelsky výrazně dostupnější, přesto je nutné, aby pro kvalitnější aplikace ***uživatelé ovládali alespoň základní principy analytických metod***, tj. především dimenzionálního a datového modelování.
- Je zde zvýšené riziko ***špatných rozhodnutí*** na základě ***„špatných dat“*** od byznys uživatelů.
- Mění se ***pozice IT,*** z producenta výstupů do pozice tzv. „supervizora“. Je na něj tedy kladena tíha řízení a kontroly celého prostředí a produkovaných výstupů.
- Pouhou instalací SSBI nástrojů se požadované přínosy nezískají. Je potřeba, ***aby uživatelé pochopili přínos*** a možnosti daných nástrojů a začali je využívat.

6.4 Závěry k řešení Self-Service Business Intelligence

- Vznik Self-Service Business Intelligence, SSBI znamenal významný mezník v rozvoji podnikové analytiky. Vytvořil ***předpoklady pro samostatnou práci uživatelů*** v návrhu a využití analytických a plánovacích aplikací.
- Návrhy SSBI ***na straně uživatelů*** vytvářejí nároky na ***zvyšování znalostí***, nejen pokud jde o provozované produkty, ale i ***metody*** uplatňované při analýze a návrzích, zejména multidimenzionálního modelování a datového modelování.
- Nasazení aplikací SSBI se ***liší obvykle podle velikosti a potřeb firem***. U menších firem může pokrývat jejich potřeby v celém rozsahu a nahradit komplexní BI systémy. U větších firem tomu tak často není a aplikace SSBI se stávají doplňkem komplexních BI aplikací poskytující uživatelům větší operativnost při jejich práci.
- Na trhu je již adekvátní ***nabídka produktů SSBI i souvisejících služeb***. To vytváří i odpovídající nároky na kvalifikovaný výběr těchto produktů a služeb.
- Současné kvalitní SSBI produkty poskytují ***možnosti vstupů z nejrůznějších zdrojů***, tedy databází, internetových zdrojů, typů souborů atd.
- Systémy IT ve firmě, jak zdrojové, tak BI procházejí trvale ***vývojem a řadou změn***. To při provázanosti SSBI na zdrojové databáze a aplikace vytváří vysokou ***potřebu kvalitního řízení, právě na rozhraní zdrojů a SSBI***.
- I pro řešení úloh SSBI je účelné definovat ***sadu analytických otázek***, které by měly být předmětem řešení a diskusí mezi manažery, analytiky a uživateli.

7. Mobilní BI

Účelem kapitoly je:

- vymezit základní ***charakteristiku mobilních BI aplikací,*** jejich pozitiva a problémy,
- vytvořit podklad pro ***analýzu jejich možností*** při analytické podpoře úloh zejména transakčního charakteru, např. řízení prodeje, nákupu, skladů, majetku.

Mobilní Business Intelligence slouží k zobrazování a ***práci s BI nástroji na mobilních telefonech*** (smartphones) a tabletech.

7.1 Podstatné charakteristiky mobilního BI

Typickým uživatelským scénářem mobilního BI je ***zobrazení dashboardů, reportů*** a možnost s nimi pracovat a využívat základní analytické funkce. Vede to ***k efektivnímu využití informace a možnost flexibilního rozhodování*** založeného na relevantních informacích. ***Cílem mobilního BI*** není nabízet kompletní funkcionalitu, jako desktopové verze, ale využít dostupné aktuální informace k okamžitému rozhodování (Ševčík, 2020).

„Mobilní BI je trendem, kdy byznys uživatelé přistupují ke svým datům a dashboardům pomocí mobilních telefonů a tabletů.“ [Logi Analytics, 2020]. Představuje ***jednoduchý přístup*** k firemním datům, reportům, klíčovým ukazatelům výkonu či dashboardům, pomocí mobilních zařízení. Zároveň některé aplikace umožňují ***sdílení získaných dat*** s kolegy, okomentování dashboardů pro zbytek týmu či dotazování dat v přirozeném jazyce. [MICROSOFT, 2020].

K těmto datům lze ***přistupovat*** v rámci mobilního BI řešení pomocí ***aplikace stažené přímo v konkrétním zařízení***. Tyto aplikace by měly být navrženy speciálně pro tato zařízení a práce s nimi by měla být pro uživatele jednoduchá a poměrně intuitivní. Správně vytvořené aplikace mobilního BI si přizpůsobí zobrazení dashboardu tak, ***aby byl efektivně využit prostor***, ale zároveň bylo vše pro uživatele čitelné a srozumitelné.

Druhou možností je ***přístup přes internetový prohlížeč*** v mobilním zařízení, pomocí kterého se lze dostat na webovou aplikaci, kterou je třeba zpřístupnit. To umožňuje používat ***stejné funkce, jak na počítači, tak i na mobilu***. [Klipfolio, 2020] Problém tohoto řešení může být v tom, že pokud se přistupuje k datům pomocí ***webového prohlížeče***, může se stát, že ***grafy nebudou vypadat správně***, popisky nebudou sedět nebo budou nečitelné nebo prostě jen se bude s webovou aplikací špatně pracovat, protože byla primárně navrhována pro počítačové využití. Webová aplikace musí tak být ***kompatibilní i s mobilním prohlížečem***, a ne pouze s počítačovým prohlížečem. [Logi Analytics, 2020]

Velké úskalí mobilního BI se podle společnosti BARC nachází v nutnosti zabezpečení mobilního zařízení i samotné aplikace. ***Bezpečnostní opatření*** by podle mnohých expertů a společností měla být na vysoké úrovni, kvůli citlivosti dat, se kterými se v aplikaci pracuje a ke kterým by při ztrátě mobilního zařízení mohl získat kdokoliv přístup. [BARC, 2020].

7.2 Efekty mobilního BI

- Existují možnosti ***distribuce obsahu*** do další početné množiny zařízení.
- ***Dostupnost*** informací umožňuje využití hodnoty informace okamžitě po jejím získání. Jinak hodnota informace po jejím získání rychle klesá, čím dřív je využita, tím větší efekt přinese.
- Nabízí se vysoká ***jednoduchost ovládání*** a popisnost poskytnutých dat.
- Umožňuje využití ***funkcí*****,** které ***smartphone nebo tablet*** nabízí – GPS senzor, kamera, WiFi.
- Mobilní BI je v plném souladu s cloud paradigmatem – ***pracovat odkudkoli***, kde je dostupný internet.

7.3 Omezení, problémy, předpoklady mobilního BI

- Návrh uživatelsky ***přívětivého rozhraní*** znamená nutnou optimalizaci pro smartphones a tablety. Uživatelské rozhraní a design aplikace se zcela liší od desktopové verze – je použita jiná technologie interakce (dotyková obrazovka) a jiné uživatelské scénáře.
- Zvyšují se nároky na ***bezpečnost*** – aplikace pracuje s citlivými podnikovými daty. Organizace musí řešit různé úrovně zabezpečení a kontrol přístupu.
- Je obvykle nutné vyvíjet aplikaci ***pro každou platformu zvlášť*** – Android, iOS, Windows Phone.
- ***Internetová služba*** založená na standardu HTML5 řeší předchozí bod a nabízí multiplatformní řešení, avšak HTML5 hrozí rychlé zastarání z důvodu nízké flexibility.

7.4 Závěry k řešení mobilního Business Intelligence

- Nasazení aplikací mobilních BI se ***liší obvykle podle charakteru firem***. Na významu nabývají tam, kde jde o vysokou dislokaci jednotek a aktivit uživatelů, např. v řízení prodeje, nákupu, detašovaných skladů, v řízení investic apod.
- Na trhu je již adekvátní ***nabídka produktů mobilních BI***. To vytváří i odpovídající nároky na kvalifikovaný výběr těchto produktů a služeb.
- I pro řešení úloh mobilních BI je účelné definovat ***sadu analytických otázek***, které by měly být předmětem řešení a diskusí mezi manažery, analytiky a uživateli.

8. Competitive intelligence, CI

Účelem kapitoly je:

- vymezit základní ***charakteristiku aplikací competitive intelligence,*** jejich pozitiva a problémy,
- vytvořit podklad pro ***analýzu jejich možností*** při řešení úloh strategického řízení firmy, marketingu, nebo řízení prodeje.

Competitive Intelligence (CI) je ***systematická a etická disciplína*** zabývající se sběrem, zpracováním a ochranou informací ***s cílem získat konkurenční výhodu*** a zlepšit tak celkové postavení podniku na trhu. K tomu využívá jednak ***primárních informačních zdrojů*** – lidí, tak ***sekundárních artefaktů*** – dokumentů, audiovizuálních záznamů, tiskových zpráv, dat ze sociálních sítí atp.

8.1 Podstatné charakteristiky CI

CI pracuje výhradně ***s legálními informačními zdroji***, nejedná se o průmyslovou špionáž. Disciplína má primární **vazbu na strategické řízení** podniku a podnikové IT, ale ***podporuje analytickou, plánovací a rozhodovací činnost organizace na všech úrovních*** a ve všech oblastech podnikového řízení, tj. prodeje, nákupu, marketingu, finančního řízení, controllingu, majetku, řízení lidských zdrojů, výroby a dalších.

Konkurenční zpravodajství (CI, Competitive Intelligence) je, na rozdíl od průmyslové špionáže, ***etická a legální činnost*** využívající veřejně dostupných zdrojů s cílem zvýšit konkurenceschopnost společnosti za pomoci podpory rozhodování, analýzy trhu, identifikace rizik a příležitostí, ať už současných, nebo budoucích, a to v nekončícím systematickém procesu.

8.2 Efekty CI

- ***Tradiční forma „aktivního CI“*** se primárně zaměřuje na ***externí informační zdroje a tržní okolí*** podniku (odběratelé, dodavatelé, partneři, konkurence, legislativní rámec, profesní a zájmové komunity, zpravodajské portály, sociální sítě apod.) s tím, že ***racionálně umožňuje*****:**
  - zmapovat tržní prostředí a aktivity konkurence,
  - zhodnotit pozici firmy vůči konkurentům,
  - odhalit případné vnější hrozby,
  - identifikovat možné příležitosti pro další růst a rozvoj.
- ***Doménou tzv. „pasivního CI“ (counter-CI****) je pak ochrana vlastních interních zdrojů* před CI aktivitami konkurence a dále snaha zmapovat a vytěžit tyto interní zdroje. Nabízí tyto možnosti:
  - omezit efekt CI aktivit konkurence,
  - upevnit / zlepšit stávající pozici firmy na trhu,
  - identifikovat klíčové nositele znalosti z řad zaměstnanců,
  - odhalit a využít možný skrytý potenciál uvnitř vlastní firmy.
- Podporuje ***strategické plánování a řízení*** firmy díky aktuálním informacím o stavu trhu a aktivitách konkurence.
- ***Zkracuje reakční doby*** v případě nenadálé události či hrozby díky její včasné identifikaci.
- ***Proaktivně vyhledává příležitosti pro inovace*** a růst a tím i zlepšování pozice podniku na trhu.

8.3 Omezení, problémy, předpoklady CI

- ***Efekty CI jsou obtížně měřitelné*** a mají spíše měkký / podpůrný charakter.

- Aktivity CI ***nelze plně automatizovat***, je nutné je neustále přizpůsobovat okolnostem a aktuálním podmínkám.
- ***Úspěšnost CI*** řešení silně závisí na použitých informačních zdrojích – dostupnost kvalitních a spolehlivých informačních zdrojů je zcela klíčová pro efektivní CI.
- CI přináší ***největší efekt v dlouhodobém horizontu*** a vyžaduje trvalou investici do zdrojů (lidských i finančních).
- Pro úspěšnost CI je podstatné zvolit ***vhodný přístup k celému řešení a navrhnout správně architekturu*** s ohledem na další části firmy a dostupné informační zdroje – a to jak koncepčně, tak i z hlediska samotné použité technologie.

8.4 Závěry k řešení Competitive Intelligence

- V rámci Competitive Intelligence se stále uplatňují především ***individuální řešení na míru*** marketingovým a obchodním specialistům a manažerům.
- Analýza a návrh aplikací CI musí vycházet z detailního pochopení ***aktuálních i budoucích obchodních aktivit a potřeb firmy***.
- S ohledem na individuální charakter CI je prakticky ***nezbytná účast manažerů a specialistů firmy*** na jejich řešení.
- Analytická příprava aplikací je zde vesměs založena na ***specifikaci externích zdrojů a jejich charakteru a kvality.***
- Pro řešení úloh CI je rovněž účelné definovat ***sadu analytických otázek***, které by měly být předmětem řešení a diskusí mezi manažery a analytiky.

D) Komponenty podnikové analytiky

[11] Datové báze v podnikové analytice		
[11.1] Datový sklad	**[11.2] Datové tržiště**	**[11.3] Operační datový sklad**
[11.4] Data Lake	**[11.5] Analytické funkce a datový sklad**	
[12] Transformace dat		
[12.1] Dočasné úložiště dat	**[12.2] Extract, Transform, Load (ETL)**	
[12.3] Změny ve zdrojích	**[12.4] Historizace dat**	
[14] Analytika		
[14.1] OLAP databáze	**[14.4] Tabulární model**	
[14.2] In-Memory Analytics	**[14.3] Sandbox**	
[16] Real-Time Data Warehouse		

Oddíl D zahrnuje podstatné ***charakteristiky jednotlivých dílčích komponent a konceptů*** zejména BI a SSBI řešení rozdělených do několik skupin. Kromě ***funkčního a technologického vymezení*** zahrnují jednotlivé kapitoly, resp. podkapitoly i hlavní poskytované ***efekty*** na jedné straně a případná ***omezení*** nebo ***potenciální problémy*** na straně druhé. V návaznosti na předchozí schéma jsou ***v dalším přehledu*** v pracovním členění uvedeny ***komponenty a koncepty pouze s nejstručnější charakteristikou***, detailnější popis je v následujících kapitolách nebo podkapitolách:

- ***Datové báze podnikové analytiky***:

- ***Datový sklad, DWH*** je integrovaný, konsolidovaný, subjektově orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu (INMON, 2002).
- ***Datové tržiště, DMA*** je specifický datový sklad určený pro omezený okruh uživatelů (oddělení, divize, pobočka, závod)
- ***Operační datový sklad, ODS*** je místo datové integrace aktuálních dat z primárních systémů podporující relativně jednoduché dotazy nad menším množstvím aktuálních dat.
- ***Data Lake*** je úložištěm veškerých podnikových dat (strukturovaných i nestrukturovaných) i těch v takzvané surové podobě.
- ***Analytické funkce datového skladu*** představují komplex analytických nástrojů, které využívají data z datových skladů.

- ***Transformace dat***:
  - ***Dočasné úložiště dat, DSA*** představuje dočasné uložení extrahovaných dat z produkčních databází s cílem zajistit jejich přípravu a potřebnou kvalitu před vstupem do datového skladu
  - ***Extract Transform Load (ETL / ELT),*** resp. datová pumpa data ze zdrojových systémů vybírá (Extract), upravuje do požadované formy a uspořádání (Transform) a nahrává je do datových schémat datového skladu nebo tržiště (Load).
  - ***Zachycení změn ve zdrojových systémech, Change Data Capture (CDC)*** reprezentuje schopnost zachytávat změny v datech ve zdrojových systémech a získávat tyto změny efektivně, a pokud je to možné, dosáhnout co nejmenšího zatížení zdrojových systémů.
  - ***Historizace dat*** představuje možnost určit, kdy, která data byla či jsou platná, resp., že všechna data v datovém skladu jsou určitým způsobem identifikována prvky dimenze času.
- ***Analytika:***
  - ***OLAP databáze*** představují jednu nebo několik souvisejících a vzájemně propojených OLAP kostek, které na rozdíl od datových skladů již zahrnují předzpracované agregace dat podle definovaných hierarchických struktur dimenzí a jejich kombinací.
  - ***In-Memory Analytics*** je analytická funkcionalita realizovaná v operační paměti, představuje součást produktů umožňující realizovat multidimenzionální analýzy, a to i s detailními daty, uloženými v operační paměti.
  - ***Sandbox*** představuje možnosti pro uživatelské zkoušení, prototypování a případně testování nových nástrojů a analytických úloh.
  - ***Tabulární model*** představuje sadu metadat, zahrnující tabulky, vztahy, metriky, KPI, skupiny kalkulací (calculation groups), hierarchie, překlady, bezpečnostní role a mnoho dalších prvků.
- ***Real-Time Data Warehouse*** je založen na dvou hlavních principech, to integraci dat v reálném čase a teda proces extrakce, transformace a následného nahrání dát (ETL) v reálném čase a zadruhé na inkrementální údržbě multidimenzionálního modelu datového skladu.

Detailněji se uvedeným komponentám věnují ***další kapitoly***.

11. Datové báze v podnikové analytice

Účelem kapitoly je:

- vymezit základní ***charakteristiky databází,*** majících již analytický charakter, tj. datových skladů, datových tržišť a dalších, jejich pozitiva a problémy,
- vytvořit podklad pro ***analýzu jejich možností a nároků,*** zejména při řešení architektury BI systémů a výběru nejvhodnějších databázových produktů,
- do této kapitoly ***nejsou zařazeny zdrojové databáze***, jejichž otázky jsou řešeny např. s řízením kvality dat.

11.1 Datový sklad, DWH

Snítil, J. (KIT, VŠE), Vaněk, J. (KIT / DHL)

Technologie datových skladů **(*Data Warehouse, DWH*)** představuje v současné době již běžnou součást podnikových IT. Dle (INMON, B.: Building the Data Warehouse. Indianopolis, John Wiley and Sons 2002) je ***datový sklad vymezen*** takto - "Datový sklad je integrovaný, konsolidovaný, subjektově orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu." Tyto pojmy ***lze interpretovat*** takto:

- ***subjektově orientovaný*** – data jsou rozdělována podle jejich typu, ne podle aplikací, ve kterých vznikla,
- ***konsolidovaný*** – data jsou konsolidována z různých zdrojů, struktur a forem do jedné výsledné formy (do jedné verze pravdy),
- ***integrovaný*** – data jsou ukládána v rámci celého podniku, a ne pouze v rámci jednotlivých útvarů,
- ***stálý*** – datové sklady jsou koncipovány převážně jako pouze pro čtení (read only), až na výjimky se zde žádná nová data nevytvářejí ani neaktualizují,
- ***časově rozlišený*** – do datového skladu je uložena i historie dat, tedy obsahují dimenzi času.

11.1.1 Efekty a přínosy datového skladu

- Disponuje ***konsolidovanými, konsistentními, subjektově orientovanými a historizovanými daty,***
- DWH poskytuje ***optimální informace*** podporující manažerské rozhodování na strategické, taktické a operativní úrovni řízení,
- Umožňuje získat ***strukturovaný přehled*** z velkého množství dat uchovávaných v různorodých databázích,
- Nabízí rychlý a ***komplexní přístup*** k velkému množství dat,
- Disponuje schopnostmi pro ***modelování a prognózy*** podle potřeb řízení firmy,
- Funguje na principu ***„read only“,***
- Data jsou do DWH ***nahrávána*** obvykle ***v definovaném časovém úseku***, méně často průběžně (on line),
- Disponuje schopností aktivního ***využití externích datových zdrojů*** třetích stran (např. dlouhodobé zachycení vývoje produktu),
- Disponuje ***operativními dotazy*****,** analýzami časových řad, multidimenzionální analýzou a finančními analýzami,
- Disponuje informacemi uloženými ***v metadatech***,
- DWH představuje datový ***zdroj pro analytické aplikace a reportingové nástroje***.

11.1.2 Otázky a problémy datových skladů

- Vysoká ***technologická a finanční náročnost*** plynoucí z možných kombinací zdrojových systémů a zákaznických požadavků nutných pro realizaci DWH,
- Více než 50 % DWH projektu ***překročí rozpočet***, nebo není dokončena,
- Při zavádění nového podnikového informačního systému je ve valné většině případů nutné ***realizovat nový DWH***,
- Pří zásahu do DWH je v případě změn dimenzí nutné provést ***opětovnou agregaci dat***,
- Oproti klasickým relačním databázím je ***paměťově náročnější*** a dochází k delší doby odezvy,
- Řešení představují komplikované ***načítací procedury***,
- Obecná ***neznalost funkcionality*** a možností datových skladů ze strany podnikových manažerů.

11.1.3 Realizace datových skladů

V současné době datové sklady obvykle ***obsahují normalizovaná data*****,** zatímco denormalizace, např. na bázi STAR schémat se realizuje až na úrovni datových tržišť.

Datové sklady (obdobně datová tržiště) jsou ***realizované v prostředí relačních databázových systémů***, např. MS SQL Server, Oracle, DB/2, Teradata a další.

11.1.4 Požadavky na datový sklad

Požadavky na datový sklad lze formulovat do těchto bodů:

- Informace z DWH ***musí být jednoduše dostupné*** – DWH musí poskytovat informace jednoduše, obsah DWH musí být **pochopitelný** pro uživatele, musí nabízet informace v nejrůznějších kombinacích (slicing and dicing), zajišťovat co nejkratší dobu odezvy,
- Informace z DWH musí být ***presentovány konsistentně*** – musí poskytovat věrohodné informace, tj. data musí být shromážděny z různých zdrojů, pečlivě **kontrolovány a čištěny** a poskytovány až tehdy, pokud jsou v pořádku.
- Data musí být postavena na ***jasném vymezení obsahu*** a jasném odlišení jejich identifikátory, resp. názvy,
- DWH musí být ***flexibilní vzhledem ke změnám*** – musí být adaptabilní ke změnám v uživatelských požadavcích, podnikovému prostředí, datovým zdrojům, technologiím,
- DWH musí ***zajišťovat bezpečnost dat*** – DWH obsahuje kompletní a často citlivé informace, musí být zajištěno řízení přístupů k jednotlivým částem DWH,
- DWH musí tvořit ***základ pro zkvalitňování řídících a rozhodovacích procesů*** – pro DWH musí být deklarovány efekty, které přinese (i když nemusí být vždy exaktně vyjádřeny a ve finančních ukazatelích),
- Podniková ***komunita*** musí DWH ***akceptovat adekvátním způsobem*** – využití DWH (a BI) není, na rozdíl od transakčních systémů, nezbytné a je závislé na ochotě a invenci uživatelů. Proto je také nezbytné ***naplnit všechny předchozí požadavky***.

11.1.5 Úrovně řešení datového skladu

DWH má obvykle 3 úrovňovou architekturu. Dále využívá 2 databáze pro potřeby logování a správy metadat. Celkově se tak skládá z 5 databází:

- ***DWH_Stage (L0):***
    - Vstupní databáze.
    - Obsahuje objekty se strukturou 1:1 oproti objektům ze zdrojových systémů. Je jedno, z jakého zdroje objekt, potažmo data v objektu přicházejí, všechny proudí do této

databáze. Uživatel může jasně rozpoznat, z jakého zdroje objekty jsou pomocí dvou ukazatelů. První je atribut v samotné tabulce, který má defaultní hodnotu jako název zdrojového systému. Druhý je schéma, ve kterém je objekt uložen. Tedy počet schémat je roven počtu zdrojových systémů.

- Tato vrstva se neudržuje. Před každým iterativním nahráváním dat se všechny tabulky promažou.

- ***DWH (L1):***
  - Přestupní vrstva mezi vstupní, a výstupní.
  - Obsahuje více schémat, ovšem nejdůležitější je PreComputed a dbo. PreComputed schéma obsahuje stejné struktury jako L0, ovšem s tím rozdílem, že po každém iterativním nahrání dat se velikost dat pouze rozšiřuje a nemaže. Tedy slouží jako persistentní (stálá) vstupní vrstva. Schéma dbo již obsahuje data transformovaná do podoby dle potřeb budoucích dimenzních a faktových tabulek na úrovni L2. Objekty a atributy se zde řídí dle jmenné konvence, získávají optimální datové tipy a ukládají pouze ty atributy, které se nyní využívají. Pokud některý atribut bude v budoucnu potřebný pro reporting, dá se celá jeho historie transformovat ze schématu PreComputed.
  - Všechny cizí klíče se ve schématu dbo transformují na Surrogate Keys (úspora prostoru, transformace složených klíčů na jednoduché).
- ***DWH_Mart_Service:***
  - Výstupní vrstva.
  - Obsahuje již konečnou podobu dimenzí a faktů tak, aby odpovídaly potřebám reportingu.
- ***DWH_Meta:***
  - Ukládá data o mapování objektů pro potřeby ETL a všechny procedury pro tvorbu procedur z těchto metadat.
- ***DWH_Log:***
  - Zde se ukládají informace o průběhu procedur. Je zdrojem pro monitoring a monitorovací reporty.

11.2 Datové tržiště, DMA

Princip datových tržišť (Data Mart, DMA) je obdobný, jako v případě datových skladů.
Rozdíl je v tom, že datová tržiště jsou určena pro ***omezený okruh uživatelů*** (oddělení, divize, pobočka, závod). Podstatou jsou tak ***decentralizované „datové sklady“,*** které se pak mohou postupně integrovat do celopodnikového řešení.

Datové tržiště je tak ***problémově orientovaný datový sklad, určený pro pokrytí konkrétní problematiky*** vymezeného okruhu uživatelů a umožňující flexibilní „ad-hoc“ analýzy dat. ***Výsledkem*** vytváření datových tržišť je ***zkrácení doby návratnosti investic***, snížení nákladů a podstatné zmenšení rizika při jejich zavádění.

11.2.1 Efekty a přínosy datového tržiště

- DM poskytuje ***informace podporující rozhodování*** manažerů na strategické, taktické a operativní úrovni řízení,
- Nabízí uživatelům ***flexibilní přístup*** k datům,
- V porovnání s DWH ***lze DM vyvinout levněji*** a rychleji,
- V porovnání s DWH poskytuje ***kratší dobu odezvy*** při dotazování,

- Disponuje konsolidovanými, konsistentními, agregovanými, subjektově orientovanými ***daty*** a historizovanými daty,
- Nabízí schopnost získat ***strukturovaný přehled*** z velkého množství dat uchovávaných v různorodých databázích,
- Umožňuje rychlý a ***komplexní přístup*** k rozsáhlému množství dat,
- Přístup je zajištěn pouze ***oprávněným uživatelům*** právě v takovém rozsahu, který jim přísluší v návaznosti na konkrétní právní politiku příslušné organizace.

11.2.2 Otázky, problémy a omezení spojené s datovým tržištěm

- Při zavádění nového podnikového informačního systému je ve většině případů nutné ***realizovat nové DM,***
- Pří zásahu do DM je v případě změn dimenzí nutné provést ***opětovnou agregaci dat***,
- Oproti klasickým relačním databázím je DM ***paměťově náročnější*** a dochází ve většině případů k delší době odezvy,
- Data uložená v DM ***nelze modifikovat***,
- Uživateli je zpřístupněn ***pouze omezený soubor dat***, což může u zaměstnance vést ke špatnému pochopení a k desinformaci kontextu podnikových dat

11.3 Operační datový sklad, ODS

Operační datový sklad (***Operational Data Store, ODS***) je komponenta datové vrstvy, která je charakteristická následujícím ***využitím***:

- Jedná se o ***jednotné místo datové integrace*** aktuálních dat z primárních systémů,
- Podporuje ***relativně jednoduché dotazy*** nad menším množstvím aktuálních dat.

11.3.1 Efekty a přínosy operačního datového skladu pro kvalitu řízení podniku a IT

- Představuje jednotné místo ***datové integrace*** aktuálních dat,
- Nabízí možnost ***rychlého dotazování***,
- Představuje zdroj pro sledování konsolidovaných, agregovaných dat ***s minimální dobou odezvy*** po zpracování,
- Podporuje ***dotazování*** nad omezeným množstvím aktuálních dat,
- Obsahuje ***aktuální záznamy*** vybraného množství dat,
- Podporuje ***interaktivní komunikaci*** se zákazníkem,
- Umožňuje schopnost napojení na ***integrační platformy***,
- Je zdrojem pro ***reportingové nástroje***,
- Nejedná se o ***povinnou*** komponentu,
- Obsahuje ***konzistentní*** a v některých případech agregovaná data.

11.3.2 Problémy a omezení spojené s operačním datovým skladem

- Zvyšuje ***náklady*** na realizaci DWH,
- Neposkytuje pohled na data ***v časovém srovnání***,
- Často náročné na ***integraci*** se zdrojovými systémy

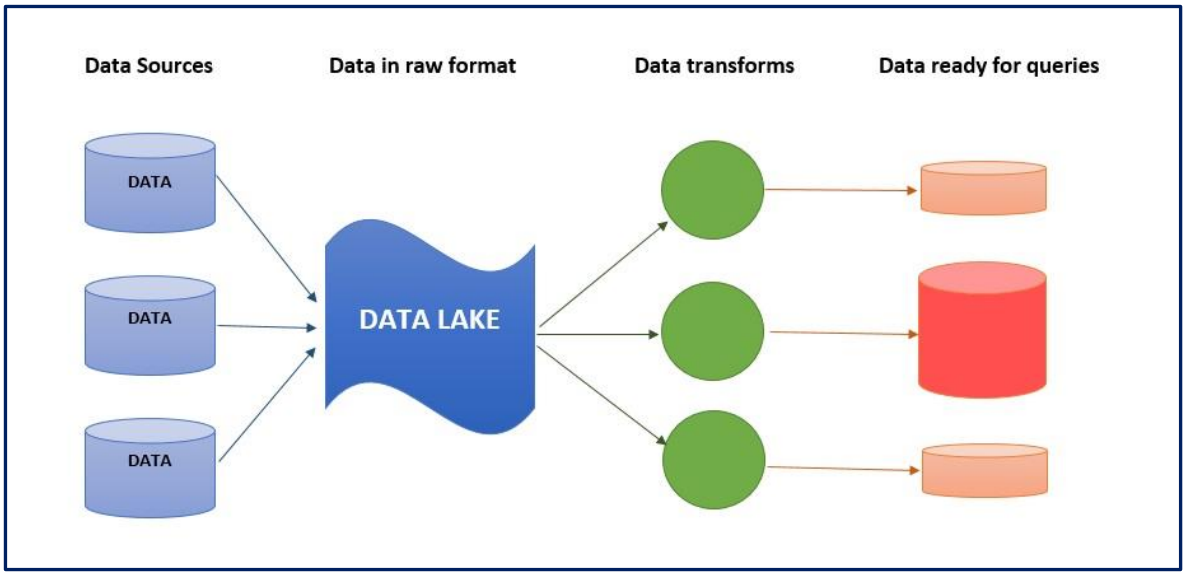
11.4 Data Lake

Vzhledem k ***množství nestrukturovaných dat*** a jejich nárůstu v poslední době, bylo potřeba umět s těmito daty efektivně pracovat a nakládat. Právě z tohoto důvodu ***vznikl Data Lake*** (Ševčík, 2020)

Toto řešení zatím není úplně dokonalé a očekává se zlepšení do budoucnosti, aby nedocházelo k takzvaným ***data swamps***, ale také, aby práce s daty v uloženými v Data Lake byla jednodušší a efektivnější. [DARKWAH, 2020]

Data lake je ***úložištěm veškerých podnikových dat*** i těch ***v takzvané surové podobě***. Jedná se tedy o data, která nebyla zatím nijak zpracována a jejich další účel nemusí být zatím znám nebo určen. Jedna společnost může mít v tom samém repositáři ***uložena strukturovaná data***, například ze svých relačních databází, ***semistrukturovaná data***, jako jsou například soubory vytvořené pomocí Microsoft Excel, ale ***i nestrukturovaná data***, u kterých si jako příklad můžeme uvést obrázky nebo videa. [Talend, 2020]. Data lake se tím stává ***centrem datové integrace*** podniku, které ale zároveň poskytuje schopnost dotazování a získání tím námi požadovaných informací nad velkým množstvím dat v reálném čase.

Další obrázek dokumentuje, že při ukládání dat do Data Lake je ***využíván převážně proces ELT*** (Extract, Load, Transform), díky kterému by mělo být umožněno nejprve nahrání a ***uložení dat s následnou možností provedení více nezávislých transformací*** nad stejnými daty. Data, která jsou ukládána nemají nijak určený specifický formát pro jejich uložení a můžou vstupovat v reálném čase, stejně tak jako v dávkových vstupech. [DARKWAH, 2020]



Obrázek 11-1: Způsob práce s daty v data lake (Zdroj: [30])

Je nutné zajistit, že Data Lake bude vhodně spravován a surová data budou mít požadovanou kvalitu. ***Nároky na správu dat*** jsou poměrně vysoké, protože se v Data Lake pracuje s obrovským objemem dat, která mají nejrůznější struktury. Pokud by došlo k tomu, že by data nebyla správně spravována nebo nebyla spravována vůbec a Data Lake by tím ***přestal poskytovat námi požadovanou a zamýšlenou hodnotu*** informací, respektive by nebylo možné jej nadále efektivně využívat, tak se jedná o Data Swamp. [KNIGHT, 2020]

Nejzákladnější rozdíly mezi Data Warehouse a Data Lake jsou následující:

- ***Data Lake*** obsahuje ***surová data***, u kterých často ***není určen účel***, za kterým jsou ukládána. Může tak dojít k tomu, že budou uložena data, která ani nikdy nebudou reálně použita. Zatímco ***Data Warehouse*** je určen pro ***strukturovaná data***, která již byla anebo jsou používána a zpracována [Talend, 2020].
- Rozdílem je také ***míra nahrání a uložení dat***. Jsou využívány zejména ***dva procesy***. Pro Data Lake je to již zmiňovaný proces ***ELT*** a pro Data Warehouse proces ***ETL*** (Extract, Transform, Load). [DARKWAH, 2020]. Základním rozdílem je zde to, že při ETL procesu dochází k transformaci dat do určitého jasně daného formátu ještě před nahráním do Data Warehouse. U ELT dochází po extrakci dat k jejich nahrání do Data Lake a až poté dochází k případné transformaci. Zároveň proces ***ELT umožňuje provádět více různých transformací*** využívajících stejná data. [PEARLMAN, 2020]

- Jelikož ***Data Lake*** nemá prakticky jasnou strukturu je ***mnohem snazší jej měnit*** a jakékoliv změny dat mohou být provedeny velmi rychle a nenáročně. U ***Data Warehouse*** je tomu právě naopak. Každá změna provedená na datech je velmi náročná a nákladná.
- Poslední základní rozdíl je ***spektrum uživatelů*** schopné využít a porozumět datům uloženým v jednom či druhém úložišti. Data uložená v ***Data Lake*** jsou určena především ***pro informační specialisty*** z důvodu náročnosti porozumění nezpracovaných surových dat. Na druhou stranu datům uloženým v Data Warehouse by měl být schopen porozumět uživatel. Jedná se například o grafy nebo tabulky. [Talend, 2020]

11.4.1 Efekty a přínosy Data Lake

- Poskytuje podstatně ***větší flexibilitu*** při zpracování analytických úloh,
- Představuje jednotné místo ***datové integrace*** z původních oddělených komponent dočasného i operativního úložiště dat,
- Nabízí možnost ***rychlého dotazování***, prakticky v reálném čase s využitím kvalitních uživatelsky orientovaných nástrojů,
- Zahrnuje řízení a ***zpracování velkých objemů*** z různých typů a zdrojů dat,
- Představuje zdroj pro sledování konsolidovaných, agregovaných dat ***s minimální dobou odezvy*** po zpracování,
- Zajišťuje požadovanou ***úroveň kvality*** komplexu dat pro podnikovou analytiku, s odpovídajícími nástroji a funkcemi pro čištění a konsolidaci dat,
- Umožňuje ukládání, zpracování a využívání ***i nestrukturovaných dat***,
- Podporuje jak ***dávkové vstupy*** dat, jak ***vstupy v reálném čase***,
- Podporuje ***interaktivní komunikaci*** se zákazníkem,
- Umožňuje schopnost napojení na ***integrační platformy***,
- Je zdrojem pro ***reportingové nástroje***,
- Obsahuje ***konzistentní*** a v některých případech agregovaná data.
- Zahrnuje i analytické nástroje a nástroje ***pro Data Governance***,
- V jistých případech může Data Lake ***nahrazovat i datový sklad*** a integrovat tak všechny tři původní komponenty a vytvořit tak Logický datový sklad.

11.4.2 Problémy a omezení spojené s Data Lake

- Představuje relativně vysoké ***nároky na Data Governance*** s ohledem na rizika spojená s velkými objemy dat a jejich složitými strukturami,
- Při nezvládnutém řízení Data Lake může ***docházet k problémům s jeho efektivním využitím***,
- Musí být jasně nastavená pravidla pro ***řízení změn v produkčních zdrojích***, včetně zodpovědností za poskytování informací o plánovaných a realizovaných změnách,
- Vysoké nároky na kvalitní ***provázání s celým systémem řízení dat*** v podniku.

11.4.3 Reference:

- DARKWAH, Johnon. Rozdíly mezi Data Lake a Data Warehouse. In: Gauss Algorithmic [online]. 2018-06-20 [cit. 2020-03-13]. Dostupné z: https://www.gaussalgo.com/datova-integrace/rozdily-mezi-data-lake-a-data-warehouse
- Data Lake vs Data Warehouse: Key Differences. In: Talend [online]. [cit. 2020-03-13]. Dostupné z: https://www.talend.com/resources/data-lake-vs-data-warehouse/

- KNIGHT, Michelle. Data Lake vs. Data Swamp: Leveraging Enterprise Data. In: Dataversity [online]. 2018 [cit. 2020-03-14]. Dostupné z: https://www.dataversity.net/data-lake-vs-data-swamp-leveraging-enterprise-data/
- PEARLMAN, Shana. ELT vs. ETL: Defining the Difference. In: Talend [online]. 2019-08-09 [cit. 2020-03-14]. Dostupné z: https://www.talend.com/resources/elt-vs-etl/
- Slánský, D.: Data and Analytics for the 21st Century: Architecture and Governance, Professional Publishing, 2018. ISBN 978-80-88260-16-5.
- Ševčík, O.: Současné trendy v business intelligence. BP, VŠE, 2020

11.5 Analytické funkce datového skladu

Snítil, J. (JSN, s.r.o.)

Účelem využití je zajištění analytických funkcí přímo v datovém skladu.

Koncept se zabývá ***analytickými funkcemi a souvisejícími možnostmi datového skladu***. Analytických nástrojů, které ***využívají data z datových skladů,*** je celá řada. Tyto nástroje svým uživatelům poskytují rozsáhlé možnosti analýzy dat.

V praxi se lze setkat s častým použitím těchto nástrojů ve formě ***tlustých klientů*** případně ***dedikovaných serverů***. Toto použití je vzhledem k vývoji a způsobu využívání těchto analytických nástrojů logické. ***V začátcích při malých datových množinách*** je možné tyto nástroje ***použít přímo na klientských stanicích***, kdy data jsou prvně načtena z datového zdroje (často z datového skladu) do tohoto nástroje a poté jsou v tomto nástroji analyzována.

S rostoucí množinou dat začíná být toto použití ***problematické***, a to především z důvodů nedostatečných výpočetních zdrojů na straně klientské stanice. Jedním z nabízejících se řešení je ***vytvoření dedikovaného analytického serveru***. Tento server je daleko ***lépe škálovatelný*** a náročnost výpočtu je na tento server přenesena z klientské stanice.

S dále rostoucí množinou dat nutnou k analýzám se objevují ***problémy*** i při použití dedikovaného analytického serveru, například ***s migrací dat***. Pokud množství dat začne přesahovat velikost stovek gigabitů, stává se migrace obtížná. Typicky se jedná o ***zatížení přenosové infrastruktury, zdrojového systému i cílového systému***.

Toto zatížení ***nemusí být kritické***, obvykle ale dochází ke značným ***zpožděním***. V běžné praxi také nelze očekávat, že všechny ***tři zmíněné prvky budou plně vyhrazené analytickým potřebám***. Tyto prvky musí plnit svoji ***obvyklou funkcionalitu*** a nelze je tedy dedikovat pro migraci dat. To vede buď ***k ještě větším prodlevám***, nebo ***k přesunutí migrací dat do časového okna***, kdy tyto systémy jsou méně využívané. V obou případech je uživatel nucen ***důkladněji plánovat harmonogram práce*** a to, jaká data budou kdy potřeba na analytickém serveru.

V případě nutnosti ad-hoc analýzy nad jinými, než připravenými daty na analytickém serveru ***mohou být doby dodání této analýzy velmi dlouhé***. Dalším možným ***problémem***, který se objevuje při velké množině dat určené k analýzám, je samotná ***škálovatelnost analytického serveru***. Může se jednat o problémy spojené s ***hardwarovým škálováním***, cenou za ***dodatečné licence***, které pokryjí výkonové nároky nebo i o samotný ***software, který musí umožnit daný hardware efektivně využít***, například tím, že je schopný danou úlohu vhodně paralelizovat.

Dalšími nevýhodami může být samotné ***uložení takto velké množiny dat*** na analytickém serveru, kdy v podstatě dochází k jejich ***duplikaci.*** Jedním z možných ***řešení*** těchto problémů může být *přenesení výpočtu přímo do datového skladu*, kdy je uživateli nabídnuta podobná funkcionalita a prostředí jako v případě analytických aplikací. Toto použití datového skladu klade ***nároky na použitou technologii datového skladu***, která musí toto použití vhodně podporovat tak, aby samotný výpočet mohl probíhat přímo v datovém skladu.

Výhodou tohoto použití datového skladu je, že ***nedochází k nutnosti migrovat data,*** a tedy není nutné vytvářet celá řešení kolem této migrace. Veškerá ***data*** uložená v datovém skladu také mohou být ***okamžitě dostupná pro analytické potřeby***. V případě hlubších analýz a například při vytváření různých modelů může analytikům poskytnout ***větší efektivitu práce a výrazně zkrátit dobu*** potřebnou pro dodání výsledků.

11.5.1 Efekty a přínosy analytických funkcí

- Jedná o sdílený výkon, kdy ***není potřeba výkonově škálovat dvě technologické platformy a řešit integraci*** mezi nimi.
- Je možné takto ušetřené ***finanční prostředky investovat právě do jedné technologie,*** což může přinést flexibilitu v rámci přesunů výpočetního výkonu podle aktuálních priorit, kterými může být například zpracování dat, reporting či právě pokročilá analytika. Pokud nastane mimořádná situace a je potřeba akutně zpracovat určitá data, lze výkon alokovat právě na tuto činnost. V případě dedikovaného analytického serveru by datový sklad při takového situaci nemusel mít dostatečný výkon na poskytnutí aktuálních dat dedikovanému analytickému serveru, což by mohlo způsobit, že jeho výkonová kapacita by nemohla být využitá.
- Všechna data v datovém skladu jsou ***ihned k dispozici*** pro analytické použití. Není potřeba provádět migraci či jejich úpravu.
- Za možnou důležitou výhodu tohoto konceptu lze označit relativně ***jednoduchou přenositelnost ze stávajících řešení***, které jsou již v organizaci používané pro analytické úlohy.
- Pokud analytická platforma datového skladu podporuje tyto technologie nebo jazyky je jejich ***převod relativně jednoduchý***. V podstatě se jedná jen o napojení na dané rozhraní poskytované danou platformou. S relativně nízkou pracností lze tak řešení přesunout do technologie datového skladu a využívat tak již zmíněné výhody tohoto rozšiřujícího konceptu.

11.5.2 Otázky a problémy analytických funkcí

- Při použití tohoto konceptu v reálném prostředí je potřeba ***zvážit dodatečné požadavky na výpočetní výkon***, které vzniknou při použití technologie datového skladu i pro analytické potřeby. Při volbě konkrétní technologie a jejího výpočetního výkonu je tedy potřeba do odhadu zahrnout kromě obvyklých výkonových požadavků, které jsou kladeny na datový sklad i požadavky související se specifickými analytickými úlohami, které zde budou zpracovávané.
- Lze doporučit při použití tohoto konceptu navrhnout a ***provozovat vhodné řízení výpočetních zdrojů (tzv. Database Resource Management)*** právě s ohledem na analytické potřeby a jejich dopady do vytížení celého systému.
- Při zvažování tohoto přesunu nebo při výběru nové dodatečné technologie v rámci datového skladu pro analytické použití je možné v praxi téměř vždy doporučit provést ***podrobné vyzkoušení na konkrétních příkladech***. Zejména se jedná o případy, kdy je již například dané analytické oddělení specializováno na používání určitého analytického nástroje. Přeučení a adaptace zcela nové technologie by poté mohla být velmi časově i finančně náročná. Pokud ale nová technologie umožňuje vysokou míru možné integrace, přechod a kombinace daných technologií by mohla být vysoce přínosná.
- Při praktickém použití tohoto konceptu je potřeba ***prověřit jeho chování na konkrétní zvolené technologii*** a případně zařadit čištění dočasných objektů mezi pravidelně prováděnou údržbu.

11.6 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řešení a hodnocení databázových systémů jako součásti komplexních BI řešení.

.

- Analytické databázové systémy se ***liší svojí mohutností a pozicí*** v architekturách BI řešení.
- Jednou z klíčových otázek je, zda bude systém postaven na ***jednom centralizovaném datovém skladu*** nebo na ***datových tržištích***, resp. ***kombinaci*** datového skladu a datových tržišť

- Další aktuální variantou může být ***vytvoření „Data Lake“*** integrující všechny analytické databáze nebo jejich části.
- Analytické databáze mají řadu svých specifik a pak je účelné disponovat ***specialisty*** pro tato řešení.
- S ohledem na různé účely je nutné dobře rozlišit řešení ***normalizace dat*** na úrovni datového skladu i datových tržišť.

12. Transformace dat

Účelem kapitoly je:

- vymezit základní ***charakteristiky těch součástí BI systémů,*** které se podílejí na transformacích dat mezi zdrojovými a analytickými databázemi, resp. mezi analytickými zdroji navzájem,
- vytvořit podklad pro ***analýzu možností a nároků*** jednotlivých typů produktů pro datové transformace, zejména při řešení architektury BI systémů a výběru nejvhodnějších technologií a řešení,
- poskytnout základní informace o funkcích související i procesem transformace, tj. o možnostech ***efektivního zachycování změn ve zdrojových systémech*** a o řešení ***historizace dat***..

12.1 Dočasné úložiště dat, DSA

Úkolem dočasného úložiště dat (***Data Staging Area, DSA***) je dočasné uložení extrahovaných dat z produkčních databází s cílem ***zajistit jejich přípravu a potřebnou kvalitu*** před vstupem do datového skladu. Pro data v dočasném úložišti je podstatné, že jsou to ***data detailní, neagregovaná, často nekonsistentní, bez časové dimenze***. DSA obsahuje ***pouze aktuální data***, tj. po jejich zpracování v DSA a přenosu do datového skladu nebo tržiště se z DSA odstraní.

12.1.1 Efekty a přínosy DSA pro kvalitu řízení podniku a IT

- ***Uvolnění výkonu*** primárních systémů,
- Urychluje proces ***extrakce*** dat,
- Nejedná se o ***povinnou*** komponentu,
- Snižuje celkové ***náklady*** na provoz DWH.

12.1.2 Problémy a omezení spojené s DSA

- Data ***nejsou agregována*** ani historizována,
- Data jsou ***nahrazena*** po jejich zpracování další dávkou dat,
- Data v DSA ***nejsou kontrolována***,
- Zvyšuje ***náklady*** na realizaci DWH

12.2 Extract Transform Load (ETL/ELT)

ETL (Extract, Transform, Load) je jednou z nejvýznamnějších komponent celého komplexu business intelligence. Běžným označením pro prostředky ETL je rovněž ***datová pumpa***. Jejím ***úkolem*** je data ze zdrojových systémů ***vybrat (Extract), upravit*** data do požadované formy a uspořádání ***(Transform) a nahrát*** je do specifických datových struktur, resp. datových schémat datového skladu nebo tržiště ***(Load).*** Tyto nástroje lze tedy použít pro ***přenos dat mezi dvěma (či více) libovolnými databázemi, nebo datovými soubory*** (textovými, tabulkovými, XML soubory atd.).

ETL nástroje pracují obvykle ***v dávkovém režimu***, data jsou tedy přenášena najednou v určitých časových intervalech, např. denních, týdenních apod. Právě transformace dat spojené s ETL jsou ***pracovně, časově i finančně nejnáročnější*** a obvykle představují cca 60 % vynaložených pracovních kapacit. Pro úspěšné BI řešení však znamenají zcela nezbytný předpoklad.

12.2.1 Efekty a přínosy ETL/ELT

- Díky ETL lze z množiny nestrukturovaných dat ***vyjmout*** právě ta ***data*****,** která mají být ***relevantní*** vzhledem k zákaznickým požadavkům na podporu řízení, resp. k řešení datového skladu (DWH),
- Velké množství ETL nástrojů umožňuje ***flexibilní přístup*** a schopnost identifikovat právě nástroj, který nejvíce odpovídá řešení konkrétní úlohy,
- Pomocí ELT metody lze rozdělit DWH procesy na ***diverzní části***, které následně dělí projekt na subprojekty a nabízí tak transparentní pohled na celkový proces,
- Obvykle je ELT implementace všech zdrojových dat nahrávána do DWH jako ***část extrakčního a loadovacího procesu.*** Tato situace v kombinaci s isolovaným transformačním procesem znamená, že možné změny mohou být v DWH struktuře relativně snadno realizovány,
- ***Čištění dat*** - odstranění duplicitní dat, chyb způsobených lidským faktorem, atd.,
- ETL procesy v běžném případě probíhají ***ve stejné časové frekvenci*** a ve stanoveném čase (většinou v noci). Tento způsob nasazení je zcela vyhovující pro většinu operativních analýz,
- Podporuje ***speciální typy analýz***, vyžadujících okamžité vyhodnocení, př. fraud systémy,
- ***Centralizovaná správa*** transformací bývá uložena v repozitáři,
- Nabízí možnost vytváření a ***správy metadat***,
- Řízené ***plánování a spouštění transformací***,
- Správně navržený ETL proces ***umožňuje načíst data*** s neúplnou, nebo ***porušenou referenční integritou***,
- Správně navržený ETL systém ***zaznamenává všechny chyby*** do chybového žurnálu.

12.2.2 Problémy a omezení spojené s ETL/ELT

- Modifikace původní' ETL procedury představuje ***nárůst nákladů a času*** na její řešení,
- V případě změny ETL technologie je nutná ***znalost aktuálního i původního skriptovacího jazyka,***
- Pro řešitele je nezbytně nutná ***znalost transformačních technik***, metodik a funkcionalit,
- Při ***nedodržení pravidel*** uplatňovaných při ETL (tj. funkcionalita a metodika), je princip ETL prakticky nevyužitelný a stává se při opětovném využití nepřehledným (nečitelná dokumentace způsobuje nepřehledný systém ETL procedur),
- Vysoká ***míra závislosti na analýze*** primárního systému,
- ***Technologické bariery*** (např. systém není schopen transformovat zdrojový datový formát).

12.2.3 Podstatné charakteristiky ETL/ELT

Pro ETL jsou **podstatné** následující **charakteristiky:**

- ze zdrojových databází musí být vybrána pouze taková ***data,*** která jsou určena pro ***analytické, plánovací a rozhodovací aktivity*** podniku (tedy nikoli všechna) a to je jedním z prvních úkolů analytiků business intelligence,
- data jsou transformována do nových datových struktur ***analytických databází***, které musí být předem navrženy tak, aby nejlépe ***odpovídaly potřebám řízení podniku*****.** To souvisí s možnostmi využití multidimenzionality a granularity dat,
- data do business intelligence ***vstupují z různých zdrojových databází*** (ERP, e-Business, CRM atd.), přičemž v těchto různých zdrojích, mohou být jedna a ***tatáž data uložena vícekrát*** (a k tomu ještě různě), např. různé databáze prodejců, zákazníků apod. Ale do analytických databází musí vesměs vstoupit pouze jednou. V transformační vrstvě (s využitím ETL) ***musí dojít ke konsolidaci dat,*** tj. určení vstupujících dat s vyloučením duplicit, či multiplicit,

- se zajištěním konsolidace dat úzce souvisí i **dosažení potřebné kvality dat,** tj. vyloučení chyb, nepřesností atd.

12.3 Zachycení změn ve zdrojových systémech, CDC

Účelem fungování ***CDC je schopnost zachytávat změny v datech ve zdrojových systémech.*** Snahou použití CDC je získávat tyto změny efektivně, a pokud je to možné, dosáhnout co nejmenšího zatížení těchto zdrojových systémů. Získání zdrojových dat by tedy nemělo mít ***žádný anebo pouze zanedbatelný dopad do běžného používání zdrojových systémů,*** a to zejména tak, aby nebyla narušena jejich primární funkce.

Zachycení změn ve zdrojových systémech, resp. ***Change Data Capture (CDC)*** se v architektuře datového skladu obvykle nachází ***mezi zdrojovými systémy*** (zejména jejich databázemi) ***a dočasným uložištěm*** dat.

CDC ***nabízí alternativu k použití ETL***, které vždy ***zatíží zdrojovou databázi***, a to z důvodu, že se do zdrojové databáze připojuje ***přes konektor***, z kterého čte data (nehledě na to, zda se jedná o nativní konektor konkrétní databáze nebo o generický konektor, jakým je například ODBC či JDBC). Naproti tomu ***CDC*** se dokáže vyhnout tomuto zatížení, protože ***dokáže změny provedené v datech vyčítat z databázových logů***, obvykle se tedy nevytváří žádné databázové připojení, databáze nemusí tvořit exekuční plán a provádět načítání dat z disků.

Jedním z nástrojů, které je možné použít pro implementaci tohoto rozšiřujícího konceptu je například *IBM InfoSphere Change Data Capture*. Ve většině případů tedy ***není potřeba nahrávat data z těchto databází a z jejich diskových úložišť,*** ale pomocí znalosti databázových logů lze odvodit, jaká data byla, jak změněna a na základě toho vyvodit potřebný stav v cílovém uložišti.

12.3.1 Efekty a přínosy zachycení změn ve zdrojových systémech

- Uplatněním konceptu se lze ***vyhnout*** výraznému ***zatížení zdrojových systémů*** a zároveň je možné potřebná data zpřístupnit pro další zpracování.
- Efektem je relativně ***rychlé zachycení změn*** ve zdrojových systémech, a tedy tyto změny mohou být velmi rychle zapisovány do cílového systému a zpřístupněny tak k dalšímu zpracování.
- V případě nasazení specializovaného CDC nástroje v konkrétním prostředí je velkou výhodou zejména v případě, kdy zvolený ***nástroj umí změny získávat ze všech zdrojových systémů***, které jsou provozované v daném prostředí.
- Nasazení specializovaného nástroje vede ke ***zjednodušení používaného řešení,*** kdy místo různých způsobů získávání dat ze zdrojových systémů by existoval jednotný způsob získávání zdrojových dat. Správa takového řešení a jeho provoz by tak mohl být efektivnější.

12.3.2 Otázky a problémy zachycení změn ve zdrojových systémech

- Uvedený způsob získání datových změn ze zdrojových systémů ***na první pohled nepřináší výrazné nevýhody***.
- Určitým rizikem a omezením jsou případné ***změny ve strukturách tabulek na zdrojovém systému.*** V takovém případě může dojít k zastavení přenášení změn ze zdrojového systému do cílového a je nutné provést znovu mapování dat.

12.4 Historizace dat

Účelem historizace je zajistit ***potřebu v cílovém systému udržovat historickou kolekci dat*** (v datovém skladu) a podle toho i data vhodně transformovat.
Koncept Historizace se zabývá splněním části ***definice*** datového skladu Williama Inmona v tom smyslu, že *datová kolekce je historická*. To znamená, že je možné ***určit, kdy, která data byla či jsou platná.*** Z jiného úhlu pohledu lze také říci, že ***všechna data*** v datovém skladu jsou určitým způsobem ***identifikována prvky dimenze času***.

Jelikož se jedná o z***ákladní funkčnost*** datového skladu (ať v užším či širším pojetí) ***existuje mnoho řešení,*** jak splnit tuto definiční podmínku. V tomto kontextu jsou podstatné ***možností historizace po***

použití konceptu CDC, tj. jaký dopad bude mít pro datový sklad situace, kdy přicházejí ze zdrojových systémů ***pouze provedené změny***. Tím tento koncept logicky navazuje na rozšiřující koncept CDC.

V takovém případě ***lze výstup z CDC považovat za první vrstvu datového skladu***, kterou je možné označit jako první nultou vrstvu ***L0 (z anglického slova layer)***. Z této vrstvy je potřeba ***data transformovat do další vrstvy L1***, která již obsahuje historickou kolekci dat. Pokud zanedbáme přenosové a technologické zpoždění mezi zdrojovým systémem a vrstvou ***L0***, tato vrstva obsahuje ***identický stav dat jako je na zdrojovém systému***. Naopak ***vrstva L1*** obsahuje ***poslední známý stav zdrojových systémů*** provedený v rámci ***poslední transformace***. Tím, že nemáme k dispozici informaci, která data byla změněna v rámci vrstvy L0, musí se během této transformace L0 do L1 provést ***porovnání dat L0 vůči poslednímu dostupnému stavu dat ve vrstvě L1.***

Toto ***porovnání nutné pro identifikaci změn*** ve zdrojovém systému vůči historické kolekci ve vrstvě L1 je výkonově poměrně náročné a je spojené s dalšími možnými nevýhodami, které jsou podobné jako v případě použití celkového exportu dat ze zdrojových systémů.

Další možností je ta, kdy každý ***přenesený záznam obsahuje časovou značku*** určující čas změny na zdrojovém systému, ke které se daný záznam váže. Oproti předchozím dvěma mapováním přenáší do cílového systému ***pouze změny ze zdrojového systému***. Při použití tohoto mapování ***nelze jednoduše zjistit stav zdrojového systému,*** ale tento stav se musí ze získaných změn dopočítat.

12.4.1 Efekty a přínosy historizace

- Celý koncept je použitelný a ***umožňuje zpracování dat z výstupů CDC***, kde jsou obsaženy pouze změny provedené na zdrojovém systému, do další vrstvy, která reprezentuje celkový obraz dat na zdrojovém systému včetně historie.
- Je možné dosáhnout značné ***univerzálnosti celého konceptu***, kdy nehledě na zdrojový systém je možné vytvářet relativně jednoduše historické kolekce dat. Tyto historické kolekce dat poté mohou sloužit jako vstup pro následné transformační úlohy, které jsou obvykle pracnější a časově náročnější na vytvoření. Zvláště v případech, kdy tyto transformační úlohy jsou teprve vyvíjené, lze jim po jejich vytvoření a otestování nabídnout historická data a doplnit tak i zpětně nové datové struktury v dalších konsolidovaných vrstvách řešení datového skladu.
- ***Možnost zpracování dat k určeným časům***. Pokud je pro specifické potřeby nezbytné vycházet z dat platných k určeným časovým bodům například z důvodů porovnatelnosti reportů, nebo z potřeb sledovat vývojové trendy, toto řešení takového zpracování umožňuje, a to i zpětně, kdy s využitím auditních dat lze dopočítat stav zdrojového systému k danému okamžiku.
- Data zpracovaná pomocí tohoto konceptu lze také ***použít pro důkladné a/nebo velmi složité analýzy***. Tato data totiž obsahují historii dat zdrojových systémů a zároveň jsou ve stejné podobě jako data zdrojových systémů. Je zde tedy minimální riziko možné ztráty určité informace, které existuje v obvyklých transformačních a agregačních datových úlohách. Také je zde minimalizované riziko možných chyb, které mohou být omylem způsobené transformačními úlohami.

12.4.2 Otázky a problémy historizace

Při použití tohoto rozšiřujícího konceptu je ***potřeba zvážit četnost zpracován***í vzhledem k výpočetní náročnosti. I když v tomto případě nedochází k porovnání dvou plných snímků, tedy celkovému stavu dat ve zdrojových systémech a poslednímu dostupnému stavu dat v historické kolekci, ale jsou v tomto případě zpracovávané pouze provedené změny na zdrojovém systému, určitá ***výpočetní náročnost na udržování historické kolekce*** zde existuje, zvláště pak v porovnání s řešeními, které historickou kolekci dat zdrojových systémů neudržují. V každém případě lze doporučit vhodně využívat optimalizace, které zvolená technologie pro zpracování dat nabízí.

Určitým rizikem, které je spojeno s tímto rozšiřujícím konceptem, je zejména v dlouhodobém horizontu ***nárůst objemu uchovávaných dat.*** Před zavedením tohoto konceptu je vhodné provést analýzy za účelem odhadnutí datového objemu historické kolekce a tempa jejího růstu. Řešením tohoto problému může být ***přesouvání starších dat z historické kolekce do jiného systému,*** který nabízí relativně levné uložení těchto dat. V případě potřeby je možné i historickou kolekci dat odmazávat a udržovat v ní například pouze poslední rok, tak aby bylo dosaženo optimálního poměru mezi cenou datového uložiště a přidanou hodnotou v ní obsažených dat.

klíčovou roli pro transformace dat hrají ***ETL nástroje***, resp. datové pumpy, ale i ty mohou být nahrazovány speciálně vyvinutými aplikacemi.

12.5 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řešení transformací dat a možnostem, které se zde nabízejí.

.

- Transformace dat patří k pracovně i časově nejnáročnějším aktivitám v rámci řešení podnikové analytiky. Je tomu tak nutné věnovat ***v harmonogramech dostatečný časový prostor***.
- V průběhu transformací dat se řeší i nezbytné ***kontroly jejich kvality*** a odpovídající ***čištění*** dat, neboť datová analytika je na kvalitu dat značně citlivá, např. na úrovni dočasného úložiště dat.
- Jednou z otázek je, ***jakými prostředky*** bude efektivní transformace realizovat, tj. integrovanými prostředky v rámci databází (např. MS Integration Services), specializovanými prostředky (např. Informatica), případně vytvořením vlastních nástrojů, např. na bázi SQL skriptu.
- V současnosti se nabízejí i prostředky pro ***automatické zachycování změn*** ve zdrojových systémech označených jako CDC, Change Data Capture.
- V souvislosti s transformacemi dat je nutné posoudit i nároky na řešení ***historizace dat***, tj. zachycení jejich hodnot v průběhu časového vývoje.

14. Analytika

Účelem kapitoly je:

- vymezit základní ***charakteristiky součástí BI systémů,*** které se vztahují přímo k realizaci analytických funkcí, kde obvyklou základní komponentou jsou OLAP databáze,
- současně i poskytnout základní informace o technologiích ***tabulárních modelů*** a ***in-memory analytics*** jako o jedněch z výrazných trendů podnikové analytiky,
- vytvořit podklad pro ***analýzu možností a nároků*** jednotlivých typů technologií a produktů pro analytické funkce, zejména při řešení architektury BI systémů a výběru nejvhodnějších technologií a řešení,
- poskytnout základní informace o funkcích a možnostech ***uživatelského testování analytiky, tzv. sandbox***..

14.1 OLAP databáze

OLAP databáze představují jednu nebo několik souvisejících a ***vzájemně propojených OLAP kostek.*** Ty většinou, na rozdíl od datových skladů již ***zahrnují předzpracované agregace dat*** podle definovaných hierarchických struktur dimenzí a jejich kombinací.

- ***Varianty OLAP technologie:***
- ***MOLAP (Multidimensional OLAP)*** - speciální uložení dat v multidimenzionálních – binárních OLAP kostkách,
- ***ROLAP (Relational OLAP)*** - řeší multidimenzionalitu uložením dat v relační databázi,
- ***HOLAP (Hybrid OLAP)*** - kombinace předchozích přístupů - detailní data v relační databázi a agregace v OLAP kostkách,
- ***DOLAP (Desktop OLAP)*** - umožňuje připojit se k centrálním úložišti OLAP dat a stáhnout si podmnožinu kostky na lokální počítač. Analytické operace - prováděny nad lokální kostkou - výhodné pro mobilní aplikace a podporu mobilních uživatelů,
- ***WOLAP (WEB based OLAP)*** – kombinace OLAP a web technologií,

14.1.1 Efekty a přínosy OLAP

- Nabízí využití ***multidimenzionálních analýz***,
- Výhodou je transparentní pohled na ***aktuální data***, sloužící jako podklad při rozhodování,
- Umožňuje ***podporu reportingu***,
- Představuje vstup pro ***realizaci metod*** (např. BSC, dynamický BSC atd.),
- Podporuje automatizaci ***plánovacích procesů***,
- Umožňuje ***uvolnění lidských zdrojů***, kontrolu vývoje reality a porovnání s plánem,
- Disponuje funkcemi ***drill down, drill up, dril through***,
- Umožňuje předdefinovat ***jednotnou obchodní logiku***, jako například způsob výpočtu ceny produktu apod.,
- Disponuje ***agregovanými ukazateli***,
- Nabízí ***předdefinovanou funkcionalitu*** (např. identifikaci KPI),
- ***MOLAP :*** navržený pro rychlé přijímání dat, podporuje slice & dice operace, obsahuje předpřipravené kalkulace a dále je specifický rychlou odezvou při dotazování,
- ***ROLAP:*** Je specifický podporou rozsáhlého množství dat, kdy rozsah závisí na velikosti zdrojové databáze, umožňuje využití funkcionality zdrojové databáze,

- ***HOLAP*:** Zahrnuje kumulaci a možné kombinace výše popsaných výhod.

14.1.2 Problémy a omezení spojené s OLAP databázemi

- ***MOLAP* :** Je charakteristický prací s omezeným rozsahem dat, kdy dochází ke generování kalkulací již při samotné tvorbě OLAP kostky,
- ***ROLAP* :** Omezení ROLAP vychází z SQL funkcionality zdrojové databáze, oproti MOLAP disponuje horší dobou odezvy při SQL dotazování,
- ***HOLAP*** : představuje kumulaci a kombinace výše popsaných nevýhod

14.2 In-Memory Analytics

Analytická funkcionalita v paměti, resp. ***analytické aplikace realizované v operační paměti*** (In Memory Analytics) je nabízena jako technologie, resp. součást nových produktů umožňující realizovat multidimenzionální analýzy, a to i s detailními daty, uloženými v operační paměti. Je to pojetí konceptu BI, které ***pro zvýšení výkonu*** ve své architektuře používá ***komponentu in-memory databází***.

Existuje několik přístupů k in-memory koncepci. Prvním přístupem je ***kompletní in-memory řešení,*** nebo ***integrace in-memory databáze*** do stávajícího řešení BI.

14.2.1 Efekty a přínosy In-Memory Analytics

- Umožňuje díky výkonu in-memory databáze analyzovat data doslova ***v reálném čase***, s dobou odezvy pouhých několika sekund. Hodí se pro zpracování velkých objemů dat a pro složité výpočty nad těmito daty.
- Pomáhá ***odstínit zátěž vedlejších i zdrojových systémů***. Různé systémy disponují různou dobou odezvy, která se navíc mění se stupněm zátěže těchto systémů. Přidáním komponenty in-memory databáze do řešení BI se tak značně odlehčí zátěž vedlejších systémů, včetně systémů zdrojových.
- Nastartování procesu ***zlepšení datové kvality*.** Dostupnost informací je základním předpokladem datové kvality. Díky zvýšení výkonnosti analýzy dat uživatel získá informace v požadovaném čase (nebo alespoň výrazně aktuálnější informace).
- ***Jednodušší provoz a údržba*** systému. Jelikož analytické aplikace pracují nad obrovským množstvím dat, vyžadují vysoký výpočetní výkon. S kompletním in-memory řešením není již potřeba složitě budovat architekturu BI po výše uvedených vrstvách (nákup HW, SW, implementace, integrace aplikací a dat, ladění systému). In-memory řešení je možné získat od jediného dodavatele v mnohem kratším čase a tím snížit náklady na provoz a údržbu systému.
- ***Standardizované BI aplikace*** – S příchodem koncepčně „jednoduchého“ in-memory řešení BI není třeba vyvíjet aplikace šité zákazníkovi na míru, nebo vyvíjet kompletní řešení na „zelené louce“. Již dnes několik dodavatelů nabízí hotové řešení BI pro standardizované systémy typu ERP, CRM a další. Ty obsahují předdefinované komponenty pro jednotlivé vrstvy (ETL, vlastní datový model, sady reportů a dashboardů).
- ***Masové rozšíření BI***. Díky jednoduššími nasazení, nárůstu výkonu i novému typu BI aplikací postupně dojde k postupnému rozšíření řešení BI mezi širší okruh uživatelů. Vzhledem k flexibilitě škálování řešení, si ho budou moci v budoucnu dovolit i malé a střední podniky.

14.2.2 Problémy a omezení spojené s In-Memory Analytics

- ***Cena nejvýkonnějších řešení*** zůstává díky vysoké ceně operačních pamětí prozatím velmi vysoká. Díky tomu si mohou in-memory BI pro zpracování opravdu velkých objemů dat dovolit jen velké společnosti. (Na druhé straně se však in-memory používá i pro self service BI a tedy i pro menší řešení),

- Pro úspěšnou implementaci řešení je potřeba zvolit ***vhodný přístup pro danou situaci***, ve které se podnik nachází. Na základě potřeb podniku je klíčové vybrat komponenty řešení tak, aby plně vyhovovaly jeho potřebám,
- Využití in-memory databází ***bez použití energeticky nezávislých operačních pamětí*** typu NVRAM s sebou nese zásadní nedostatek takového řešení. Tím je potřeba nahrání dat do in-memory databáze ze zdrojových systémů nejen při plánované aktualizaci, ale i při výpadku systému.

14.3 Sandbox

Poskytuje prostor pro ***uživatelské zkoušení, prototypování*** a případně testování nových nástrojů a analytických úloh.

14.3.1 Efekty a přínosy Sandbox

- Nabízí možnost ***rychlého zkoušení a dotazování***, s využitím nových uživatelsky orientovaných nástrojů,
- ***Nejedná se o povinnou*** komponentu,
- S větším využitím podporuje efektivně ***zvyšování kvalifikace a znalostních předpokladů*** uživatelů pro další běžný provoz v podnikové analytice.

14.3.2 Problémy a omezení spojené se Sandbox

- Sandbox ***nemůže být využíván v běžném*** provozním režimu,
- ***Nepodporuje obvyklou dokumentaci*** ani běžně předpokládanou datovou kvalitu dat,
- ***Nepodporuje backupy*** a obnovu dat po výpadcích,
- ***Nenabízí obvyklou úroveň zajištění bezpečnosti*** zpracování a využívání dat,
- Je otázkou, ***jaká data a s jakými zodpovědnostmi*** do Sandboxu ukládat.

14.3.3 Reference:

- Slánský, D.: Data and Analytics for the 21st Century: Architecture and Governance, Professional Publishing, 2018. ISBN 978-80-88260-16-5.

14.4 Tabulární model

Stanovská, I., Pátková, N. (KIT, VŠE)

14.4.1 Tabulární model, principy

Tabulární model lze definovat jako „sadu metadat, zahrnující tabulky, vztahy, metriky, KPI, skupiny kalkulací (calculation groups), hierarchie, překlady, bezpečnostní role a mnoho dalších prvků, které tvoří tzv. sémantický model, který zajišťuje systém navigace v klientských nástrojích, jako je realizace reportů v Power BI nebo v Excelu.

Model odpovídá datové sadě ve službě Power BI anebo databázi ve službě Analysis Services (Azure Analysis Sevices anebo SQL Server Analysis Service). Datové sady a databáze jsou modely naplněné daty. Definice modelu bez jakýchkoli dat je „tabulární model“. (Russo 2022)

Datové sady anebo databáze založené na tabulárních modelech jsou speciální ***analytické databáze***, které pro ukládání dat mohou používat 2 druhy řešení:

- ***In-memory***, což znamená, že vykonávají dotazy na data v paměti aplikace nad celou databází,

- ***DirectQuery***, což znamená, že data nejsou kopírována do aplikace, přes kterou k nim přistupujeme (například Power BI Desktop, Microsoft SQL Server Analysis Services, Azure Analysis Services), ale dotazy jsou předávány přímo do zdrojových databází, kde probíhá zpracování a do aplikace je předáván až výsledek dotazu. V tomto případě je tedy tabulární model založen v příslušném nástroji, ale data nejsou jeho součástí.

Tabulární modely ***kombinují technologii MOLAP a relačních databází***. (Rabeler et al. 2018)

Zatímco ***in-memory*** modely jsou ***výchozí***, ***DirectQuery režim se používá pro*** příliš velké objemy dat na to, aby se vešly do paměti, nebo pokud volatilita dat vylučuje rozumnou strategii zpracování. ***DirectQuery*** dosahuje shody s in-memory modely prostřednictvím ***podpory široké škály datových zdrojů***, schopností zpracovávat kalkulované tabulky a sloupce v DirectQuery modelu, zabezpečení na úrovni řádků prostřednictvím výrazů DAX, které se dostanou do databáze typu back-end, a optimalizace dotazů, jejichž výsledkem je rychlejší průchodnost (propustnost). (Rabeler et al. 2018)

Tabulární modely jsou definovány strukturami a protokoly dokumentovanými Microsoftem (“. (Russo 2022):

- MS-SSAS-T: SQL Server Analysis Services Tabular Protocol
- Tabular Object Model (TOM)
- Tabular Model Scripting Language (TMSL)
- XML for Analysis (XMLA) Reference

Sama definice tabulárního modelu má díky TOM strukturu JSON souboru, známá jako model.bim. Díky tomu lze model vytvářet prostřednictvím:

- Power BI
- Microsoft Visual Studio (Analysis Services Porject) ,
- Tabular Editoru,

Tabulární modely mohou být ***nasazeny do služby Azure Analysis Services*** nebo na instance služby ***SQL Server Analysis Services (SSAS)*** nakonfigurované na režim Tabular server anebo do služby Power BI Premium. Nasazené tabulární modely je možné spravovat přes SQL Server Management Studio. (Rabeler et al. 2018)

Nástroje tabulárního původně podporovaly mezi tabulkami ***pouze vazby 1:1 nebo 1:N s jednosměrnou i obousměrnou filtrací***. Od roku 2020 jsou implementovány i vazby M:N, i když je doporučeno tento typ vazeb používat pouze ve specifických scénářích užití, stejně jako obousměrné vazby 1:N.

14.4.2 Engine

Tato část pojednává o ***výpočetních engine***, které jsou ***součástí tabulárního modelu***. Dotazy na tabulární databáze jsou v analytických databázích (SSAS, AAS, PowerBI) zpracovávány ***ve 2 vrstvách výpočetních engine, formula engine a storage engine***.

Existují ***2 druhy storage engine***,

- VertiPaq,
- DirectQuery.

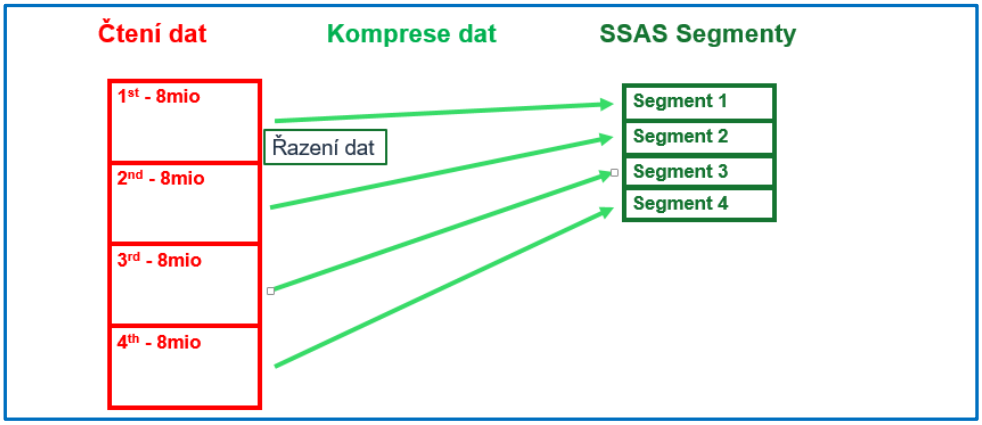
14.4.3 Vertipaq engine

Datový model oficiálně využívá ***xVelocity in-memory Analysis Services***, avšak z toho důvodu, že in-memory engine byl dlouhou dobu pojmenován jako Vertipaq, je známější pod tímto názvem. Navíc interně je stále zobrazován název Vertipaq a dotazovací engine vykonává Vertipaq dotazy. (Haman, 2018). Vertipaq je implementován jak SSAS, AAS tak i v Power BI desktop

Vertipaq engine při načítání dat nejprve ***data přečte*** a následně je ***transformuje do sloupcové struktury*** VertiPaq, tedy tabulku si rozdělí na sloupce. ***Pro jednotlivé sloupce*** tabulky si vytvoří ***slovník jedinečných hodnot***, které dále nahrazuje ***indexy***. Sloupce tak obsahují pouze číselné datové typy, konkrétně ***integer***. (Haman, 2018)

V dalším kroku si VertiPaq engine ***seřadí duplicitní výskyty indexů*** a zpracuje je podle algoritmu RLE (= Run Length Encoding). ***RLE*** je algoritmus ***pro komprimaci dat***. Aby bylo dosaženo lepšího kompresního poměru, RLE tedy data nejprve seřadí tak, aby řádky následující ve sloupci obsahovaly

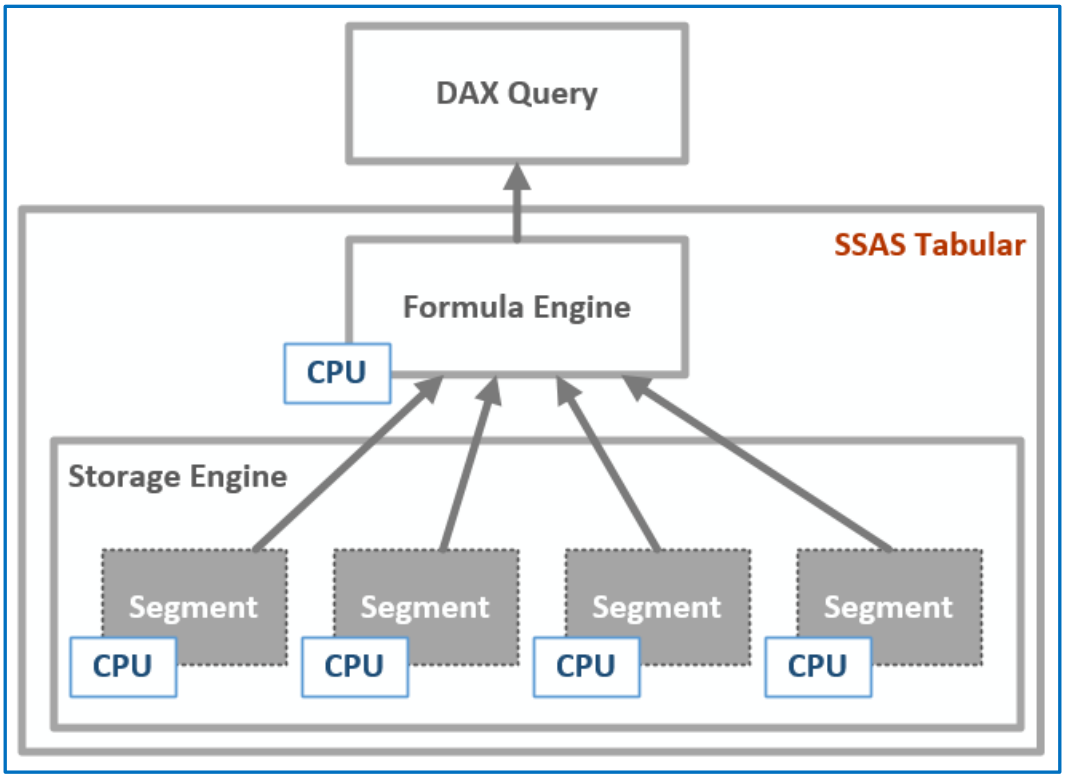
stejnou hodnotu. Data jsou čtena a ***zpracovávána po dávkách***, v angličtině *batch*. Jeden batch může obsahovat až 8 milionů řádků. ***Data jsou tedy rozdělena do dávek, zkomprimována a uložena do segmentů,*** jak je znázorněno níže na obrázku (Obrázek 14-1). Výsledné řešení je závislé na poměru počtu hodnot a počtu unikátních hodnot, ale především je paměťově úspornější a rychleji prohledatelné, neboť je uložen pouze počet opakování stejných hodnot. (Joyful Craftsmen, 2018) Znalost tohoto principu je důležitá, při návrhu tabulárních modelů. Čím větší počet sloupců v tabulkách s velmi rozdílnými dlouhými textovými řetězci mají vliv na budoucí problémy s velkostí výkonností modelů, protože nelze data dost dobře komprimovat.



Obrázek 14-1: Komprese dat ve VertiPaq engine (Joyful Craftsmen, 2018)

Komplexnější kalkulace vykonává formula engine po získání výsledků z VertiPaq engine. (NisalBI 2017).

Na obrázku (Obrázek 14-2) je nastíněno schéma zpracování dotazů, kde lze vidět, že Formula engine zpracovává dotazy od jednoho uživatele pouze na jednom vlákně, zatímco VertiPaq engine používá více vláken paralelně.



Obrázek 14-2 Schéma zpracování dotazů v SSAS Tabular (Joyful Craftsmen, 2018)

Pro in-memory engine je nutné, aby RAM na serveru, kde je datový model umístěn, byla dvakrát až třikrát větší, než je velikost datové kostky.

Při návrhu struktury tabulárního modelu je doporučeno, aby výsledná model měl podobu „star“. A vazby mezi faktovými a dimenzionálními tabulkami realizovat pomocí atributů typu integer. Takové modely jsou výkonnější. (Russo 2021)

14.4.4 DirectQuery engine

DirectQuery mód se ***dotazuje na data přímo na zdroj***, tedy posílá SQL dotazy do zdrojové relační databáze, které SQL podporují. Je vhodnější použít directquery pro velké datové objemy, jelikož ukládá pouze metadata, a navíc poskytuje nejaktuálnější data. (Duncan 2019)

Typ engine je možné změnit i po nasazení datové kostky do produkce.

14.4.5 Perspektivy

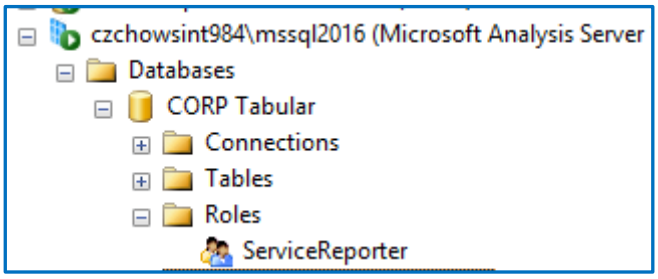
Perspektivy v tabulárních modelech ***umožňují definovat podmnožiny modelu***, které budou zobrazené ***specifikovaným uživatelům či skupinám*** uživatelů, a zobrazují různé pohledy na konkrétní model tabulární databáze. Primární záměr vytváření perspektiv není zabezpečení tabulárního modelu, nýbrž ***usnadnění přístupu*** k datům pro uživatele. Veškeré ***zabezpečení pro konkrétní perspektivu je zděděno*** ze zabezpečení celého modelu. Není tedy možné poskytnout uživateli pomocí perspektiv přístup k objektům modelu, ke kterým přístup nemá. Perspektivy je možné vytvářet v návrháři modelu pomocí dialogového okna Perspectives. Následně, například v Excelu připojeném na tabulární model v AAS nebo SSAS si uživatel může vybrat pouze ty perspektivy, které mu jsou nabízeny.

14.4.6 Security handling v tabulárním modelu

Pro zabezpečení tabulárního modelu se používá ***definice rolí uživatelů v databázi***. Zabezpečení modelu pomocí rolí lze řešit ***až na úroveň řádků (tzv. RLS – Row Level Security)***, které již vyžaduje použití jazyku DAX.

14.4.7 Databázové Role

Jak lze vidět na obrázku (Obrázek 14-3), zabezpečení analytické databáze je ***obdobné jako zabezpečení relační databáze***. V databázi nebo přímo v tabulárním modelu jsou vytvářeny ***role, pod které jsou přidáváni uživatelé***. Těmto rolím jsou pak ***udělována práva***, a to buď plná kontrola nad databází (Administrator), práva na procesování databáze (Process database), práva pouze na čtení dat z databáze (Read) či práva na čtení a procesování databáze (Read and Process).



Obrázek 14-3 Pohled na role v databázi (vlastní zpracování)

Vytvořené role lze zobrazovat a vytvářet filtry jak v aplikaci SSDT, tak v aplikaci SSMS. Filtry lze vytvářet pomocí výrazů v jazyku DAX. K této úrovni zabezpečení lze také využít možnosti proměnných, kdy lze nastavit podmínky za využití těchto proměnných, jako je například Username(), která vrací uživatelské jméno přistupujícího uživatele ze služby Active Directory.

14.4.8 ACL tabulky

Access Control List (ACL), v překladu ***seznam řízení přístupu***, je seznam obsahující informace o oprávnění a přístupech k objektu, k němuž se váže. ***Jednotlivé záznamy*** v ACL seznamu lze označit jako ***Access Control Entries (ACE)***, v překladu položky řízení přístupu, kdy jednotlivé ***položky označují jednotlivé uživatele či skupiny uživatelů a jejich oprávnění***, která se mohou týkat nejen přístupu, ale také práce se soubory či daty. Atributy, které se mohou v seznamu ACL objevit, se odvozují od objektu, ke kterému je ACL vázán.

Mohou být 2 typy ACL pro zabezpečitelný objekt:

- Discretionary Access Control List (DACL)
- System Access Control List (SACL).

DACL, v překladu volitelný seznam řízení přístupu, pracuje se zabezpečením tak, že jakmile se proces pokusí o přístup k objektu, ***systém nejprve zkontroluje, zda tento objekt má DACL***. Pokud objekt DACL nemá, přístup k objektu je tomuto procesu udělen v plné míře. Pokud však objekt má DACL, systém v sekvenci prochází veškeré záznamy, tedy ACE, v tomto seznamu, dokud nenajde alespoň jeden záznam ACE nebo pokud neprojde celý seznam. Pokud je DACL objektu prázdné, tedy neobsahuje žádné záznamy, systém odepře přístup veškerým procesům. Pokud systém nalezne v seznamu hledané ACE, zachová se následně dle obsahu tohoto záznamu v seznamu.

SACL, v překladu seznam řízení přístupu systému, slouží k zaznamenávání pokusů o přístup k objektu, ke kterému je SACL vázáno. Pokusy o přístup jsou zaznamenávány do ***Security event logu***, tedy do protokolu událostí zabezpečení, kam je možné zaznamenat veškeré přístupy.

14.4.9 Zdroje

DUNCAN, Owen. 2019. *What is Azure Analysis Services?* [online]. Microsoft Docs. [cit. 2019-04-30]. Dostupné z: https://docs.microsoft.com/en-us/azure/analysis-services/analysis-services-overview

DUNCAN, Owen, HAMILTON, Bruce, CAI, Saisang, DUNN, James, KUMAR, Sudeep, MILENER, Gene a RABELER, Carl. 2018. *About SQL Server Analysis Services* [online]. Microsoft Docs. [cit. 2019-04-30]. Dostupné z: https://docs.microsoft.com/en-us/sql/analysis-services/analysis-services?view=sql-server-2017

DUNCAN, Owen, GUYER, Craig a RABELER, Carl. 2018. *Tabular model roles* [online]. Microsoft Docs. [cit. 2019-04-14]. Dostupné z: https://docs.microsoft.com/en-us/sql/analysis-services/tabular-models/tabular-model-roles-ssas-tabular?view=sql-server-2017

GÁLA, Libor, POUR, Jan a ŠEDIVÁ, Zuzana. 2009. *Podniková informatika.* 2. přepracované vydání. Praha: Grada Publishing. ISBN 978-80-247-2615-1.

GARTNER, 2017. *Gartner IT Glossary – Active Directory* [online]. Gartner. [cit. 2019-04-27]. Dostupné z: https://www.gartner.com/it-glossary/active-directory/

GARTNER, 2017. *Gartner IT Glossary - Self-Service Business Intelligence*. Gartner [online]. [cit. 2019-04-14]. Dostupné z: https://www.gartner.com/it-glossary/?s=self+service+business+intelligence

GUYER, Craig, MILENER, Gene, KUMAR, Sudeep, HAMILTON, Bruce, ROTH, Jason, LAUDENSCHLAGER, Douglas a RABELER, Carl. 2017. *Master data services Overview (MDS)* [online]. Microsoft Docs. [cit. 2019-04-30]. Dostupné z: https://docs.microsoft.com/en-us/sql/master-data-services/master-data-services-overview-mds?view=sql-server-2017

HAMAN, Martin. 2018. *Data Day 2018* [online]. Dostupné z: https://www.gopas.sk/getattachment/47219b58-3c18-427e-b304-c940604185a8/Martin_Haman_all.pdf.aspx;.php;.html;;

JOYFUL CRAFTSMAN. 2018. *Hitchhiker's Guide to the SSAS Tabular* [online]. [cit. 2019-04-14]. Dostupné z: https://www.sqlsaturday.com/779/Sessions/Details.aspx?sid=85530

KENNEDY, John a SATRAN, Michael. 2018. *Access Control Lists.* Microsoft Docs [online]. [cit. 2019-04-23]. Dostupné z: https://docs.microsoft.com/en-us/windows/desktop/secauthz/access-control-lists

LACKO, Luboslav. 2003. *Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL Serveru a Oracle*. Vydání první. Brno: Computer Press. ISBN 80-7226-969-0.

MBI tým. *AQ001A*: *Aplikace: Power Pivot* [online]. MBI. [cit. 2019-04-30]. Dostupné z: https://mbi.vse.cz/mbi/index.html#obj/APPLICATION-156

POUR, Jan, MARYŠKA, Miloš, STANOVSKÁ, Iva a ŠEDIVÁ, Zuzana. 2018. *Self Service Business Intelligence: Jak si vytvořit vlastní analytické, plánovací a reportingové aplikace.* Praha: Grada Publishing. ISBN 978-80-271-0616-5.

RABELER, Carl, GUYER, Craig a DUNCAN, Owen. 2018. *Tabular Models*. Microsoft Docs [online]. [cit. 2019-04-14]. Dostupné z: https://docs.microsoft.com/en-us/sql/analysis-services/tabular-models/tabular-models-ssas?view=sql-server-2017

RABELER, Carl, GUYER, Craig a DUNCAN, Owen. 2018. *Perspectives in tabular models* [online]. Microsoft Docs. [cit. 2019-04-29]. Dostupné z: https://docs.microsoft.com/en-us/sql/analysis-services/tabular-models/perspectives-ssas-tabular?view=sql-server-2017

ŘEHOŘ, F. *F455: BI: Self Service BI* [online]. MBI. [cit. 2019-04-14]. Dostupné z: https://mbi.vse.cz/mbi/index.html#obj/FACTOR-110

RUSSO, M. 2022 *Development tools for Tabular models in 2021*. [cit. 2022-05-22]. Dostupné z:

https://www.sqlbi.com/articles/development-tools-for-tabular-models-in-2021/

RUSSO, M *, FERRARI, A. Header/Detail vs Star Schema models in Tabular and Power BI* . 2021 [cit. 2021-10-11]. Dostupné z: https://www.sqlbi.com/articles/header-detail-vs-star-schema-models-in-tabular-and-power-bi/

STANOVSKÁ, Iva. *AQ002A: Aplikace: Power BI – Principy, funkce* [online]. MBI. [cit. 2019-04-30]. Dostupné z: https://mbi.vse.cz/mbi/index.html#obj/APPLICATION-161

SUPPORT OFFICE, 2019. *Přehled technologie OLAP (Online Analytical Processing)* [online]. Support Office. [cit. 2019-04-14]. Dostupné z: https://support.office.com/cs-cz/article/p%C5%99ehled-technologie-olap-online-analytical-processing-15d2cdde-f70b-4277-b009-ed732b75fdd6#top

TEJADA, Zoiner. 2018. *Technologie OLAP (Online Analytical Processing)* [online]. Microsoft Docs. [cit. 2019-04-14]. Dostupné z: https://docs.microsoft.com/cs-cz/azure/architecture/data-guide/relational-data/online-analytical-processing

VertiPaq engine in SSAS Tabular Database [online]. 2017. [cit. 2019-04-30]. Dostupné z: http://nisalbi.blogspot.com/2017/11/vertipaq-engine-in-ssas-tabular-databse.html

What is SharePoint? [online]. Microsoft Support Office. [cit. 2019-04-30]. Dostupné z: https://support.office.com/en-us/article/what-is-sharepoint-97b915e6-651b-43b2-827d-fb25777f446f

ZHANG, Janine, GUYER, Craig, MILENER, Gene a MABEE, Dan. 2018. *SQL Server Integration Services* [online]. Microsoft Docs. [cit. 2019-04-30]. Dostupné z: https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-2017

14.5 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řešení a realizaci vlastních analytických operací.

.

- Uvedené prostředky a přístupy představují ty, které bezprostředně ***realizují analytické operace*** s daty.
- Jádrem prostředků analytických operací jsou ***OLAP databáze*** a jejich různé formy (viz výše)

- Efektivním řešením analytiky, zejména ve vazbě na produkty self service business intelligence jsou ***tabulární modely***, snižující nároky na zdroje při potřebné provozním výkonu.
- Velmi častým problémem je zajištění ***přípravy uživatelů*** s možností odzkoušení si některých analytických funkcí, k tomu se využívá prostředí označované jako „***sandbox***".

16. Real-Time Data Warehouse

Zálepa, M. (KIT, VŠE)

Účelem kapitoly je:

- vymezit ***podstatu datového skladu fungujícího v reálném čase,***
- definovat všechny další charakteristiky a součásti řešení, zejména ***možnosti transformací dat*** z datových zdrojů v prostředí jejich zpracování v reálném čase.
- vytvořit podklad pro ***analýzu možností a nároků*** využití zpracování v reálném čase ve vazbě na funkce datového skladu.

16.1 Zpracování v reálném čase

Dátový sklad označujú (FAROOQ, SARWAR, 2010) ako čoraz viac používanú technológiu na zlepšenie Business Intelligence v podniku. Tradičné dátové sklady zvyčajne poskytujú na analýzu jeden deň staré historické dáta, zatiaľ čo aktuálne transakčné dáta zostávajú v operačných, respektíve zdrojových systémoch. V dnešnej rýchlej dobe a vysoko konkurenčnom prostredí potrebujú biznis užívatelia mať dáta k dispozícii čo najskôr, ideálne v reálnom čase a preto vzniká ***veľký dopyt po implementácii real-time dátových skladov,*** alebo skratkou aj nazývaných RTDW (Real-Time Data Warehouse).

RTDW prináša do sveta dátových skladov ***dve výzvy***. Prvou z nich je ***integrácia dát v reálnom čase*** a teda proces extrahovania, transformácie a následného nahranie dát (ETL) a druhou výzvou je ***inkrementálna údržba DW multidimenzionálneho modelu (DMM***).

ETL nástroje, ktoré slúžia na prenos dát v reálnom čase majú za úlohu ***identifikovať, zachytiť, transformovať a následne nahrať zmenu dát okamžite do dátového skladu***. Statický, historický a agregovaný model dátového skladu ja potom potrebné prepočítať, alebo postupne udržiavať. Pod ***inkrementálnou údržbou*** je myslené, že do DMM sa šíria ***iba aktualizácie údajov***. Voľby DMM (to znamená štruktúrovaná, alebo polo štruktúrovaná) však veľmi výrazne ovplyvňuje rýchlosť prírastkovej údržby. Žiadna ETL technológia na prenos dát v reálnom čase nie je užitočná, ak latencia je spôsobená modelovaním zmien dát.

Taktiež (SANTOS, BERNARDION, 2008) potvrdzujú, že dopyt po čerstvosti dát v dátovom sklade neustále rastie. ***Aktualizácia dát tradične prebieha v off-line režime***, to znamená, že v čase keď prebieha aktualizácia dát, tak pre užívateľov dáta nie sú dostupné. Preto sa tieto procesy vykonávajú v jasne daných časových oknách, kedy sa neočakáva, že užívatelia budú k dátam pristupovať. Na druhú stranu je stále viac a viac podnikov, ktoré fungujú na báze 24x7 a preto je žiadúce aby dáta boli čo najčerstvejšie a neustále dostupné. Ako záver svojho reportu, ktorý prezentoval T. B. Pedersen, uviedol že čoraz viac podnikov potvrdzuje to, že dáta v reálnom čase a neustála dostupnosť dát sú priority pre podniky, ktoré sú dátovo orientované. IT manažéri čelia pred výzvou či ich dátové sklady už nie sú zastaralé a keď nebudú dáta spracovať dáta v reálnom čase či im neutečie ich konkurencia. V niektorých špecifických prípadoch, ak budú dáta aktualizované s odozvou väčšou než niekoľko minút, či dokonca sekúnd, môže byť ohrozená použiteľnosť celého systému.

Dosiahnutie takmer nulovej odozvy medzi OLTP a OLAP systémami pozostáva v ***zaistení nepretržitej dátovej integrácie*** zo zdrojových systémov do cieľových, teda do dátového skladu. Aby bolo možné toto dosiahnuť je potreba vziať do úvahy nasledujúce:

- OLTP systémy sú navrhnuté tak, aby dokázali naplniť krátky čas na odpoveď s cieľom dosiahnuť to, aby bol ***systém čo najviac dostupný***. To znamená, že RTDW by sa musel vyrovnať s úplne identickou situáciou.
- ***Faktové tabuľky*** v dátovom sklade sú zvyčajne veľkých rozmerov a preto ***pridanie nových údajov*** a s tým spojených úkonov ako aktualizácia indexov, alebo kontrola referenčnej integrity by určite mali ***vplyv na výkonnosť OLAP*** systémov a dostupnosť údajov.

16.2 ETL prístupy

ETL technológie na prenos dát (FAROOQ, SARWAR, 2010) klasifikujú ***do štyroch nasledujúcich kategórii:***

- Tradičné ETL
- ETL takmer v reálnom čase (NR-ETL = Near Real-Time ETL)
- ETL s priamym vstupom dát (ETL-DDF = ETL with direct data feed)
- ETL s dátovými cache v reálnom čase (ETL-RDC = ETL with real time data caches)

Tabuľka 1 zhrňuje aký sú hlavné rozdiely v jednotlivých prístupoch. V stĺpci ***Typ ETL*** sa nachádza určenie o aký typ prístupu ide. Stĺpec ***Mód*** určuje ako dochádza ku spracovaniu záznamov, teda či sa klasicky spracováva v jednotlivých dávkach, alebo ide o inkrementálne spracovanie. Stĺpec ***Frekvencia*** nám určuje ako často k tomuto spracovaniu dochádza. Stĺpec ***Veľkosť dát*** hovorí o tom, aké množstvo dát sa spracuje v jednej dávke, alebo v jednom inkrementálnom naliatí dát. Stĺpec ***Odozva*** určuje to, ako sú dáta v dátovom sklade v podstate neaktuálne voči skutočnosti, respektíve voči zdrojovému systému.

Tabuľka 1: Porovnanie ETL prístupov (FAROOQ, SARWAR, 2010)

Typ ETL	Mód	Frekvencia	Veľkosť dát	Odozva
Tradičné ETL	Dávka	Mesačne, týždenne, denne	Veľká	Hodiny
NR-ETL	Inkrementálne	Denne, Dva krát za deň, Každú hodinu, Každú minútu	Nízka	Minúty
ETL-DDF	Inkrementálne	Každú minútu, každú sekundu	Nízka	Minúty, sekundy
ETL-RDC	Inkrementálne	Každú minútu, každú sekundu	Nízka	Minúty, sekundy

16.3 Požiadavky na RTDW

Ako tvrdia (SANTOS, BERNARDION, 2008), tak pre podniky nie je problém vygenerovať obrovské množstvo dát, ktoré s ľahkosťou dosahujú aj niekoľko megabajtov či gigabajtov denne. Tým pádom ako plynie čas a neustále zbierame dáta do dátového skladu, tak jeho veľkosť môže narásť až na terabajty či dokonca petabajty. To znamená, že pri každom dotaze, ktoré do dátového skladu pošleme, tak pristupujeme k obrovskému množstvu dát s tým že nad tými dátami vykonávame rôzne operácie ako spájanie tabuliek, radenie, zoskupovanie záznamov a rôzne analytické funkcie, tým pádom sa reakčná doba stále viac a viac zhoršuje. Aby dátový sklad dokázal túto ***reakčnú dobu znížiť*** tak používame napríklad ***indexy, alebo partitions*** (slúžia na rozdelenie tabuľky na menšie časti), ktoré taktiež zaberajú dosť miesta a vyznačujú sa taktiež svojou zložitosťou. Toto všetko jasne potvrdzuje, že aktualizácia dát v reálnom čase nie je úplne jednoduchá úloha. V závislosti na frekvencii a objeme dát môže veľmi ľahko nastať situácia, kedy server úplne preťažíme. Nehovoriac o tom aký negatívny vplyv to môže spôsobiť na výkonnosti OLAP.

V skratke sa ***Real-Time Data Warehouse snaží o skrátenie času, ktorý je potrebný na spravenie rozhodnutia a pokúša sa dosiahnuť nulovú latenciu*** medzi príčinou a efektom tohto rozhodnutia.

Pre RTDW je potrebné sa vysporiadať minimálne s dvoma výraznými zmenami. Jednou z nich je ***neustále vykonávanie aktualizácii*** do dátového skladu, ktoré by sa malo z väčšej časti týkať najmä pridávania záznamov. A druhou je, že tieto ***aktualizácie by mali prebiehať paralelne*** s vykonávaním OLAP dotazov. Teda hlavné požiadavky na RTDW by sa dali zhrnúť ako:

- Dosiahnuť ***čo najvyššiu aktuálnosť dát*** pomocou efektívnej a rýchlej aktualizácie dát z OLTP systémov do dátového skladu
- ***Minimalizovať odozvy na OLAP*** dotazy počas toho ako súčasne prebieha neustála integrácia dát
- ***Maximalizovať dostupnosť dátového skladu*** pomocou toho, že je zredukované časové okno, ktoré slúži na aktualizáciu dát v dávkach, počas ktorého sú OLAP aplikácie off-line

16.4 Prístupy k spracovaniu dát v reálnom čase

Ako potvrdzujú (GUERRA, ANDREWS, 2011), tak kľúčovým termínom pri ***dosahovaní synchronizácie v reálnom čase je Change Data Capture, označovaná aj ako CDC***. Jedná sa vlastne o sledovanie zmien nad danou tabuľkou, respektíve databázou. V priebehu niekoľkých rokov bolo odskúšaných mnoho metód ako dosiahnuť zachycovanie zmien CDC, ale iba jedna z nich sa ukázala ako najvhodnejšia a najlepšia. Táto metóda je založená na čítaní dát z log súborov, ktoré sú vytvorené systémom pre riadenie databáze. Jedine táto technika dokázala zaručiť, aby sa všetky dáta dostali zo zdrojového systému do dátového skladu ***– podrobně také kapitola 11.1.5***.

16.5 Change Data Capture

Zapisovanie do logov ako tvrdia (GUERRA, ANDREWS, 2011) nebolo vymyslené kvôli tomu aby mohli vznikať dátové sklady v reálnom čase, ale toto zapisovanie slúžilo skôr ako zálohovanie. Všetky systémy na riadenie databáze (DBMS) majú v sebe aj mechanizmus, ktorý slúži na ukladanie dát, ktoré slúžia ako záloha. Všetky tieto systémy zvolili veľmi podobný prístup, ktorý je založený na tvorbe log súborov. Všetky keď sa na zdrojovom systéme udeje nejaká zmena, tak DBMS slúži k tomu aby táto ***zmena bola zaznamenaná aj v log súbore***. Tieto súbory sa zvyknú ukladať na iné zariadenie než na to na ktorom sú ukladané dáta, aby v prípade výpadku jedného zariadenia bola k dispozícii záloha. V log súboroch sa nachádzajú všetky zmeny, ktoré sa odohrali na zdrojovom systéme a sú zapísané v zdrojovej databáze a každá zmena v tomto súbore je zapísaná spolu ***s časovou značkou*** kedy sa táto zmena uskutočnila. Vďaka všetkým týmto skutočnostiam sú naplnené ***podmienky pre to, aby mohol vzniknúť RTDW***.

Všetky podniky, ktoré majú dátový sklad, tak už majú implementovaný ETL proces. Tento ETL proces sa uskutočňuje pomocou vybraného nástroja. Toto môže byť obmedzením, pretože ***nie všetky ETL nástroje podporujú načítanie dát z log súborov***, ale podporujú iba čítanie priamo zo zdrojovej databáze. V takomto prípade musíme tento ETL proces obohatiť o novú technológiu, ktorá je schopná čítať logy a pracovať s týmito dátami.

Je celkom zjavné, že tradičný ETL proces sa veľmi nelíši od tohto ETL procesu v reálnom čase. Vo fáze extrakcie dát zo zdrojového systému ***v tradičnom ETL dochádza ku čítaniu dát priamo zo zdrojovej databáze*** pričom pri používaní ***ETL v reálnom čase dochádza k čítaniu log súborov***, ktoré obsahujú všetky zmeny, ktoré sa na zdroji odohrali. Toto je hlavný rozdiel medzi týmito prístupmi. Následne nastáva ***proces transformácie, kde nie je žiadny rozdiel***, ak neberieme do úvahy používanú technológiu. A v poslednej fáze nahrávania dát je ***jediným rozdielom to, ako často k tomuto nahrávaniu dochádza***. Kde v tradičnom ETL to väčšinou býva jeden krát denne, ale v ETL v reálnom čase k tomuto dochádza kontinuálne.

16.6 Alternatívy ku CDC založenej na log súboroch

CDC založené na čítaní logov označili (GUERRA, ANDREWS, 2011) ako najlepšie, ale taktiež definovali ***tri alternatívne prístupy ako dosiahnuť RTDW***:

- ***Databázový trigger*** – funkcionalita DBMS spúšťa určenú akciu vždy, keď sú naplnené dané podmienky, napríklad ako pridanie, alebo aktualizácia záznamu v databáze. Kód, ktorý trigger spustí môže mať za následok nahranie transakcie do databázovej tabuľky a ETL proces môže tieto záznamy nahrávať do dátového skladu v pravidelných časových intervaloch. V praxi ale tento spôsob nie je veľmi spoľahlivý, pretože triggre sú pravidelne mazané/ blokované a následne znova založené/ povolené.

- ***Fronty správ*** – existuje aj niekoľko produktov, ktoré fungujú ako middleware, a umožňujú zachytiť zmeny už v aplikácii, nie v databázy, a tieto zmeny ďalej predávať ETL nástroju. Veľkou nevýhodou je cena licencii na tieto produkty. A hlavne fronty správ nie sú dôveryhodným zdrojom informácii o zmene, pretože tieto informácie posielajú priamo z aplikácie. To znamená, že nereflektujú žiadne rutinné, alebo manuálne aktualizácie databáze.
- ***Dátumové a časové značky*** – mnoho zdrojových aplikácii pri jednotlivých záznamoch uchováva aj časové značky, kedy naposledy došlo na danom zázname k zmene. Takže tento prístup opakovane číta zdrojové dáta a hľadá nedávne zmeny. Hlavným problémom je, že tento prístup sa spolieha na programy, ktoré upravujú dáta, že bezchybne budú vždy aktualizovať aj toto pole. Ďalším problémom tohto prístupu je aj to že stratí prehľad o zmazaných záznamoch, pretože záznamy sú mazané celé aj s časovou značkou o ich poslednej zmene.

16.7 Zdroje

- FAROOQ, Farrah, SARWAR, Syed, Mansoor. Real-time data warehousing for business intelligence. *Proceedings of the 8th International Conference on Frontiers of Information Technology - FIT '10* [online]. 2010 [cit. 2020-01-25]. Dostupné z: doi: 10.1145/1943628.1943666.
- SANTOS, Ricardo, Jorge a BERNARDINO, Jorge. Real-time data warehouse loading methodology. *Proceedings of the 2008 International Symposium on Database Engineering & Applications - IDEAS '08* [online]. 2008 [cit. 2020-01-26]. Dostupné z: doi: 10.1145/1451940.1451949.
- GUERRA, Joseph a ANDREWS, David. *Creating a Real Time Data Warehouse* [online]. 2011 [cit. 2020-01-28]. Dostupné z: https://s3.amazonaws.com/academia.edu.documents/31661949/Creating_RealTime_DW.pdf.

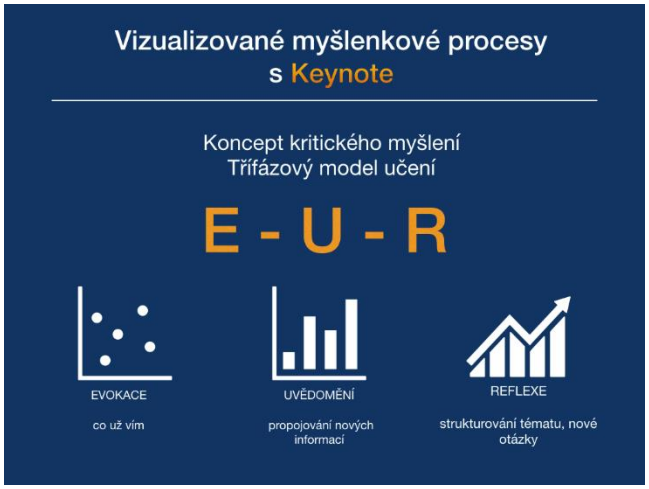
16.8 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řešení a využití datových skladů v prostředí reálného času.

.

- Spoločnosti budú čoraz viac nútené k implementácii technológii na prenos dát v reálnom čase, pretože momentálne to ***poskytuje konkurenčnú výhodu***, ale časom sa to stane neodmysliteľnou súčasťou. Kto bude vedieť rýchlejšie a flexibilnejšie reagovať na zmeny na trhu bude vždy o krok vpred.
- Možnosti akým spôsobom vyberať dáta zo zdrojových databáz je viacero, ale momentálne je ***jasnou voľbou CDC***. Čítanie dát z logov je najbezpečnejší spôsob ako sa dostať k overeným, bezpečným a aktuálnym dátam.

E) Reporting a vizualizace dat



[18] Reporting
[19] Dashboardy

Oddíl se zabývá dvěma součástmi na prezentační úrovni úloh podnikové analytiky:

- ***Část věnovaná reportingu*** zahrnuje ***členění uživatelů*** reportů a jejich potřeby, ***kategorizaci*** reportů podle vybraných hledisek podnikového řízení a vymezení ***potenciálních efektů***, které kvalitní reporting přináší firmě a na druhé straně případných ***problémů***, které jsou s řešením reportů a jejich organizací spojeny.
- ***Dashboardy*** lze charakterizovat jako ***grafické vyjádření sledovaných metrik*** významných pro řízení firmy. Jsou prezentovány formou grafů, tabulek, budíků a jiných grafických ukazatelů a využívají se např. při řízení podnikové výkonnosti (CPM). Kapitola obsahuje ***podstatné charakteristiky*** dashboardů, jejich ***klasifikaci*** podle úrovní řízení, kterým slouží, ***možnosti zobrazení*** a rovněž hlavní ***efekty*** pro firmu a specifikace případných ***problémů*** s jejich řešením.

18. Reporting

Účelem kapitoly je:

- nastínit ***klasifikaci*** různých možností a forem reportingu,
- analyzovat funkce a možnosti reportingu z pohledu jejich ***potenciálních efektů*** a na druhé straně ***omezení*** jako podklad pro efektivní návrhy reportingových řešení,
- definovat a vyhodnotit možnosti ***vizualizace dat*** v rámci reportů pro zkvalitňování jejich formální úrovně.

Reporting slouží jako ***systematický zdroj informací a ukazatelů*** vypovídající o činnosti společnosti pro své uživatele, kterých v podniku mohou být desítky, stovky nebo i tisíce. Za tímto zdrojem informací pro uživatele stojí většinou ***dotazy na data*** do databází firmy.

Každá skupina uživatelů nebo dokonce každý jednotlivý uživatel mají ***různé požadavky*** týkajících se obsahu, formy i času distribuce výkazů, jež ke své úloze v podniku potřebují. Důležitou součástí je zajištění takové ***distribuce***, která pracovníkům ve firmě zajistí ***přístup pouze k jim relevantním údajům*** a bude chránit důvěrná data před nežádoucím šířením.

18.1 Uživatelé reportingu

Uživatele lze rozdělit na ***dvě základní skupiny***:

- ***Interní uživatelé*** zahrnují manažerské a další interní pozice ve firmě.
- ***Externí uživatelé*** přistupují pouze k externím informacím – tedy k informacím, které podnik zveřejňuje. Zpravidla se jedná o externí výkazy finančního účetnictví. Patří sem např. spolupracující firmy – dodavatelské a odběratelské společnosti, banky a další, státní orgány, například finanční úřady, správy sociálního zabezpečení, krajské úřady, zastupitelské orgány obcí a měst, vlastníci podniku a další.

Vlastník podniku v případě externích uživatelů představuje speciální případ. Zde se mluví o společnostech, které rozlišují roli vlastnickou a řídící; v takovém modelu ***majitel podniku nemůže využívat podrobné informace*** o činnostech firmy, protože je „pouhým“ investorem, jenž do společnosti vložil kapitál a očekává ideálně maximální zhodnocení jeho poskytnutých prostředků. ***Prostřednictvím statutárních orgánů*** může uplatňovat své právo na ovlivnění rozhodnutí, jež jsou statutárnímu orgánu vyhrazeny ve stanovách podniku.

Druhý model definuje vlastníka jako osobu, která ***vykonává řídící funkci***, rozhodnutí a další manažerské úlohy, tedy patří mezi interní uživatele. V praxi mnohdy ***oba modely splývají*** do sebe.

18.2 Kategorizace reportů

Tato část se zabývá kategorizací reportů, jež může pomoci poznat jejich účel a obsah podrobněji.

18.2.1 Interní reporting

Reporting z časového hlediska se dělí na standardní a mimořádný (Fibírová, Šoljaková 2010):

Standardní reporting se vyznačuje zprávami dodávaných ***v pravidelných časových úsecích*** – obvykle jednou za měsíc, za čtvrt let nebo za celý rok. ***Struktura*** zprávy z hlediska obsahu informací, výpočtů nebo analýz a podobně je ***předem stanovená***. Výjimkou nejsou reporty týdenní nebo i denní. Pracovníci reportingu by však měli u každého periodického reportu ***zvážit nákladovou náročnost*** generování takového reportu.

Mimořádný reporting představuje zprávy ***generované na požadavek*** nebo reporty, které se běžně negenerují – například analýza rizika, analýza sortimentních skupin a podobně.

Každý podnik může mít jiný ***systém interních reportů***, lze však často vysledovat ***dvě elementární kategorie reportů – základní souhrnné*** (či přehledové) výkazy a ***dílčí*** zprávy (Fibírová a Šoljaková 2010).

- ***Souhrnný reporting*** podává přehled o činnostech podniku ***za určité období***; do jeho obsahu tradičně patří ***základní finanční ukazatelé*** včetně srovnání těchto hodnot s hodnotami plánovanými nebo s hodnotami minulé časové periody. Fibírová a Šoljaková (2010) zde zdůrazňují důležitost doplňujícího komentáře, jež se zabývá rozdíly mezi plánem a skutečností a analýzou, která hledá odpovědi na to, proč odchylky nastaly.
- ***Dílčí reporting*** se věnuje ***konkrétnějším informacím***, struktura více odpovídá daným požadavkům pracovníků podniku. Jejich potřeby se navíc liší od odboru k odboru, ve kterém svou práci vykonávají (obchod, výroba, marketing a podobně).

18.2.2 Externí reporting

Povinnost ze zákona předkládat výsledky hospodaření formou ***auditovaných finančních výkazů*** zajišťuje velkou část externího reportingu. Jde o standardní ***periodické výkazy o hospodaření***, které vychází jednou ročně, tedy o rozvahy, výsledovky a výkazy peněžních toků. Důležitým prvkem externích výkazů je úplnost, srovnatelnost a spolehlivost údajů.

18.2.3 Další kategorie reportů

Samotné reporty se dělí na ***standardní a ad hoc***. Podobné dělení již bylo zmíněno, zde se však neklade důraz na časovou dimenzi reportů, nýbrž na dimenzi uživatelskou; přesněji řečeno na míru, se kterou může uživatel ovlivňovat datový obsah výpisu.

Ad hoc zprávy si může vytvořit sám uživatel vytvořením konkrétního jednorázového dotazu nad databázemi. Dotazům se tato práce věnuje podrobněji dále.

Standartní reporty ještě lze dále členit na statické a dynamické. Uživatel si ***statický výpis*** již nemůže upravovat, kdežto s daty ***v dynamickém reportu*** může manipulovat podle svých potřeb. Je však vždy omezen nástroji, které mu byly ve zprávě vyhrazeny. Nejčastěji se používají filtry. Uživatel si tak může například vymezit jiné časové období, než je v reportu defaultně nastaveno, může si nechat vypsat data se společnými parametry.

Podle obsahové stránky dat, jež předurčuje i uživatele zprávy, ***se rozlišují kategorie datové domény***. Mezi ně patří *finance, prodej, marketing, zákaznický servis, nákup, sklady, lidské zdroje, výroba, informační technologie a další*. Některá oddělení, například výrobní, vyžadují ve výstupu i ***přehledy sledovaných norem***.

18.3 Efekty a přínosy reportingu

- Umožňují ***monitoring procesů*** ve firmě a s tím souvisí i vyhodnocování takových dat, na jejichž základě lze ***identifikovat problém a jeho příčinu***. Díky těmto informacím se vylepšují a ***optimalizují podnikové procesy***. Zkvalitnění podnikových procesů vede ke ***zlepšení rozhodování*****,** optimalizace výkonnosti a určení správného budoucího směru fungování společnosti.
- Kvalitní reporting přináší také ***odhalení příčin problémů*** pomocí včasných a správných informací s možností pohlížení z různých perspektiv a úrovně detailu.
- Reporting také znamená ***zvýšení transparentnosti*** ve firmě. Nejde v rozporu s nežádoucím šíření citlivých informací. Jde o poskytování ***úplných relevantních informací*** všem, kteří by přístup k nim měli mít, k nimž může přistupovat veřejnost nebo zájmové skupiny *(stakeholders)*. To přináší i ***zlepšení prestiže podniku***.
- ***Jednotný pohled*** na podnikovou realitu na základě konsolidovaných reportů vede ke ***zkvalitnění komunikace*** mezi zaměstnanci a manažery na všech úrovních řízení.
- ***Podpora efektivity*** – zaměstnanci mají více času na hlubší analýzy a akce, nemusí ručně konsolidovat informace do pravidelných reportů, pracovníci neřeší tvorbu reportů, ale jejich obsah.

- Na druhé straně možnost pro zaměstnance vytvářet ***vlastní reporty bez podpory IT útvarů***, radikálně snižuje dobu dodání informací k příjemci.
- Nabízejí možnosti ***předdefinované logiky***, díky které nedochází k situacím, kdy výstup dvou reportů tvoří odlišná čísla na základě stejné vypočítávané veličiny,
- ***Integrace*** reportingu s kancelářskými aplikacemi umožňuje zefektivnění tvorby reportů nejrůznějšího druhu, včetně ***exportu reportů*** do standardních aplikací pro další zpracování a sdílení. S tím souvisí i ***navázání dalších aplikací*** na report (např. kliknutím na jméno dodavatele se otevře mapa s jeho sídlem atd.).

18.4 Problémy a omezení spojené s reportingem

- Vysokou pozornost je třeba věnovat kvalitní ***strukturalizaci reportů*** tak, aby uživatel se např. mohl efektivně pohybovat po různých úrovních detailu obsažených dat a mohl se v nich rychle orientovat.
- Velmi častým problémem je, že reporty se vytvářejí překotně a ***ve svém obsahu se různě překrývají***, pracovníci firmy jsou zahlceni někdy i zbytečnými reporty. Příprava každého reportu by se neměla obejít alespoň ***nezbytnou analýzou*** v uvedeném smyslu.
- V řadě případů nemají pracovníci zejména ve větších firmách souhrnný, centralizovaný přehled o existujících reportech, což vede k ***vytváření nových reportů čistě duplicitních***, nebo s částečnou duplicitou.
- Výstup reportingu je vysoce ***závislý na kvalitě*** zdrojových dat.
- Některé reporty, např. u poboček zahraničních společností, musí ***dodržovat mezinárodní standardy***, např. ***IFRS*** (*International Financial Reporting Standards*) nebo ***US GAAP*** (*United States General Accepted Accounting Principles*). Díky standardům lze zajistit nejen celistvé vymezení obsahu reportů, ale i takovou strukturu, která bude srozumitelná pro všechny.
- ***Finanční náročnost*** sofistikovanějších reportingových nástrojů vede někdy k jejich odmítnutí bez vyhodnocení jejich potenciálních efektů.
- Pro efektivní využití reportů je podstatná ***kvalitní vizualizace dat***, která však musí dodržovat základní principy, např. při výběru grafů, schémat apod.

18.5 Zdroje

FIBÍROVÁ, J., ŠOLJAKOVÁ, L., WAGNER, J., PETERA, P. 2015. *Manažerské účetnictví – nástroje a metody.* 2. aktualizované a přepracované vydání. Praha: Wolters Kluwer. ISBN 978-80-7478-743-0

FIBÍROVÁ, J., ŠOLJAKOVÁ, L. 2010. *Reporting.* 3. vydání. Praha: Grada Publishing, a.s. ISBN 978-80-247-2759-2

18.6 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řešení podnikových reportů a jejich organizaci.

.

- Účelem reportingu je data vybírat a ***poskytovat vzhledem k uživatelským problémům*** a požadavkům.
- Reporty musí nabízet i vhodné ***formy prezentace*** založené na celé škále možností ***vizualizace*** dat.
- V některý případech se ale věnuje ***větší pozornost vizualizaci*** před kvalitním obsahem, a to může být problém.
- Reporty se různě člení, s tím, že základní rozdělení je na ***interní a externí***.

- S reportingem souvisí i nezbytná jejich ***ochrana*** vzhledem k oprávněním jednotlivých uživatelů.
- Častým problémem je i ***neúměrné množství*** různě se překrývajících, duplicitních reportů, což zatěžuje zejména orientaci uživatelů s jejich využitím.

19. Dashboardy

Účelem kapitoly je prezentovat základní informace k dashboardům a zejména shrnout jejich klíčové efekty a omezení. Další informace jsou v přiřazeném dokumentu.

Dashboard je ***grafickým vyjádřením sledovaných metrik***, které jsou důležité z pohledu byznysu organizace. Metriky jsou prezentovány formou grafů, tabulek, budíků a jiných grafických ukazatelů. Dashboardy jsou ***využívány*** např. při ***řízení podnikové výkonnosti (CPM).*** Dashboardy by měly respektovat **principy** názornosti a viditelnosti všech sledovaných metrik, umožnit posouzení plán vs. skutečnost, poskytnout přehled o zlepšujících se a zhoršujících se metrikách a sledovat výkonnost zaměstnanců.

19.1 Efekty a přínosy dashboardů

- Poskytování aktuálních informací, které ***umožňují uživatelům zlepšit rozhodování, optimalizaci procesů a efektivitu práce***. Umožňují tedy všem uživatelům (od vedoucích manažerů až po běžné zaměstnance) ***monitorovat klíčové aktivity*** a procesy potřebné k dosažení podnikových cílů.
- ***Definování podnikové strategie*** pomocí ***cílů, opatření a iniciativ***, které jsou určené jak skupinám v podniku, tak i některým jednotlivcům. Uživatelé mají přehled, co přesně musí udělat ve svých oblastech k dosažení podnikových cílů.
- Existence dashboardů má pozitivní vliv ***na ladění a provedení firemní strategie*** s daleko menšími výkyvy než bez nich.
- Zvýšení ***predikce pomocí KPI ukazatelů***, pomocí nichž manažeři mohou lépe predikovat prognózu na základě minulé činnosti. Tento proces pomáhá společnosti vyhnout se tak nečekaným výpadkům.
- Zlepšení ***koordinace a komunikace*** mezi manažery jednotlivých úseků či oddělení, aby více spolupracovali mezi sebou.
- Zvýšení ***motivace zaměstnanců*** pomocí vyplácení odměn za vyšší efektivitu práce podloženou jednoznačně lepším KPI. Využitelnost při ***sledování výkonnosti***, hodnocení a motivaci zaměstnanců.
- ***Jednoznačný pohled na podnikání***, jelikož se udržuje pouze jedna společná verze, tudíž nedochází k nedorozumění mezi IT manažery a analytiky v organizaci, jaká verze dat je správná.
- ***Snížení nákladů a redundance*** pomocí konsolidace a standardizace informací.
- Možnost pracovníkům ***měnit si dashboardy podle svých požadavků*** bez podpory IT oddělení a analyzovat informace z různých úhlů pohledů.
- Dodání ***včasných informací*** potřebných k nápravě problému a přijetí opatření.
- Vysoce názorné, ***grafické zobrazení*** sledovaných metrik.
- Okamžité posouzení sledovaných metrik z pohledu ***plán / skutečnost***.

19.2 Problémy a omezení dashboardů

- Řešení dashboardů vyžaduje ***specifické analytické přístupy*** s důrazem na adekvátní výběr relevantních ukazatelů (KPI) pro jednotlivé role v řízení podniku,
- ***Přeceňování analýz*** a výstupů z dashboardů, zejména jejich grafické stránky.
- Pro grafické zobrazení metrik je potřebný relevantní ***soubor dat za určitý časový úsek***.

- Problémy ***související s kvalitou*** a uložením dat. Při špatné kvalitě dat prezentují dashboardy špatné výsledky a poskytují nekvalitní podklady pro rozhodování a řízení podniku.
- ***Nevhodně zvolené nástroje*** pro prezentaci informací, které nepodporují kvalitní funkce pro analýzu a práci s daty.
- ***Aktualizace dashboardu*** neprobíhá zcela automaticky z důvodu velkého množství informací, byznys uživatelé stráví velké množství času, většinou několik hodin v týdnu aktualizací stavu namísto jeho včasného vyhodnocení, což není udržitelné v dlouhodobém horizontu.
- Příliš ***velká izolovanost a granularita dashboardu*** oproti zbytku řešení. KPI ukazatele poskytují příliš úzký pohled na byznys vůči celému podniku.

19.3 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řešení dashboardů a jejich využití.

.

- Dashboardy jsou obvykle ***grafickým vyjádřením ukazatelů podle klíčových dimenzí***, a to formou grafů, tabulek, budíků a podobně.
- Řešení dashboardů má ***specifické nároky na jejich analýzu*** a řešení, neboť musí přesně vystihnout hlavní souhrnné informace o dané oblasti řízení podle potřeb příslušného manažera, kterému je určen.
- Základní ***členění dashboardů*** je na tyto typy:
  - strategické,
  - taktické,
  - provozní.
- Každý z uvedených typů dashboardů má ***odlišné určení, funkcionalitu i způsob využití*** v praxi.
- Dashboard musí současně nabízet ***cesty k detailním informacím***, z nichž jsou vytvořeny souhrnné hodnoty.
- Charakter dashboardů se rozlišuje ***podle úrovní řízení*** a manažerů, kteří na nich pracují.

F) Pokročilá podniková analytika

[20] Data Science,
[20] Data mining, Text mining, Machine Learning
[23] Predictive Analytics, PA

Oddíl věnovaný pokročilé analytice, tj. ***analytickým úlohám*** a postupům kombinujícím analytickou základnu (metriky, datové zdroje, základní technologie) s využitím ***pokročilých matematických a statistických metod***. Pokročilá analytika je založena na celém spektru analytických funkcí, jejichž pouze velmi stručný přehled je uveden v souvislosti s vymezením úloh pokročilé analytiky v ***kapitole 1.2.4***

Oddíl obsahuje:

- ***Data Science*** zahrnuje: porozumění business logiky dat, přípravu dat, modelování / optimalizace / simulace, vyhodnocení a nasazení analytického modelu.
- ***Data Mining,*** dolování dat jako proces extrakce relevantních, předem neznámých nebo nedefinovaných informací z velmi rozsáhlých databází.
- ***Text Mining,*** resp. textová analytika představuje analýzu textových zdrojů a získávání z nich nových informací, kde zdroje mohou být velmi různorodé od knižních nebo novinových publikací, přes blogy a další.
- ***Machine Learning*** obsahující vymezení podstaty a kroků řešení úloh ML, vymezení možností produkcionalizace a governance řešení.
- ***Prediktivní analytika, PA*** představuje typ analýzy využívající data a prediktivní modely pro předpověď jevů na mikroekonomické úrovni. Učí se ze zkušeností (dat) a předvídá budoucí chování jedinců, zahrnuje v sobě množství statistických a analytických technik. Kapitola obsahuje podstatné charakteristiky a modely prediktivní analytiky, doporučený postup řešení, výhody a nevýhody a současně i příklady z oblasti využití prediktivní analytiky v řízení údržby a v řešení pojistných podvodů.

20. Data Science

Účelem kapitoly je:

- poskytnout podklady pro ***konceptů*** a nástrojů data science a formulovat doporučení ***pro jejich uplatnění v praxi***,
- definovat podstatné ***charakteristiky*** těchto nástrojů a jejich souvislosti,
- vymezit významné ***efekty a případná omezení*** dále uvedených typů aplikací pokročilé analytiky jako podstatné vstupy pro jejich způsob využití.

Data Science je přístup k řešení IT, který ***z velkého množství komplexních dat, které navíc i velmi rychle narůstá, získává informace a znalosti*** (dále podle F. Provost a T. Fawcet v publikaci „*Data Science for Business*“, 2013). Z pohledu tohoto dokumentu a jeho uspořádání představuje Data Science také ***společnou nadstavbu*** nad řadou dílčích a rozdílných metod, aplikací a nástrojů pro získávání znalostí z datových zdrojů.

20.1 Data Science jako koncept

V praxi je možné definovat sadu základních konceptů, které jsou podstatou pro formování základních přístupů získávání informací a znalostí z dostupných dat. Tyto koncepty pokrývají celý proces od pochopení byznys problémů, přes využití technik data science až po poskytování výsledků manažerům a analytikům s cílem zkvalitňovat jejich rozhodování. Koncepty zahrnují tyto tři typy:

- ***Koncepty ukazující***:
  - jak data science začlenit do firmy a organizace v rámci konkurenčního prostředí,
  - jak efektivně sestavovat a připravovat týmy pro data science,
  - jak formovat způsob analytického myšlení zaměření na získávání konkurenčních výhod,
  - jak efektivně postupovat při řešení projektů orientovaných na data science.
- ***Koncepty zaměřené na způsob analytického myšlení***. Ty umožňují:
  - identifikovat vhodná data a metody pro řešení úloh data science,
  - aplikovat různé procesy data miningu a dalších přístupů a efektivně přispívat k řešení úloh.
- ***Základní koncepty pro získávání znalostí z dat*** jako součásti řešení úloh data science.

Příkladem takového konceptu je ***určování podobnosti*** dvou entit, k nimž se vztahují určitá data. Takový koncept ***vytváří základnu pro řešení nejrůznějších úloh***, např. pro hledání podobnosti zákazníka s jinými zákazníky. To pak vytváří podklady pro prediktivní úlohy zaměřené na řešení cíleného marketingu, hodnocení zákazníka z hlediska obchodních možností (tržeb), pravděpodobnosti, že právě on využije presentovaných nabídek firmou.

Tyto koncepty jsou také základnou pro ***úlohy clusteringu*** sdružující objekty (např. zákazníky, zboží apod.) do skupin na bázi jejich sdílených vlastností. S tím souvisí i ***využívání vzorů („pattern“)*** pro analýzy vlastností zkoumaných objektů, a to v různém kontextu. Cílem je poskytovat doporučení pro zaměření úloh marketingu, prodeje, zásobování apod.

Je zřejmé, že objasňování principů Data Science na základě uvedených konceptů výrazně ***usnadní komunikaci*** mezi specialisty v Data Science na jedné straně a manažery a firemními analytiky na straně druhé. To znamená, že se vytváří předpoklady pro vzájemné pochopení a ***podstatně přesnější formulaci problémů*** firmy řešitelných na bázi úloh a aplikací Data Science.

20.2 Analytické myšlení

V celé řadě případů se řešitelé i uživatelé úloh spojených s Data Science zaměřují logicky primárně na metody a techniky které se k řešení takových úloh váží. Na druhé straně je ale dobré zdůraznit, že ***jedním z klíčových předpokladů*** je kvalita a ***úroveň analytického myšlení*** jak na straně řešitelů, tak uživatelů. To znamená vidět byznys problémy z datové perspektivy a pochopit principy, jak efektivně

získávat z těchto dat znalosti pro řešení definovaných byznys problémů. To je základní struktura datově analytického myšlení. Vedle tohoto základního principu je třeba brát v úvahu i další ***podstatné aspekty*** kvalitního analytického myšlení a těmi jsou ***intuice, kreativita, selský rozum a velmi dobrá znalost byznys obsahu*** řešených problémů (F. Provost a T. Fawcet, 2013). Ze zkušenosti je dobré i přidat „***schopnost řešitelů vžít se do myšlení druhé strany***" jako nezbytný předpoklad pro kvalitní, kvalifikovanou a konstruktivní komunikaci a kooperaci obou stran.

Datová perspektiva a ***schopnost analytického myšlení*** vytvoří předpoklady, struktury a vzory pro systematické řešení byznys problémů, a to i vysoce komplexních. V tomto smyslu je dobré doplnit i tyto ***klíčové charakteristiky analytického myšlení***:

- ***analytický pohled na obsah*** byznys problémů a řízení firmy je třeba formovat ***podle jednotlivých komponent řízení*** (úlohy, resp. procesy, ukazatelé, dokumenty atd.), který vedle jejich standardních manažerských charakteristik bere v úvahu i ty, které jsou pro analýzu významné, např. realizace procesů a podmínky jejich úspěšnosti, zdroje dat pro ukazatele (metriky) a souvislosti s analytickými dimenzemi apod.,
- ***důraz na kontext***, tedy pochopení a řešení značného množství vazeb a souvislostí, na nichž je řízení firmy založeno,
- pochopení a řešení ***firmy jako celku*** bez ohledu na to, zda je předmětem řešení jen její část, nebo celá firma – pro analytika platí, že i když řeší např. sklady musí stále firmu vidět a vnímat jako jeden celek. V souvislosti s Data Science to také znamená, že ř***ešení pokrývá většinu oblastí a útvarů řízení firmy,*** a to tak vyvolává potřebu širokého pochopení analytických principů, racionální komunikace a kooperace těchto různých útvarů se členy Data Science týmu,
- ***pochopení toho,*** jak firma funguje a schopnost identifikovat příčiny, když nefunguje,
- ***schopnost jasně formulovat analytické otázky*** směřují k identifikaci byznys problémů, pochopení jejich podstaty, možných cest řešení i priorit, které jim budou v rámci celého řešení přiřazeny,
- schopnost ***vyhodnocení, jak racionálně využívat IT produkty a služby,*** jak je nejlépe nasadit pro jednotlivé oblasti a komponenty řízení firmy, jaké jsou vzhledem k byznys obsahu potenciální možnosti a na druhé straně omezení IT produktů a celých řešení (*business – IT alignment*).

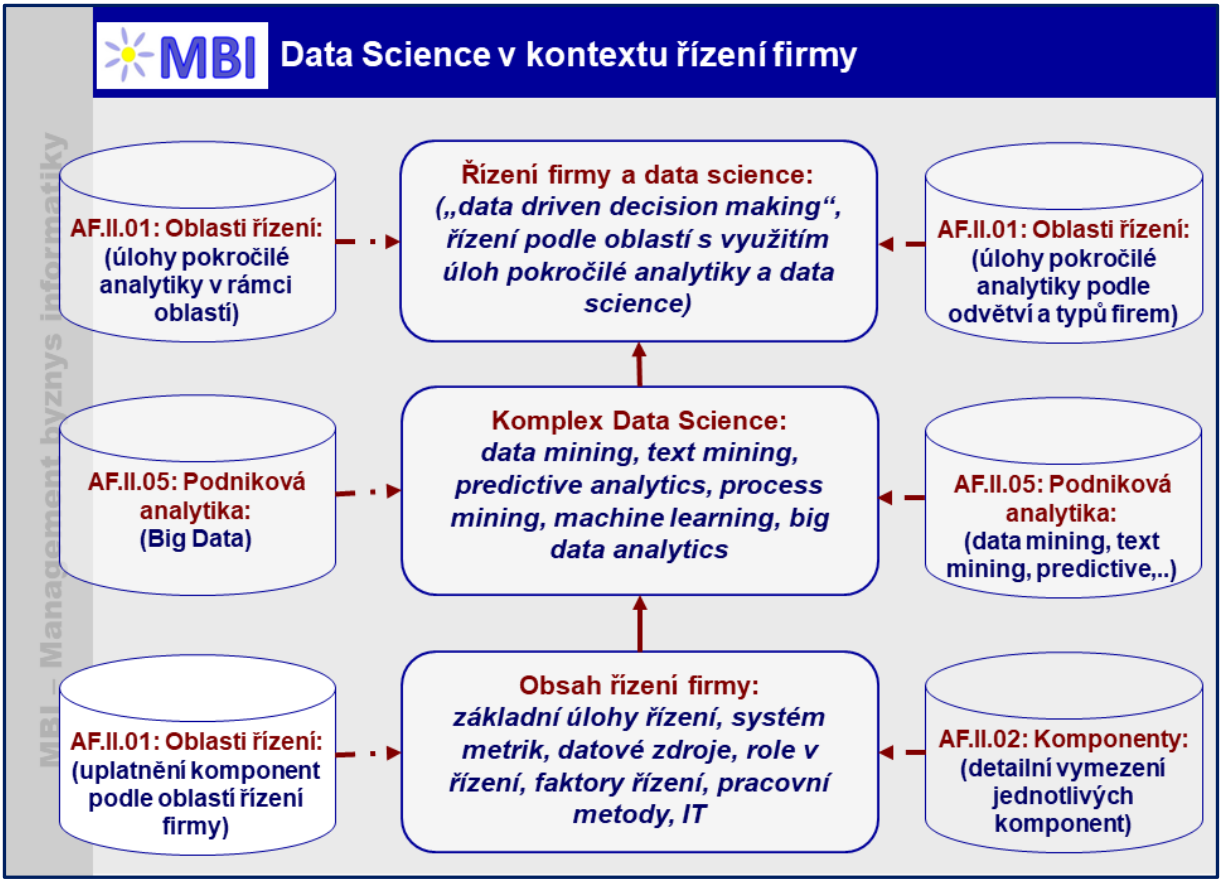
Je zcela evidentní, že schopnost a kvalita analytického myšlení je ***významným předpokladem pro řešení jakýchkoli úloh a problémů*** v rámci IT (i mimo ně), v případě řešení úloh Data Science to platí zřejmě dvojnásob.

20.3 Data Science v kontextu řízení firmy

Zásadním a výchozím ***předpokladem pro efektivní využívání celého komplexu*** data Science, jeho principů, přístupů a metod je jeho ***zasazení do kontextu řízení celé firmy***, to znamená:

- kvalifikované pochopení ***zasazení*** Data Science do celé ***sady definovaných komponent řízení***, tj. základních úloh, systému metrik, datových zdrojů, rolí, faktorů, metod a technologií,
- pojetí Data Science jako ***soustavy typů aplikací a odpovídajících nástrojů***, které se vzájemně doplňují, ale i překrývají, tj. respektování takových souvislostí a jejich efektivní využití vzhledem k potřebám dané firmy,
- pochopení efektivního využití aplikací, nástrojů a funkcí Data Science ***v úlohách pokročilé analytiky v rámci jednotlivých oblastí řízení*** firmy a následně i v oblastech řízení respektujících odvětvová specifika.

Uskutečnění uvedených předpokladů i ***s využitím dokumentů na portále MBI-AF*** dokumentuje Obrázek 20-1.



Obrázek 20-1: Data Science v kontextu řízení firmy

K uvedenému schématu, které využívá základní konstrukci v (Provost, Focett, 2013) a rozšiřuje ji, jsou v dalších podkapitolách ***doplněny některé poznámky*** (ve směru odzdola nahoru). V podbodech jsou zdroje v rámci dokumentů MBI-AF:

20.3.1 Obsah řízení firmy

Pokud se cíleně zaměříme na reálné využití Data Science v řízení firmy, pak ***východiskem musí být zmapování a kvalitní pochopení obsahu řízení*** firmy, a to ve vztahu k řešení úloh spojených s Data Science. Sem patří:

- Vymezení ***obsahu řízení firmy podle jednotlivých oblastí*** a na bázi definovaných typů úloh, tj. mimo úlohy pokročilé analytiky. To znamená, kde podklad pro analýzu, bude vhodné implementovat úlohy Data Science, s jakými nároky a s jakými prioritami:
  - *AF.II.01: Oblasti řízení* – vymezení obsahu všech úloh podle oblastí.
- Detailní ***charakteristika jednotlivých metrik***, tedy ukazatelů rozdělených podle oblastí řízení a ***analytických dimenzí***, které jsou předmětem analýz a mohou být i předmětem úloh Data Science:
  - *AF.II.02: Komponenty a souvislosti* – vymezení obsahu a dílčích charakteristik metrik (ukazatelů), kapitola 2 a relevantních dimenzí (pokud budou také předmětem řešení), kapitola 3.
- Detailní obsahová ***charakteristika datových zdrojů*** jako podklad pro hodnocení jejich dostupnosti, rozsahu a kvality vzhledem k potřebám úloh Data Science. Otázka dostupnosti a kvality vstupních datových zdrojů je de facto pro všechny úlohy zařazené do Data Science klíčovým problémem:
  - *AF.II.02: Komponenty a souvislosti* – vymezení obsahu datových zdrojů rozdělených podle oblastí řízení, kapitola 4.

- Definování ***funkční náplně a požadovaných znalostí rolí***, pracovníků spojených s podnikovou analytikou, a tedy i s Data Science. Jaké má být personální vybavení týmů pro řešení Data Science:
  - Funkční náplň a znalosti specifických rolí ve vztahu k podnikové analytice jsou náplní kapitoly 2 tohoto dokumentu.
  - *AF.II.02: Komponenty a souvislosti* – vymezení náplně dalších rolí, které do řešení Data Science mohou vstupovat, kapitola 5.
- ***Faktory ovlivňující řešení*** podnikové analytiky a Data Science, jako např. byznys prostředí, úroveň řízení a organizace firmy a další definují podmínky pro posuzování a prvotní rozhodování, zda a kde řešení úloh data Science realizovat, případně od něj upustit.
  - Vymezení jednotlivých faktorů jsou obsahem kapitoly 3 tohoto dokumentu.
- ***Pracovní metody a metodiky***, tedy nejen ty, které se váží k jednotlivým úlohám Data Science, ale i manažerské, případně metody řízení IT a další podklady, které je rovněž při řešení úloh účelné zvažovat:
  - Metody a metodiky, které se přímo váží k úlohám Data Science jsou součástí příslušných kapitol tohoto oddílu.
  - *AF.II.02: Komponenty a souvislosti* – kapitoly 8, 9, 10 obsahují stručné charakteristiky dalších metod a metodik, které v souvislosti s Data Science mohou přicházet v úvahu.
- ***IT aplikace*** je účelné zvažovat zejména v kontextu přípravy a zpracování zdrojových dat pro úlohy Data Science, tj. jejich kvalitu, efekty, omezení, perspektivy.
  - *AF.II.04: IT aplikace* – podstatné charakteristiky IT aplikací, především transakčních, jejich výhody a omezení.

20.3.2 Komplex Data Science

Prostřední vrstvu na schématu (Obrázek 20-1) představuje ***vymezení celého komplexu*** úloh, aplikací a nástrojů spadajících do Data Science, a to:

- ***Data mining*** – jako extrakce relevantních, předem neznámých nebo nedefinovaných informací z velmi rozsáhlých databází.
  - kapitola 21 v tomto dokumentu,
- ***Text mining*** – představuje analýzu textových zdrojů a získávání nových informací.
  - kapitola 22.1 v tomto dokumentu.
- ***Predictive analytics*** – je typ analýzy využívající data a prediktivní modely pro předpověď jevů na mikroekonomické úrovni.
  - kapitola 24 v tomto dokumentu.
- ***Machine Learning*** obsahující vymezení podstaty a kroků řešení úloh ML.
  - kapitola 22.2 v tomto dokumentu
- ***Big Data Analytics*** – umožňují hledat odpovědi na otázky, které byly v minulosti nerealizovatelné.

20.3.3 Řízení firmy a Data Science

Nejvyšší vrstva na schématu (Obrázek 20-1) dokumentuje ***uplatnění Data Science*** a jednotlivých úloh, aplikací a nástrojů ***v rámci definovaných oblastí řízení***. Pro tyto účely je prakticky ve většině oblastí řízení vymezena úloha „Pokročilé analytiky“. Je v nich určeno, jak se úlohy Data Science konkrétně mohou aplikovat v dané oblasti řízení. Postupně jsou takové konkretizace promítány i do dokumentů pro odvětvová řešení – AF.III.nn.

- *AF.II.01: Oblasti řízení* – úlohy „Pokročilé analytiky“ mají standardní číslování x.9, např. „*7.9 Pokročilá analytika v řízení prodeje*“.

20.4 Funkce Data Science

Je účelné ***základní typy funkcí*** Data Science ve vztahu k problémům byznysu, neboť to podporuje systematičnost řešení a zvyšuje jeho kvalitu. To také zdůrazňují F. Provost a T. Fawcet. K hlavním funkcím, podle uvedených autorů, patří:

- ***Klasifikace a hodnocení pravděpodobnosti tříd*** *(Classification and class probability estimation)*:
  - predikuje pro každý objekt, do které sady tříd patří,
  - obvykle jsou třídy vzájemně disjunktní,
  - data mining vytváří model, který určuje, do které třídy objekt patří a s tím související scoring, tj. odhad pravděpodobnosti správnosti takového zařazení.
- ***Regrese – „hodnocení hodnot“*** *(Regression – „value estimation“):*
  - odhaduje nebo predikuje pro každý objekt numerickou hodnotu určité proměnné,
  - např. „Nakolik určitý zákazník bude využívat určitou službu?“ – proměnná je „využití služby“.
- ***Přiřazování podobností*** *(Similarity matching):*
  - identifikuje podobné objekty na základě o nich známých dat,
  - v této souvislosti se využívá tzv. „*firmographic*“, což jsou podstatné charakteristiky firem a od nich odvíjených objektů,
  - např. „Jaké firmy jsou obdobné našim nejlepším zákazníkům?“
- ***Klastrování*** *(Clustering):*
  - seskupuje objekty na základě jejich podobnosti, ale ne odpovídající nějakému specifickému účelu,
  - např. „Představují naši zákazníci nějaké přirozené skupiny nebo segmenty?“
- ***Seskupování podle výskytů*** *(Co-occurence grouping):*
  - pokouší se najít asociace mezi entitami na základě transakcí, které je zahrnují,
  - patří sem také „Odhalování asociačních pravidel“ nebo „Analýza nákupního košíku?“,
  - např. „Jaké položky se obvykle nakupují dohromady?“
- ***Profilování*** *(Profiling):*
  - charakterizuje typické chování jednotlivce nebo skupiny,
  - označuje se také jako „Popis chování“,
  - např. „Jaké je typické použití mobilního telefonu v tomto segmentu zákazníků?“
- ***Predikce vazeb*** *(Link prediction)*:
  - predikuje vazby mezi datovými položkami a obvykle navrhuje, že taková vazba by měla existovat a jaká je síla této vazby,
  - např. „Jestliže úloha analýza prodeje je ovlivněna faktorem úroveň managementu, pak by měla být ovlivněna i faktorem kultura firmy, a to s obdobnou sílou tohoto vztahu“.
- ***Redukce dat*** *(Data reduction)*:
  - redukuje velké objemy dat pro analýzy pouze na ty nejpodstatnější informace,
  - menší objemy vybraných dat jsou snáze využitelné, nakonec i lépe poskytují právě relevantní informace.
- ***Náhodné modelování*** *(Casual modeling)*:
  - pokouší se pochopit, které události nebo akce aktuálně ovlivňují ty ostatní,
  - často je založené na náhodně vybraných a řešených experimentech.

Jeden z fundamentálních principů pokročilé analytiky je podle (PROVOST, F., FAWCETT, T., 2013) správně vybrat, kombinovat a využívat uvedené úlohy pro řešení konkrétních problémů byznysu.

20.5 Metody spojené s Data Science

Kulhavý, L., KIT VŠE, upraveno

V dalším textu jsou dokumentovány metody spojené s úlohami Data Science a spadají tak do celkového komplexu (data mining, predictive analytics a další).

Z řešení vystupují znalosti, jež mohou mít mnoho podob, především (Berka, 2003, str. 85).

- ***reprezentativní příklady,*** například metody založené na analogii,
- **funkce reprezentující shluky** (metody subsymbolické), kde shluk v prostoru atributů reprezentuje navzájem si podobné příklady, například neuronové sítě, bayesovská klasifikace, evoluční algoritmy
- **rozdělení prostoru atributů na snadno popsatelné, pravidelné útvary** (metody symbolické), například rozhodovací stromy, rozhodovací a asociační pravidla.

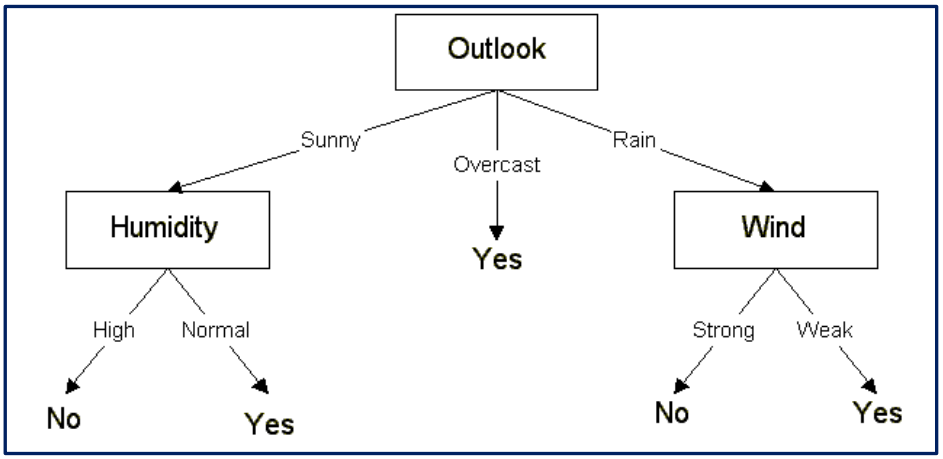
Významným rozdílem mezi metodami symbolickými a subsymbolickými je ***srozumitelnost prezentovaných znalostí pro uživatele***. Například porovnání reprezentace formou rozhodovacího stromu, jakožto seznamu jednoduchých otázek vedoucích k odpovědi na uvažovaný problém, oproti často složité reprezentaci formou neuronových sítí.

20.5.1 Rozhodovací stromy

Rozhodovací strom je grafickou reprezentací logického vývoje časově na sebe navazujících alternativních rozhodnutí. (Friebelová, 2006). Jde o zvláštní případ grafu, kdy rozhodovací strom je tvořen z: (Vomlelová, 2009)

- ***kořene a vnitřních uzlů*** – označených atributem, ze kterých vede jedna hrana pro každou možnou hodnotu tohoto atributu,
- **listů,** které jsou označeny predikovanou hodnotou cílového atributu G, za předpokladu, že ostatní atributy nabývají hodnot na cestě od kořene do listu (pokud se některé atributy na cestě nevyskytují, pak na jejich hodnotě v rámci klasifikace nezáleží).

Klasifikaci příkladů na základě rozhodovacích stromů lze jednoduše provést tak, že z výchozího uzlu postupujeme přes vnitřní uzly (z uzlu volíme hranu odpovídající hodnotě daného atributu) až do listu, který je klasifikací daného příkladu.



Obrázek 20-2: Rozhodovací strom, úloha, zda jít hrát golf v závislosti na předpovědi počasí. (zdroj: http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm)

Náročnějším úkolem v případě metod rozhodovacích stromů, je ***jejich konstrukce***. Postupuje se metodou „***rozděluj a panuj***“ (Divide and Conquer). Tato metoda znamená, že jsou trénovací data postupně rozdělována na stále menší podmnožiny, s cílem nalézt tu podmnožinu, ve které převládají příklady jedné třídy (hodnoty cílového atributu).

Základní algoritmus vycházející z této metody se nazývá ***TDIDT (Top down induction of decision trees***), indukce probíhá v prostoru hypotéz směrem shora dolů). Hlavním cílem je nalézt strom, jakožto reprezentaci získané znalosti, konsistentní s trénovacími daty. Konsistentních s trénovacími daty

může být více různých stromů, v tom případě je dávána přednost stromům menším a jednodušším (optimalizuje následné využití rozhodovacího stromu v procesu klasifikace).

Tabulka 20-1: Algoritmus TDIDT (Berka, 2003 str. 86)

1. zvol jeden atribut jako kořen dílčího stromu
2. rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a přidej uzel pro každou podmnožinu
3. existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak skonči

Algoritmus TDIDT tak má snahu ***vytvořit strom co možná nejlépe reprezentující trénovací data***. Existují však situace, kdy dosažení takové reprezentace není zcela možné (např. v případě existence šumu[1] v datech), či tato snaha vede k příliš rozsáhlým stromům a tím i ke zvýšenému riziko přeučení.

Proto je v praktických algoritmech často aplikováno takzvané ***prořezávání stromů***, kdy dojde k nahrazení určitého podstromu jediným listem. Volba, ve kterém uzlu je vhodné uvažovat o aplikaci prořezávání stromů, vychází buď ze statistického testu na trénovacích datech, či jsou využita tzv. validační data testující chování prořezaného stromu při klasifikaci nových případů v porovnání se stromem neprořezaným. Výsledkem je ***jednodušší a srozumitelnější podoba rozhodovacího stromu***, strom méně zatížený případným šumem v datech, avšak většinou za cenu zhoršeného chování při klasifikaci nových příkladů.

Klíčovou otázkou procesu tvorby rozhodovacích stromů je, ***který atribut zvolit pro větvení v kořenovém uzlu*** (následně každý vnitřní uzel je kořenovým uzlem pro podstrom, tvořící danou podmnožinu). Snahou je zvolit takový atribut, který od sebe nejlépe oddělí příklady patřící do různých tříd. Kritériem pro volbu atributu může být:

- ***Entropie*** – vyjadřuje míru neuspořádanosti nějakého systému. Volen je atribut s minimální entropií.
- **Informační zisk** je definován jako rozdíl entropie pro cílový atribut a pro zvolený atribut (atribut, u kterého uvažujeme větvení). Vybíráme zisk s maximální hodnotou.
- **Poměrný informační zisk** vychází z informačního zisku, avšak zohledňuje i počet hodnot atributu. Stejně jako u informačního zisku je volena největší hodnota poměrného informačního zisku.
- **χ^2** vyhodnocuje vzájemnou souvislost mezi dvěma atributy, volí se atribut nejvíce související atributem cílovým, tedy vztah, kde χ^2 má největší hodnotu.
- **Gini index** využívá znalost o relativním počtu příkladů každé třídy. Obdobně jako u Entropie je vybrán atribut s minimální hodnotou Gini indexu.Pro další informace na téma volby atributů lze čerpat například z knihy Dobývání znalostí z databází (Berka, 2003, stránky 86-93).

Při aplikaci metod rozhodovacích stromů je třeba uvažovat ***omezení v práci s numerickými a chybějícími hodnotami atributů.*** Pro numerické hodnoty atributů je omezením nekonečný počet možných hodnot, a tedy i nekonečná větvení v uzlech stromu. Řešením je ***diskretizace dat*** do konečného (malého) počtu intervalů. Tuto diskretizaci lze provádět v rámci předzpracování dat, alternativou je volba algoritmu, který dokáže s numerickými daty pracovat (především se jedná o možnosti online diskretizace v rámci běhu algoritmu, pokud tuto možnost algoritmus podporuje).

V případě numerického cílového atributu hovoříme o takzvaných ***regresních stromech***, které namísto klasické klasifikace objektů do tříd (klasifikační stromy), provádějí odhad numerické hodnoty cílového atributu.

[1] Pod pojmem šum se rozumí především tzv. odlehlé hodnoty, tedy hodnoty značně vzdálené od hodnot pro daný atribut běžných, či hodnoty zjevně nesmyslné.

20.5.2 Rozhodovací pravidla

Použití rozhodovacích pravidel je podobné jako u rozhodovacích stromů. Rozhodovací pravidla jsou ***zapisována ve tvaru „IF Ant THEN Class“*** a umožňují klasifikovat kategorii cílového atributu z kombinace kategorií vstupních atributů. ***Ant*** je nazýván jako ***antecedent neboli předpoklad***, ***Class*** reprezentuje ***třídu neboli kategorii cílového atributu***. Stejně jako v případě rozhodovacích stromů jsou pro rozhodovací pravidla problematické numerické atributy a chybějící hodnoty.

Podobnost metod se odráží i v možnosti ***tvorby rozhodovacích pravidel z rozhodovacího stromu*** (teoreticky existují i postupy, jak získat z rozhodovacích pravidel rozhodovací strom). Každé pravidlo reprezentuje cestu stromem od kořene až do listu. Jednotlivé uzly (včetně kořenového, avšak bez listového uzlu) tvoří předpoklad pravidla (pro každý uzel jako atribut a jeho hodnotu jako příslušnou hranu), závěrem pravidla je listový uzel. ***Celý strom lze reprezentovat jako sadu rozhodovacích pravidel.*** Důvodem pro tento převod může například být snadnější zpracovatelnost rozhodovacích pravidel v rámci automatizovaných systémů.

V případě sady rozhodovacích pravidel chápeme ***jednotlivá pravidla jako navzájem nezávislá***, kdy při klasifikaci hledáme pravidlo odpovídající příkladu. Jinou možností jsou takzvané ***rozhodovací seznamy,*** které jsou uspořádaným seznam pravidel. Uspořádaným ve smyslu zápisu

IF Ant_1 THEN $Class_i$,
ELSE IF Ant_2 THEN $Class_j$,
ELSE IF Ant_3 THEN $Class_k$…

V případě rozhodovacího seznamu každé další ***pravidlo již nezohledňuje příklady, které by splnily pravidla předcházející*** („Uspořádání spočívá v tom, že v každé podmínce ELSE IF se implicitně skrývá negace všech podmínek předcházejících pravidel.“ (Berka, 2003, str. 134)).

Pro rozhodovací pravidla platí ***obdobná omezení jako pro rozhodovací stromy***. I řešení těchto omezení jsou podobná. Chybějící hodnoty je třeba nahradit (nabízí se ignorování objektu s chybějící hodnotou, vytvoření hodnoty specifické pro chybějící hodnoty, nahrazení některou z existujících hodnot.

20.5.3 Asociační pravidla

V minulé kapitole byla uvedena pravidla zvaná rozhodovací. Pravidla zapisovaná ve tvaru ***IF-THEN***. ***Tento tvar mají asociační a rozhodovací pravidla shodný***, liší se však možnostmi v závěru pravidla. ***U asociačních pravidel není*** na rozdíl od pravidel rozhodovacích definován ***cílový atribut pro zařazení příkladu do určité třídy*** (v rámci nastavení konkrétního algoritmu lze nastavit omezení na vybrané atributy jako možné závěry pravidla), naopak nás zajímají ***vazby mezi různými atributy v předpokladu a v závěru***. S tím souvisí problematičtější vyhodnocování těchto pravidel.

Asociační pravidla jsou ***zapisována jako „Ant => Suc“,*** kde ***Ant*** značí ***předpoklad pravidla***, takzvaný antecedent a ***Suc značí závěr pravidla***, takzvaný sukcedent. Antecedent i sukcedent jsou kombinací kategorií (tedy pravidlo může obsahovat více kategorií více atributů jak na straně předpokladu, tak i na straně závěru pravidla).

Hodnocení jednotlivých asociačních pravidel vychází z kontingenční tabulky (též nazývané ***čtyřpolní tabulka***) reprezentující četnosti objektů, kde:

Tabulka 20-2: Čtyřpolní tabulka

	Suc	**¬ Suc**	∑
Ant	a	b	r
¬ Ant	c	d	s
∑	k	l	n

a je počet objektů, splňujících předpoklad i závěr pravidla, tedy splňují celé pravidlo
b je počet objektů, splňujících předpoklad, avšak nesplňujících závěr
c je počet objektů, nesplňujících předpoklad, avšak splňujících závěr
d je počet objektů, které nesplňují ani předpoklad ani závěr
r,s je počet objektů splňujících, resp. nesplňujících předpoklad

k,l je počet objektů splňujících, resp. nesplňujících závěr
n je počet všech objektů

Z hodnot uvedených ve čtyřpolní tabulce pak vycházejí charakteristiky asociačních pravidel, zmíním základní dvě, podporu (support) a spolehlivost (confidence):

- **Podpora** P(Ant ∧ Suc) = $\frac{a}{a+b+c+d}$

  tedy relativní počet objektů splňujících předpoklad i závěr

- **Spolehlivost** P(Ant | Suc) = $\frac{a}{a+b}$

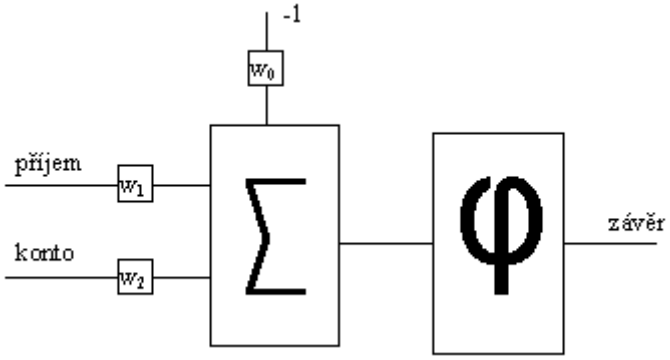
  tedy podmíněná platnost předpokladu, pokud platí závěr

Další charakteristiky lze nalézt například v knize Dobývání znalostí z databází (Berka, 2003, stránky 104-105).

20.5.4 Neuronové sítě

Myšlenka „umělých neuronových sítí" vychází ze znalostí o biologických neuronových sítích. Umělé neuronové sítě ***se skládají z jednotlivých navzájem propojených neuronů***. ***Neuron*** je jednotka, která přijímá na vstupu podněty, vstupující do součtového členu (provádí vážený součet skalárních součinů podnětu a jeho váhy). ***Váhy*** přiřazené každému vstupu jsou využívány v procesu učení (v okamžiku počátku procesu učení jsou váhy nastaveny na náhodná, malá čísla, blízká nule (Berka, 2003, str. 160) a učením se modifikují).

Hodnotu vzniklou v součtovém členu následně aktivační funkce transformuje ***na výstup***, pokud jeho minimální hodnota překročí práh. Příkladem aktivačních funkcí mohou být: sigmoidální funkce, funkce hyperbolického tangensu či tzv. skoková funkce (vrací „1" pro hodnotu vyšší než stanovený práh, jinak vrací „0").

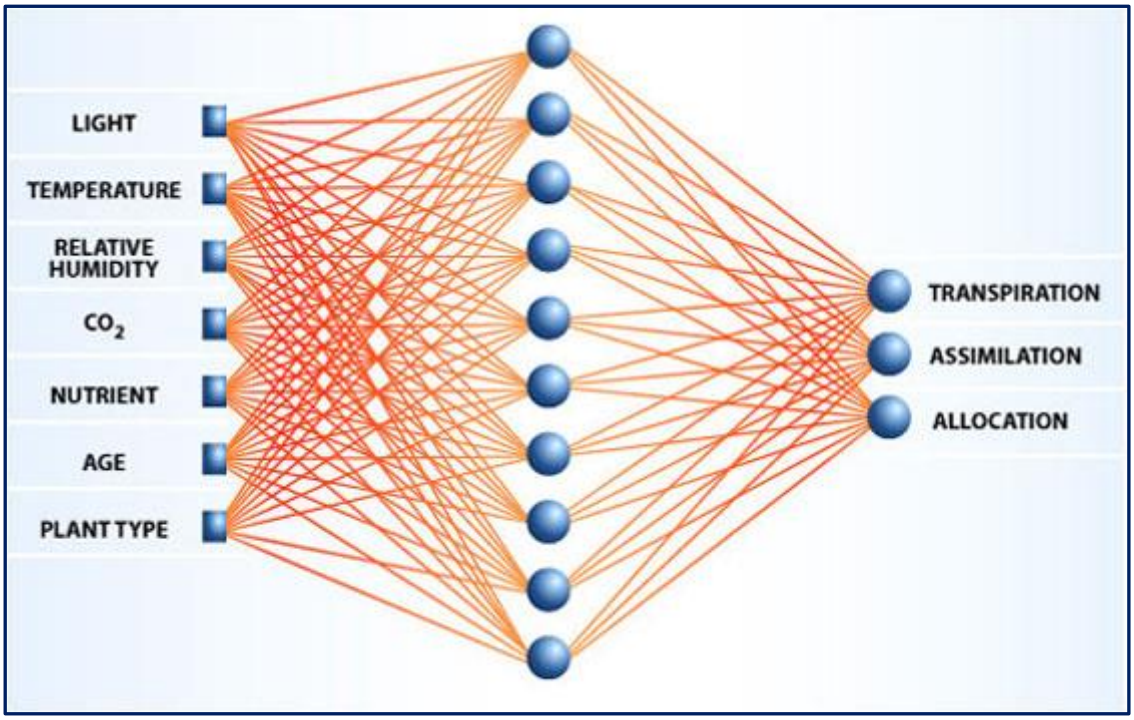


Obrázek 20-3: Schéma neuronu, kde w1 a w2 jsou vstupy, w0 je prahem. Suma reprezentuje součtový člen,φ aktivační funkci. (Berka, 2003, str. 158)

Uspořádání jednotlivých neuronů v sítí může mít více podob. První neuronová síť Perceptron (rok 1957, Frank Rosenblat), byla původně navržena jako model zrakové soustavy. Tato síť má 3 vrstvy. Později vznikly další topologie sítí, například vícevrstvé sítě a Kohonenova mapa. Vícevrstvá síť je tvořena jednotlivými vrstvami neuronů, přičemž neurony jedné vrstvy mezi sebou nejsou nijak propojeny, avšak každý neuron je propojen se všemi neurony vrstvy sousední (topologie o 1 skryté vrstvě, tedy 3 vrstvá síť, je zobecněním jednoduchého perceptronu). Kohonenova mapa, nebo také označovaná jako SOM (Self Organizing Map), je sítí dvouvrstvou, tvořenou vstupní vrstvou a vrstvou zvanou Kohonenova mřížka (vrstva uspořádaná do čtvercové matice vzájemně spojených neuronů, zároveň je vstupní vrstva propojena na každý neuron v mřížce) (Berka, 2003, str. 169). V rámci procesu klasifikace pak vítězí vždy pouze jeden neuron v Kohonenově mřížce (ten neuron, jehož váhy jsou nejblíže vstupnímu příkladu), výstupy ostatních neuronů jsou potlačeny.

V současnosti asi ***nejpoužívanějším algoritmem*** mezi metodami neuronových sítí je ***algoritmus SVM (Support Vector Machine).*** „Hlavní myšlenkou je, převést pomocí vhodné datové transformace

úlohu klasifikace do tříd, které nejsou lineárně separovatelné, na úlohu klasifikace do tříd lineárně separabilních“ (Berka, 2003, str. 171). Následně v prostoru transformovaných atributů, zvaných příznaků, je hledána nadrovina rozdělující příklady z trénovacích dat.



Obrázek 20-4 Příklad neuronové sítě s jednou skrytou vrstvou, reprezentuje chování rostlin (zdroj http://aemc.jpl.nasa.gov/activities/bio_regen.cfm)

Neuronové sítě patří mezi ***metody subsymbolické***, možnou nevýhodou může být ***horší čitelnost*** v nalezených znalostech pro člověka. Tento fakt však nemusí být překážkou v okamžiku, kdy jsou neuronové sítě využívány pro klasifikaci či predikci pomocí počítačových systémů. Zároveň se neuronové sítě liší oproti dříve zmíněným metodám v problematice zpracování numerických hodnot.

Neuronové sítě naopak ***využívají numerické atributy a problematické pro ně jsou atributy kategoriální*** (řešením může být binarizace). Neuronové sítě neumí pracovat s chybějícími daty a vyřešení této problematiky je opět směřováno do fáze přípravy dat (ačkoli existují i myšlenky práce například s několika různými sítěmi v rámci jedné úlohy či možnosti predikce chybějících hodnot v rámci běhu algoritmu (Sharpe & Solly, 1995)).

20.6 Efekty Data Science

- ***Využití netriviálních dat*** z různých zdrojů ke strategickým rozhodnutím – nalezení souvislostí napříč různými datovými zdroji.
- ***Využití výsledků*** například v těchto oblastech: CRM, optimalizace a automatizace ve výrobních podnicích, quality a risk management, analýzy a vzory zákaznického chování.
- **M*ultidisciplinarita*** – využití nejnovějších trendů informačních technologií (zpracování nestrukturovaných dat) ve spojení se statistickou analýzou a dalšími obory.
- Stále rostoucí ***počet nestrukturovaných dat***, který se nezastaví, s sebou přináší výzvu, jak tato data využít. S příchodem trendu Internet-of-Things dojde k ještě prudšímu nárůstu dat.

20.7 Omezení, problémy, předpoklady Data Science

- Obor je ***náročný na znalosti*** z konkrétního oboru a kooperaci s ostatními odborníky.

- ***Jeden z fundamentálních principů Data Science*** – správně vybrat, kombinovat a využívat uvedené úlohy pro řešení konkrétních problémů byznysu.
- ***Nekvalifikované analýzy*** mohou vést ke špatným rozhodnutím.
- Velké množství dat přináší otázku, ***jak tato data řídit*** uvnitř organizace a jak je zabezpečit před zneužitím.
- ***Vysoké náklady*** – je potřeba nejnovější technologie a vysoké personální náklady dané dosaženou kvalifikací pracovníků.

20.8 Zdroje

- PROVOST, F., FAWCETT, T.: Data Science for Business. What You Need to Know About Data Mining and Data-Analytic Thinking. O´Reilly Media. Sebastopol. 2013. ISBN: 978-1-449-36132-7
- KULHAVÝ, L., *Praktické uplatnění technologií Data Mining v pojišťovnictví,* DP, VŠE, 2011.
- BERKA, P.: *Dobývání znalostí z databází.* Praha : Academia, 2003. str. 366. ISBN 80-200-1062-9.
- FRIEBELOVÁ, JANA: Rozhodovací stromy. *Osobní stránka - Jana Friebelová.* [Online] 28. 12 2006. [Citace: 10. 11 2010.] http://www2.ef.jcu.cz/~jfrieb/rmp/data/teorie_oa/STROMY.pdf.
- SHARPE, P. K. and SOLLY, R. J.: Dealing with missing values in neural network-based diagnostic systems. *NEURAL COMPUTING & APPLICATIONS.* 1995, pp. 73-77.
- VOMLELOVÁ, MARTA: Rozhodovací stromy. *Výuka - Marta Vomlelová.* [Online] 10 20, 2009. [Cited: 11 10, 2010.] http://kti.mff.cuni.cz/~marta/slistromy.pdf.

21. Data mining, DMI

Kulhavý, L. (KIT, VŠE), upraveno

Účelem kapitoly je:

- poskytnout celkový obraz o principech a možnostech využití aplikací a nástrojů data miningu a formulovat hlavní doporučení pro jejich řešen,
- účelem zde není (obdobně jako v kapitole Data Science) uvádět definice podle autorů a charakterizovat do detailu a formalizovaného vyjádření jednotlivé metody a algoritmy, k tomu slouží další publikace, zejména ***Berka, Petr. 2003*****.** *Dobývání znalostí z databází.* Praha : Academia, 2003. str. 366. ISBN 80-200-1062-9. Hlavním smyslem je ***dokumentovat postavení data miningu v kontextu řízení*** celé firmy,
- vymezit významné ***efekty a případná omezení*** data miningu jako podstatné vstupy pro jejich způsob využití.

Dolování dat (Data Mining, DMI) je proces ***extrakce relevantních, předem neznámých nebo nedefinovaných informací*** z velmi rozsáhlých databází. Důležitou vlastností dolování dat je, že se jedná o ***analýzy odvozované z obsahu dat***, nikoli předem specifikované uživatelem nebo implementátorem, a jedná se především o odvozování prediktivních informací.

21.1 Podstatné charakteristiky data mining

Původní, a velmi často citovaná, definice dobývání znalostí pochází z článku „Knowledge Discovery in Databases: An Overview“ kolektivu autorů Frawley, Piatelsky-Shapiro a Matheus:

„Dobývání znalostí je netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat.“ (Frawley & Piatetsky-Shapiro, 1991, str. 58)

Později, roku 1996, přichází autoři Fayyad a Piatelsky-Shapiro s přepracovanou definicí v článku „From Data Mining to Knowledge Discovery: An Overview“. Tato definice již pracuje s pojmem dobývání znalostí z databází a měla by podle autorů více odrážet poslední vývoj v tomto oboru (poslední z pohledu vzniku článku, tedy rok 1996):

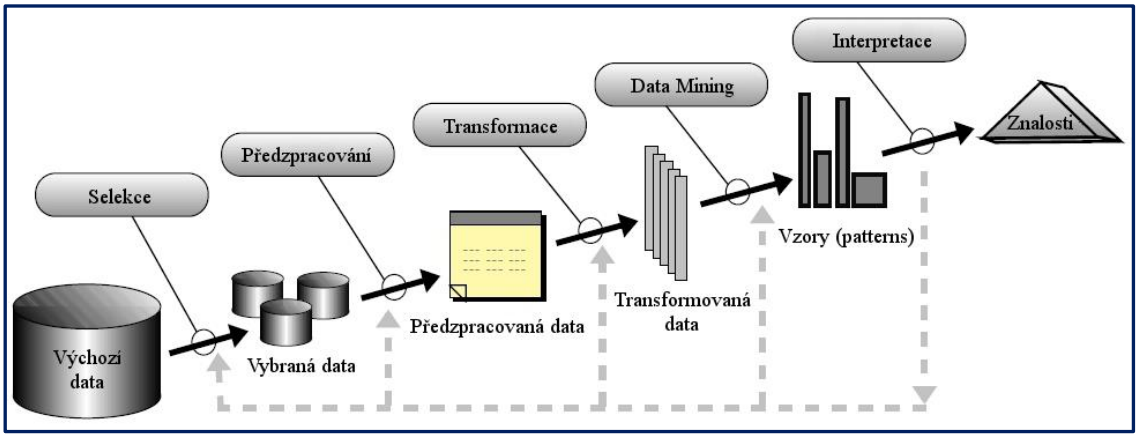
„Dobývání znalostí z databází je netriviálním procesem identifikace platných, nových, potencionálně užitečných, reálně srozumitelných vzorů v datech.“ (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996, stránky 1-34)

Proces dobývání znalostí se sestává ***z 5 fází***, které jsou prováděny opakovaně. Každá fáze tak může (a většinou je) prováděna opakovaně, s cílem nalezení nejlepšího možného výsledku dané fáze, a jeho předání jako vstupu fáze následující. Tyto fáze jsou (Berka, 2003):

1. ***Selekce*** vytvoření či shromáždění cílových dat, tj. těch dat, která poslouží pro dobývání znalostí
2. ***Předzpracování*** „čištění“ špatných dat, zpracování chybějících hodnot, příprava odvozených atributů apod. Předzpracování dat je pokládáno za nejdůležitější část celého procesu dobývání znalostí. Na jeho kvalitě závisí výsledek celého procesu dobývání znalostí, protože z nevhodných (či dokonce špatných) dat nelze získat kvalitní a správné výsledky.
3. ***Transformace*** převedení dat do formátů vhodných pro různé algoritmy data miningu či různé softwarové nástroje. Tato fáze zároveň může obsahovat různé agregace a výběry vzorků dat.

4. ***Data mining*** aplikace analytických metod. Výstupem této fáze jsou nalezené vzory a modely.
5. ***Interpretace*** fáze s cílem vyhodnocení nalezených vzorů a modelů. Tato fáze vyžaduje znalost odborníka na řešenou oblast, neboť identifikuje výsledky obsahující znalosti nové, předem známé, samozřejmé, nezajímavé apod. Výstupem se pak stávají analytické zprávy, vizualizace či rozhodnutí provést vhodné akce.

(Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996, stránky 37-54) (Berka, 2003, stránky 16-18)



Obrázek 21-1: Přehled fází tvořících proces Dobývání znalostí z databází (Fayyad, a další, 1996 str. 41)

V pátém kroku, interpretaci, byla zmíněna ***znalost odborníka na řešenou oblast***. Jsou to právě odborné znalosti z dotčené oblasti, které jsou nezbytné pro kvalitní výstupy. Jedná se především o znalosti odborníka na řešenou oblast (obchodní problematika, vědecký projekt apod.) a ***odborníka na zdrojová data*** (co znamenají které hodnoty, vztahy ve zdrojových relačních databázích atd.). ***Není možné generovat smysluplné vzory a modely, bez chápání jejich zdrojových dat***, natož je pak interpretovat bez znalosti řešené problematiky. K této otázce se vrátíme i v souvislosti s prediktivní analytikou,

21.2 Metodika CRISP-DM

CRISP-DM značí „***Cross-Industry Standard Process for Data Mining***“ (CRISP-DM, n.d.). ***Cílem*** je vytvoření standardního ***procesního modelu, bez vazby na konkrétní prostředí***, zdarma a veřejně dostupného, se zacílením na praktickou aplikaci data miningu (povětšinou komerční). Mimo jiné jde o předvedení data miningu jako perspektivní a dostatečně vyspělé technologie.

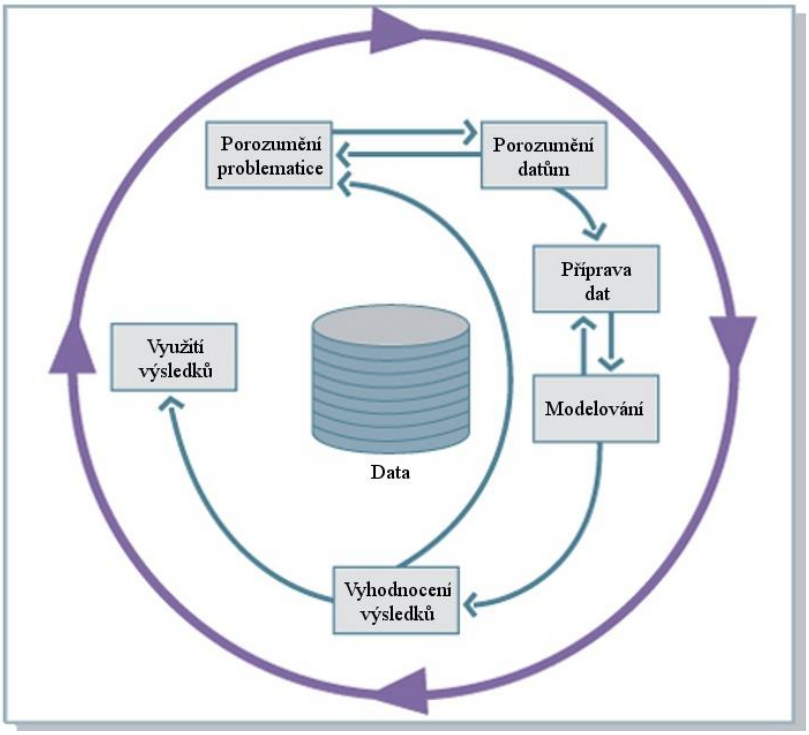
Metodika CRISP-DM se ***skládá z 6 fází***:

1. ***Porozumění problematice***
   tato fáze spočívá v identifikaci cílů úlohy, porozumění zadání formulovaného manažery (vycházející z business cílů). Důležité je takovéto zadání transformovat do zadání vhodného pro úlohu data miningu. Souvisejícím úkolem tohoto kroku je zajištění vhodných zdrojů (především datových, ale i výpočetních či lidských) a stanovení předběžného plánu práce.
2. ***Porozumění datům***
   fáze zabývající se získáním dat a jejich následnou analýzou. Jsou sledovány charakteristiky atributů jako četnosti hodnot, extrémní hodnoty, rozdělení četností hodnot apod. Je vhodné pro tuto činnost využívat přímo produktů pro data mining a vizualizačních nástrojů. Výsledkem je základní představa o datech, která máme k dispozici. Důležitá je identifikace chybějících hodnot a celková analýza kvality dat.
3. ***Příprava dat***
   úkolem této fáze je připravit finální datový soubor, který bude zpracováván konkrétními algoritmy (fáze čtvrtá). Tento soubor je připravován s přihlédnutím ke konkrétnímu zvolenému algoritmu (schopnost algoritmu zacházet s chybějícími hodnotami, vyžadované

datové typy na vstupu, rozsah datového souboru). Dochází zde k *„selekci dat, čištění dat, transformaci dat, vytváření dat, integraci dat a formátování dat“* (Berka, 2003, str. 26). Je třeba si uvědomit, že se jedná o časově náročnou operaci (uváděno, že se jedná dokonce o nejnáročnější operaci z hlediska času v rámci celého cyklu CRISP-DM) a že na kvalitě této fáze závisí ve značné míře úspěch celého projektu. Poměrně často také dochází k návratu do této fáze z kroku modelování, ať již z důvodů odhalených nepokrytých chyb v datech, či v případě nutných změn souvisejících s konkrétním algoritmem.

4. ***Modelování***
   je fází odpovídající výrazu data mining z pohledu KDD. V této fázi jsou aplikovány analytické metody (algoritmy pro úlohy data miningu), dochází k hledání optimálních nastavení parametrů pro jednotlivé algoritmy. Vzhledem k faktu, že neexistuje optimální metoda pro všechny úlohy, doporučuje se hledat vhodnou metodu a vhodné nastavení parametrů, a pro definitivní výběr nejlepší metody porovnat výsledky různých nastavení a různých metod. Součástí této fáze je i ověřování získaných výsledků.
5. ***Vyhodnocení výsledků***
   je fází, kdy dochází ke kontrole dosažených výsledků oproti manažerskému zadání. V této fázi již máme výsledné znalosti (modely, vzory) dosažené analytickými metodami. Je důležité zpětně zkontrolovat celý proces dobývání znalostí v souvislosti s manažerským zadáním a případně identifikovat nedostatečně zohledněné obchodní (respektive vědecké) aspekty. Metodika CRISP-DM taktéž považuje jako součást tohoto kroku rozhodnutí o využití získaných znalostí.
6. ***Využití výsledků***
   fáze umožňující další využití získaných znalostí. Výstup této fáze může mít podobu sahající od seznamu pravidel, přes formulace získaných znalostí či analytické zprávy, až po praktická nasazení získaných modelů (např. pro klasifikaci zákazníka) do produkčních systémů podniku.

(CRISP-DM, n.d.) (Berka, 2003, stránky 24-28)



Obrázek 21-2: Fáze procesního modelu CRISP-DM (CRISP-DM)

CRISP-DM zahrnuje přímo kroky věnující se ***formulaci zadání společně s pochopením problematiky a finálnímu využití výsledku***.

21.3 Aplikace data miningu

Úlohy data miningu lze jistě rozdělit dle mnoha kritérií a do různých skupin, např. nejprve jsou uvedeny dva primární ***cíle data miningu***: predikce a deskripce:

- ***Predikce*** „je funkcí, která využívá některých proměnných k předpovězení neznámých či budoucích hodnot jiných proměnných, které nás zajímají“ (těch, které nejsou vstupy funkce,
- ***Deskripce*** „se zaměřuje na hledání lidem srozumitelných vzorů popisujících data“.

V oblasti podniků jde samozřejmě především o aplikace technologií data mining ***zaměřených na plnění business cílů*** (finanční cíle, nárůst počtu zákazníků, identifikace nových obchodních příležitostí, minimalizace rizik…). Jedny z nejčastějších komerčních aplikací jsou (Máša, 4IZ560 - Data mining - praktické aplikace, 2008):

- ***Segmentace***
  Cílem segmentace je rozdělení objektů do skupin, které mají podobné charakteristiky. Nejčastěji se jedná o segmentaci zákazníků, a to dle různých charakteristik. Lze tak vytvářet segmentace demografické, behaviorální (dle chování), geografické apod. Máša (Máša, Adastra Czech - Segmentace – umění nebo věda?, 2008) pak uvádí rozdělení na „segmentace založená na výzkumu trhu“ a „segmentace založená na firemních datech“. Rozdíl je zřejmý, zatímco druhá jmenovaná segmentace využívá interní data podniku o klientech, první využívá průzkumy trhu. Segmentace založená na výzkumu trhu pak představuje řešení v situaci, kdy existují omezení (např. nepřítomná datová struktura v podniku – absence datových skladů apod.) či dokonce nedostupnost vhodných vlastních interních dat. Jiným důvodem může být zaměření na potenciální budoucí zákazníky na trhu. Důvodem pro aplikace segmentace jsou především cílené marketingové akce a vývoj nových produktů na základě specifik jednotlivých segmentů.
- ***Predikce odchodu zákazníků***
  Predikce odchodu zákazníků vychází z myšlenky, že *„náklady na získání nového klienta překračují náklady na udržení stávajícího klienta“* (Máša, 4IZ560 - Data mining - praktické aplikace, 2008). Cílem této aplikace data miningu je předpovědět klienty, kteří hodlají odejít ke konkurenci, přestat využívat určitého produktu apod. Vhodné je uvážit hodnotu konkrétního klienta pro podnik při vyhodnocování výsledků predikce odchodu zákazníků.
- ***Cílený marketing***
  Jedná se o marketingové akce specializující se pouze na určitý okruh klientů, takových, u kterých je oproti ostatním klientům vyšší pravděpodobnost, že zareagují na konkrétní kampaň. Díky tomu lze dosáhnout úspory v rámci kampaně (oslovíme menší počet klientů, s vyšším procentem úspěšnosti). Souvisí s pojmy „cross selling“ (nabídnutí souvisejícího zboží ke koupi) a „up selling“ (nabídnutí lepšího produktu přinášejícího podniku vyšší zisk).
- ***Credit scoring***
  Credit scoring je ohodnocením klienta dle úvěrového rizika. Tedy rizika, zda budou splácet např. svůj úvěr bance. Podnik se snaží za pomoci credit scoringu zjistit, jaké riziko má podstupovat. Příliš velké riziko může mít za následek nedobytné pohledávky, naopak riziko příliš nízké má za následek ztrátu zisku z potenciálních zákazníků.
- ***Fraud detection***
  Jedná se o aplikace, kdy jsou hledány podvody (ať již v rámci pojistných událostí, bankovních transakcí či v systémech telekomunikačních operátorů).
- ***Analýza nákupního košíku***
  zjišťuje souvislosti mezi produkty, které kupují zákazníci společně. (např. párky společně s hořčicí). Díky tomu lze zákazníkovi vhodně nabízet kombinace produktů (cross selling), pracovat s rozmístěním produktů v rámci prodejní plochy prodejny či stránky s produktem v rámci eshopu.

21.4 Efekty DMI

- Umožňuje ***využívat speciální algoritmy***, pomocí kterých lze v datech ***vyhledat strategické informace.***
- Umožňuje provádět ***proces rozsáhlé extrakce relevantních a předem neznámých*** dat z předem neznámých databází.
- Poskytuje informace na ***prediktivní a deskriptivní bázi***.
- Představuje ***silný analytický nástroj managementu*** v případě dobré znalosti DMI.
- Podporuje detekci založenou ***na fraudových systémech***.
- DMI napomáhá ***detekovat faktory***, které mají vliv na ***zvýšení příjmu a snížení nákladů*** společnosti.
- V oblasti Marketing / Retailnig nabízí managementu užitečné a přesné ***trendy chování zákazníků.***
- V oblasti Banking / Crediting představuje podporu finančních institucí v oblasti ***credit reportingu*** a dále v oblasti zabývající se úvěrovou politikou.
- V oblasti vymáhání trestného práva napomáhá ***odhalovat trestné činy a pachatele***, tím, že pomocí algoritmů zkoumá trendy, zvyky a jiné chování v geograficky ohraničené oblasti.
- Ve výzkumu ***napomáhá vědcům v oblasti analýz***, kdy výrazně urychluje jejich proces.

21.5 Omezení, problémy, předpoklady DMI

- Vzniká ***hrozba potencionálního zneužití osobních údajů*** ze strany společností vlastnících citlivá data (např. společnost American Express prodala jiné společnosti výpisy kreditních karet svých zákazníků).
- V praxi je velmi častá ***absence vhodného bezpečnostního systému***.
- Problém ***zneužití dat ze strany realizátora***, popřípadě majitele DMI analýzy.
- V praxi ***špatné povědomí a neznalost DMI*** ze strany managementu.
- Často příliš ***složitá problematika pro pochopení*** zákazníka.
- ***Vysoké náklady*** spojené s implementací.

21.6 Zdroje

Kulhavý, L. 2011. *Praktické uplatnění technologií Data Mining v pojišťovnictví,* DP, VŠE, 2011.

Abraham, Ajith P. 2008. *2nd European Conference on Data Mining (ECDM 2008).* Lisbon : IADIS Press, 2008. 978-972-8924-63-8.

Andrássyová, Eva and Paralič, Ján. 1999. Knowledge Discovery in Databases - a comparison of different views. *FEI Fakulta.* [Online] 1999. [Cited: Květen 31, 2010.] neuron.tuke.sk/~andrassy/publikacie/varazdin/andrassy.doc.

Berka, Petr. 2003. *Dobývání znalostí z databází.* Praha : Academia, 2003. str. 366. ISBN 80-200-1062-9.

CRISP-DM. CRISP-DM - Home. *CRISP-DM - Home.* [Online] CRISP-DM. [Cited: 11 06, 2010.] http://www.crisp-dm.org/.

Dalkilic, Mehmet M. Introduction & Motivation to Data Mining. *Computer Science Program | School of Informatics and Computing—Bloomington | Indiana University.* [Online] Indiana University. [Cited: Květen 31, 2010.] https://www.cs.indiana.edu/~dalkilic/.

Execution-MIH. What is knowledge discovery in databases- data mining-KDD? *Discipline of Making-it-Happen.* [Online] Execution-MIH. [Cited: Květen 30, 2010.] http://www.executionmih.com/data-mining/what-is-kdd.php.

Fayyad, Usama M., a další. 1996. *Advances in Knowledge Discovery and Data Mining.* Cambridge : AAAI Press/MIT Press, 1996. str. 560. ISBN: 0-26-256097-6.

Fayyad, Usama M., Piatesky-Shapiro, Gregory a Smyth, Padhraic. 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magzine.* 1996, stránky 37-54.

Frawley, William J. a Piatetsky-Shapiro, Gregory. 1991. *Knowledge Discovery In Databases.* Cambridge : AAAI Press/MIT Press, 1991. str. 539. ISBN 0-26-266070-9.

Friebelová, Jana. 2006. Rozhodovací stromy. *Osobní stránka - Jana Friebelová.* [Online] 28. 12 2006. [Citace: 10. 11 2010.] http://www2.ef.jcu.cz/~jfrieb/rmp/data/teorie_oa/STROMY.pdf.

Hájek, Petr, Havel, I. and Chytil, M. 1966. Metoda GUHA automatického vyhledávání hypotéz. *Kybernetika.* 2, 1966, pp. 31-41.

kddnuggets.com. 2007. Poll: Data Mining Methodology. *Data Mining Community's Top Resource.* [Online] kddnuggets.com, Srpen 2007. [Cited: Květen 23, 2010.] http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.

Máša, Petr. 2008. *4IZ560 - Data mining - praktické aplikace.* 2008.

—. 2008. Adastra Czech - Segmentace – umění nebo věda? *Adastra Czech.* [Online] Adastra, 6 2008. [Cited: 11 7, 2010.] http://www.adastra.cz/644_segmentace-e28093-umeni-nebo-veda.aspx.

Nguyen, Hung Son. 2006. DATA MINING COURSES. *Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw.* [Online] Únor 2006. [Cited: Květen 31, 2010.] http://www.mimuw.edu.pl/~son/datamining/DM/1-intro.pdf.

Novotný, Ota, Pour, Jan a Slánský, David. 2005. *Business Intelligence.* Praha : Grada Publishing, 2005. str. 256. 80-247-1094-3.

Parker, Kathleen. 2009. Kathleen Parker - Too Much Information, Too Little Understanding - washingtonpost.com. *Washington Post - Politics, National, World & D.C. Area News and Headlines - washingtonpost.com.* [Online] Washington Post, 4 1, 2009. [Cited: 11 7, 2010.] http://www.washingtonpost.com/wp-dyn/content/article/2009/03/31/AR2009033103318.html.

Rauch, Jan. 2009. *4IZ460 - Pokročilé přístupy k Dobývání znalostí z databází.* 2009. 4IZ460 02 29 zari.pptx.

SAS. SAS | SEMMA. *SAS | Business Analytics and Business Intelligence Software.* [Online] SAS. [Cited: 11 6, 2010.] http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html.

Vomlelová, Marta. 2009. Rozhodovací stromy. *Výuka - Marta Vomlelová.* [Online] 10 20, 2009. [Cited: 11 10, 2010.] http://kti.mff.cuni.cz/~marta/slistromy.pdf.

Witten, Ian H. a Frank, Eibe. 2005. *Data Mining: practical machine learning tools and techniques.* 2nd ed. Boston : Morgan Kaufmann, 2005. str. 525. ISBN 0-12-088407-0.

Yang, Qiang and Wu, Xindong. 2006. 10 Challenging Problems In Data Mining Research.

22. Text mining, machine learning

22.1 Text Mining

Text Mining, resp. textová analytika představuje analýzu textových zdrojů a získávání nových informací.

Zdroje mohou být ***velmi různorodé*** od knižních nebo novinových publikací, přes blogy, pracovních poznámek zaměstnanců, např. o kontaktech se zákazníky, o situaci na trhu apod. Text mining se velmi silně ***uplatňuje v aplikacích Competitive Intelligence***. Do určité míry s text mining souvisí i oblast analýzy obsahu, ***Content Analytics***, která rozšiřuje zdroje o fotografie, multimédia, hlas, případně další.

22.1.1 Efekty text mining

- Text mining i analýzy obsahu umožňují např. ***identifikovat významné nebo naopak problémové zákazníky,*** problémy spojené s dodávanými produkty nebo poskytovanými službami.
- Umožňují kvalitně a komplexně ***vyhodnocovat aktivity konkurence*** a případná ohrožení konkurencí.
- Vyhodnocují a připravují ***varování související s legislativními změnami***.
- Vyhodnocují ***bezpečnostní problémy*** nebo ohrožení.

22.1.2 Omezení, problémy, předpoklady text mining

- Je účelné provést ***komplexní analýzu dostupných zdrojů nestrukturovaných dat***, ale i jejich provázání nebo kombinace se strukturovanými daty, které mohou být předmětem analýzy jinými nástroji.
- V souvislostí s dostupností datových zdrojů pro text mining je účelné ***vyhodnotit jejich technickou i organizační dostupnost a ekonomickou náročnost*** vzhledem k požadavkům a potenciálním efektům.

22.2 Machine Learning, ML

Žydyk, R (KIT, VŠE)

Když se objeví problém, pro jehož řešení se rozhodneme použít machine learning, typicky následují tyto kroky:

1. ***Získání dat*** – pro výběr a naučení modelu budou potřeba podkladová data.

2. ***Explorace a validace*** – prvotní prozkoumání a tzv. profiling dat – zjišťuje se, jak data vypadají jako celek, co v nich je, co by mohlo chybět, jaká je struktura atp.

3. ***Čištění (transformace, předzpracování) dat*** – nejnáročnější operace procesu – data je třeba řádně připravit (vyřešit chybějící či odlehlé hodnoty, provést kódování textových proměnných, vytvořit nové prediktory atp.).

4. ***Rozdělení dat*** – když jsou data připravena, rozdělí se na trénovací a testovací množinu pro učení a ověření kvality modelu.

5. ***Výběr a trénování modelu*** – dle povahy úlohy je třeba vybrat druh modelu (regrese, klasifikace…) a provést trénování (učení).

6. ***Vyhodnocení výkonu modelů*** – pro dosažení co nejlepšího výsledku se zpravidla testují různé druhy modelů s různými hyperparametry.

7. ***Uložení nejlepšího modelu*** – nejlepší model je uložen a zaverzován pro další použití.

8. ***Převedení do produkce*** – model je třeba zpřístupnit pro použití v konzumujících aplikacích.

9. ***Monitoring (správa) modelu v produkci*** – model v produkčním prostředí je třeba dále monitorovat (chyby, rychlost, zátěž…

22.2.1 Kontinuální povaha životního cyklu ML modelu

Z předchozího vyplývá ***kontinuální povaha*** celého životního cyklu ML modelu. Když se změní nějaká část (data, model, aplikace) je ***třeba projít částí, nebo celým procesem*** znova. Pokud je celý proces vykonáván manuálně, lze si asi představit, že zapracovávat změny může být velmi časově i finančně náročné, navíc je zde riziko lidských chyb, jejichž počet se může v čase zvyšovat (únava, rutinní činnost…).

22.2.2 Produkcionalizace

Jedním z největších dnešních problémů v oblasti data science je převedení machine learning modelu do produkce, aby dlouho vyvíjený model dokázal skutečně generovat přidanou hodnotu pro business, která se od něj očekává.

Dle výzkumu společnosti Algorithmia z roku 2021 pouze 10 % z dotázaných společností dokáže dostat ML model do produkčního prostředí do týdne. Naopak 64 % společností trvá tento úkon měsíc a více.

22.2.3 Proč je produkcionalizace takový problém?

1. ***Nedostatečná expertíza*** – data scientisté, experti ve své oblasti, vynikají ve vývoji ML modelu – vědí jaký typ modelu použít pro jaký typ úlohy, jak nastavovat parametry, jak modely vyhodnocovat. Avšak problematika infrastruktury, kde má být model skutečně obsluhován je naprosto odlišná disciplína. Jednoduše to neodpovídá pracovní náplni data scientistů.

2. ***Komplexita ML úloh*** – obdobně jako data scientista není odborník na infrastrukturu, DevOps inženýři (případně programátoři) nejsou odborníci na machine learning úlohy.

3. ***Regulatorika*** – povaha ML úloh (business se na základě výstupů rozhoduje) vyžaduje bezvadnost, důsledný monitoring a governance.

4. ***Škálovatelnost*** – některé ML úlohy mohou být velice výpočetně náročné, a tak se nabízí využití paralelizovaného zpracování – jeho nastavení však opět není triviální.

5. ***Povaha využívání ML modelu*** – architektur nasazení ML modelu je několik (dávkové zpracování vs. zpracování v reálném čase vs. předpočítané hodnoty…), což situaci dále komplikuje.

6. ***Nabídka podpůrných nástrojů*** – aktuálně není na trhu příliš mnoho nástrojů, které by tyto činnosti zjednodušovaly.

7. ***Zakomponování ML modelu do aplikace*** – samotný model uložený někde na uložišti, případně už i zpřístupněný např. jako REST API endpoint je dobrá věc, nicméně pro business uživatele poměrně stále nepoužitelná – většinou bude existovat potřeba model obalit nějakou aplikací s grafickým uživatelským rozhraním. ¨

22.2.4 Governance

S převedením modelu do produkce souvisí ***i jeho správa***. Když už se model používá, je velmi důležité mít nad ním dohled a kontrolu.

V čase se mohou trendy v reálném světě měnit a model, který byl natrénován na určitých historických datech tak ***nemusí odpovídat skutečnému chování***. Může tak například začít docházet ke snížení výkonnostních metrik modelu či, mnohem hůře, business, který se na základě výstupů z modelů rozhoduje začne vykazovat velmi špatné výsledky.

Kromě business dopadů je třeba také ***sledovat celé prostředí***. Monitoruje se tedy jaké modely v produkci běží, ***na jakých datech jsou trénované***, kdo je schválil, jaká je vytíženost či doba odezvy.

S governance souvisí také (povinná) ***regulatorika*** – např. finanční ***instituce mají povinnost své modely řádně dokumentovat, monitorovat a verzovat***, aby nedošlo k fatálním chybám (např. obrovský úvěr pro jasně nesolventního klienta by mohl mít katastrofální důsledky) a aby bylo jasně doložitelné, na základě, jakých dat bylo dané business rozhodnutí učiněno.

22.3 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k uplatnění pokročilé analytiky, tj. data science, data mining a text mining.

- V rámci pokročilé analytiky převažují vesměs sofistikovaná ***individuální řešení*** založená na využití matematického a statistického aparátu.
- Vysoká náročnost řešení pokročilé analytiky většinou vyžaduje ji realizovat ***v kooperaci 2 specialistů***. Jeden by měl být orientován na znalost předmětu řešení, resp. obsahu řízení, a druhý zaměřený na znalosti metod a nástrojů data miningu, text miningu analytiky.
- Analýza a návrh všech forem pokročilé analytiky musí vycházet z kvalifikovaného posouzení ***aktuálních i budoucích potřeb sofistikovaných analýz***.
- Jedním z podstatných úkolů řešení těchto aplikací je zajistit jejich ***efektivní a relativně nenáročné zpřístupnění*** běžným uživatelům. Pokud se toho nedosáhne je často vynaložená práce znehodnocena.
- Před zahájením řešení aplikací pokročilé analytiky je účelné dobře ***vyhodnotit faktory firmy*** a firemního prostředí (kapitola 3) a podle toho rozhodnout, zda v daných podmínkách má taktové řešení smysl.
- Zapojení manažerů a specialistů, tedy lidí s omezenými časovými možnostmi, současně předpokládá, že již na počátku řešení budou ***jasně deklarovány jejich ekonomické i mimoekonomické efekty***, které řešení může firmě přinést.
- Vstupem pro analytickou přípravu analytických aplikací by měla být ***sada analytických otázek k potřebám analytiky v rámci jednotlivých oblastí řízení***. Podklad je k dispozici v textu „*AF 2.1 Oblasti řízení*“, podkapitola x.8 v rámci každé kapitoly, např. „*7.8 Analytické otázky pro řízení prodeje“.*
- Aplikace pokročilé analytiky mají výrazné nároky na ***dostatečný rozsah a kvalitu datových zdrojů***, což je třeba posoudit také jako jeden z faktorů řešení.

23. Prediktivní analytika, Predictive Analytics, PA

Účelem kapitoly je:

- vymezit hlavní ***charakteristiky*** prediktivní analytiky,
- analyzovat funkce a možnosti prediktivní analytiky z pohledu jejich ***potenciálních efektů*** a na druhé straně ***omezení*** jako podklad pro určení jejího uplatnění v praxi.

Schopnost úspěšné predikce je ***využita ke zlepšení rozhodnutí***, které je tak více postavené na faktech (vztazích, trendech) nalezených v datech než na intuici. Postupně se realizuje ***promítání aktuálních událostí do prediktivních modelů*** v reálném čase.

Prediktivní analytika (Praus, 2013) je ***typ analýzy využívající data a prediktivní modely pro předpověď jevů na mikroekonomické úrovni.*** PA využívá technologie, která se učí ze zkušeností (dat), aby předvídala budoucí chování jedinců k lepšímu rozhodování (SIEGEL, 2013). PA v sobě zahrnuje množství statistických a analytických technik, které využívá k předpovědi. (NYCE, 2007).

Předpovědi poskytnuté prediktivní analytikou se týkají převážně ***mikroekonomických efektů***, vyhodnocuje se chování jednoho člověka, a ne masy lidí. PA například předvídá, kdo si, s jakou pravděpodobností koupí nějaký produktů (SIEGEL, 2013).

Vstupem prediktivní analytiky jsou data se ***znalostí cílové hodnoty (targetu)***, která má být v budoucnu predikována. Data se v rámci PA zorganizují, pročistí a vytvoří se nový datový zdroj. Na něj jsou poté ***aplikovány metody data miningu***, především různé typy prediktivních modelů, které jsou schopné strojového učení z velkého množství dat.

Výstupem je nejúspěšnější prediktivní model, schopný s určitou pravděpodobností cílovou hodnotu target předpovídat.

→ Detailněji v dokumentu „*AF_II_05_05_Prediktivni_Analytika.pdf*“

G) Data pro podnikovou analytiku

[25] Správa dat	
[25.1] Metadata	**[25.2] Master data**
[25.3] Data profiling	**[25.4] Data governance**
[26] Zajištění kvality a čištění dat	

Oddíl se orientuje na řešení a zajištění datových zdrojů pro podnikovou analytiku. Zahrnuje tyto hlavní části:

- ***Řešení správy dat***, kam obvykle spadá:
  - ***Řízení a správa metadat***, která představují klíčové charakteristiky veškerých informačních systémů a jejich aplikací, tj. byznys procesů, resp. úloh, transformací dat, analytických a plánovacích úloh, jejich funkcionality a výstupů a dalších.
  - ***Řízení master dat***, která charakterizují business entity jako lidi (zákazníci, dodavatelé), místa (prodejny, výrobní závody), věci (produkt, aktiva), které jsou klíčově pro fungování organizace.
  - ***Data Profiling,*** což je použití technik analýzy dat na existující datový zdroj, jejichž účelem je odhalení jeho skutečného obsahu, struktury a kvality dat, které jsou v něm obsažené.

- ***Data governance*** je programem zaručující patřičnou kvalitu dat, tak aby jednotliví zaměstnanci dodržovali jasně nastavená pravidla při jakékoliv práci s daty. [Profisee, 2020].

- ***Zajištění kvality a čištění dat*** je zde důležité proto, že jedním z nejčastějších důvodů neúspěchu projektů podnikové analytiky je nízká kvalita dat a zejména produkčních datových zdrojů.

25. Správa dat

Účelem kapitoly je:

- vymezit ***podstatu metadat***, možnosti a efekty jejich využití, případné nároky a problémy,
- charakterizovat ***Master Data Management*** jako sadu procesů a postupů pro zajištění klíčových dat firmy a jejich kvality,
- vymezit ***datové profily*** zahrnující charakteristiku jejich obsahu, struktury a kvality dat,
- ve všech případech definovat ***podstatné efekty a omezení*** jednotlivých přístupů jako vstup do analýz datových zdrojů.

25.1 Řízení a správa metadat

Metadata jsou definována jako data o datech a v této souvislosti slouží pro dokumentaci konkrétních implementací informačních systémů podniku. Jsou ***popisem veškerých informačních systémů i součástí podnikové analytiky***, tj. byznys procesů, resp. úloh, transformací dat (ETL / ELT), analytických a plánovacích úloh, jejich funkcionality a výstupů a dalších.

Z pohledu podnikové analytiky zahrnují zejména ***datové modely, popisy funkcí, business a transformačních pravidel, reportů či*** požadavků na reporty apod.

25.1.1 Efekty a přínosy

- Metadata vnáší do celého informačního systému a jeho řízení ***větší pořádek*** a vysokou úroveň dokumentace,
- Metadata slouží pro efektivnější ***identifikaci a řešení případných chyb a problémů*** v informačním systému. Metadata přesněji vymezují i zodpovědnosti a kompetence za jednotlivé části informačního systému.
- Využití metadat výrazně přispívá k ***řešení souladu potřeb byznysu a IT***, tyto vztahy mohou být jasně definované,
- Význam a využití metadat ***se zvyšuje i s architekturou zaměřenou na služby*** (SOA),

25.1.2 Problémy a omezení

- ***Zvyšuje se pracnost*** spojená s rozvojem a provozem IT.
- Metadata pro racionální jejich využití vyžadují i jejich ***průběžnou aktualizaci na bázi prakticky denně realizovaných procesů*** a s nimi spojených rozhodování, což při běžném provozu podniku a jeho IT není často dosaženo,
- Je nezbytné jasně ***identifikovat existující zdroje metadat***, včetně jejich vazeb a využití v nástrojích pro řízení dat a jejich integraci.

25.2 Master data a jejich řízení

Wejskrab, J. (Solitea, a.s.)

Úvodem této části je vymezen ***pojem master data*** společně s dalšími kategoriemi dat. Následuje definování pojmu řízení master dat, jeho benefity, výzvy a důvody. Jsou diskutovány architektury systémů na řízení master dat. Nakonec se shrnuje současný stav ***řízení master dat ve veřejném sektoru***.

25.2.1 Master data

Pro vymezení pojmu „master data“ je nejprve důležité definovat ***rozdělení dat do kategorií*** včetně avizovaných master dat. V literatuře se data dělí na transakční data, referenční data, metadata a

master data. Někteří autoři pojmy master data a referenční data zaměňují. (McGilvray, 2008; Allen a Cervo, 2015). V následujících bodech jsou definovány všechny čtyři kategorie dat a rozdíly mezi nimi:

Transakční data – Popisují ***interní či externí transakci*** (reakci na událost), která souvisí s činností organizace. Transakční data bývají doplněna dalšími daty v podobě referenčních dat nebo master dat, které dodávají transakčním datům kontext (McGilvray, 2008; Allen a Cervo, 2015). Transakční data ***zřídka obsahují všechny atributy master dat***, jelikož jsou optimalizovány pro zápis. Obsahují pouze nezbytné popisné atributy (Mundy, Thorthwaite a Kimball, 2008). Příkladem transakčních dat jsou objednávky, prodeje, platby apod.

Master data – (v češtině též označována jako kmenová data (Patrný, 2011)) ***popisují business entity*** jako lidi (zákazníci, dodavatelé), místa (prodejny, výrobní závody), věci (produkt, aktiva), které jsou klíčově pro fungování organizace. V porovnání s transakčními daty je ***jejich objem nižší a jsou relativně konstantní***. Pro master data je typické, že jsou ***používána napříč organizací*** rozličnými aplikacemi a procesy. Jejich synchronizace v organizaci je klíčová pro správné fungování. Master data jsou seskupována v entitách a doménách. Mohou obsahovat i referenční data (McGilvray, 2008; Allen a Cervo, 2015).

Referenční data – představují seznam záznamů, které ***jsou referovány aplikacemi, systémy, procesy a transakcemi***. Oproti master datům nemusí být specifická pro danou organizaci. Typicky referenční data pochází ***z ISO standardů***, od státní správy či regulátorů. Avšak organizace mohou také mít svá ***vlastní referenční data***, jako kódy produktů, které jsou v organizaci standardizované. Příkladem jsou zkratky států, pohlaví, města apod. (McGilvray, 2008; Allen a Cervo, 2015). Česky mohou být také označována jako číselníky nebo ve státní správě často jako registry.

Metadata – jsou data o datech. Dle obsáhlejší definice jsou to ***data, která nesou strukturovanou businessovou či technickou informaci*** o jiných datech, například pravidla, omezení, statistické údaje, časové značky, vztahy, logickou a fyzickou strukturu dat (DAMA International, 2017). V literatuře se dále rozdělují na ***metadata technická, businessová, auditová***. Technická metadata popisují typ souboru, jméno, původ nebo jména sloupců. Příkladem businessových metadat jsou definice sloupců, statistiky o datech apod. Auditová metadata ukládají informace jako kdo, kdy a jak naposledy četl, upravil, či nahrál data. Například to může být datum s časem či identifikátor uživatele. (McGilvray, 2008; Allen a Cervo, 2015).

Charakteristikou master dat je stabilita, znovupoužitelnost a vysoká hodnota pro organizace (Berson a Dubov, 2007). Navíc kvalita transakčních dat částečně závisí právě i na kvalitě master dat (Kokemüller a Weisbecker, 2009). Všechny tyto body jsou imperativem pro řízení master dat.

25.2.2 Řízení master dat

Řízení master dat (anglicky ***master data management***; dále také zkráceně ***MDM***) definujeme jako ***soubor procesů a technologií pro vytvoření a údržbu spolehlivého, autoritativního, udržitelného, přesného a bezpečného zdroje master dat*** a vztahů mezi nimi. Cílem je vytvoření ***autoritativní verze master dat***, která jsou používána napříč procesy, aplikacemi, systémy a business jednotkami (Berson a Dubov, 2007).

Loshin (2013) charakterizuje MDM jako ***soubor technik na řízení dat organizovaný business účastníky***, klíčovými stakeholdery a business klienty. Otto (2012) pak zdůrazňuje, že řízení master dat nejen obsahuje samotný software, ale také komplikované organizační procesy na řízení dat v MDM systému.

Iniciativa vytvoření MDM programu v organizacích často přichází buďto z businessu, nebo IT oddělení. Častou motivací businessu jsou četné problémy s master daty či snaha o zkvalitnění služeb pro zákazníky. IT iniciativa často nemá dostatečnou podporu kvůli chybějícímu business problému (Berson a Dubov, 2007).

25.2.2.1 Důvody k řízení master dat

Od 90. let minulého století se drasticky snížila cena úložiště, internet zvýšil všudypřítomnost dat. Mnohdy mají organizace obrovský počet dat, často duplicitních a neexistuje jedna definitivní verze dat. Kvůli tomu je ***práce s daty*** v organizacích ***méně efektivní*** a může dokonce vést k chybným výstupům. Organizacím tudíž vznikla potřeba řízení master dat a jejich kvality (DAMA International, 2017).

Organizace na data více spoléhají a činí na jejich základě klíčová rozhodnutí. Navíc data mohou být nejen ***interní, ale i externí***. V praxi se pak používají rozličné aplikace a databáze, které uchovávají

data. Často duplikovaná, nekvalitní, nekonzistentní, uložena separátně v oddělených silech (Lucas, 2010).

Aspirací řízení master dat je ***vytvoření jednoho důvěryhodného zdroje dat***, konkrétně kritických, často používaných entit. Příkladem takových entit je zákazník, produkt apod. Řízení master dat znamená ***nejen software*** na manipulaci s daty, ale také ***souhrn řídích procesů, pravidel, infrastrukturu***. Řízení master dat je technickým projektem, který však vyžaduje minimálně stejné organizační snahy (Cheung a Chung, 2014). Řízení master dat ***zahrnuje nasazení master data systému, data governance, řízení kvality master dat či tvorbu hierarchií*** (Cervo a Allen, 2011; Loshin, 2013).

Zmíněnými koncepty při řízení master dat jsou ***role datových stewardů a data ownerů*** (česky vlastníci dat). ***Data owneři*** jsou jednotlivci či ***skupina jednotlivců z businessu***, kteří vytváří, získávají a mohou i kontrolovat data. Příkladem jsou obchodní zástupci, kteří mají data o svých klientech. ***Datový steward*** je role ***na pomezí businessu a IT***. Stewardi nejsou vlastníky dat, ale mají za úkol sledovat kvalitu dat, jejich dostupnost, a spolupracovat s technickými týmy jako datoví architekti, databázoví administrátoři, BI tým. Zároveň musí rozumět dané doméně dat (Berson a Dubov, 2007)

Podle Plotkina (2014) se ***datoví stewardi mohou dělit*** do několika rolí, primárně na ***business a technické data stewardy***. Business stewardi s daty běžně pracují, dělají návrhy k jejich zkvalitnění, vytvářejí pravidla. Typickou rolí jsou business analytici. Naopak techničtí data stewardi pochází z IT oddělení. Mají přehled o ETL procesech, úložištích dat a hlavně vědí, kde jsou data generována, uložena, transformována.

25.2.3 Efekty a přínosy

- ***Proaktivní řízení master dat*** má na organizaci pozitivní vliv. Nejdůležitějším efektem je pro organizace ***snížení nákladů***. Řízení dat decentralizovaně v silech má celou řadu negativních účinků zvyšujících náklady na řízení kvality stejných master dat v různých informačních systémech. Rozhodování či použití chybných či neúplných master dat také způsobuje zvýšené náklady. Navíc master data pro uživatele nejsou snadno dostupná na jednom místě.
- Kvalitní master data mají navíc ***pozitivní vliv i na transakční data***. Transakční data bývají doplněna master daty a v případě jejich nízké kvality dochází také k poklesu kvality transakčních dat. Kombinací transakčních a master dat získávají organizace analytický pohled na data, na jehož základě se pak rozhodují. Vysoká kvalita dat je proto klíčová.
- Neméně důležitým benefitem je ***zvýšení kvality služeb pro zákazníky***. Iniciativy řízení master dat jsou často spojeny se snahou sjednocení zákaznických dat a zkvalitnění služeb navázaných na data. V některých případech se jedná i o naplnění regulatorních požadavků (Berson a Dubov, 2007).
- ***Dalšími efekty*** jsou zejména:
  - ***Zajištění jednoznačnosti, přesnosti, jasného určení***, sémantické konsistence a efektivní sdílení dat jako vlastního majetku firmy,
  - ***Poskytování komplexních informací o zákazníkovi*** – úložiště master dat pro zákaznická data poskytuje jediný zdroj pro konsolidaci všech zákaznických aktivit v konzistentní formě a využívá se v transakčních i analytických aplikacích,
  - MDM ***snižuje komplexitu integrace nových dat a systémů*** v podniku a umožňuje mu rychle i s omezenými zdroji reagovat na nové obchodní příležitosti,
  - Konzistence dat podporovaná MDM ***snižuje nedůvěru k datům*** a zefektivňuje rozhodovací procesy,
  - ***Zajištění souladu s normami a předpisy*** jako např. Sarbanes-Oxley apod.,
  - ***Zjednodušení vývoje aplikací*** – s konsolidací master dat do centrálního úložiště se zvyšuje i úroveň konsolidace aplikační funkcionality.

25.2.4 Problémy a omezení

- Implementace řízení master dat ***vyžaduje vytvoření jednoho referenčního zdroje dat***, deduplikaci, obohacení, zajištění konzistence, redukci nekonzistencí a chyb. Vyžaduje proto import dat nebo přímou integraci zdrojových systémů do MDM systému. Migrace dat, tvorba modelu a další související aktivity mohou způsobovat prodloužení implementační fáze MDM (Pansara, 2021).

- Neméně důležitou součástí jsou ***organizační změny***. Časté je zavedení nových rolí a procesů v organizaci. Stanovují se vlastníci dat (data owners), stewardi, definují se business pravidla, standardy, oprávnění (Ibrahim, Mohamed a Satara, 2021).
- ***Organizační problémy často způsobují nepřijetí***, nebo špatné řízení master dat. Pro splnění očekávání je doporučeno nejprve stanovení cílů. Jelikož se jedná o aktivitu s vlivem na celou organizaci, pro vytvoření master dat je ***nezbytná spolupráce napříč odděleními*** a jednotkami (Ibrahim, Mohamed a Satara, 2021).
- Neméně důležitá je ***úroveň datové maturity v organizaci a kompetence*** zodpovědných pracovníků. Nejen pro pochopení řízení master dat, ale také k pochopení příbuzných konceptů (Ibrahim, Mohamed a Satara, 2021).
- Specifickým problémem ***ve veřejném sektoru je legislativa***, která často nabádá k rozdělování dat do sil (Vilminko-Heikkinen a Pekkola, 2012).
- V neposlední řadě je důležité zmínit, že ***master data ve větších organizacích spadají do různých domén.*** Identifikace všech domén a jejich prioritizace je další výzvou pro MDM týmy (Allen a Cervo, 2015).
- Musí být nastaven ***efektivní systém motivace mezi pracovníky*** uživatelské sféry i IT.
- Musí být jasně deklarována ***pravidla pro hodnocení kvality dat***.
- MDM musí být efektivně ***začleněno do řízení datových zdrojů v rámci řízení IT***, a to se všemi podstatnými souvislostmi, včetně ekonomických.

25.2.5 Doménové řízení master dat

Obvykle jsou master data v organizacích ***rozdělena do několika domén***. Domény organizují ***klíčové entity pro organizaci dle oblastí*** a zaměření (Allen a Cervo, 2015). Typickými entitami jsou produkty, zákazníci, dodavatelé, které se mohou skládat z několika dalších entit, atributů i hierarchií (Allen a Cervo, 2015).

Organizace pracují ***s jednotkami až s desítkami domén***. Vyšší počet domén je charakteristikou velkých organizací. Řízení každé z domén v programu master dat však nemusí být ekonomicky výhodné. ***Identifikace domén*** k řízení by měla být dle Allena a Cerva (2015) učiněna ***na základě pěti faktorů***:

- ***Hodnota*** – čím vyšší hodnotu má doména pro organizaci, tím důležitější je její řízení. Zákazníci jsou obvykle jednou z prvních domén, které jsou organizacemi řízené.
- ***Objem*** – vysoký objem master dat v doméně zvyšuje komplexitu a způsobuje potřebu jejich řízení.
- ***Volatilita*** – častá úprava master dat je problematická, obzvlášť pokud jsou data používaná napříč aplikacemi, systémy a uživateli.
- ***Znovupoužitelnost*** – data používaná napříč organizací mají vysokou prioritu pro jejich řízení. Selhání jejich řízení může způsobovat rozdílné verze master dat.
- ***Komplexita*** – Některé domény jsou značně komplexní a vyžadují jejích řízení. Například produktová doména se může skládat z několika entit (např. produkty, kategorie, barvy, velikosti) (Allen a Cervo, 2015).

25.2.6 Architektura MDM systému

Na MDM lze pohlížet z různých perspektiv. ***Manažerská perspektiva*** diskutuje cíle, road mapu, governance, nastavení procesů a reporting. Druhá ***informační architektura*** se zabývá tvorbou datového modelu, tokem dat a kontrol dat. ***Kvalita master dat*** se zabývá očištěním dat, jejich analýzou, implementací konzistenčních pravidel. ***Technologická*** se zabývá databázemi, MDM aplikacemi, dashboardy a souvisejícími nástroji. Nakonec ***proces vývoje***, který se zabývá vývojem a implementuje všechny ostatní perspektivy (Dahlberg, Heikkla a Heikkla, 2015).

Architektura MDM systémů může být pojata různě komplexně. V minulosti existovaly MDM aplikace, které ukládaly pouze referenci na master data v jiných systémech. Jiné architektury mohou pojímat master data jako registry bez napojení na zdrojové systémy, kde hlavním úkolem je poskytovat master data dalším aplikacím a systémům. ***Komplikovanější architektury*** MDM systémů jsou ***napojené na zdrojové systémy***, vyrovnávají konflikty mezi daty a dále slouží jako zdroj dat pro jiné aplikace a systémy (Berson a Dubov, 2007).

Využití MDM systémů se v organizacích může lišit. Berson a Dubov (2007) dělí využití na ***tři typy***:

- ***Analytické*** – MDM systém je využíván procesy a systémy na analýzu dat a reporting, například jako zdroj dimenzí pro datový sklad. Obvykle dochází ke čtení dat z MDM. Zároveň MDM systém nemění data ve zdrojových systémech.
- ***Provozní*** – Master data jsou využívána pro provozní úlohy, například jako zdroj pro transakční/provozní aplikace. Cílem je zvýšení kvality dat v provozních systémech.
- ***Kolaborativní*** – MDM systém umožňuje uživatelům spolupracovat na tvorbě a údržbě master dat.

Z hlediska architektury lze MDM systémy dále ***rozdělit na komponenty*** (Berson a Dubov, 2007):

- ***Úložiště dat*** – úložiště master dat, například databáze.
- ***Zdrojové systémy*** – zdroj pro master data (externí nebo interní).
- ***Poskytovatelé dat třetích stran*** – zdroje pro master data od třetích stran.
- ***ETL*** – extrakce, transformace a nahrávání dat ze zdrojových systémů či do koncových systémů. Připojuje se na zdrojová data pomocí různých prostředí, například skrze dotazování databáze či webové služby.
- ***Koncové systémy*** – systémy používající master data (např. datový sklad, aplikace).

Po stanovení architektury MDM systému následuje jeho ***implementace***. Joshi (2007) navrhuje ***osm kroků*** k implementaci MDM:

1) Definování toku dat.
2) Identifikace zdrojů master dat a uživatelů.
3) Sběr dat.
4) Definice modelu.
5) Definice požadovaných funkcí MDM nástroje.
6) Spojení zdrojových dat do úložiště master dat.
7) Vytvoření a údržba business pravidel.
8) Publikace master dat a integrace do aplikací.

25.2.7 Řízení master dat ve veřejném sektoru

Nicméně veřejný sektor na rozdíl od podniků poskytuje veřejnou službu. Tudíž potřeby státních institucí v oblasti řízení master dat jsou rozdílné.

25.2.7.1 Specifika řízení master dat ve veřejném sektoru

Z hlediska veřejných institucí rozhodování na základě špatných dat implikuje ***nejen špatné hospodaření***, ale ***i poskytování chybných dat veřejnosti*** a špatné rozhodování, které může ovlivňovat chod celé země. Zároveň veřejné instituce mnohdy nesou důkazní břemeno. Z toho důvodu je důležité mít jeden pravdivý, důvěryhodný zdroj master dat (Cheung a Chung, 2014).

Zahájení iniciativ na řízení master data ***ve veřejném sektoru má svá specifika***. Je nejen nezbytná podpora vysokého managementu, ale v některých případech také ***potřebná legislativa a politická podpora*** (Haneem et al., 2019).

Studie (Haneem et al., 2015; Haneem, Kama a Azmi, 2016) se zaměřuje na první fázi ***identifikace a prioritizace master dat ve veřejném sektoru***. Autoři poukazovali, že během rozhovorů s reprezentanty státní správy evidovali nízkou datovou maturitu. Respondenti často zaměňovali referenční a transakční data. Navíc z uvedené studie vyplývá, že ***veřejný sektor prioritizuje data o právnických osobách, fyzických osobách,*** či více specificky například o studentech.

Neobvyklá je také ***prioritizace domén*** oproti běžné podnikové praxi. Dle studií (Haneem et al., 2015; Haneem, Kama a Azmi, 2016) byli ***prioritizováni poskytovatelé služeb***[2] oproti zákaznickým

[2] Entity poskytující veřejné služby, např. služby úřadu práce

profilům[3] v podnikové praxi. Dalšími dvěma nejzásadnějšími oblastmi master dat byly zákaznické profily a služby[4].

Yoon, Joo a Kwan (2018) upozorňují, že každá státní organizace schraňuje data do svých databází. Ve veřejném sektoru jsou ***master data ukládána v silech dle organizací***. Každá z nich má své databáze a data jsou duplikována napříč organizacemi. Například profil jednoho občana je v několika registrech (vzdělání, zdravotnictví, daně) (Haneem, Kama a Azmi, 2016). ***Obtížnost řízení*** master dat může proto být ještě ***vyšší než v soukromém sektoru***, jelikož agenda státních institucí je širší a počet přidružených organizací často vyšší.

Následkem jsou ***vyšší náklady*** na řízení dat, ***problémy s propojováním dat*** napříč organizacemi a ***nižší kvalita dat***. V reakci na tyto problémy Yoon, Joo a Kwan (2018) doporučují vytvoření sdílených národních master dat pro často užívané číselníky.

Identifikování master dat ve veřejném sektoru ***podporuje kolaboraci organizací na datech, sdílení zdrojů*** a vyšší úroveň služeb pro občany. Pro organizace veřejného sektoru je však identifikace master dat problematická, jelikož veřejnému sektoru je věnována nízká pozornost ve výzkumu.

Vilminko-Heikkinen a Pekkola (2017) identifikovali ***specifické problémy s implementací*** řízení master dat na projektu ve veřejném sektoru:

- Identifikace relevantních procesů, účastníků a expertních skupin.
- Zapojení všech organizací do MDM projektu.
- Meziorganizační spolupráce a komunikace (spolupráce více organizací je časově a organizačně náročná).
- Porozumění konceptu MDM a dalším souvisejícím konceptům.
- Určení rolí a zodpovědností.
- Sjednocení pojmů (např. zákazník má rozdílné definice).
- Konsenzus na cílech MDM projektu.
- Identifikace vlastníka MDM projektu.
- Implementace organizačních změn (např. změna interních procesů).
- Legislativní výzvy (fungování veřejného sektoru je přímo závislé na legislativě).

25.2.7.2 Příklady řízení master dat ve veřejném sektoru

Berson a Dubov (2007) diskutují řízení master dat v kontextu veřejné správy. V master datech spatřují ***potenciál ke zkvalitnění poskytování služeb*** jak k fyzickým (G2C), tak právnickým osobám (G2B). Podobně jako jiní autoři však zmiňují především holistická master data o zákaznících (např. občanech). V rámci statní správy pak zdůrazňují ***přínos master dat v sociálních službách a státní bezpečnosti***. Dále poukazují na to, jak jsou číselníková data řízena na několika úrovních, například státní, krajské a obecní úřady, což vede k nekonzistencím v datech.

Dalším příkladem ***řízení master dat ve státní správě je ruský registr PMSR*** (Public and Municipal Services Register), který byl jedním z digitalizačních projektů ruské administrativy. Autoři Styrin a Zhulin (2011) popisují implementaci MDM systému pro back office registr veřejných služeb. Systém uchovával master data o veřejných službách jako například registrace vozidel, výměna cestovního pasu, či zápis do katastru nemovitostí.

Tang et al. (2021) popisují implementaci ***HR master dat ve veřejných institucích***. Některé další studie se pak ve veřejném sektoru zaměřují na maturitu řízení master dat (Krismawati, Ruldeviyani a Rusli, 2019; Rahman et al., 2019).

V kontextu státní správy České republiky je nutné zmínit ***Informační systém základních registrů*** (dále ISZR), jenž byl zřízen na základě zákona č. 111/2009 Sb. o základních registrech. Čtyři základní registry zřízené zákonem jsou: ***registr obyvatel, osob, územní identifikace, práv a povinností***. Zmíněný informační systém zajišťuje sdílení dat napříč základními registry, sdílí data s agendovými informačními systémy a spravuje přístupy k datům. Registry pak ukládají referenční údaje například o fyzických, právnických osobách (Ministerstvo vnitra, 2022).

[3] Např. profily fyzických a právnických osob

[4] Např. formuláře, online aplikace apod.

25.2.8 Zdroje

ALLEN, Mark a Dalton CERVO, 2015. *Multi-domain Master Data Management: Advanced MDM and Data Governance in Practice*. Waltham, MA (USA): Morgan Kaufmann. ISBN: 978-0-12-800835-5

BERSON, Alex a Larry DUBOV, 2007. *Master data management and data governance*. New York, NY (USA). McGraw-Hill. ISBN: 978-0-07-174458-4

DAHLBERG, Tomi, Jukka HEIKKILÄ a Marikka HEIKKILÄ, 2011. Framework and Research Agenda for Master Data Management in Distributed Environments. In: *Proceedings of IRIS 2011* [online]. Finsko: Turku Centre for Computer Science, s. 82-90 [cit. 2022-05-16]. ISBN: 978-952-12-2648-9. ISSN: 1797-8831. Dostupné z: https://jyx.jyu.fi/handle/123456789/37203

DAMA International, 2017. *DAMA – DMBOK: Data Management Body of Knowledge*. 2nd ed. Basking Ridge, NJ (USA): Technics Publications. ISBN: 9781634622349

HANEEM, Faizura, Nazri KAMA, Azri AZMI, Azizul AZIZAN a Mohd Sam SURIANI, 2015. Prioritization Of Master Data Domains In Public Sector Organisations. In: *ICT for Transformation 2016* [online] [cit. 2022-05-27]. Dostupné z: https://www.researchgate.net/publication/320584614_Prioritization_Of_Master_Data_Domains_In_Public_Sector_Organisations

HANEEM, Faizura, Nazri KAMA a Azri AZMI, 2016. Master Data Identification in Public Sector Organisations. *Advanced Science Letters* [online]. 22, 2999–3003. [cit. 2022-05-27]. DOI: 10.1166/asl.2016.7115. Dostupné také z: https://www.researchgate.net/publication/316599013_Master_data_identification_in_public_sector_organisations doi: 10.1166/asl.2016.7115

HANEEM, Faizura, Nazri KAMA, Nazim TASKIN, David PAULEEN a Nur Azaliah ABU BAKAR, 2019. Determinants of master data management adoption by local government organizations: An empirical study. *International Journal of Information Management* [online]. **45**, 25–43 [cit. 2022-05-24]. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2018.10.007. Dostupné také z: https://www.sciencedirect.com/science/article/abs/pii/S0268401218304468

CHEUNG, Kenneth a Tera CHUNG, 2014. Master data management fundamental and integrated solution. In: *Proceedings of 24th Annual International Conference on Computer Science and Software Engineering* [online]. USA: IBM Corp., s. 317–319 [cit. 2022-06-25]. CASCON '14. DOI: 10.5555/2735522.2735570. Dostupné také z: https://dl-acm-org/doi/abs/10.5555/2735522.2735570

IBRAHIM, Azira, Ibrahim MOHAMED a Nurhizam Safie Mohd SATAR, 2021. Factors Influencing Master Data Quality: A Systematic Review. *International Journal of Advanced Computer Science and Applications* [online]. **12**(2), 181–192 [cit. 2022-05-18]. ISSN: 2158-107X. DOI: 10.14569/IJACSA.2021.0120224. Dostupné také z: https://pdfs.semanticscholar.org/bbc7/122958e2311a62e7d533294651a5d6e81b1a.pdf

JOSHI, Aparna, 2007. MDM governance: A unified team approach. *Cutter IT Journal* [online]. **20**(9), s. 30–35 [cit. 2022-05-15]. ISSN 1522-7383. Dostupné také z: https://www.cutter.com/article/mdm-governance-unified-team-approach-415056

KOKEMÜLLER, J. and WEISBECKER, A., 2009. Master data management: Product and research. In: *Proceedings of the Fourteenth International Conference on Information Quality* [online]. Německo: ICIQ, s. 8-18 [cit. 2022-05-24]. DOI: 10.1.1.462.1292. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.1292&rep=rep1&type=pdf

KRISMAWATI, Dewi., Yova RULDEVIYANI a Rinaldi RUSLI, 2019. Master data management maturity model: A case study at statistics business register in statistics Indonesia. In: *2019 International Conference on Information and Communications Technology, ICOIACT 2019* [online]. s. 931–936 [cit. 2022-05-18]. ISBN: 978-1-72811-655-6. DOI: 10.1109/ICOIACT46704.2019.8938482. Dostupné také z: https://ieeexplore.ieee.org/document/8938482

LOSHIN, David, 2013. Data Governance for Master Data Management and Beyond. *SAS Institute* [online]. Cary, NC (USA): SAS Institute [cit. 2022-05-16]. Dostupné z: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-governance-for-MDM-and-beyond-105979.pdf

LUCAS, Ana, 2010. Corporate Quality Management: From theory to practice. In: *Iberian Conference on Information Systems and Technologies 2010* [online]. Santiago de Compostela (Španělsko): CISTI, s. 524-548 [cit. 2022-05-18]. Dostupné z: https://www.researchgate.net/publication/224169693_Corporate_data_quality_management_From_theory_to_practice

MCGILVRAY, Danette, 2008. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Waltham, MA (USA): Morgan Kaufmann. ISBN: 978-0123743695

MUNDY, Joe, Warren THORNTHWAITE a Ralph KIMBALL. *The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset*. 2nd ed. Indianapolis, IN (USA): Wiley Publishing Inc. ISBN: 978-0-470-64038-8

OTTO, Boris. 2012. How to design the master data architecture: findings from a case study at Bosch. *International Journal of Information Management* [online]. **32**(4), s. 337-346 [cit. 2022-05-28]. ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2011.11.018. Dostupné z: https://www.sciencedirect.com/science/article/abs/pii/S0268401211001435

PANSARA, Ronak, 2021. Master Data Management Challenges. *International Journal of Computer Science and Mobile Computing* [online]. **10**(10), s. 47-49 [cit. 2022-05-26]. ISSN: 2320-088X. DOI: 10.47760/ijcsmc.2021.v10i10.008. Dostupné také z: https://www.academia.edu/60017914/Master_Data_Management_Challenges_

PATRNÝ, Vojtěch, 2011. Master Data Management. *IT systems* [online]. Brno: SystemOnline, 11/2021 [cit. 2022-06-26]. ISSN: 1802-615X. Dostupné z: https://www.systemonline.cz/business-intelligence/master-data-management-1.htm

STYRIN, Evgeny a Andrey ZHULIN, 2011. Public and municipal services register: a back office system for one stop shop e-services portal in Russian federation. In*: Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times* [online]. New York, NY, USA: Association for Computing Machinery, s. 347–348 [cit. 2022-06-25]. dg.o '11. ISBN 978-1-4503-0762-8. DOI: 10.1145/2037556.2037619. Dostupné také z: https://dl.acm.org/doi/abs/10.1145/2037556.2037619

TANG, Li, LIQINGLI, Yang XIE, Jilan ZHANG, Gongliang LI a Shuping LIU, 2021. Construction Scheme of Human Resources Master Data Standard System. In: *Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering* [online]. New York, NY, USA: Association for Computing Machinery, s. 1309–1313 [cit. 2022-06-25]. EITCE 2021. ISBN 978-1-4503-8432-DOI: 10.1145/3501409.3501640. Dostupné také z: https://dl.acm.org/doi/abs/10.1145/3501409.3501640

VILMINKO-HEIKKINEN, Rikka a Samuli PEKKOLA, 2017. Master data management and its organizational implementation: An ethnographical study within the public sector. *Journal of Enterprise Information Management* [online]. **30**(3), 454–475 [cit. 2022-05-18]. ISSN: 1741-0398. DOI: 10.1108/JEIM-07-2015-0070. Dostupné také z: https://www.emerald.com/insight/content/doi/10.1108/JEIM-07-2015-0070/full/html

VILMINKO-HEIKKINEN, Rikka a Samuli PEKKOLA, 2012. Organizational issues in establishing master data management function. In: *Proceedings of the Seventeenth International Conference on Information Quality* [online]. Německo: ICIQ, s. 1-13 [cit. 2022-05-25]. Dostupné z: https://bit.ly/3yk7gOy

YOON, Sang-Pil, Moon-Ho JOO a Hun-Yeong KWON, 2018. Role of law as a guardian of the right to use public sector information. In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* [online]. New York (USA): ACM, s. 1-10. [cit. 2022-06-25]. dg.o '18. ISBN: 978-1-4503-6526-0 DOI: 10.1145/3209281.3209297 Dostupné také z: https://dl.acm.org/doi/abs/10.1145/3209281.3209297

25.3 Data Profiing

Data Profiling je použití technik analýzy dat na existující datový zdroj, jejichž účelem je odhalení jeho skutečného obsahu, struktury a kvality dat, které jsou v něm obsažené. Data Profiling je spojen ***s následujícími vlastnostmi***:

- předpokládá se, že stávající metadata jsou chybná nebo nekompletní a dosažení jejich požadované kvality je cílem Data Profilingu,
- využívají se principy reverzního inženýrství,
- s ohledem na značné objemy dat v datových zdrojích se Data Profiling realizuje na vybraném vzorku dat, na základě předem specifikovaného objemu vzorku dat a na základě určených metod pro výběr dat do vzorku,

- použité analýzy vedou k odhalení typických charakteristik uložených dat, které jsou pak posuzována analytikem s cílem zjištění jejich využitelnosti,
- cílem je i vytvoření znalostní databáze obsahující přesná metadata,
- na základě přesných metadat a business pravidel se identifikují data, která jim nevyhovují.

Postup řešení úloh Data Profilingu na vybraném vzorku dat ***se člení do těchto kroků***:

- analýza vlastností jednotlivých polí (*Column Property Analysis*),
- analýza struktury tabulek (*Structure Analysis*),
- analýza jednoduchých pravidel (*Simple Data Rules Analysis*),
- analýza komplexních pravidel (*Complex Data Rules Analysis*),
- analýza hodnot (*Value Rule Analysis*).

25.3.1 Analýza vlastností jednotlivých polí (Column Property Analysis)

Tento krok ***se realizuje u každého sloupce tabulky*** s cílem specifikovat jeho typické vlastnosti, tj. doménu dat (tedy obor akceptovatelných hodnot ve sloupci nebo skupině sloupců), typický tvar dat ve sloupci, počty prázdných hodnot apod. ***Cílem*** tohoto kroku je:

- vymezení akceptovatelných hodnot v každém sloupci (definování přesných metadat pro sloupce),
- nalezení hodnot, která toto pravidlo porušují.

Specifikace obsahuje znakovou sadu, minimální a maximální počet znaků a povolený rozsah hodnot na základě určeného intervalu, resp. seznamu hodnot nebo znakové šablony.

Obvyklým problémem řešeným v tomto bloku je tzv. nadužívání sloupce, tedy využívání jednoho sloupce pro více účelů. Typickými příklady je obsazení textového sloupce více hodnotami oddělenými oddělovacím znakem, podmíněné použití sloupce závislé na hodnotě jiného sloupce, tj. význam daného sloupce se liší podle hodnoty (kódu) jiného sloupce apod.

Metriky užívané v tomto kroku jsou:

- úplnost = počet nenulových vět / počet všech vět
  - měření je realizované pomocí příkazu SQL a funkce count, nenulové věty jsou v tomto případě ty, které v daném sloupci mají hodnotu rozdílnou od NULL hodnoty,
- doménová konsistence = počet doménově konsistentních vět / celkový počet nenulových vět
  - doménově konsistentní věta obsahuje ve sloupci hodnotu, která je validní pro doménu, v níž sloupce leží.

25.3.2 Analýza struktury tabulek (Structure Analysis)

V rámci tohoto kroku se sledují možná ***strukturní omezení***. Vychází z principu funkční závislosti, tj. kdy právě jeden výskyt hodnot v jedné skupině sloupců odpovídá právě jedné skupině hodnot v druhé skupině sloupců (např. *Jméno* a *Příjmení* pracovníka je funkčně závislé na sloupci *Identifikace pracovníka*.

Cílem analýzy v tomto kroku je omezit počet definujících sloupců na minimum a rozbor vztahů primární klíč – cizí klíč, neboť s funkční závislostí souvisí i pravidla normalizace databází. Výsledkem i tohoto kroku jsou upřesněná metadata a přehled těch záznamů, která pravidla struktury tabulek v metadatech porušují.

Metrikou je:

- referenční integrita = počet vět neodkazujících se na platný primární klíč / počet všech vět
  - má smysl měřit pouze na sloupcích, které jsou nositelem cizího klíče.

25.3.3 Analýza jednoduchých pravidel (Simple Data Rules Analysis)

Analýza ve třetím kroku je založena na ***posuzování vztahů mezi pravidly o datech a business objekty*** (např. objektem zákazník, objednávka, faktura) s cílem zpřesnit dokumentaci pravidel a vytvořit

přehled odchylek. ***Dokumentace business objektu*** musí zahrnovat alespoň jeho název, textový popis jeho významu pro organizaci a seznam pravidel, podle nichž se v organizaci chová.

Pravidlo představuje ***skupinu podmínek*** určující, zda jsou ***data o objektu přesná***. Pro jednoduché pravidlo platí, že pracuje s daty vztahujícími se pouze k jednomu business objektu. Dokumentace pravidla musí zahrnovat název pravidla, název business objektu, textový popis pravidla a formalizovaný popis pravidla, např. v SQL.

Rozlišují se ***tvrdá pravidla*****,** jejichž porušení znamená jednoznačně nepřesnost dat (např. stav zásob na jednom skladu je větší než v celém podniku), a měkká pravidla, která znamenají pouze varování (např. podnik nabízí zboží mimo standardní katalog). Výstupem třetího kroku je zpřesněná dokumentace jednoduchých pravidel, přehled odchylek od tvrdých i měkkých pravidel.

Metrikou je:

- kompatibilita s jednoduchým pravidlem = počet nekompatibilních záznamů / počet všech záznamů

25.3.4 Analýza komplexních pravidel (Complex Data Rules Analysis)

Analýza komplexních pravidel je založena na obdobných principech, jako analýza jednoduchých pravidel s tím rozdílem, že ***komplexní pravidlo se vztahuje na více business objektů najednou***.

Příkladem komplexního pravidla je „jeden produkt nebo služba v podniku může mít v jednom okamžiku pouze jednoho vlastníka“ nebo „počet klientů s dispozičním právem k účtu je vždy nižší než 2“ apod. Výstupem čtvrtého kroku je zpřesněná dokumentace komplexních pravidel, přehled odchylek od tvrdých i měkkých pravidel.

Metrikou je:

- kompatibilita s komplexním pravidlem = počet nekompatibilních záznamů / počet všech záznamů

25.3.5 Analýza hodnot (Value Rule Analysis)

Pravidlo o hodnotách dopředu ***určuje očekávanou hodnotu v analyzovaných, resp. profilovaných datech*** a je obvykle založeno na expertních odhadech. Do těchto pravidel lze např. zařadit pravidla o extrémních hodnotách v datových záznamech, o obvyklém rozdělení hodnot v datovém zdroji (např. rozdělení objemů objednávek) apod. Každé takové pravidlo ***musí být dokumentováno*** názvem a popisem pravidla, způsobem měření hodnot, specifikací očekávaných hodnot profilovaných dat.

Metrikou je:

- kompatibilita s pravidlem o hodnotách = počet nekompatibilních záznamů / počet všech záznamů

25.3.6 Efekty a přínosy

- Ekonomické efekty, např. v úsporách za ztráty při marketingových kampaních,
- Mimoekonomické efekty ve zvýšení celkové úrovně řízení a pověsti firmy vzhledem k zákazníkům,
- Snížení časových ztrát při dohledávání chyb v rámci podnikového reportingu.

25.3.7 Problémy a omezení

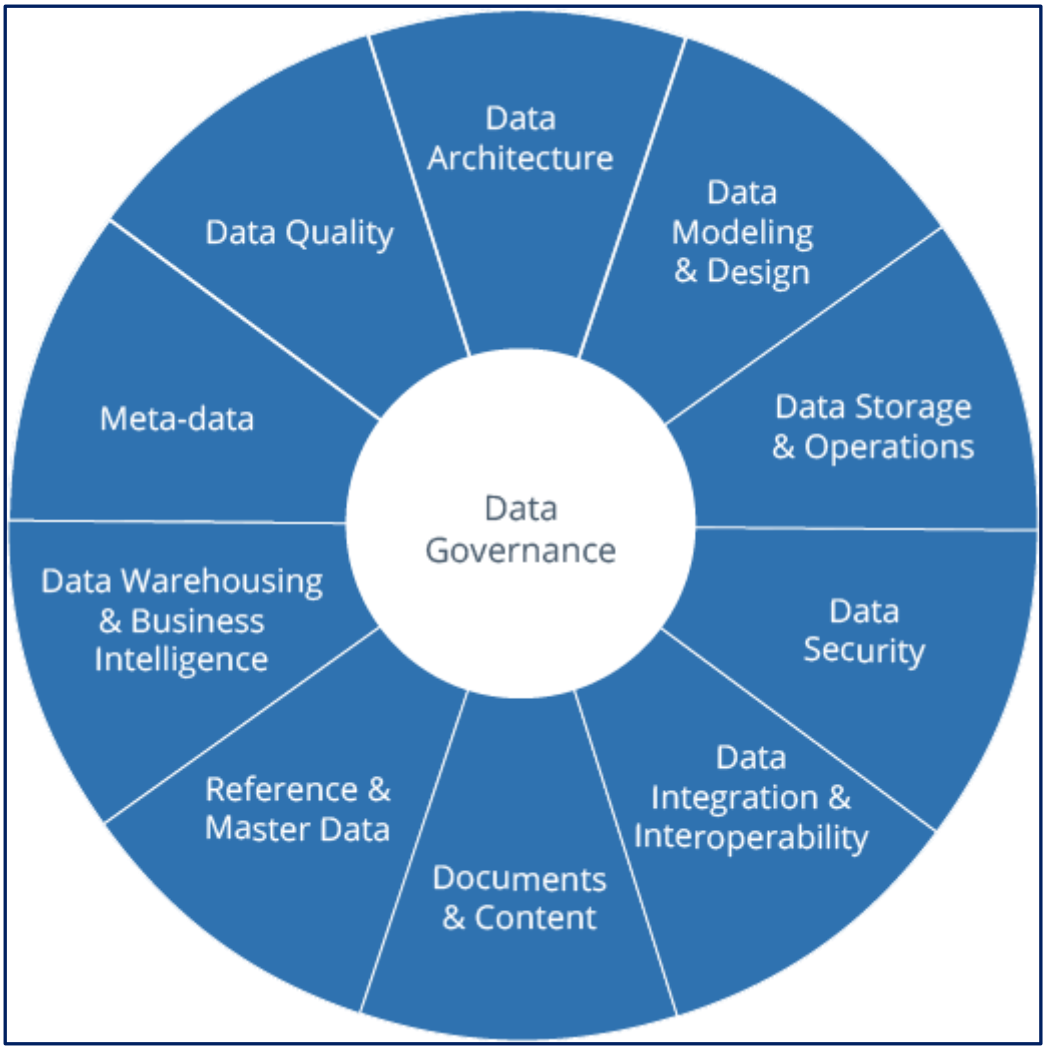
- Ztráty z příjmů za produkty a služby, ve zvýšených nákladech na informační služby i ostatní podnikové aktivity,
- Snížení počtu zákazníků,
- Snížení kvality služeb pro zákazníky, v chybovosti obchodních dokumentů a s tím spojených ekonomických a časových ztrátách,

- Problémy v neefektivitě a neúměrném času běžných řídících a provozních procesů, v nekvalitní vnitrofiremní komunikaci, v nízké kvalitě komunikace s externími partnery a působení na vnější prostředí firmy.

25.4 Data Governance

Společnost ***Profisee*** uvádí k Data governance následující ***definici***: „*Data governance je sadou principů a pokynů zaručujících vysokou kvalitu dat po celou dobu jejich životního cyklu.*“. Data governance je tedy ***programem zaručující patřičnou kvalitu dat*** v celé organizaci. Aby došlo ke skutečnému posunu při práci s daty a jejich ***kvalitou či bezpečností***, tak je důležité, aby jednotliví zaměstnanci dodržovali jasně nastavená pravidla při jakékoliv práci s daty. [Profisee, 2020]

Kromě kvality dat postihuje data governance ještě ***další oblasti spojené s daty***. První touto oblastí je ***bezpečnost***. Dalšími oblasti jsou zaměřené na ***architekturu, integraci*** nebo například na ***datové sklady*** BI.



Obrázek 25-1: Oblasti postihované data governance (Zdroj: JANOSCHEK, 2019)

25.4.1 Úkoly data governance

Na dalším obrázku je uvedeno, kolik oblastí data governance pokrývá. Několik alespoň ***základních bodů spojených a data governance*** je v následujícím přehledu:

- určení ***jednoznačných pravidel*** týkající se metadat, bezpečnosti dat, ale také požadované kvality a přístupu k datům,

- určení ***majitelů a osob zodpovědných za data***. Data jsou využívána na různých úrovních a každé oddělení užívající nebo produkující data by se mělo podílet na jejich správnosti a nést určitou zodpovědnost za tato data. Nemělo by se tedy stát, že majitelem dat bude pouze jedno oddělení,
- ***specifikace kontroly dat***, určení časového úseku, kdy se budou data testovat, zda data odpovídají požadované kvalitě, případně jak budou probíhat změny v datech a zápisy o těchto změnách,
- ***realizace datových auditů*** a zjišťování, na jaké úrovni se data nachází a jak je dodržována nastavená politika.

25.4.2 Zdroje

- Data Governance - What, Why, How, Who & 15 Best Practices. In: *Profisee* [online]. 2020 [cit. 2020-02-25]. Dostupné z: https://profisee.com/data-governance-what-why-how-who/
- JANOSCHEK, Nikolai. Data Governance: Definition, Challenges & Best Practices. In: *BI Survey* [online]. [cit. 2020-03-01]. Dostupné z: https://bi-survey.com/data-governance
- Ševčík, O.: Současné trendy v business intelligence, BP, VŠE, 2020

25.5 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řízení a správě dat v rámci podnikové analytiky.

- Řešení úloh a aplikací podnikové analytiky přináší a současně si vynucuje ***vnesení pořádku do dat***. K tomu slouží i řada nástrojů pro správu dat, které jsou součástí podnikové analytiky a dobře se v ní uplatňují.
- Pro plánování a řešení podnikové analytiky musí být k dispozici ***přehled o existujících datových zdrojích***, jejich dostupnosti, kvalitě, zodpovědnosti za ně atd. To by měla v tomto kontextu zajišťovat správa dat.
- Základním předpokladem pro řešení otázek správy dat jsou ***kvalitní metadata*** a systém nástrojů pro práci s nimi.
- Zvláštní pozornost se věnuje ***master datům***, tj. těm na který stojí základní procesy firmy. Nastavení pravidel pro práci s nimi, řízení zodpovědností za jejich kvalitu apod. obsahuje „***Master Data Management***“.

26. Zajištění datové kvality a čištění dat

Kvalita dat představuje ve většině řešení podnikové analytiky ***klíčový faktor*** ovlivňující jejich úspěšnost či neúspěšnost. To je i jedním z důvodů, proč se tomuto tématu věnuje v současné praxi tak velká pozornost.

Účelem kapitoly je:

- vymezit ***podstatu kvality dat***, tj. její podstatné parametry, zdroje možných problémů a způsob jejich řešení,
- vymezit ***obsah, resp. funkce***, činností spojených s řízením kvality a čištění dat,
- definovat ***podstatné efekty a omezení*** postupů v řízení kvality dat pro nastavení celého systému pro tyto účely.

Jedním z nejčastějších důvodů neúspěchu projektů business inteligence je nízká kvalita dat a zejména produkčních datových zdrojů.

Slovník Kalifornské státní univerzity ***definuje datovou kvalitu*** takto: „Datová kvalita je stupeň dokonalosti dat. Faktory, které přispívají ke kvalitě dat zahrnují (DWGlossary):

- Data jsou uložena podle datových typů.
- Data jsou v konzistentním stavu a nejsou duplicitní.
- Data splňují definovaná pravidla a odpovídají vytvořeným doménám.
- Data jsou aktuální, úplná a srozumitelná.
- Data odpovídají potřebám podniku a uživatelů."

26.1 Podstatné atributy datové kvality

Kvalita dat je vždy dána řadou vlastností, které jsou předmětem i procesu čištění dat. **Vlastnosti kvality** dat jsou podle TDWI (*The Data Warehouse Institute*) souhrnně vyjádřeny v dalším přehledu:

***Správnost*:**

- ***Obsah dat:***
  - přesnost (*Accuracy*) – sleduje, zda všechna data odpovídají realitě a jsou ve správném kontextu (např. zda odpovídá dané PSČ zadané adrese),
  - úplnost (*Completeness*) – určuje, do jaké míry jsou k dispozici všechny informace pro daný kontext, např. existuje správné PSČ, ale celá adresa je neúplná,
  - vyrovnání (*Balancing*) – agregovaná data souhlasí s celkovými hodnotami,
  - konzistence (*Consistency*) – specifikuje možné problémy v porušení standardů nebo vazeb mezi daty. Např. databáze zákazníků v různých lokalitách mají pro stejného zákazníka různá identifikační čísla.
- ***Úroveň detailu:***
  - preciznost (*Precision*) – data jsou zaznamenávána na potřebné úrovni přesnosti,
  - granularita (*Granularity*) – data jsou definována na potřebné úrovni detailu.
- ***Časové charakteristiky dat:***
  - validita (*Currency*) – je odpovídající aktuálnost dat,
  - kontinuita (*Continuity*) – historická data neobsahují mezery nebo překrývání.

Integrita – strukturální správnost:

- pravidla identity (*Identity rules*) - znamenají omezení hodnot primárních klíčů,
- referenční pravidla (*Reference rules*) - vyjadřují referenční integritu, omezení hodnot cizích klíčů,
- základní pravidla (*Cardinal rules*) - jsou omezení řídící počty výskytů dat,

- pravidla pro sady hodnot (*Value set rules*) - řídí přípustné hodnoty pro každý datový element.

Psychologické charakteristiky dat:

- presentace (*Presentation*) – určují, do jaké míry jsou data poskytnuta v použitelné a pochopitelné formě,
- využití (*Application*) ukazují reálné využití dat v praxi.

Další často uváděné charakteristiky kvality dat jsou následující:

- dostupnost (*Availability*) – vyjadřuje možnosti, jak může uživatel přistupovat k informacím v okamžiku, když je právě potřebuje, a zahrnuje:
  - dostupnost v čase,
  - dostupnost v místě, lokalitě, kde mají být informace využity,
  - dostupnost v požadované struktuře,
  - dostupnost v požadovaném formátu,
- srozumitelnost (*Understandability*) - vyjadřuje úroveň pochopení vzhledem ke konkrétní skupině jejich uživatelů,
- relevance (*Relevance*) – udává použitelnost dat v rozhodovacím procesu z hlediska jejich vztahu k objektu, který popisují, tj. jak jsou pro daný objekt významná,
- důvěryhodnost (*Credibility*) – je dána ochotou uživatelů s danými daty pracovat a využívat je.

26.2 Zdroje chyb v datech

Významným aspektem zajištění kvality dat je určení a ***zjišťování možných zdrojů chyb, nepřesností, neúplností a porušených konsistencí*** v datech. Při analýze je proto nutné identifikovat nejen syntaxi dat (tabulky, atributy, klíče), ale i jejich sémantiku (význam jednotlivých atributů, kombinace různých příznaků apod.).

Největším zdrojem chyb a různých poruch v datech jsou obvykle ***transakční aplikace a jejich provoz***, či jejich využití koncovými uživateli. V souvislosti s tím lze uvést následující situace způsobující ***obvyklé chyby v datech***:

- chyby při ***manuálních vstupech*** dat, např. prohození číslic, pravopisné chyby, špatně zadané kódy, hodnoty zapsané do nesprávného pole,
- data pocházejí ***z různých zdrojů*** (různých zdrojových aplikací), s tím jsou spojeny problémy konsolidace dat, tj. sjednocení číselníků, dodržení formálních standardů v datech apod.,
- ***neoprávněné zkoušení*** a testování aplikací uživateli na „ostrých“ datech,
- ***rozdíly v interpretaci*** dat, kdy různí uživatelé chápou sémantiku datové položky různě. To je dáno většinou nepřesným vymezením datových objektů, nedostatky v dokumentaci k jednotlivým datovým polím,
- ***zásahy do zdrojových systémů***, které jsou vesměs i oprávněné, (např. administrátor přidává do struktury nové pole). Problém je ale v komunikaci, kdy to zapomene oznámit správcům ostatních aplikací nebo databází,
- chyby při ***migracích a konverzích*** dat, např. v důsledku špatně interpretované logiky a struktur dat na vstupu a výstupu migračních a konverzních operací,
- chyby přicházející ***z existujících externích datových zdrojů***, obsažené přímo v těchto zdrojích, nebo vznikajících v důsledku špatně nastavených vazeb k externím datům v interních datových zdrojích,
- chyby vznikající v důsledku přímých vstupů do databází ***externími partnery*** (zákazníky, dodavateli atd.) v rámci sdílených nebo otevřených aplikací pro vnější subjekty, například přes webová rozhraní,
- ***chyby v návrhu databází***, zejména nedodržovanou doménovou nebo referenční integritou, kdy jsou povoleny nekorektní vstupy, aniž by to uživatel věděl nebo mu takové operace byly odmítnuty,
- data nebývají ***uložena dle relačních zásad***, neexistují primární klíče, není vynucována povinnost atributů, data nejsou v třetí normální formě, jedna tabulka odpovídá více reálným entitám, atributy nejsou závislé jen na klíči, jeden sloupec v tabulce zahrnuje více atributů (vlastností entity),

- existují ***problémy s konverzí datových typů*** (zejména datum), znakových sad (LATIN2, CP 852, CP 1250, EBCDIC).

Přesné specifikace a analýzy zdrojů chyb jsou velmi důležitým základem pro ***nastavení procesů řízení kvality***, určování místa kontrol a řešení organizačních i technologických opatření. Procesy řízení kvality dat musí definovat ***obsahovou stránku kontrol*** a transformačních operací s daty, ale i ***organizační aspekty***, tj. kde, kdo a kdy má data kontrolovat a případně spouštět další navazující procesy.

Ponechání kontrolních a opravných operací až na vlastní datový sklad, ETL, ne-li dokonce na uživatelské aplikace přináší s sebou mnoho negativních dopadů. Chyby či neúplnost dat, které jsou v transakčních operacích "pouze" nepříjemné, mohou však ***na úrovni podnikové analytiky*** mít zcela ***zásadní význam*** (např. díky kumulacím chyb). Uživatelé pak logicky k takovým výstupům a aplikacím ztrácejí důvěru a hledají jiné zdroje, často i za cenu vyšší pracnosti. Kromě toho, pokud chyby se kontrolují a opravují až např. ***v BI vrstvách, neodstraní se tím často jejich zdroj***, tzn. transakční aplikace, provoz a kvalita produkčních databází. To znamená, chyby se budou i nadále opakovat. Řešíme-li otázky řízení BI, pak to není jenom otázka BI, ale řízení informatiky jako celku. Potřeba řízení kvality dat na úrovni transakčních aplikací a produkčních databází to jasně dokumentuje.

26.3 Dopady nekvality dat

- ***Dodatečné náklady*** – podle Data Warehousing Institutu (TDWI) je to např. pouze z dodatečných nákladů na tisk, poštovné a práci zaměstnanců.
- ***Ztráty času*** – dohledávání správných dat a následné opravy jsou obvykle velmi časově náročné.
- ***Problémy s rozhodováním*** – manažeři nemohou činit správná rozhodnutí, protože přesně neví, co se v jejich společnosti či na trhu odehrává, musí pak spoléhat na intuici, což je při dnešních velmi rychlých tržních změnách značně nebezpečné.
- ***Ztráta důvěry***, a to:
  - ***Ztráta důvěry zaměstnanců v IS*** např. ve zvýšení zaměstnanecké nespokojenosti nebo i v to, že pracovníci firmy nebudou chtít IS používat.
  - ***Ztráta důvěry zákazníků*** či obchodních partnerů ve firmu, např. poškození dobrého jména společnosti, negativní publicita, přímé ztráty zákaznéků.

26.4 Principy řešení kvality dat

Do řešení kvality je nutné zařadit celý komplex činností, procesů a metod, které se v rámci systému řízení kvality dat realizují. ***Systém řízení kvality dat*** (*DQMS, Data Management Quality System*) představuje zahrnuje tyto hlavní funkce:

- opravu chyb a jiných nedostatků v datech hlášených organizaci zákazníkem a jinými subjekty,
- periodické čištění datových zdrojů a databází,
- čištění aktuálně používaných dat,
- prevenci chyb.

Řešení otázek zajištění kvality dat lze rozdělit do ***dvou základních rovin***:

- ***analytické***, tj. v rovině analýzy datových zdrojů zjišťující zdroje a příčiny chyb, strukturu a četnost chyb a definující nároky na organizační i technologická řešení a obsahuje:
  - měření kvality dat,
  - detailní analýzu dat (Data Profiling),
- ***návrhové a implementační***, která se promítá do návrhu kontrolních a opravných procedur a implementace procesů ETL a zahrnuje:
  - správný návrh datových struktur, normalizaci záznamů dat,
  - odstranění duplicitních záznamů a porovnání záznamů dat.

Analýza datových zdrojů probíhá v rámci řízení IT v několika oblastech. Jednak tvoří součást průběžného řízení datových zdrojů na úrovni celého informačního systému. Analýza datových zdrojů pak musí proběhnout i při formulaci strategie a plánování projektů BI, jako součást nezbytné analýzy stavu celého informačního systému a jeho připravenosti na tyto aplikace.

Souhrnně lze definovat následující ***principy řešení a zajištění kvality dat***:

- řešit datovou kvalitu jako ***záležitost celého podniku***, nejen IT útvarů, tj. v rámci všech podnikových funkcí a oblastí řízení,
- definovat ***odpovědnosti za kvalitu*** dat u jednotlivých manažerů a zajistit i jejich potřebnou motivaci,
- řešit ***změny v definování podnikových procesů*** s ohledem na kvalitu dat,
- pokud možno zajistit ***automatické kontroly a opravy***,
- jasně ***označovat zjištěné chyby*** a, pokud to lze, nahradit je default hodnotou a důsledně protokolovat, co a kde bylo opraveno (např. primární klíč, původní chybná i nová hodnota apod.),
- pokud je to možné, nemazat záznamy v databázi,
- v okamžiku výskytu chyb ***zaslat co nejdříve chyby vlastníkům*** dat podle priority chyby (SMS nebo mail) na základě definovaných pravidel, kdo může opravit data
- definovat ***specifické testy*** pro určité typy úloh s tím, že každý test má svou prioritu závažnosti možné chyby,
- ***logovat průběh*** testů do databáze a presentovat reporty aktuálního stavu systému na www,
- je účelné, zejména u větších podniků, stanovit ***specialisty a speciální týmy pro datovou kvalitu*** (data stewards).

Jednou ***z opravných operací je standardizace***, která zajistí jednotnou reprezentaci informací pro další zpracování. Po prvotním manuálním sjednocení dat může tento proces probíhat již automaticky a zahrnuje, např.:

- ***transformace*** datových položek podle předdefinovaných pravidel,
- ***formátování*** položek do konzistentního tvaru,
- ***normalizaci*** datových elementů, s ohledem na jazykové a národnostní specifikace a další.

26.5 Efekty a přínosy

- Dosažení ***ekonomických efektů***, např. v úsporách za ztráty při marketingových kampaních,
- Ve zvýšení ***důvěryhodnosti*** firmy vzhledem k zákazníkům,
- Snížení ***časových ztrát a nákladů*** při dohledávání chyb v rámci podnikového reportingu.

26.6 Problémy a omezení

- Problémy ***ve ztrátách z příjmů*** za produkty a služby, ve zvýšených nákladech na informační služby i ostatní podnikové aktivity,
- Problémy ***ve ztrátách zákazníků***, ve snížení kvality služeb pro zákazníky, v chybovosti obchodních dokumentů a s tím spojených ekonomických a časových ztrátách,
- Problémy ***v neefektivitě a neúměrném času*** běžných řídících a provozních procesů, v nekvalitní vnitrofiremní komunikaci, v nízké kvalitě komunikace s externími partnery a působení na vnější prostředí firmy

26.7 Zdroje

DRAŽIL, M.: Datová kvalita, integrita a konsolidace dat v BI, DP, VŠE, 2009.

POTANČOK, M., POUR, J., CHRAMOSTOVÁ, V. Podniková analytika pro manažery, Oeconomia, Praha, 2021,

POUR, J., STANOVSKÁ, I., NOVOTNÝ, O., SLÁNSKÝ, D.: IT a anatomie firmy – principy. Praha, Professional Publishing, 2021. ISBN 978-80-88260-56-1.

POUR, J., MARYŠKA, M., STANOVSKÁ, I., ŠEDIVÁ, Z.: Self Service Business Intelligence, Praha. Grada, 2018. ISBN 978-80-271-0616-5

POUR, J., NOVOTNÝ, O., MARYŠKA, M.: Business Intelligence v podnikové praxi. Praha, Professional Publishing, 2012. ISBN 978-80-7431-065-2

26.8 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k zajištění kvality a čištění dat v rámci podnikové analytiky.

- Řízení kvality dat představuje do jisté míry speciální disciplínu, ale současně i část ***řízení datových zdrojů*** v rámci řízení IT, resp. Data Governance.
- Řešení kvality dat a jejich čištění je rovněž ***součástí datových transformací*** v rámci podnikové analytiky a jejich prostředků (DSA, ETL apod.). Analýza a návrhy datových transformací tak musí otázky řešení kvality dat také zahrnovat.
- Řešení kvality dat je v mnohých firmách zásadní otázkou, a proto se řeší i vznikem ***specializovaných týmů nebo útvarů*** a specializovaných rolí.

H) Big Data a podniková analytika

[27] Big Data – principy	
[28] Big Data Analytics	[29] Data Driven Company

„Big Data“ je výrazný fenomén současné doby, který je dán extrémním nárůstem dat v souvislosti s uplatňováním IoT, zejména IIoT, sociálními sítěmi a dalším aspekty dnešní informatiky. S Big Data souvisí i uplatňování principů podnikové analytiky na těchto objemech dat. Oddíl tak obsahuje tyto body:

- ***Principy Big Data***, tj. jejich vymezení na základě různých zdrojů, architektury Big Data, technologie a technologické přístupy k jejich řešení a další aspekty.
- ***Uplatnění analytiky nad Big Data***, resp. „***Big Data Analytics***“ obsahující systém doporučení a dokumentující různé možnosti a oblasti uplatnění těchto úloh.
- Koncept označovaný jako “***Data Driven Company***“ zahrnuje jeho vymezení a zejména specifikaci její organizace a charakteristiky jednotlivých specifických pozic nebo rolí ve firmě.

27. Big Data - principy

Halama, J., Miloš, M. (KIT, VŠE), upraveno

Účelem kapitoly je:

- vymezit ***podstatu***, velkých dat, „*Big Data*“ a jejich nejvýznamnější charakteristiky,
- definovat ***podstatné efekty a omezení*** systémů pracujících v prostředí Big Data,
- definovat ***hlavní zdroje dat*** pro systémy Big Data,
- charakterizovat obvyklé ***architektury*** aplikované v prostředí Big Data, se zaměřením na systém Hadoop,
- definovat charakteristiky vybraných ***dotazovacích nástrojů***.

Mezi základní faktory změny přístupu ke sběru, správě a zpracování dat patří:

- ***Zvyšující se objem dat*****:** Je to např. digitalizace výrobních procesů, řízení vztahů se zákazníky nebo sběr dat ze sociálních sítí. Takový objem dat ***roste úměrně s velikostí společnosti a počtem zákazníků***. Globálně se pak o „datovou explozi“ starají hlavně mobilní telefony a zařízení IoT. Podle společnosti IDC (International Data Corporation) jsou dnes tato zařízení schopna generovat objem dat v řádech zettabytů (1 zettabyt = 1 000 000 000 terabytů).
- ***Cloudové služby*****:** S příchodem cloudových řešení mají společnosti možnost ***správu infrastruktury přesunout na poskytovatele*** Cloudových služeb (Google, Amazon, Microsoft aj.). Díky tomu společnost může manipulovat s daleko vyšším objemem dat, než tomu bylo dříve.
- ***Změna očekávání*****:** S příchodem ML (strojového učení) a AI (umělé inteligence) společnosti a jiné subjekty od dat očekávají ***možnosti prediktivní analýzy***. Ta dokáže na základě dat určit budoucí vývoj např. podnikových ukazatelů. Predikce se užívají např. v marketingu, kde modely strojového učení dokáží s vysokou přesností provádět segmentaci zákazníků nebo předpovídat nákupní chování zákazníků. K vytvoření takovýchto modelů je třeba vlastnit velký objem relevantních a označených dat.

Za autora původní definice Big Data se považuje Doug Laney, který ve svém reportu „*3D Data Management: Controlling Data Volume, Velocity and Variety“* z roku 2001 popsal tři základní V. Mezi základní definované V's však patří *Volume, Velocity a Variety*.

- ***Volume, objem či množství***. Volume popisuje to, že objem dat je velký, ***vetší než standartně zpracovávané záznamy***. Exaktní objem dat však nikdo nikde nepopsal. Někteří popisují tento objem v terabytech či petabytech v kontextu společností.
- ***Velocity, rychlost***. Vyjadřuje charakteristiku dat z pohledu rychlosti zpracovávání, streamování a ukládání dat.
- ***Variety*****,** ***rozmanitost***. Data, která zpracováváme a uchováváme, jsou rozmanitá a nemají stejný formát ani charakter. Společnosti získávají cenné informace z nejrůznějších dat z webových stránek, sociálních sítí, backendových systémů ale i obrázků či emailů.

“Big data is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making, and process automation.“ (Gartner, 2021)

Společnost ***IBM*** v roce 2016 model rozšířila o jedno V – ***Veracity***. V tomto reportu však zmiňuje i páté potenciální V – **Value**. (OpenSistemas, 2016)

- ***Veracity*****,** ***věrohodnost***. Jde o věrohodnost dat, která se mají využívat k analýze či rozhodování. Data mohou obsahovat různé záznamy generované chybou, duplicity či nekonzistenci v datech (chybějící či irelevantní údaje).
- ***Value*****,** ***hodnota***. IBM však toto V označuje jako potenciální. Hodnota dat není vždy zaručena. Společnost musí data uchovávat, zpracovávat a řídit jejich kvalitu, aby mohla dosáhnout požadovaných výsledků – hodnoty.

27.1 Efekty a přínosy

- ***Otevření dosud nevyužitých informačních kanálů,*** sociální sítě, multimediální data či senzory výrobních zařízení, které mohou v kombinaci s původními datovými zdroji poskytovat prostor pro analýzy dosud nedosahovaného rozsahu,
- Využití tzv. ***Dark Data*****,** tedy dat, která společnost shromažďovala, ale bylo nákladově neefektivní tato data analyzovat, a které na základě současných technologií mohou přinést transparentní informace využít je při rozhodovacích procesech,
- Uložená transakční data obsahují ***přesné a detailní informace o výkonnosti*** podnikových procesů. Tyto informace mohou sloužit pro zlepšování výkonnosti procesů a podávají více informací managementu při rozhodování za nejistoty. Toto platí pro všechny stupně managementu: strategický, taktický, operativní,
- ***Optimalizace výkonnosti organizace,*** ve vztahu k podnikovým procesům lze lépe monitorovat vliv změn na výkonnost, zjišťovat příčiny výkyvů ve výkonnosti a efektivně rozdělovat podnikové zdroje,
- ***Transparentnost,*** snadnější a rychlejší přístup k relevantním datům, dostupným napříč odloučenými odděleními,
- ***Využití analytických nástrojů*****,** mohou výrazně zlepšit proces rozhodování, minimalizovat riziko nebo odhalit cenné informace v rozsáhlých objemech dat, které by jinak zůstaly skryté,
- ***Úzká segmentace zákazníků*** a vytvoření produktu, služby přesně podle jejich požadavků,
- ***Inovativní postupy pro modelování experimentů*** s možností testovat hypotézy a analyzovat jejich výsledky pro důležitá investiční rozhodnutí, komplexní podpora vývoje nových produktů a služeb, jejichž vlastnosti a funkce pomáhá přesněji definovat analýza velkých objemů různorodých dat.

27.2 Problémy a omezení

- Jestliže společnost ***zanedbala správu dat*** v rámci databází a BI výstupy tak byly nekvalitní, Big Data není okamžitým a spásným řešením,
- Big Data nesou s sebou nové a podstatně ***vyšší nároky na kvalifikaci*** odborníků, a to jak na straně managementu nebo business analytiků (ti, co budou klást otázky hypothesis-driven), tak na straně lidí, kteří budou data zkoumat,
- Data i ve velkých objemech a různorodosti dostávají smysl až ve chvíli, kdy jsou ***konfrontovány s nějakým modelem nebo hypotézou***. Pokud hypotéza nebo model nejsou relevantní (business požadavky jsou špatně interpretovány), potom Big Data nepřináší přidanou hodnotu, naopak, spotřebovávají čas, prostředky a náklady,
- Nutnost ***využití clusterů*** (množin zařízení) ke správě a zpracování dat, což může (ale ne nutně) znamenat vyšší pořizovací náklady na HW (platí jen v případě on-premises řešení),
- Big Data, která jsou získávána prostřednictvím sociálních sítí, vyhledávání atd., mohou přinést cenné informace jedině za předpokladu ***reálně nastavených hypotéz*** (hypothesis-driven analytics) a k nim relevantních dat,
- K problémům Big Data patří především ***hrozby související se soukromím, bezpečností a vlastnictvím*** dat,
- Big Data ***není náhradou relačních databází,*** spíše se odhaduje, že technologie budou jedna od druhé přebírat best-practices až se nakonec sloučí.

27.3 Technologické definice

Dvě definice, které původní definice doplňují o důležité aspekty podnikového vnímání Big Data:

„If data is varied in its structure, and its size, timing, processing needs and scaling requirements have expanded to such an extent that a single computing

system is unable to meet those demands, then this data can be called Big Data."
(Goswami, Dāša and Mukherjee, 2019)

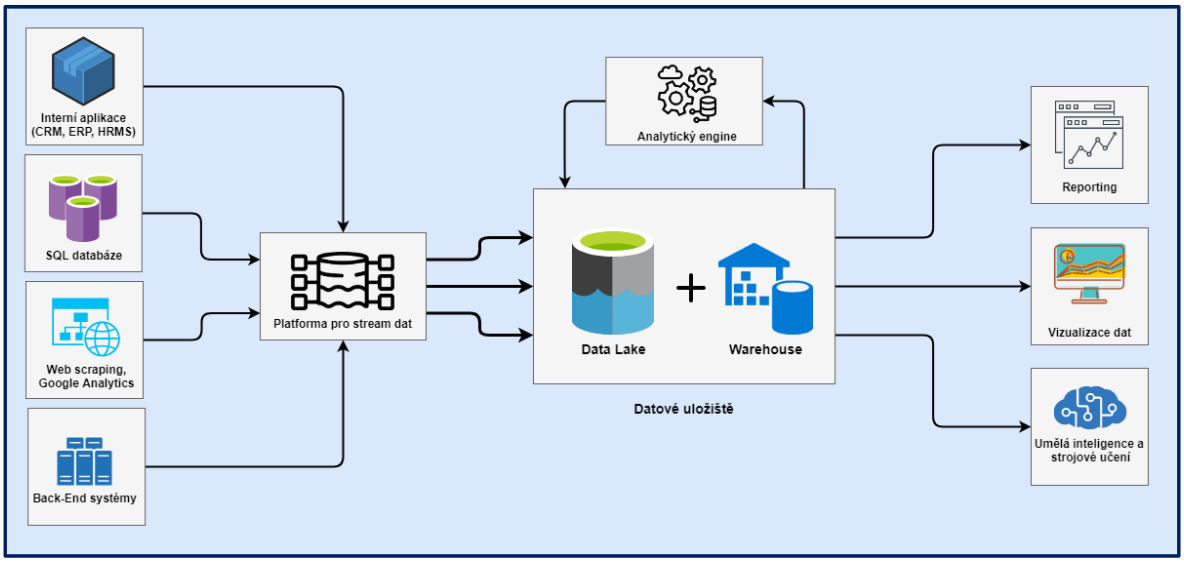
Zásadní ***doplnění*** se skrývá v druhé polovině definice. Pokud data nejsou zpracovatelná jedním výpočetním systémem, označujeme je, při splnění i předešlých podmínek, za Big Data. Autor zde zmiňuje často užívaný koncept ***DFS (Distributed File System)*** a výpočetního clusteringu počítačů či serverů.

"Big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before." **(Oracle, 2021)**

Ve druhé části definice Oracle popisuje že tyto ***datasety jsou tak objemné***, že je ***tradiční software nedokáže zpracovat***

27.4 Big Data jako proces

Mnoho společností vysvětlují Big Data jako kontinuální proces sběru charakteristicky velkých a komplexních dat, jejich transformace a analýzy. Každá společnost tento proces upravuje dle své infrastruktury, potřeb a hlavně poskytovatele, který upravuje rámec nabízených služeb (Obrázek 27-1).



Obrázek 27-1: Příklad Big Data jako proces. (Zdroj: Halama, 2021)

Proces kopíruje tok dat infrastrukturou a úkony s nimi. Začíná sběrem dat z ***rozmanitých*** zdrojů. Zdrojem může být téměř cokoli, co ve společnosti generuje data. V uvedených příkladech jsou zejména interní aplikace, ostatní SQL databáze, databáze 3. stran (Google Cloud Platform, Amazon AWS Cloud Storage) nebo zprávy z interních back-end systémů.

Tento ***sběr je realizován pomocí platforem pro stream dat***. Ty mají za úkol dopravit data ze zdroje do datového uložiště. V mnoha společnostech je potřeba tato data monitorovat v **reálném čase**, proto se využívá vysoce komplexních a výkonných řešení. Tato řešení také musí obsahovat ochranu před ztrátou dat např. z důvodu náhlé nedostupnosti datového uložiště.

Architektura datového uložiště je vždy upravena dle potřeb společnosti a ***objemu*** dat. Data mohou být uložena ***v interních*** datových uložištích nebo ***na preferovaném cloudovém uložišti***. Užívá se zde ***konceptu DFS (Distributed File System)***. To znamená, že data jsou rozprostřena mezi několik datových uložišť, které se uživateli při dotazování jeví jako jedno. Architektura realizovaná na obrázku (Obrázek 27-1) rozděluje datové uložiště na ***Data Lake*** a ***přidružený Data Warehouse*** k analytickým účelům. V konkrétních řešeních může být celé datové uložiště realizováno ***jedním Data Lakem***, který je spravován jedním z mnoha ***Data Warehousing systémů*** a dokáže v uložišti připravit potřebnou strukturu a přístup k jednotlivým Data Warehouse a Data Martům. Na druhou stranu může být realizován ***několika uložišti propojenými mezi sebou ETL*** nástroji. Jak již bylo zmíněno, architektura je silně individuální dle potřeb společnosti.

Aby bylo možné s daty v datovém uložišti manipulovat i jinak než prostým dotazem, je nad datové uložiště stavěn ***analytický engine***. Ten zajišťuje rychlé dotazování, transformaci a ukládání dat v uložišti.

Využívá výpočetní síly ***clusteru*** a dokáže alokovat prostředky mezi všemi uživateli, kteří pracují s datasety v uložišti.

Poslední částí procesu jsou již ***finální výstupy***. Reporting může být realizován ukládáním agregovaných dat do příslušných databází v uložišti. ***Vizualizace*** slouží k lepší ***interpretaci dat*** a jsou realizovány připojením dat k nástrojům vizualizace (PowerBI, Grafana, Kibana).

Poslední částí je využití dat pro potřeby ***strojového učení a umělé inteligence***. To z pravidla probíhá ***na oddělených platformách***, které jsou orientovány na rychlé zpracování dat, ne na jejich transformaci.

27.5 Zdroje pro Big Data

Podstata a potřeba řešení Big Data tak vychází zejména ***z nových zdrojů a typů dat*****,** které jsou často automaticky ***generovány a vytvářeny v obrovských objemech*** nejrůznějšími monitorovacími zařízeními nebo mimořádně velkými počty uživatelů jako je tomu u sociálních sítí.

Konzultantská společnost ***CloudMojo*** uvádí že existují ***tři primární zdroje Big Data dat***. Mezi ně patří ***sociální data***, ***data generované stroji a transakční data***. Dále existují i ***hybridní zdroje*** dat, které ***pracují s metadaty*** kvůli interpretaci a tagování. Společnosti užívají takové zdroje, které jsou typické pro její produkty či business model (CloudMojo, 2020)

27.5.1 Sociální data

Za sociální data považujeme taková data, která jsou ***generována lidskou působností***, ať už vědomě či nevědomě. V informační společnosti člověk vytváří data ***jakoukoliv interakcí s nejrůznější elektronikou***, sociálními sítěmi, pořizováním digitálních záznamů či procházením webových stránek. Mezi ***hlavní zdroje*** patří právě sociální sítě a webové stránky. U nich můžeme využívat ***technologii scrapingu*** webových stránek nebo sociálních sítí.

Společnosti mohou, za pomoci speciálních ***nástrojů a scriptů, extrahovat veřejná textová data*** za účelem analýzy a vytváření znalostních grafů. ***U sociálních sítí*** s příchodem umělé inteligence se výzkum zaměřuje na ***sémantickou analýzu textů***. Tyto postupy se dají využít např. u zpětné vazby zákazníků, kde umělá inteligence rozezná obsah sdělení a její emoční zabarvení.

Druhou velkou částí zapadající do sociálních dat je ***sledování chování uživatelů na webových stránkách***. Společnosti zaznamenávají každý klik na dané webové stránce a díky nim lze získat informace, čeho si návštěvník všimne jako prvního, zda reaguje na reklamní sdělení či jak dlouho se na webu zdrží. Analogicky to funguje i ***s mobilními aplikacemi***. Na principu sběru sociálních dat funguje i nástroj ***Google Trends***.

U sociálních dat se však společnosti potýkají s obecným ***nařízením o ochraně osobních údajů (ang. GDPR, General Data Protection Regulation***). Toto nařízení bylo vyhlášeno v roce 2016 na území EU a zabývá se ochranou fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů. V České republice je GDPR doplněno ***zákonem O zpracování osobních údajů***. Ten upravuje práva a povinnosti při zpracování osobních údajů osob. Pokud je při sběru dat sbírán i *osobní identifikátor*, je třeba ho ***bezpečně*** anonymizovat či hashovat. Pokud by společnost chtěla osobní údaje i přes to uchovávat (banka, pojišťovna), je ***třeba souhlasu zákazníka*** a data uchovávat tak, aby se předcházelo úniku dat. Taková data se však nesmí užívat k cílenému marketingu ani prodeji.

27.5.2 Data generované stroji

Data generovaná stroji se v dnešní době vážou hlavně ***k termínům Industry 4.0 a Internet of Things (IoT)***. Očekává se, že objem těchto dat poroste exponenciálně v závislosti na rozšíření zařízení IoT. Do této kategorie spadá však ***širší spektrum dat***. Ke sběru se nejčastěji ***užívají senzory***, a to je v tomto případě obecné označení pro zařízení, které ***sbírá data pro řídící systém***. Řídícím systémem může být člověk, či stroj v podobě expertního systému a umělé inteligence.

Příkladem může být ***letecká doprava*** a dopad na ni. ***Autopilot*** pracuje s daty generovanými systémy a vnějšími senzory letadla. Na základě nich dokáže doporučit budoucí kroky pilotům. Řídící systém letadla data sbírá po celou dobu letu a za jeden let může vyprodukovat okolo 200 GB dat. Tato data se pak dají použít i k analýze opotřebení jednotlivých součástek a k prediktivní údržbě letadel.

27.5.3 Transakční data

Transakční data jsou ***generována každodenními transakcemi v obchodech a na internetu***. Taková data vypovídají o peněžních tocích jednotlivých společností, obchodů, jednotlivců. V době pandemie Covid-19 se drtivá většina transakcí přesunula na internet a je to jeden z důvodů nárůstu objemu těchto transakčních dat.

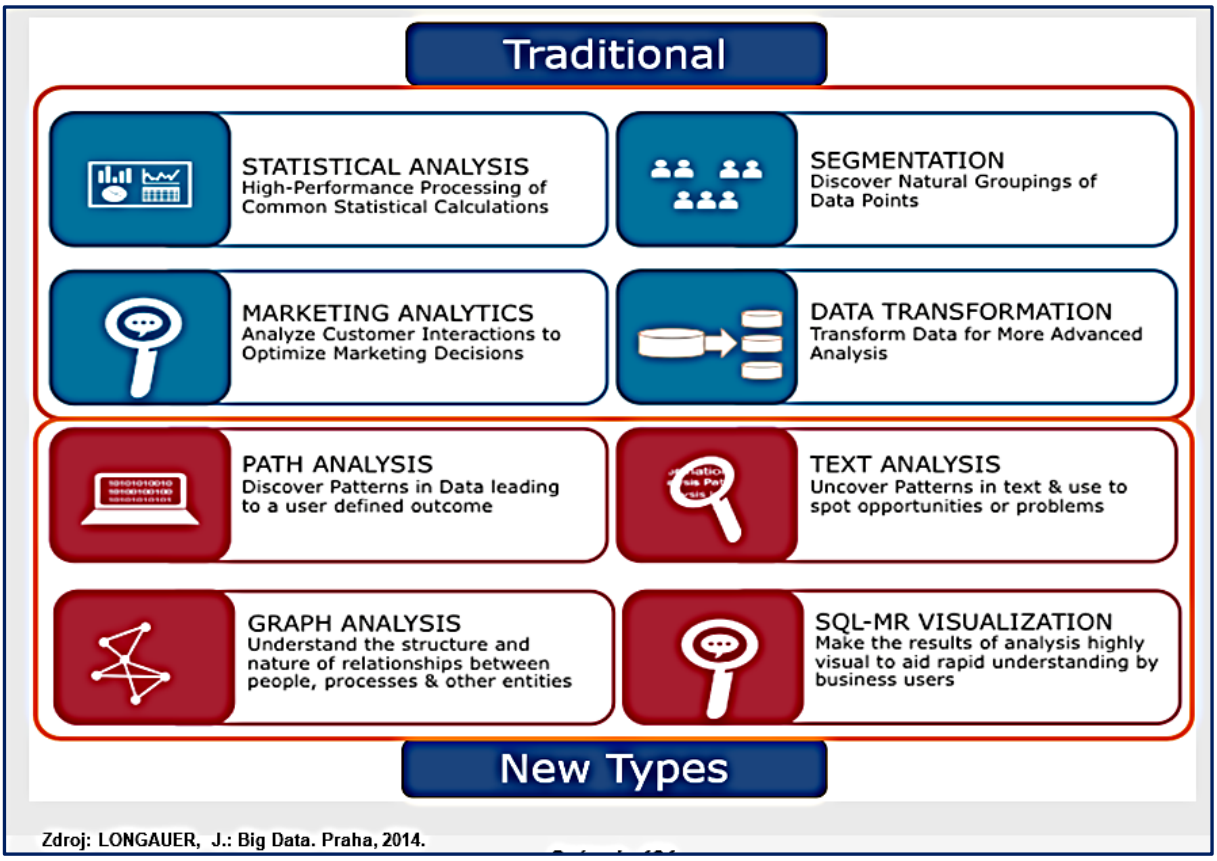
27.5.4 Hybridní data

Některá data mohou spadat do hybridní kategorie, protože ***spojují dvě a více oblastí***. Jsou to taková data, které je třeba ***interpretovat pomocí metadat z jiné kategorie***. Příkladem může být oblast ***Health Care*** a v ní data generovaná přístroji jako rentgen, magnetická rezonance nebo ultrazvuk. Snímky z těchto zařízení by byly považovány jako data generovaná stroji. Pokud se k nim přidá doménová znalost lékařů a diagnóza pacienta, jedná se již o hybridní datový zdroj. Takové data patří k nejcitlivějším osobním údajům a přístup k nim mají jenom určené osoby.

Příkladem z ***bankovnictví*** mohou být ***bankovní transakce***, které budeme používat ***ke skóringu zákazníka***, který si chce zažádat o hypotéku. Samotné transakce bez metadat v podobě výpisu z rejstříku dlužníků, pracovní pozice a manželského stavu, budou spadat pouze do transakčních dat. Po obohacení těchto dat sociálními metadaty dostáváme hybridní dataset připravený k aplikaci skóringových algoritmů.

27.6 Principy řešení Big Data

Big Data řešení jsou vystavěna ***na principu paralelního zpracování dat*** (ukládání, zpracování a analýza na větším počtu zařízení v rámci clusteru) právě z důvodu zvýšení rychlosti zpracování velkého objemu různorodých denormalizovaných dat. Nové druhy dat je ***potřeba efektivně skladovat a pracovat s nimi***, vyžadují (a umožňují) nové druhy analýz a poskytují zcela nové možnosti analýzy dat, jako např. grafové analýzy, analýzy textu či analýzy cest.



Obrázek 27-2: Nové typy analýz dat v souvislosti s Big Data

Smyslem řešení založených na Big Data je efektivně nacházet ***potřebné informace ve velkém objemu různorodých dat***. S tím ovšem souvisí významný ***rozpor***, a to jak objem a různorodost dat rychle narůstá, jejich hodnota se s časem snižuje. Tedy vyvolává se stále silnější potřeba ***zpracovávat tyto obrovské a různorodé objemy dat v co nejkratším čase.***

Přesto je potřeba s těmito daty ***efektivně pracovat a analyzovat je***. To na jedné straně vedlo **ke *vzniku specializovaných řešení pro analýzy*** nových druhů dat (např. Teradata Aster), ale i ***ke vzniku a zlepšování nových způsobů ukládání těchto dat – NoSQL databází. Velká data a NoSQL*** spolu ***úzce souvisejí. NoSQL jsou lépe připraveny*** pro práci s nestrukturovanými daty než relační databáze.

Analytické aplikace zaměřené na velká data obsahují zase ***potřebné nástroje pro efektivní analýzu*** dat. Zpočátku tato řešení nebyla vhodná pro širší použití, jelikož neexistovaly snadno použitelné nástroje, které by umožňovaly efektivní práci s nimi. Lidé znalí obchodní problematiky neměli dostatek znalostí pro psaní vlastních dotazů do NoSQL databází, a naopak technicky zaměřeným lidem chyběly znalosti obchodní problematiky. Tyto nedostatky ale v posledních letech mizí.

27.7 Architektura Big Data

Big Data a business intelligence architektury jsou si na první pohled velmi podobné. Zásadní ***rozdíly jsou zejména v charakteru zpracovávaných dat*** s důrazem na jejich transformaci a distribuované zpracování. ***Architektura Big Data se skládá ze čtyř vrstev*** (Obrázek 27-3), podobných vrstvám u business intelligence. Patří sem:

- Vrstva datových zdrojů
- Vrstva ETL
- Vrstva datového zpracování
- Vrstva pro analýzu dat

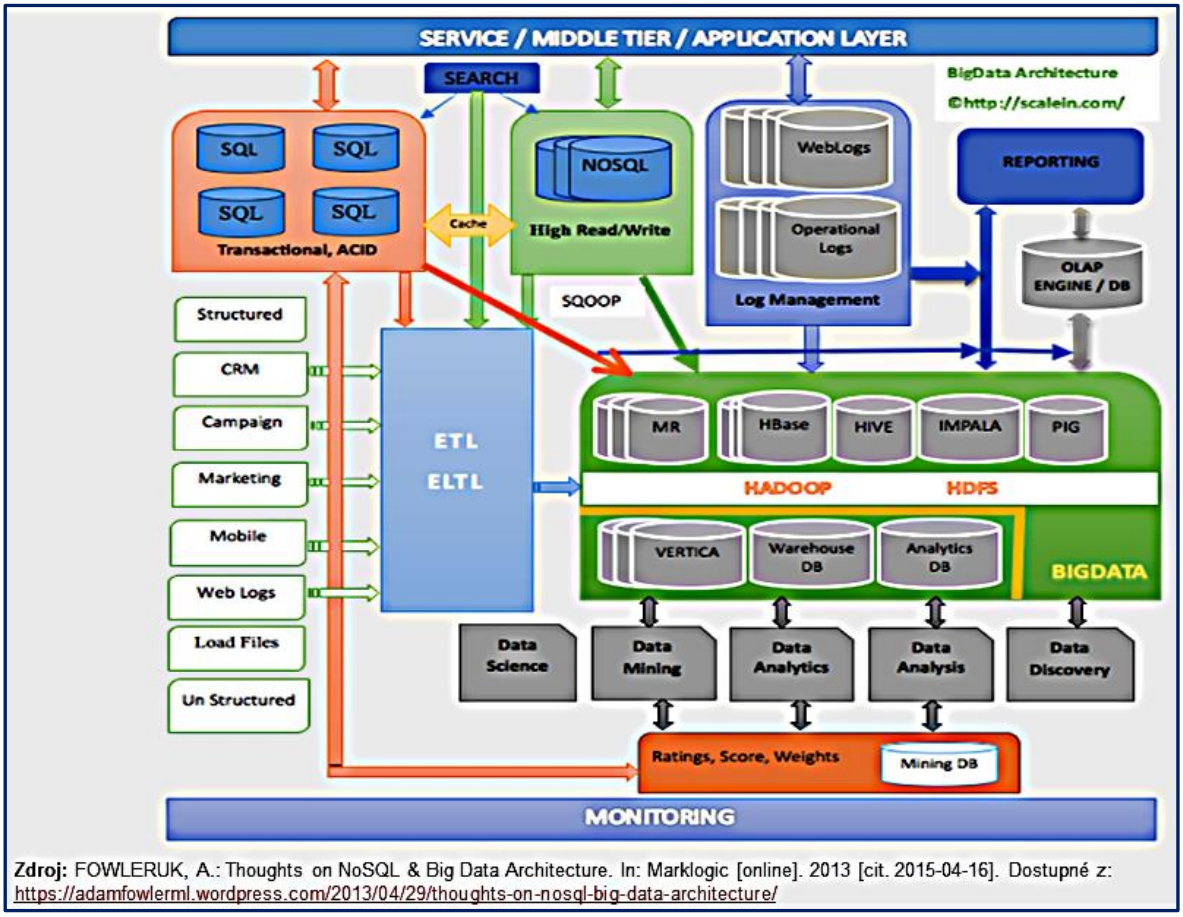
27.7.1 Vrstva datových zdrojů

V případě Big Data pochází ***data z několika heterogenních datových zdrojů***, které lze rozdělit ***do třech skupin***. ***Prvním zdrojem*** jsou převážně ***strukturovaná data*** z relačních databází (např. MySQL, Oracle) nebo NoSQL5 databází (např. MongoDB). ***NoSQL databáze*** jsou škálovatelné distribuované databáze optimalizované pro rychlé zpracovávání rozsáhlých objemů dat, jejichž charakter nevyžaduje relační model (např. párování klíče a hodnoty) a používají se zejména k analýzám v reálném čase (např. analýza sociálních médií) nebo statistickým analýzám velkých objemů dat.

Druhým zdrojem jsou ***částečně strukturovaná a nestrukturovaná*** data (CRM, marketing, kampaně).

Třetím skupinou datových zdrojů jsou automaticky generované webové logy nebo jiné logové soubory (historie požadavků, uživatelská aktivita, uživatelské kliky, atd.).

[5] NoSQL je překládáno jako Not only SQL



Zdroj: FOWLERUK, A.: Thoughts on NoSQL & Big Data Architecture. In: Marklogic [online]. 2013 [cit. 2015-04-16]. Dostupné z: https://adamfowlerml.wordpress.com/2013/04/29/thoughts-on-nosql-big-data-architecture/

Obrázek 27-3: Nové typy dat v souvislosti s Big Data

27.7.2 Vrstva ETL

Vrstva ETL se stará zejména ***o transformaci dat*** do podoby, která je vhodná pro další kroky a jejich ***přípravu pro vrstvu datového zpracování***. V této vrstvě jsou využívány nástroje pro import a export v závislosti na charakteru dat a na nástrojích pro jejich skladování (SQL/NoSQL) jako např. Sqoop. ***Sqoop*** je jeden nástrojů pro efektivní přesun dat mezi datovým zdrojem a Hadoopem (systém pro distribuované zpracování dat). Pro ***transformaci logů*** se používají nástroje pro log management (Splunk, Syslog, loggly a další).

27.7.3 Vrstva datového zpracování

Vrstva datového zpracování ***kombinuje strukturovaná i nestrukturovaná data na jednom místě*** pro jejich ***další zpracování*** (rozbor a skladování) a ***generaci dat použitelných struktur*** pro datovou analýzu a prezentaci výstupů. Mezi nástroje používané v této vrstvě patří hlavně ***Hadoop a jeho ekosystém*** (HBase a HDFS[6] pro distribuované skladování, MapReduce pro paralelní zpracování a Pig a Hive pro přístup k datům), a také nástroje pro Big Data in-memory analýzu nad Hadoopem v reálném čase jako Driud či Kognitio.

27.7.4 Vrstva pro analýzu dat

Komponenty vrstvy pro datovou analýzu ***umožňují využívat data použitelná pro datovou analýzu konečným uživatelům***. Tyto nástroje dávají uživatelům uživatelsky příjemné prostředí pro tvorbu reportů, dashboardů a datovou vizualizaci (Tableau, MicroStrategy). Dále do této vrstvy patří nástroje pro

[6] Hadoop Distributed File System

data mining a data discovery (Spotfire, Datameer) umožňující mimo jiné prediktivní analýzy, analýzy výkonu nebo doporučení na základě prognózovacích algoritmů.

27.8 Big Data technologie

Oblast Big Data v sobě zahrnuje řadu technologií a nástrojů pro datovou integraci, skladování, analýzu a vizualizaci, ale ***primárním stavebním kamenem je ve většině případů Hadoop a technologie s ním spojené.***

27.8.1 Apache Hadoop

Na trhu Big Data technologií figuruje ***řada nejrůznějších řešení od řady poskytovatelů***. Společnosti Zde je vybráno ***Open Source Big Data řešení*** v podobě Hadoopu. To bylo vybráno na základě důvodů, které zmiňuje Tom White. ***Hadoop je nejrozšířenějším Big Data řešením*** ve společnostech pro ukládání dat a jejich analýzu. Rychle se stalo synonymem pro Big Data. Toto tvrzení je podtrženo ***množstvím poskytovatelů*** Hadoop distribucí ve světe. Mezi ně patří IBM, Microsoft, Oracle, Cloudera nebo třeba Databricks. (White, 2012)

Apache Hadoop je ***open source framework*** pro zpracování, ukládání a analýzu velkého množství distribuovaných, nestrukturovaných dat (viz 24) vytvořený ***v Javě***. Hadoop je výpočetní prostředí postavené nad distribuovaným souborovým systémem navrženým ***pro operace nad objemy dat v řádu terabajtů až petabajtů***.

Hlavní ***tvůrce*** Hadoopu je ***Doug Cutting***, který také vytvořil softwarovou knihovnu pro textové vyhledávání zvanou ***Lucene*** (viz 25). Hadoop ***se vyvinul z open source webového vyhledávače zvaného Apache Nutch***, který je založený právě na Lucene.

V roce 2003 Google publikoval architekturu vlastního distribuovaného souborového systému zvaného ***Google File Systém (GFS)*** [26]. Systém založený na GFS Nutch potřeboval systém pro ukládání velkých objemů dat generovaných indexováním webu, a proto se v roce 2004 Cutting rozhodl vytvořit open source implementaci tohoto systému pod jménem ***Nutch Distributed Filesystem (NDFS)***. Google v roce 2004 ještě publikoval článek popisující model pro paralelní zpracování zvaný MapReduce, který byl později implementován i do Nutch. Implementace NDFS a MapReduce byly použitelné i mimo webové vyhledávání a indexování, a proto byly tyto technologie osamostatněny pod novým projektem, kterým byl právě Hadoop a NDFS byl přejmenován na HDFS.

Cutting byl v roce 2006 angažován internetovým vyhledávačem Yahoo a na počátku roku ***2008 pak Yahoo do produkce uvedlo implementaci Hadoopu***, jehož cluster obsahoval 10 000 jader pro indexování webu. Hadoop byl později dále optimalizován a v roce 2009 dokonce jeden terabajt dat seřadil za pouhých 62 vteřin (viz 27).

Hadoop se skládá ze dvou základních částí, jedná se o ***distribuovaný souborový systém zvaný HDFS*** a ***programovací model pro paralelní zpracování zvaný MapReduce***. Kromě těchto dvou komponent obsahuje Hadoop ***řadu přidružených projektů***:

- Apache Avro je open source framework pro vzdálené volání procedur (RPC) a serializaci s použitím JSON7 pro definici datových typů a protokolů.
- Apache Cassandra je open source distribuovaná databáze navržená pro práci s velkými objemy dat.
- HBase je open source, sloupcově orientovaná distribuovaná databáze inspirovaná systémem BigTable společnosti Google.
- Apache Chukwa je open source robustní a škálovatelný nástroj pro monitoring, kolekci a analýzu logů postavený nad HDFS a MapReduce.
- Apache Hive je open source datový sklad postavený nad Hadoopem pro jednodušší agregaci, dotazování a analýzu dat nad velkými objemy dat v HDFS pomocí jazyka, který je velmi podobný SQL.

[7] JavaScript Object Notation

- Mahout je open source škálovatelná knihovna nad Hadoopem obsahující algoritmy pro strojové učení a analýzy (cluster analýza, analýza chování, atd.).
- Pig je platforma a vyšší programovací jazyk (Pig Latin) pro dotazování a datovou analýzu nad velkými objemy dat
- Zookeeper je open source nástroj zajišťující centralizované koordinační služby (konfigurace, synchronizace, naming, a další) pro rozsáhlé distribuované systémy.

Hadoop je ***open source*** a každá společnost jej může ***bezplatně vyzkoušet na své stávající on-premise infrastruktuře***. Veškeré postupy a kód jsou zveřejněny na stránkách Apache. Hadoop je vyvíjen od roku 2002 a ***aktuální verze je Hadoop 3.2 (k březnu 2021)***. Doslovná ***definice Hadoopu*** je dostupná na stránkách Apache:

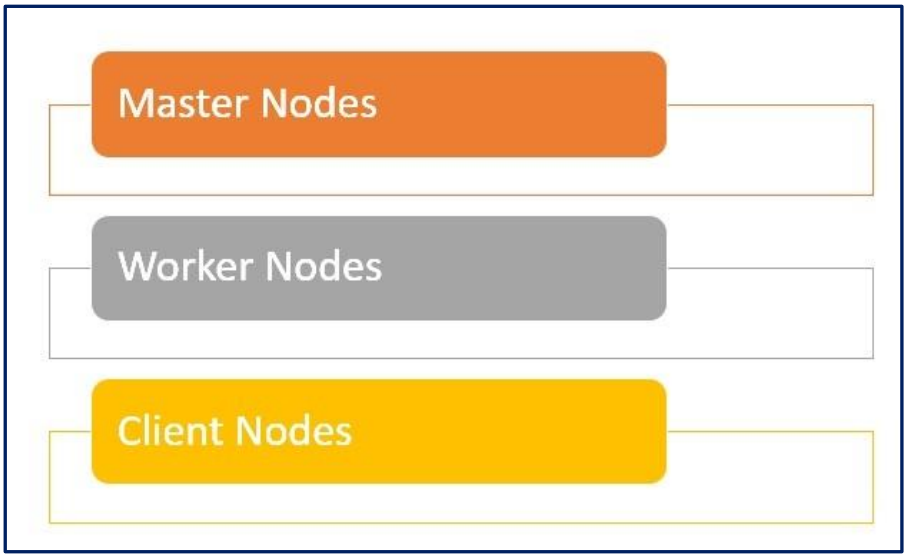
"The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. " (Apache, 2021)

Nejdůležitějším prvkem k pochopení Hadoopu je ***vnímání Hadoopu jako softwarové knihovny***, která je v podobě open source s možností škálovatelnosti. Při změnách objemu dat či potřeb, je ***možnost pružně měnit parametry celé infrastruktury*** i prvků v ní. V dalších částech je zde přímo vysvětlen princip clusteringu výpočetní síly a datového uložiště. V poslední části jsou popsány hlavní výhody Hadoopu oproti jiným řešením. Framework je zaměřen na spolehlivost manipulace s daty a nespoléhá přímo na hardware. Pokud např. dojde k poškození jednoho výpočetního serveru, je zde ***možnost zastoupení druhým bez terminace úlohy***. Tato „jistota" je hlavní výhodou Hadoopu oproti jiným řešením.

27.8.2 Hadoop Cluster

Hadoop Cluster je ***zodpovědný za veškerou manipulaci s daty včetně jejich uložení***. Koncepty Clusteru se začaly objevovat již okolo roku 1996, kdy je poprvé popsal David Balder ve svém výzkumu. Jednu z definic popsal Mark Baker z University of Portsmouth. Ten uvádí že výpočetní cluster je spojením volných počítačů v jedné lokální síti, které spolu spolupracují a navenek se klientovi jeví jako jedno zařízení. Využití Clusterů se doporučuje v místech, kde je potřeba zvýšit výpočetní sílu, nebo je zde potřeba vysoká dostupnost služeb běžících na serveru. (Baker, 2000)

Využití Cluster Computingu je pro Hadoop ***klíčové***. ***Každý server či počítač*** v Clusteru se označuje jako ***Node (uzel)***. Na obrázku (Obrázek 27-4) je ***základní dělení Nodů*** v Hadoop architektuře.



Obrázek 27-4 : Nodes Category in Hadoop. (Zdroj: Databricks, 2021)

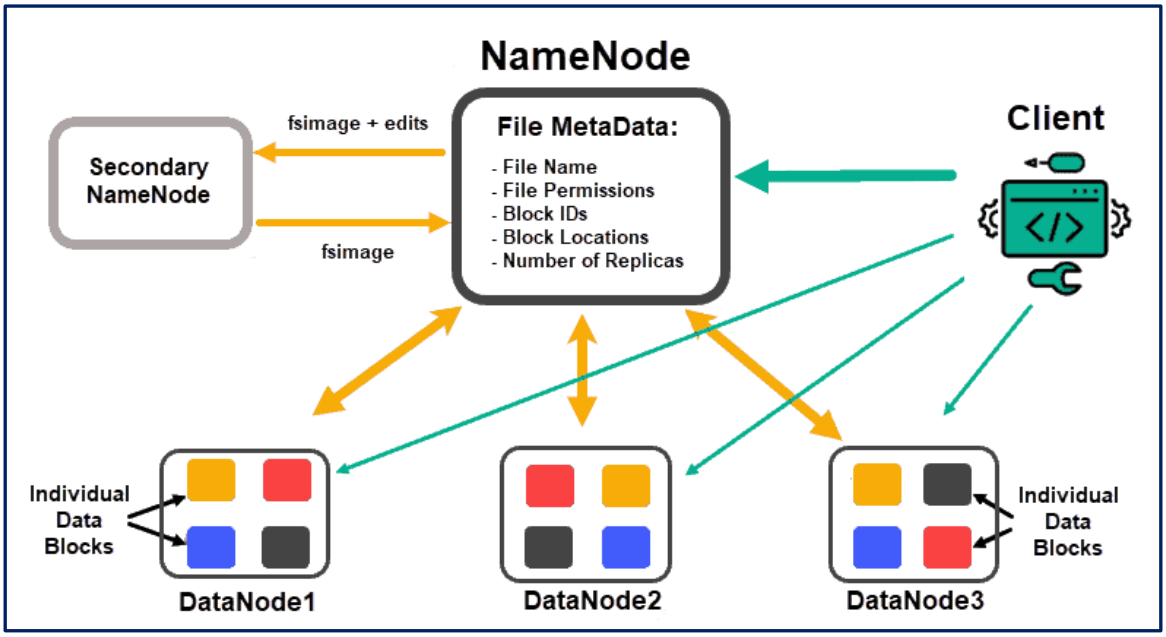
- ***Master Nodes***: Nejdůležitější Node v Clusteru. Udržuje adresy všech dat uložených v HDFS a dohlíží nad během paralelních výpočtů. Alternativní označení je ***NameNode*** a nejčastěji bývá v Clusteru ***právě jednou***. (Databricks, 2021)
- ***Worker Nodes***: Zahrnují ***většinu strojů*** (PC, Server) v Clusteru. Provádějí ***úlohu uložiště dat a výpočtů***. ***Každý Worker Node obsahuje DataNode a TaskTracker***. Worker Node výhradně přijímá pokyny z Master Node. Již ***alokované zdroje Nodu*** se označují jako ***Container***. (Databricks, 2021)
- ***Client Nodes***: Client Nodes jsou ***uzly klientů*** přistupujících k Hadoopu. Tyto uzly ***odesílají požadavky na Master Node v podobě API či Query***, ve kterých jsou popsány požadavky na manipulaci s daty. Po dokončení těchto úloh ***Client Node zpracovává výsledky výpočtů a zobrazuje je***. (Databricks, 2021)

27.8.3 Hadoop Ecosystem

Hadoop Ecosystem je architektura komponent tvořících Hadoop. Jsou to ***volně dostupné nástroje***, které běží na Clusteru a dohromady tvoří funkční architekturu pro správu a analýzu dat. Popis jednotlivých komponent je převzat z oficiální dokumentace Apache. Vedle základních komponent Hadoop jsou zde uvedeny i komponenty, které využívají největší distributoři Hadoop. (Apache, 2011)

- ***HDFS***: Naprostým základem pro ***Hadoop je HDFS (Hadoop Distributed File Systém)***. HDFS je ***distribuované datové uložiště*** s vysokou propustností dat optimalizované pro velké datasety. Data se ***distribuují mezi všemi DataNody***. Samotný framework nabízí ***zrcadlení*** dat z důvodů vyšší bezpečnosti a škálovatelnosti při dotazech. HDFS umožňuje ***ukládání dat v nejrůznějších formátech***, pro paralelní zpracování se doporučuje Apache Parquet nebo AVRO.

  Na obrázku (Obrázek 27-5) je nastíněno, jak ***HDFS využívá jednotlivé Nody v Clusteru.***



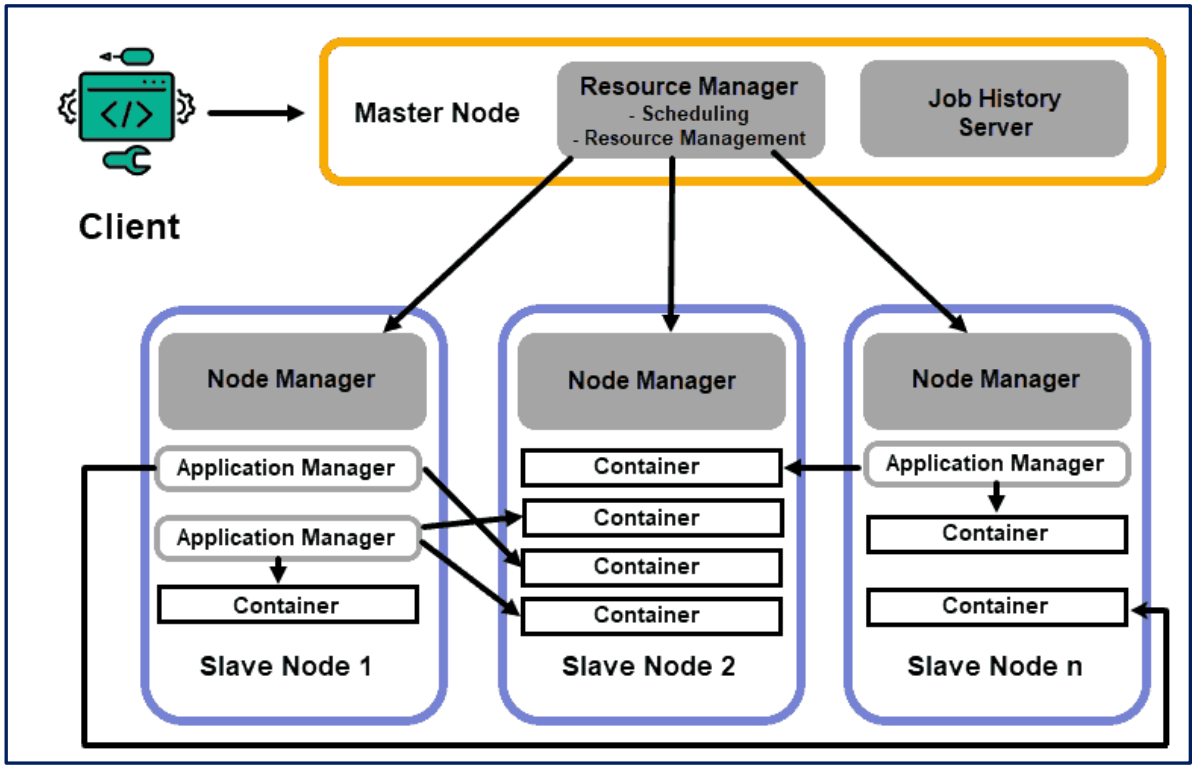
Obrázek 27-5: Architektura HDFS. (Zdroj: phoenixNAP, 2021)

NameNode je hlavním Master Nodem HDFS a obsahuje ***veškerá metadata*** o uložených souborech a jejich lokaci. ***Secondary NameNode*** spadá také do kategorie Master Nodů a obsahuje zálohu informací obsažených na hlavním NameNode.

Pod NameNode jsou umístěny ***jednotlivé DataNode*** obsahující ***bloky dat***, které ***se spojují až při konkrétním dotazu***. Ke spojování se využívá metadat uložených v NameNode. Každý ***Worker Node***

slouží k ***ukládání dat a obsahuje svůj DataNode***. Na obrázku je zobrazen i ***Client Node***, který může přes příslušné rozhraní přistupovat k datům na HDFS a konfigurovat jednotlivé Nody.

Hadoop YARN: Dalším důležitým prvkem Hadoop Ecosystem je YARN (***Yet Another Resource Negotiator***). YARN je ***nejvyšší autorita***, která ***rozhoduje o alokovaných prostředcích*** mezi všemi Nody v systému. Každý Worker Node obsahuje ***NodeManager***, který je zodpovědný za ***využití zdrojů*** (CPU, RAM, síť) pro jednotlivé úlohy. ***Samotný YARN*** se označuje ***jako ResourceManager*** a je přítomen ***v Master Node***. Konkrétní architektura YARN je nastíněna na obrázku (Obrázek 27-6).



Obrázek 27-6: Architektura YARN. (Zdroj: phoenixNAP, 2021)

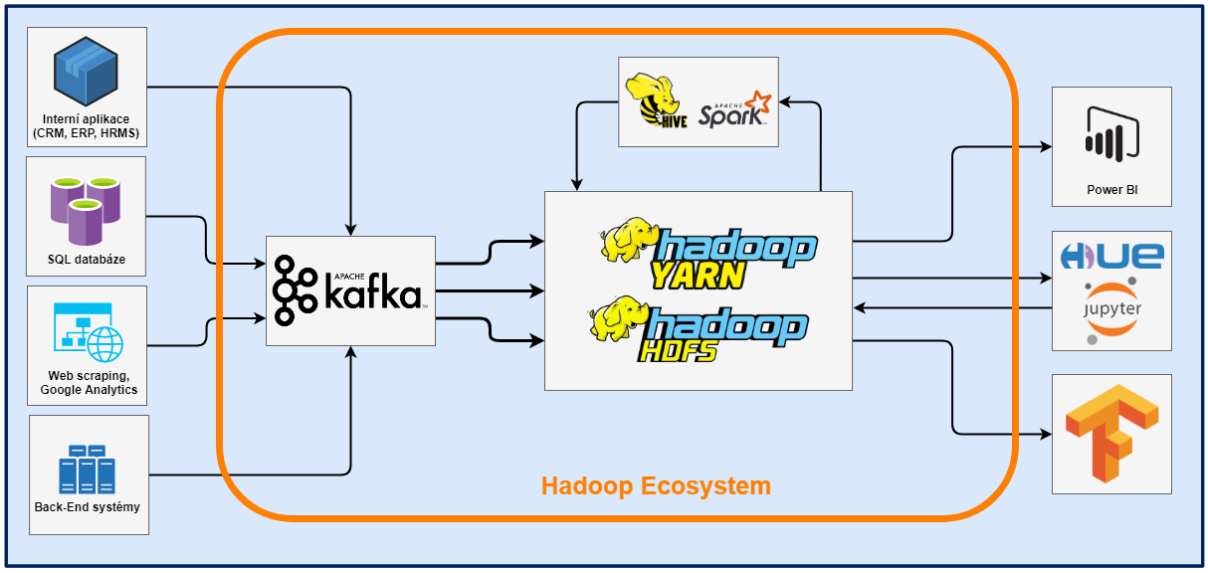
Z obrázku je patrné že ***Master Node přijímá zadání*** výpočetních úloh ***od Client Node***. Na základě požadavků ***deleguje „práci" Worker Nodům*** v Clusteru. Každý Worker Node má již zmiňovaný ***Node Manager, Application Manager*** přebírá samotné zadání úlohy od Resource Managera a Container alokovanou výpočetní sílu. Navíc je zde uveden ***Job History Server***. Ten obsahuje informace o všech dokončených úlohách na daném clusteru. Pokud úloha selže (např. přetečením paměti), uživatel nalezne detaily selhání právě zde. Tato komponenta je dostupná jako webová stránka dostupná na adrese clusteru. Komponenty mají následující charakteristiky:

- ***Hadoop MapReduce*** je jedním ze ***softwarových frameworků pro paralelní zpracovávání*** velkých objemů dat na Clusteru. MapReduce rozděluje vstupní datovou sadu na nezávislé bloky, které jsou zpracovávány úlohami Map. MapReduce třídí výstupy z Map úloh, které jsou pak vstupem do redukčních (Reduce) úkolů. Vstupy i výstupy úloh jsou uloženy na HDFS. Celá úloha je řízena pomocí YARN.
- ***Apache Spark*** je výkonný ***analytický engine*** užívaný k zpracovávání velkého objemu dat, ale i zpracovávání v reálném čase či jako ETL nástroj. Nepatří do základní knihovny Hadoopu, ale Hadoop je připraven k integraci této knihovny. V dnešní době je ***hojně užíván*** a neodmyslitelně spjat s Big Data. Spark neužívá framework MapReduce ale ***RDD (Resilient Distributed Datasets)***. RDD podporuje ***in-memory processing*** (zpracování v paměti) a pracuje zde s mezivýsledky výpočtů jako s objektem, který může předávat mezi jednotlivými nody v clusteru. Tato technologie je v typových úlohách 10x až 100x rychlejší než MapReduce a pracuje přímo s Client Nody.
- ***Hive*** je ***Warehousing systém*** umožňující správu dat na HDFS a manipulaci s nimi pomocí SQL. Umožňuje také vytvářet struktury dat nad již uloženými daty. Tím pádem lze vytvářet

celistvé databáze, ***Data Warehouse a Data Marty v rámci jednoho Data Lake*** (HDFS). Hive využívá k dotazování MapReduce framework.

- **Apache ZooKeeper** je služba pro ***centralizovanou údržbu konfigurace jednotlivých Nodů***. Tato služba je kritická pro Hadoop řešení s velkým počtem zařízení. Při majoritních změnách konfigurace či adresace ZooKeeper synchronizuje tyto informace napříč všemi Nody.
- **Apache Kafka** není přímo součástí Hadoop Ecosystem, ale neodmyslitelně je s ním spojený. Kafka je distribuovaná ***platforma pro stream dat s přímou integrací*** na HDFS. Kafka se užívá pro sběr velkých objemů dat z klientských aplikací či back-end systémů. Spark umožňuje základní manipulaci s daty tekoucími Kafkami.

Na obrázku (Obrázek 27-7) jsou znázorněny jednotlivé ***komponenty Hadoop Ecosystem*** zmíněné výše. Komponenty jsou zasazeny do obrázku interpretující Big Data jako proces na obrázku (Obrázek 27-1).



Obrázek 27-7: Konkrétní podoba Hadoop Ecosystem. (Zdroj: Hamala, 2021)

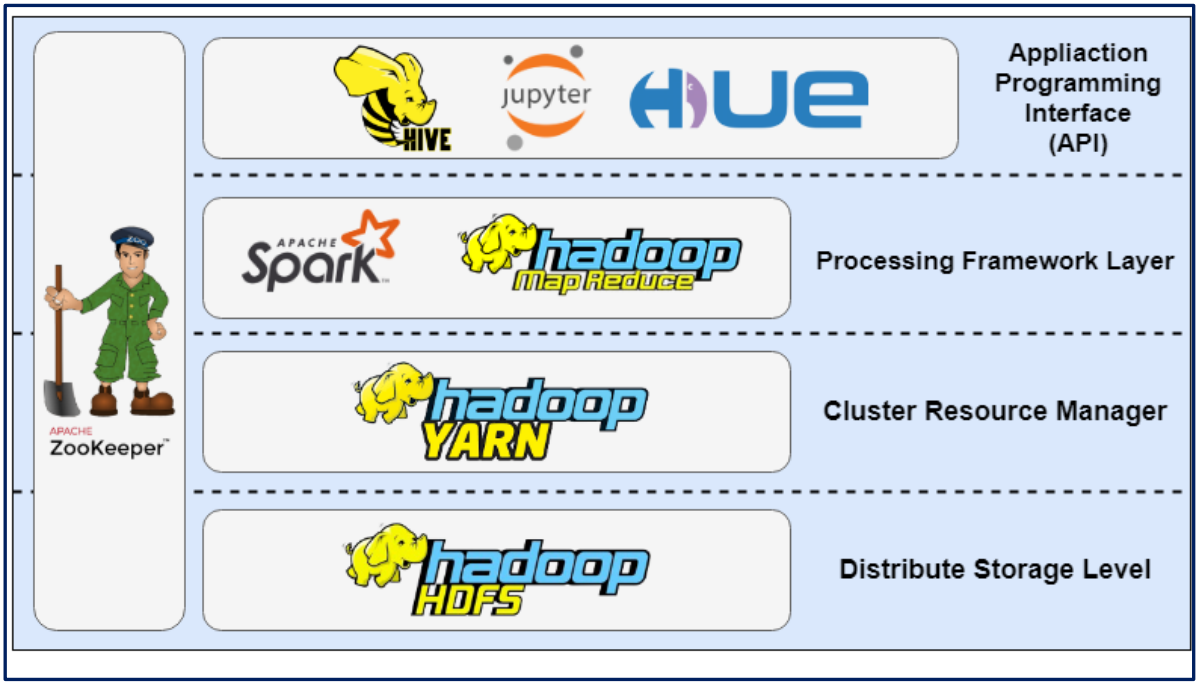
Na obrázku (Obrázek 27-7) je přesně specifikováno ***co spadá do Hadoop Ecosystem, a co je možno integrovat dále.***

V oranžovém obdélníku jsou uvedeny technologie, co přímo manipulují s HDFS a jsou ovládány YARN. Do tohoto „jádra“ Hadoopu spadají všechny Master a Worker Nody. Nad uložištěm dat je analytický engine. Hadoop od verze 2.0 umožňuje paralelní užití více analytických enginů v jednom Hadoop Ecosystem. V tomto případě zde jde o užití Apache Hive obsahujícího framework MapReduce a enginu Apache Spark.

V pravé části jsou pak uvedeny konkrétní ***příklady možné integrace***. V oblasti reportingu je to nástroj ***Power BI***, který umožňuje propojení HDFS pomocí ODBC nebo přímému přístupu k souborům. V oblasti vizualizace dat přibil tok dat směrem k HDFS a YARN. Uvedené nástroje ***Jupiter a Hue*** fungují jako uživatelské prostředí pro manipulaci s daty v Hadoopu. ***Hue je integrováno s Warehousing nástrojem Hive, Jupyter pak poskytuje API pro Spark***. Oba nástroje umí zobrazovat a vizualizovat výstupy dotazů. Poslední částí jsou pak ***nástroje pro umělou inteligenci*** a strojové učení. Hadoop podporuje integraci a užívání nástroje ***TensorFlow***, který je hojně rozšířeným řešením výzkumu umělé inteligence. Možnosti integrace dalších nástrojů jsou individuální a spočívají na případech užití jednotlivých společností.

27.8.4 Vrstvy Hadoop architektury

Hadoop Architekturu můžeme mimo jiné rozdělit i na vrstvy, které upravují vnímání a přístup k jednotlivým komponentám. Základní model má 4 úrovně a je nastíněn na obrázku Obrázek 27-8.



Obrázek 27-8: Vrstvy Hadoop architektury. (Zdroj: Halama, 2021)

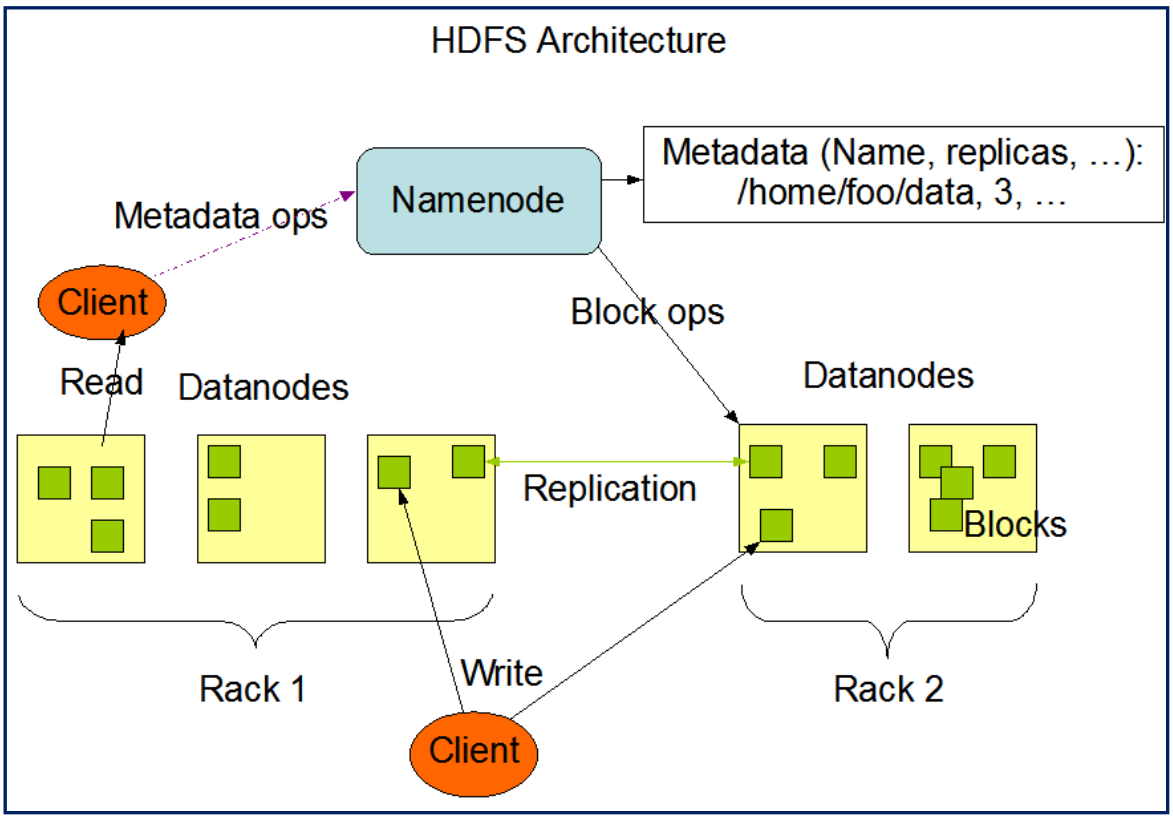
- ***Distribute Storage Level***: Nejnižší vrstvou je vrstva distribuovaného uložiště. Je to vstupní bod Hadoop architektury. Tato vrstva uchovává veškerá data a metadata potřebná k následné manipulaci a transformaci.
- ***Cluster Resource Manager***: Druhou vrstvou je správa prostředků na Clusteru. Tato vrstva umožňuje provádění zadaných úloh komputace dat z první vrstvy – HDFS. Pro tuto vrstvu je kruciální existence vrstvy distribuovaného uložiště ale i vstupu z klientských stanic.
- ***Processing Framework Layer*:** Vrstva rámce zpracování obsahuje zvolený analytický engine. Ten zpracovává zadání úlohy od klienta (Query, volání API) a určuje jaké data na jakých DataNodes se budou zpracovávat. Tyto specifikace se předávají na vrstvu správy prostředků, kde YARN tuto úlohu rozdistribuuje mezi jednotlivé Worker Nodes.
- ***Application Programming Interface (API)***: Poslední a nejvyšší vrstvou je programové rozhraní. Toto rozhraní je přístupné koncovým uživatelům Hadoop infrastruktury. Pomocí nástrojů jako je ***Hue a Jupyter*** mohou své dotazy psát pomocí ***SQL a PySpark***. Tyto dotazy se překládají do exekučního plánu, který je předán do vrstvy rámce zpracování. Díky tomuto je Cluster schopen zpracovat jednotlivé požadavky koncových uživatelů. Tato vrstva je zároveň prezenční vrstvou pro koncové uživatele a zobrazují se zde výsledky jednotlivých dotazů včetně chybových hlášek a výpisů z konzole.
- ***ZooKeeper*:** Ve všech těchto vrstvách figuruje ZooKeeper. Ten se stará o bezproblémové ***předání všech zpráv*** mezi jednotlivými vrstvami a o ***správu konfigurace*** jednotlivých Nodů spadajících do vrstev. Pokud by se stalo že bude jeden z Nodů nedostupný, ZooKeeper zajistí exekuci dotazu a dostupnost dat na jiném Nodu bez nutnosti terminace úlohy.

27.8.5 Hadoop Distributed File System (HDFS)

Když objem ukládaných dat překročí kapacity jediného databázového serveru, je nutné tento objem dat rozdělit na menší části a ty ukládat na různých vzájemně komunikujících serverech. Souborové systémy, které spravují sítě těchto serverů (a jejich data) se pak označují jako distribuované (viz 28).

HDFS je distribuovaný souborový systém navržený speciálně pro spolehlivé skladování rozsáhlých (většinou neměnných) datových souborů. Každý HDFS cluster obsahuje dva typy uzlů v modelu Master/Slave (viz výše). Uzel zvaný ***NameNode (master)*** spravuje souborový strom a skladuje metadata o všech souborech a složkách a log jejich změn v celé Hadoop instanci. NameNode navíc obsahuje informace o lokaci všech bloků, ze kterých se každý soubor skládá. ***NameNode je typicky v celé Hadoop instanci pouze jeden***, a proto je v systému považován za slabé místo, při jehož

výpadku jsou ztracena i odpovídající data. Z tohoto důvodu se doporučuje jako NameNode využívat vysoce spolehlivý a robustní server. Novější verze HDFS obsahují také zálohovací uzel zvaný ***Backup-Node*** jako rezervní server v případě selhání NameNode. Druhým typem uzlu je DataNode (slave), který skladuje a poskytuje datové bloky na příkaz NameNode nebo uživatele a pravidelně informují Name-Node o blocích, které jsou v daném uzlu uložené (Obrázek 27-9).



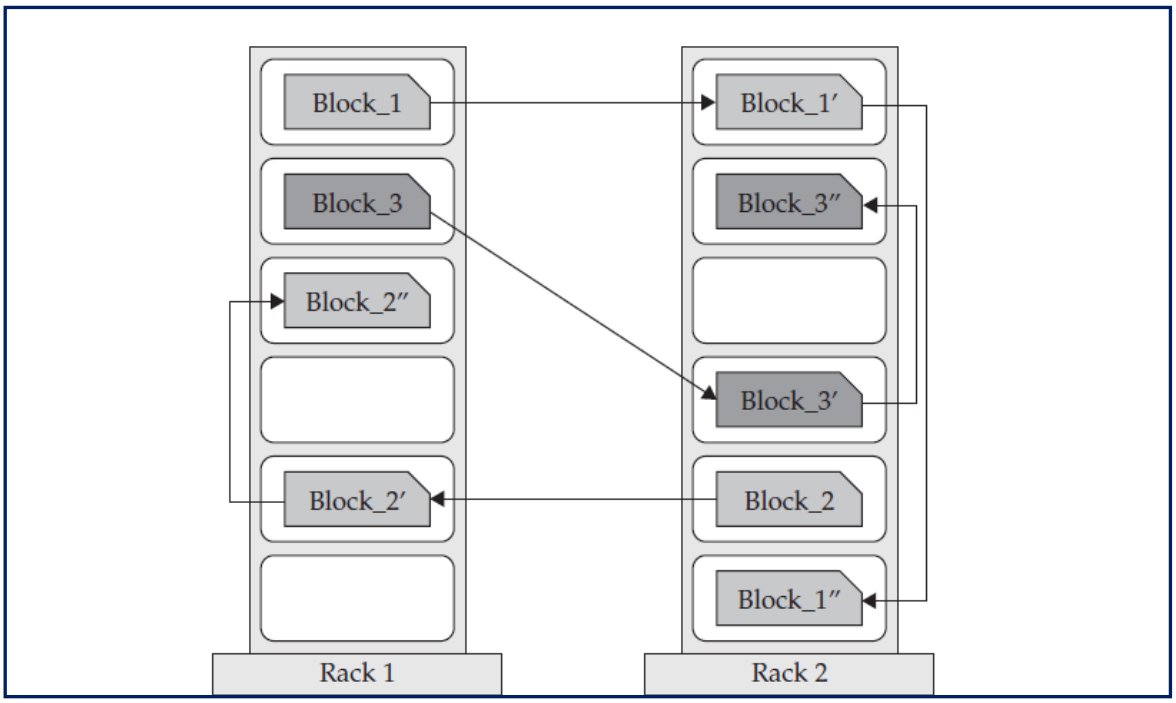
Obrázek 27-9: Architektura HDFS (Zdroj: 29)

Vstupní data v Hadoop clusteru jsou rozdělena na menší části zvané ***bloky***, typicky o velikosti (nebo násobku) ***64 MB*** a rozděleny do jednotlivých uzlů clusteru. Datové bloky v HDFS jsou výrazně větší než bloky u většiny ostatních souborových systémů, kde jsou bloky velké 512 bajtů, a relačních databází, které pracují s bloky o velikosti 4 až 32 kilobajtů. Tímto způsobem pak lze provádět MapReduce funkce na podmnožině rozsáhlých vstupních souborů, a tím dosáhnout požadované škálovatelnosti.

Cílem Hadoopu je ***využít rozsáhlé clustery složené z běžně dostupných serverů***, kde každý server obsahuje relativně levný diskový prostor.

Hadoop clustery se často skládají ***z tisíců levných serverů***, z nichž každý obsahuje několik pevných disků. Je proto nevyhnutelné, že bude docházet k hardwarovým poruchám a výpadkům. Jednou z hlavních předností je tedy fakt, že Hadoop, resp. ***HDFS s hardwarovými poruchami už ze své podstaty počítá*** a dokáže se s nimi vyrovnat (viz 30).

Pro zajištění dostupnosti dat i v případě selhání komponent HDFS ***automaticky ukládá kromě samotného datového bloku ještě jeho dvě kopie*** a ty ukládá na různých serverech v odlišném racku, než je originál, viz Obrázek 27-10. ***Úroveň redundance lze upravit*** jak pro jednotlivé soubory, tak pro celý cluster. Redundance zajišťuje vysokou úroveň spolehlivosti a přístupnosti dat v clusteru, umožňuje rozdělit výpočty do menších kousků a také zvyšuje úroveň lokálnosti dat.



Obrázek 27-10: Redundance datových bloků v HDFS, Zdroj [31]

Pro dosažení vyššího výkonu ***Hadoop automaticky přiděluje úkoly serverům, na kterých jsou data uložena***. Data jsou ***načtena ze svého úložiště přímo do procesoru*** a není tedy pro jejich zpracování nutné zatěžovat síť zbytečnými přesuny dat. Přesun výpočtů k úložišti dat zajišťuje vysokou úroveň lokálnosti dat a zároveň zvyšuje průchodnost sítě. Z těchto důvodů není vhodné pro Hadoop využívat datové úložiště jako je SAN (storage area network) nebo NAS (network-attached storage).

Hlavní principy HDFS lze tedy shrnout do několika bodů (viz 32):

27.8.5.1 Rozsáhlé soubory

Hadoop clustery v současnosti běžně obsahují terabajty i petabajty dat a jednotlivé soubory se často pohybují v řádu stovek megabajtů nebo gigabajtů. HDFS je proto optimalizovaný pro podporu práce s rozsáhlými soubory díky vysoké datové propustnosti a možností provozovat stovky uzlů v clusteru na kterém mohou být teoreticky uloženy miliony souborů.

27.8.5.2 Selhání hardwaru

HDFS je navržen pro chod na velkém množství levných serverů a hardware, u kterých je ve velkých clusterech vysoká pravděpodobnost poruchy. HDFS si s takovýmto selháním hardwaru dokáže automaticky poradit, aniž by porucha měla výrazný vliv na výkon celého systému

27.8.5.3 Write once, read many

HDFS je navržen s myšlenkou, že nejefektivnější přístup k zpracování velkého množství dat je pomocí vzoru write-once, read-many (WORM). Data tedy po nahrání do HDFS a uzavření zůstávají neměnná a nedochází k jejich úpravě (write once), zatímco analýzy mohou nad daty být prováděny opakovaně, dokud jsou data v clusteru k dispozici (read many).

27.8.5.4 Stálý tok dat

Aplikace využívající HDFS čtou data v dlouhých sekvencích. HDFS je proto optimalizován pro dávkové čtení a vysokou propustnost, což má za následek menší úroveň interaktivity a vyšší odezvy při přístupu k datům. Při provádění analýz rozsáhlých souborů je ovšem důležitější čas, za který je přečten celý soubor, než čas pro přečtení několika prvních záznamů.

27.8.5.5 Přesun výpočtů je levnější než přesun dat

Pro systém je v případě rozsáhlých souborů efektivnější provádět požadované výpočty v místě jejich úložiště než data někam přesunovat. Tento princip zabraňuje zahlcení sítě neustálými přesuny dat a zvyšuje propustnost systému. HDFS z tohoto důvodu obsahuje rozhraní, které umožňuje přemístit aplikace blíže ke zdroji dat.

27.8.6 HBase

HBase je NoSQL open source sloupcově orientovaná distribuovaná databáze inspirovaná systémem ***BigTable*** společnosti Google. HBase ***pracuje nad Hadoopem a doplňuje HDFS funkcionalitu*** o přímý přístup k velkým objemům dat v reálném čase (HDFS k datům přistupuje sekvenčně). HBase je navržen pro ***škálování v počtu řádků i sloupců*** a ***tabulky jsou automaticky horizontálně děleny a replikovány mezi uzly*** v síti. HBase se skládá z dvou typů uzlů, MasterNode má v podstatě stejnou funkci jako u HDFS a uzly na kterých jsou uložena data, se nazývají regionální servery (viz 33).

HBase se podobně ***jako relační databáze*** skládá z tabulek, sloupců a řádků ve kterých jsou uložená data (viz 34). Každá tabulka obsahuje primární klíč, přes který se k ní přistupuje a podle kterého jsou řazeny. ***Sloupce v HBase reprezentují atributy objektu*** jako například název nebo čas vytvoření. Atributy se slučují do předem specifikovaných sloupcových rodin, které jsou v dané tabulce definované při jejím vytvoření.

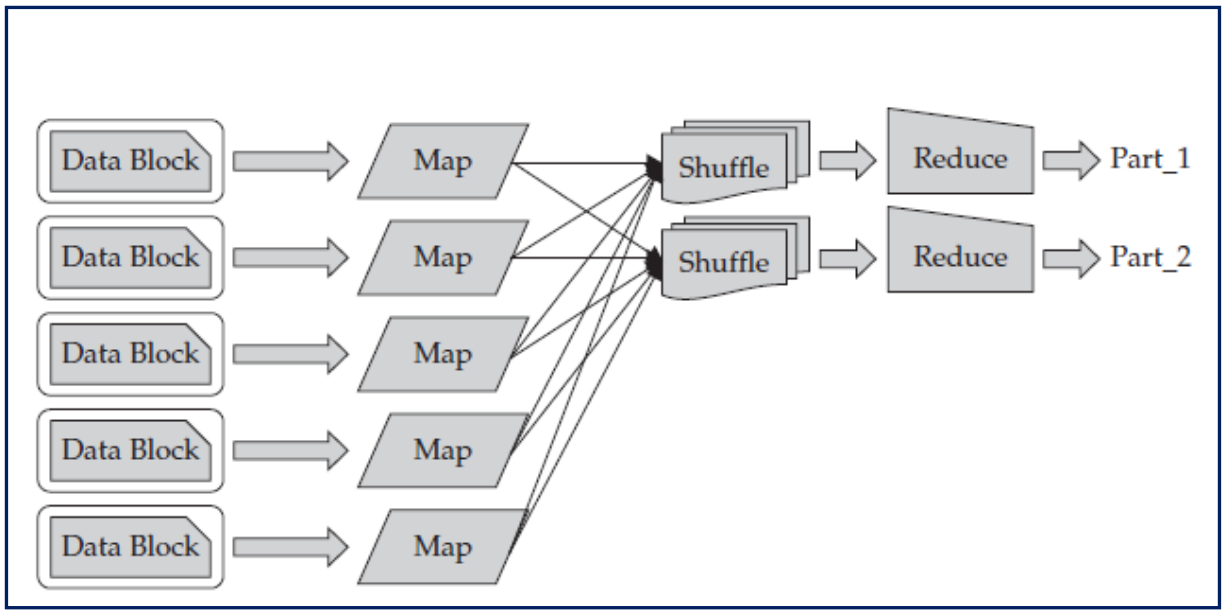
Typickým příkladem použití HBase je tabulka zvaná ***webtable obsahující atributy o webových stránkách*** jako například jejich jazyk, kódování a typ média pro danou URL. Takováto tabulka může obsahovat stovky milionů záznamů, které jsou neustále aktualizovány a jsou z nich získávány statistiky a indexy pro webový vyhledávač. Relační databáze nejsou v tomto případě vhodné kvůli rychlosti čtení a zápisu v rozsáhlých tabulkách a HDFS je zase určen pro dávkové úkoly.

HBase lze charakterizovat v těchto bodech:

- ***Žádné indexy***, řádky i sloupce jsou ukládány sekvenčně, takže nevznikají problémy s indexy a rychlost importu dat je nezávislá na velikosti tabulky.
- ***Automatické dělení,*** tabulky jsou při růstu objemu automaticky děleny do tzv. regionů a distribuovány mezi dostupné regionální servery.
- ***Automatické lineární škálování,*** při přidání nového uzlu do clusteru je zátěž systému automaticky přerozdělena tak, aby byl využit i nový uzel.
- ***Integrace s MapReduce***, s MapReduce umožňuje paralelní zpracování dat v distribuovaném prostředí,
- ***Odolnost vůči selhání,*** při výpadku jednotlivých uzlů systém bez problémů pokračuje v práci díky replikaci dat
- ***Běžný hardware,*** clustery jsou složené z velkého množství relativně levných serverů, agregovaný výkon a disková kapacita těchto serverů je efektivně vyšší než u malého počtu nákladných serverů.

27.8.7 MapReduce

MapReduce je ***programovací model umožňující paralelní zpracování velkých objemů*** díky škálování a rozdělení zátěže mezi velký počet serverů v Hadoop clusteru inspirovaný stejnojmenným modelem společnosti Google představeným v roce 2004 (viz 35). MapReduce je rozdělen na dvě ***samostatné funkce, Map a Reduce***. MapReduce rozdělí vstupní data na menší části (okolo 64 MB), které jsou paralelně zpracovány v clusteru, na kterém jsou uložena. ***Zpracování*** probíhá tak, že ***Map transformuje a seskupí jednotlivé elementy vstupních data do dočasné datové množiny (páry key/value)*** a následně je ***předá funkci Reduce***. Funkce Reduce využívá výstup z Map jako vstupní data a ***z těchto dat vytvoří ještě menší množinu*** (typicky jeden nebo žádný výstup pro každou key hodnotu).



Obrázek 27-11: Princip modelu MapReduce (Zdroj: 36)

MapReduce program je ***označován jako Job*** a dále ***se dělí na jednotlivé paralelně vykonávané úkoly*** (tasks). Uzel zvaný ***JobTracker*** komunikuje s HDFS uzlem ***NameNode***, ***aby zjistil lokaci dat*** potřebných pro daný úkol, a pak každému uzlu v clusteru přidělí konkrétní Map a Reduce úkoly, které má uzel zpracovat. JobTracker se snaží úkoly ***přidělovat uzlům, na kterých jsou požadovaná data uložena***, aby nebylo nutné data zbytečně přesunovat (viz 37). ***Průběh*** každého úkolu je ***monitorován***, a v případě selhání JobTracker daný úkol předá jinému uzlu v clusteru. MapReduce programy jsou ***vytvořené v Javě*** a jednotlivým uzlům je JobTracker předává v podobě ***Java archivu*** (jar).

MapReduce ***pracuje pouze s páry key/value***, a proto je vhodný pouze pro specifické úkoly. Nejjednodušším příkladem použití MapReduce je spočítání výskytu jednotlivých slov v textovém dokumentu. Funkce Map rozdělí obsah dokumentu na jednotlivá slova a vytvoří key/value pár pro každé slovo. MapReduce následně seřadí všechny páry a předá je funkci Reduce, která pak pouze provede jejich součet.

27.9 Nástroje pro dotazování

Tvorba MapReduce úkolů pro datovou analýzu je ***neintuitivní a náročná***, protože je nutné psát tyto funkce v Javě, která není pro tento účel vytvořena a nenabízí dostatečnou úroveň abstrakce. I jednoduché úkoly v Javě nutí psát komplexní kód, experimentování a prototypování vyžaduje kompilaci a běžné příkazy jako např. join jsou zbytečně složité a náchylné k chybám. Z tohoto důvodu bylo ***vytvořeno několik nástrojů a jazyků*** vyšší úrovně pro snadnější operace a analýzy dat ***nad Hadoopem na bázi SQL***, které následně uživatelsky čitelný kód překládají na MapReduce úkoly.

27.9.1 Pig

Pig je ***open source platforma*** explicitně vytvořena ***pro agilní analýzu rozsáhlých datových objemů***. Pig se skládá ***ze dvou komponent***, ***první*** je vysoce ***expresivní procedurální skriptovací jazyk (PigLatin)*** s vyšší úrovní abstrakce zaměřený na jednodušší a uživatelsky přístupnější analýzu dat pomocí notace (vzdáleně) připomínající SQL. ***Druhou*** komponentou je pak běhové prostředí vykonávající příkazy lokálně nebo v distribuovaně v Hadoop clusteru. Pig vznikl v roce 2006 v Yahoo jako alternativa pro psaní časově náročných MapReduce úkolů v Javě a v současnosti Pig kromě Yahoo používá například také LinkedIn, Twitter nebo Nokia.

Jednou z hlavních nevýhod MapReduce jsou dlouhé vývojové cykly i pro jednoduché úkoly. Ad-hoc datová analýza a hledání užitečných informací u velkých objemů dat vyžaduje iterativní přístup a rychlou manipulaci kódu. ***Dotazy napsané v PigLatin vyžadují oproti MapReduce v průměru dvacetkrát méně kódu a běží o 50 % rychleji*** (viz 38).

PigLatin dotaz se skládá ze série operací a filtrů inkrementálně aplikovaných na vstupní data ze kterých je následně generovaný výstup. Tyto operace popisují datový tok dotazu a v běhovém prostředí jsou

přeloženy do MapReduce úkolů, optimalizovány a zkompilovány pro spuštění v Hadoop clusteru (viz 39).

Pig se používá pro zpracování petabajtů dat nejčastěji ***v oblastech ad-hoc analýz, pravidelných reportů, strojového učení a ETL***.

Příklad jednoduchého PigLatin dotazu, který ze seznamu twitter zpráv filtruje pouze ty, které mají jako jazyk nastavenou angličtinu.

```
vstup = LOAD  'hdfs//node/twitter_data';

vystup = FILTER vstup by lang_code EQ  'en';

STORE vystup 'hdfs//node/twitter_data_en';
```

27.9.2 Hive

Apache Hive je ***open source datový sklad pro dotazování*** nad velkým objemem dat v distribuovaném prostředí postavený nad Hadoopem. Hive vznikl jako ***interní nástroj ve Facebooku*** umožňující dotazování a datovou analýzu v HDFS ***pro uživatele s SQL znalostmi***, ale s nedostatečnými zkušenostmi s programováním v Javě (pro tvorbu MapReduce úkolů). V současnosti mezi hlavní sponzory projektu kromě Facebooku patří Google, Yahoo, Microsoft a Twitter (viz 40).

Hive je navržen ***pro uživatelsky jednoduchou datovou sumarizaci***, ***ad-hoc dotazy a analýzy pomocí jazyka zvaného Hive Query Language (HQL)***, který je v podstatě mírně upravené SQL. Hive díky použití známých principů odstraňuje jinak velkou bariéru pro osvojení dotazování nad Hadoopem (viz 41).

HQL dotazy jsou stejně v jako v případě Pig dále ***přeloženy do sekvence MapReduce úkolů***. Hlavní rozdíl mezi Hive a Pig je v tom, že HQL ***využívá syntax známý z relačních databází*** a jednoduchou integraci pomocí ***JDBC***, zatímco PigLatin preferuje přehlednost a sekvenční styl dotazů vhodný zejména pro rozsáhlé a složité dotazy.

Data jsou v Hive ***uloženy do tabulek s pevným schématem*** skládajících se z řádků a sloupců. Tabulky jsou v clusteru rozdělené dle jednoho nebo více ***partition klíčů***. Data v jednotlivých partitionech lze dále rozdělit do tzv. bucketů pro efektivnější vzorkování. Hive obsahuje podporu textový souborů, sekvenčních souborů (key/value pár) a RCF8 souborů (sloupcové uspořádání relačních tabulek v clusteru).

Hive je postaven nad Hadoopem, a proto ***je navržen pro dávkové zpracování velkých objemů dat*** s vysokou odezvou (v řádu minut i u nejjednodušších dotazů). Hive není určen pro zpracovávání online transakcí a neumožňuje dotazy v reálném čase nebo úpravy jednotlivých řádků. Z těchto důvodů není vhodné Hive používat jako přímou náhradu za SQL a relační databáze (viz 42).

Podobnost s SQL je ilustrována ***na následujícím příkladu*** jednoduchého HQL dotazu, který z tabulky obsahující twitter zprávy vybere pouze ty, které mají jako jazyk nastavenou angličtinu.

```
SELECT user, tweet_text FROM twitter_data

WHERE lang_code = 'en';
```

27.10 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k zajištění zpracování a analytiky dat na velkých objemech – *„Big Data“*.

[8] Record Columnar File

- Big Data představují tak ***obrovské objemy*** dat, s takovou potřebnou ***rychlostí*** zpracování a ***variabilitou*** dat, které vyžadují ***nové přístupy, postupy a technologie*** zpracování.
- ***Zdroji pro Big Data*** jsou zejména data ze sociálních sítí, data generována technickými prostředky, zejména v rámci IoT a ve specifických případech i transakční data, např. v retailu.
- ***Architektura Big Data*** zahrnuje ***4 vrstvy***, a to vrstvu datových zdrojů, vrstvu datových transformací (ETL), vrstvu zpracování data a analytickou vrstvu.
- ***Z technologického hlediska*** je charakteristika Big Data v rámci této kapitoly orientována na systém open source ***Hadoop***.
- Řešení Big Data doplňují i ***dotazovací nástroje***, např. Pig a Hive.
- Před rozhodnutím, zda se orientovat ve firmě na Big Data, je dobré ***posoudit zejména reálný objem dat***, a to aktuálně i výhledově. Kritérium by v tomto smysl se mělo pohybovat kolem několika terabytů dat a více.

28. Big Data Analytics

Miloš, M. (KIT, VŠE)

Účelem kapitoly je:

- charakterizovat ***příklady použití*** „*Big Data*“ a jejich možnosti,
- ***analyzovat oblasti použití*** jako vstup pro konkrétní řešení.

Příklady použití Big Data Analytics lze objevit téměř v každém sektoru podnikání, kde se pracuje s rozsáhlými objemy dat. Jednou z hlavních předností Big Data Analytics je fakt, že umožňují hledat odpovědi na otázky, na které se organizace v minulosti ani nenapadlo ptát. Výsledkem mohou být nové a lepší produkty nebo optimalizace stávajících služeb. Analytik Jeff Kelly ve svém Big Data manifestu () uvedl několik příkladů použití Big Data nástrojů a technologií v současnosti v praxi již zavedených a osvědčených.

28.1 Systém doporučení

Řada elektronických obchodů využívá Big Data nástroje pro ***doporučení produktu nebo služby*** spotřebiteli na základě analýzy jeho chování, historie a profilu v reálném čase. Amazon tímto způsobem nabízí související produkty a sociální sít LinkedIn využívá Big Data nástroje pro doporučení možných přátel a známých.

28.2 Názorová analýza

Big Data nástroje umožňují pokročilé ***analýzy textu i nestrukturovaných dat např. ze sociálních médií*** jako jsou Twitter a Facebook. Jednou z pokročilých funkcí je názorová analýza, která v daném textu hledá a analyzuje slova vyjadřující názor (pozitivní a negativní indikátory) tykající se zkoumaného produktu nebo služby.

28.3 Detekce podvodů

V oblasti bankovnictví se Big Data nástroje používají k ***detekci podvodů a krádeží***. Analýzou chování uživatele a historií jeho transakcí lze odhalit transakce, u kterých je vysoká pravděpodobnost podvodu, například kompromitace účtu nebo krádež kreditní karty.

28.4 Analýza marketingových kampaní

V současnosti existuje řada nástrojů pro ***monitoring a kvantifikaci efektivnosti marketingových kampaní***. Big Data kromě standardních funkcí umožňuje tyto analýzy provádět v reálném čase a na rozsáhlejších objemech dat umožňující vyšší granularitu dat.

28.5 Retenční analýza

Analýza zákaznické věrnosti, respektive pravděpodobnosti, že ***zákazník přejde ke konkurenci*** je další z oblastí, kde mnoho podniků využívá Big Data technologie. Pravidelným rozborem zákaznického chování, využívání služeb, výdajů a dalších vzorů chování lze proaktivně zabraňovat zákaznické retenci.

28.6 Analýza sociálních vztahů

Analýza sociálních vztahů a jejich ***vizualizace pomocí sociálních grafů a sociogramů*** pomáhá určit vztahy ***uvnitř sociální skupiny*** a vliv jednotlivců na ostatní členy skupiny. Tyto informace o spotřebitelích jsou pro podniky cenné z toho důvodu, že určují nejdůležitější zákazníky jako ty, kteří nejvíce ovlivňují ostatní ke koupi, a ne ty, kteří sami kupují nejvíce zboží. Pro dosažení úspěchu produktu je tedy nutné přesvědčit právě spotřebitele s velkým vlivem na ostatní.

28.7 Analýza zákaznické zkušenosti

Big Data technologie se také používají k integraci a ***analýze řady různých datových zdrojů o interakci se zákazníkem*** jako jsou call centra, web a sociální média. Kombinace těchto zdrojů dává ucelený a

komplexní pohled na zákaznickou zkušenost a umožňuje lepší porozumění a optimalizaci zákaznické komunikace.

28.8 Monitoring logů

Big Data nástroje a technologie jsou vhodné pro skladování a ***analýzu velkých objemů automaticky generovaných strojových dat*** při provozu počítačové sítě jako jsou síťové a serverové logy a data generovaná dalším hardwarem. Analýza a monitoring těchto dat pomáhá diagnozovat problémy v síti a eliminovat slabé stránky sítě.

28.9 Věda a výzkum

Big Data technologie se často používají k analýze obrovských objemů dat ***v oblasti zdravotnictví a farmacie***, zejména pro prediktivní modelování při vývoji léků a analýze dat z klinických testů.

28.10 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k zajištění datové analytiky na velkých objemech – „*Big Data“*.

- Big Data představují základnu pro ***nejrůznější oblasti datové analytiky***, jako např. přípravu různých doporučení zákazníkům v prodejních aktivitách, detekce podvodů, např. v pojišťovnictví, analýzy zákaznické zkušenosti a další, jak je uvedeno v textu kapitoly.
- Obdobně jako v jiných kapitolách, i v tomto případě lze doporučit, aby si analytici připravili ***sady analytických otázek*** tak, aby řešení směřovalo k pokrytí hlavních problémů firmy. Náměty takových otázek jsou v dokumentu „*AF II.01: Oblasti řízení*“ v podkapitoly x.8, např. „7.8: Řízení prodeje“.

29. Data Driven Company

Halama, J., (KIT, VŠE)

Účelem kapitoly je:

- vymezit ***podstatu*** společností založených na využití velkých objemů dat,
- charakterizovat jejich ***vnitřní organizaci a významné role*** podílející se na řešení analytických úloh a aplikací v prostředí Big Data.,
- definovat ***charakteristiky pracovních týmů*** v oblasti Big Data.

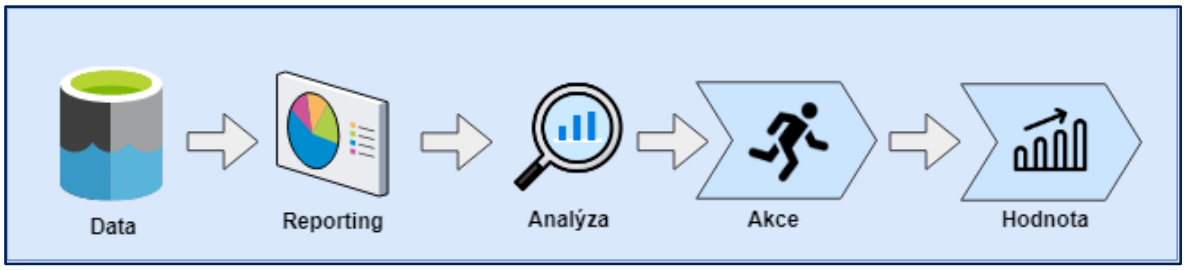
Společnosti zaměřující se na Big Data ***sami sebe označují jako Data Driven***. Velkým problémem, se kterými se takové společnosti potýkají, je však ***kvalita reportů***, chybějící analýza výstupů a ***nedůvěra managementu*** v řízení pomocí dat.

29.1 Definice Data Driven Company

Data Driven Company je taková ***společnost, která upřednostňuje rozhodnutí prováděná na základě dat*** než na osobních pocitech či intuici. Tato rozhodnutí slouží k získání přidané hodnoty pro podnik, oddělení, zaměstnance či zákazníky. U takových rozhodnutí je klíčové, že jsou zaznamenány v datech a lze je zpětně evaluovat.

*"Data-Drivenness is about building tools, abilities, and, most crucially, a culture that acts on **data**. "* (Anderson, 2015)

Samotnou definici jde opět ***znázornit procesem získávání hodnoty*** z dat na obrázku Obrázek 29-1. Tento proces je naprosto ***abstrahován od vazeb na konkrétní infrastrukturu***. Podoba uložiště i konkrétních reportingových nástrojů je vždy vztažena ke konkrétnímu podniku. Analytické postupy jsou také individuální a jsou vztaženy k charakteristice dat a požadovaným výstupům.



Obrázek 29-1: Proces získávání hodnoty z dat. (Zdroj: Dykes, 2010)

29.1.1 Data

Základní podmínkou Data Drivenness je ***vlastnictví dat.*** Průběžně sbírané kolekce dat musí být dostatečně obsáhlé a relevantní pro společnost, která data sbírá. Společnosti řízené daty nestačí pouze data sbírat, ale musí se pozastavit i nad třemi základními charakteristikami oněch dat. Tyto charakteristiky musí data splňovat, aby byla použitelná v dalších částech procesu.

- ***Spojitelné (Joinable):*** Datové kolekce by měly být spojitelné s dalšími kolekcemi v podniku. Díky tomu společnosti mohou změny v datech korelovat s ostatními událostmi v podniku. Tento aspekt dat přináší do podniku nové možnosti datové analýzy. Úplně základní podmínkou „spojitelnosti" je časová známka záznamu. Dle té můžeme provádět nejzákladnější korelaci. Dostupným nástrojem pro spojování dat je Microsoft Excel a jeho funkce VLOOKUP. Ta se hojně užívá v malých firmách vlastnících malé objemy dat. Při velkých objemech dat však Excel není dostačující a firmy vyhledávají systematičtější řešení v podobě relačních databází, NoSQL databází nebo již zmiňovaného Hadoop pro Big Data řešení.

- ***Sdílené (Shareable):*** Pokud společnost vlastní jakékoliv kolekce dat k analytickým účelům, musí být sdílené mezi všemi zainteresovanými zaměstnanci. Data by měla být sdílená v jednom uložišti s dostatečnou ochranou proti ztrátě či úniku dat. Není efektivní data uchovávat v rozdílných uložištích či v krajních případech v textových souborech na jednotlivých zařízeních. Takto sdílené datové kolekce jdou jen těžko spojovat v nástrojích k tomu určených. V dnešní době se užívá hlavně databázových systémů s vlastním uložištěm v Cloudu či distribuovaných uložišť (DFS, HDFS).

  Sdílení dat s sebou nese i řadu nebezpečí, zvlášť když jsou data uložena v cloudu nebo v interní nezabezpečené síti. Některá data v podniku mohou obsahovat utajené či osobní informace. K takovým datům může mít přístup jen úzká skupina zaměstnanců, kteří ví, jak s těmito daty nakládat a předcházet jejich úniku. Samotná bezpečnost je pak totožná se zabezpečením sítě v podniku s použitím autentizačních protokolů jako je LDAP či Kerberos.
- ***Dotazovatelné (Queryable):*** Pokud jsou v podniku kolekce dat, které jsou spojitelné a zaměstnanci k nim mají přístup, tak zde musí být možnost je dotazovat. Dotazování v databázových systémech slouží k získání konkrétních fragmentů relevantních dat k analýze a mezi základní reporting. Způsob dotazování se odlišuje dle databázového systému. Většina standartních databázových systémů využívá SQL (Structured Query Language). U Hadoop distribucí s analytickým enginem Spark je to stejnojmenný jazyk Spark či PySpark. U speciálních typů grafových databází se využívá jazyka Gremlin a Cypher. Možností je mnoho a záleží na preferencích společnosti, dovednostech zaměstnanců a infrastruktuře.

29.1.2 Reporting

Dalším krokem v procesu získávání hodnoty z dat je reporting. Reporting je proces agregace a organizace dat do souhrnů, které reprezentují minulé období a události v něm. Těžko se v čistých reportech hledají souvislosti a ***informace*** cenné k rozhodování. Marketingový ředitel SAP Institute, Jim Davis, popsal 8 stupňů analytiky. První 4 se týkají reportingu. (SAP, 2009)

- ***Standardní reporty***: Odpovídají na základní otázky. Velice často jsou generovány automaticky na pravidelné bázi. Příkladem mohou být měsíční finanční výkazy.
  Co se stalo? Kdy se to stalo?
- ***Ad-hoc reporty*****:** Konkrétnější nejčastěji jednorázové dotazy, které odpovídají na otázky managementu nebo vlastníků dat. Ad-hoc reporty se také využívají u marketingových kampaní či podezření na různé chyby či nedostatky, které by z dat mohly vyplynout. *Kolik? Jak často? Kde?*
- ***OLAP a Data Drill Down***: Pokročilý reporting. Podmínkou pro tento druh reportingu je odpovídající datový model a složitější manipulace s daty. Reporting probíhá v reportingovacích nástrojích (PowerBI, Tableu). Odpovídá na konkrétní a detailní otázky a umožňuje exploraci dat. *Kde přesně je problém? Jak najdu řešení problému?*
- ***Alerting*****:** Alerting je druh reportingu v reálném čase. Upozorňuje na náhlé změny tam, kde je alerting implementovaný a mohlo by tam dojít k problémům, které by znamenaly finanční ztrátu či výpadek systému . Nejčastěji v bankách, u operátorů či ve výrobě.
  Kdy bych měl reagovat? Jaké akce jsou potřeba?

29.1.3 Analýza

Analýza dat transformuje čísla v reportu na cenné informace, vkládá je do kontextu a interpretuje je. Je to další krok v procesu získávání hodnoty z dat. Tento krok je z celého procesu nejvíc zaměřen na kvalitu dat vystupujících z reportu a jejich správné pochopení. Analytici analyzující tato data musí mít doménovou znalost prostředí pro správnou extrakci informací z reportu. Zde jsou popsány další 4 stupně se týkají analýzy samotné.

- ***Statistická analýza*****:** Statistická analýza je základem datové analýzy. Pomocí statistické analýzy společnosti ověřují vznesené hypotézy. Mezi základní nástroje patří popisná statistika,

pravděpodobnostní rozdělení a regresní analýza. Mezi další nástroje patří data mining a shlukování.

- ***Prognostika*:** Prognostika se zaměřuje na trendy v datech a jejich opakování. Analytici tak mohou zjišťovat trendy v jednotlivých oblastech podniku a předpovídat jejich další prezenci. Prognostika se využívá např. u výrobních společností, kde se dá dle dat vyčíst v jakém období je poptávka zboží vyšší a připravit se na něj.
- ***Prediktivní modelování*:** Prediktivní modelování využívá historických dat k vytváření klasifikátorů. Je úzce spjato se strojovým učením, avšak model je vytvářen a řízen lidským faktorem. Užití může být u segmentace zákazníků nebo při zkoumání reakcí na marketingové kampaně. U prediktivního modelování je důležité mít dostatečné množství dat, které obsahují tag. Tag přidává k jednotlivým záznamům příznak, co tento záznam znamená. Tagy jsou svým způsobem metadata. Na základě nich se pak prediktivní model „učí“ klasifikovat nové záznamy.
- ***Optimalizace*:** Nejvyšším stupněm analýzy je optimalizace. Analytici na základě dat a metadat o podniku dokážou najít řešení komplexních problémů v podniku a stanovit doporučení pro jejich řešení. Tato doporučení mohou být vytvářena i pomocí expertních systémů, které jsou plně automatizovány pomocí umělé inteligence.

Mnoho společností však i přes popsané odlišnosti nedokážou rozdělit reporting a analýzu. Z tohoto důvodu Brent Dykes vytvořil tabulku klíčových atributů reportingu a analýzy dat shodnou s tabulkou, která vysvětluje rozdíly mezi těmito pojmy.

Tabulka 3 : Klíčové atributy reportingu a analýzy dat. (Zdroj: Dykes, 2010)

Reporting	Analýza
Deskriptivní popis dat	Preskriptivní popis dat
Co?	Proč?
Pohled do minulosti	Pohled do budoucnosti
Vznáší dotazy	Odpovídá na dotazy
Data -> Informace	Data + Informace -> Vhledy
Reporty, dashboardy a alerting	Zjištění, doporučení a předpovědi
Bez kontextu (či s omezeným kontextem)	Přidání kontextu

Z tabulky je patrné že proces analýzy dat na reporting přímo navazuje a využívá jeho výstupů. Klíčový je pak řádek 4, kde je popsáno, že reporting vytváří otázky na základě dat a analýza tato data interpretuje a na otázky odpovídá.

29.1.4 Akce

Předposlední částí procesu je ***akce***. Je to nejdůležitější část v procesu a je prováděna managementem společnosti či vedoucími jednotlivých oddělení. Ti hotové výsledky a interpretaci dat z analýz užívají k rozhodování. Na základě dat se např. upravuje výroba, vývoj, náklady či marketingová komunikace. Jak již bylo zmíněno, společnost řízená daty musí mít kulturu schopnou datům naslouchat a využívat je. Bez těchto předpokladů není možno společnost nazývat Data Driven a veškerý reporting i analytika je tím pádem zbytečná.

29.1.5 Hodnota

Poslední částí procesu je hodnota. Nenastává vždy, neboť ***je podmíněna správností rozhodnutí*** a je organickým sub procesem společnosti reagující na změny vyvolané akcí na základě dat. Je to přidaná hodnota získaná z dat či rozhodnutí provedeného na základě dat. Takový efekt se vyhodnocuje zpětně, kdy se předchozí stav porovnává s nynějším a evaluuje se efekt. Hodnota může být vyjádřena finančními ale i nefinančními ukazateli (spokojenost zaměstnanců nebo věrnost zákazníků).

29.2 Organizace v Data Driven Company

K Data Driven Company neodmyslitelně patří i její zaměstnanci a jejich organizace do týmů. Jedním z nejdůležitějších článků jsou datově-analytické týmy, pod které spadá značná část operativy spojené s daty. Do procesů Data Driven Company však vstupují i ostatní týmy a pozice ve společnosti. Tato část nastiňuje personální část společností řízenou daty.

Základní dělení datově-analytických pozic provedl v roce 2013 Jeanne G. Harris ze společnosti Accenture. Ten rozdělil datově-analytické pozice do 4 skupin a do nich zasadil jednotlivé pozice. Toto rozdělení je nastíněno v tabulce.

Tabulka 4 : Dělení pozic v Data Driven Company. (Zdroj: Harris, 2013)

Data businessperson	Data creative	Data developer	Data researcher
Business analysts	Data-visualization specialists	Data scientists	Statisticians
Data analysts		Data engineers	Quants
Accountants and Financial Analysts			

Ačkoli jsou zde uvedeny zdánlivě heterogenní skupiny, všechny pozice jsou úzce provázány a musí spolu komunikovat veškeré změny a události spojené s daty. ***Dovednosti a vědomosti se mezi těmito skupinami také prolínají***.

29.2.1 Data businessperson

Tato skupina zaměstnanců je přímo spojena ***s významem dat k businessu***. Business v tomto případě znamená oblast, kterou data pokrývají. Zaměstnanci v této skupině jednají s koncovými zákazníky datových analýz a reportingu. Tato skupina zaměstnanců se nachází téměř ve všech společnostech nehledě na zaměření (finanční oddělení) a užívané technologie.

29.2.2 Business analysts

Business analytici mají za úkol ***mapovat procesy společnosti a hledat způsoby jejich možné optimalizace***. Při vývoji nového produktu se zaměřují na jeho architekturu, vztah k procesům společnosti, a hlavně možné datové výstupy pro datové oddělení. Slouží jako komunikační bod mezi managementem a vývojáři.

29.2.3 Data analysts

Nejvíce rozšířená datově-analytická pozice v datově orientovaných společnostech. Datoví analytici musí mít ***dostatečný technický základ*** pro vytváření reportů a analýz (např. SQL, Python, R) ale také velmi dobrou doménovou znalost zkoumané oblasti. Ve společnostech jsou většinou datoví analytici úzce zaměřeni právě na svojí zkoumanou oblast a s tím se váží i rozdílné požadavky na technické znalosti. Některé společnosti provádějí analýzu dat v nástrojích Microsoft Excel, jiné zase využívají dotazování do Hadoop. Analytici také potřebují velice ***dobré komunikační schopnosti,*** aby dokázali závěry a výstupy svých analýz dostatečně a srozumitelně popsat.

29.2.4 Accountants and Financial Analysts

Účetní a finanční analytici jsou specifickou skupinou zaměstnanců, která se zaměřuje ***na finanční výkaznictví a ukazatele*** podnikové výkonnosti. Víc než technická znalost dotazovacích jazyků a vizualizačních nástrojů se od nich očekává ***znalost finančního výkaznictví země***, ve které společnost operuje (např. IFRS) a k němu přidružené finanční ukazatele jako je rentabilita, likvidita či marže produktového portfolia. Výstupy od finančních analytiků jsou nejčastějším požadavkem managementu společností.

29.2.5 Data creative

Datoví kreativci jsou poměrně novým pojmem, avšak zahrnují zaměstnance ***specializující se na datovou vizualizaci*** a nástroje k tomu určené. Tyto pozice jsou spojovány s nástroji pro Business Intelligence a webovou prezentaci dat. V běžné podnikové praxi společnosti zaměstnávají pouze vizualizační speciality, ojediněle vývojáře vlastních řešení pro vizualizaci. Opět se tato skupina zaměstnanců může vyskytovat i v Non-Big Data společnostech.

29.2.6 Data Visualization Specialists

Specialisté pro vizualizaci dat musí být ***obeznámeni o charakteru dat, datovém modelu, a hlavně možnostech interpretace*** oněch dat. Zásadní podmínkou je dobrá komunikace mezi zadavatelem (analytici, management) a zprostředkovatelem datového modelu. Za vstupní znalosti se předpokládá znalost vizualizačních nástrojů (PowerBI, Tableau) či knihoven zprostředkovávajících webovou prezentaci dat (D3.js).

29.2.7 Data developer

Datoví vývojáři jsou již blíže ***spjati s Big Data a infrastrukturní částí*** datově řízené společnosti. Zpracovávají data jako objekt, ale nepotřebují porozumět jejich obsahu a významu. Zajišťují datovou dostupnost a kvalitu. Rozdělují se do dvou základních podskupin.

29.2.8 Data engineers

Datoví inženýři jsou ***jedni z nejdůležitějších článků procesu transformace*** a přenosu dat mezi bodem *A* a *B*. Spravují datové pumpy, analytické nástroje v infrastruktuře a uložená data v databázovém systému či na HDFS. Jsou vybaveni velice dobrými technickými znalostmi jako jsou programovací jazyky (Java, Python, Scala) a architektonickými vzory pro ukládání dat. Zaměřují se také na nástroje pro automatizaci reportingu a jejich validaci.

29.2.9 Data scientists

Data Scientist je poměrně novým pojmem v oblasti datového inženýrství. Tato skupina zaměstnanců se zpravidla zaměřuje na ***vývoj a implementaci statistických modelů a klasifikátorů*** do produkčních systémů. Jde o datové produkty a nástroje. Data Scientisti by měli skvěle ovládat statistické metody, algoritmy pro strojové učení a vedle toho ovládat jazyky pro datové modelování (R, Python) a přidružené statistické knihovny (TensorFlow, Pandas). Zaměstnanci s tímto zaměřením se nejčastěji objevují ve společnostech zaměřujících se na umělou inteligenci.

29.2.10 Data researcher

V doslovném překladu se *datoví výzkumníci* zabývají ***aplikováním statistických metod a nových objevů*** nad daty či nástroji data zpracovávající. Tato pozice je vědeckého charakteru a zaměstnanci zkoumají a testují nové metody či přístupy k manipulaci s daty.

29.2.11 Statisticians

Statistici se na rozdíl od Data Scientistů zabývají ***aplikací statistických metod na data v celém podniku***. Vyhodnocují průzkumy, dotazníky nebo vytvářejí přehledy segmentace zákazníků. Spektrum používaných nástrojů je široké. Od vytváření kontingenčních tabulek v Microsoft Excel po využívání jazyka R. Statistici také zaštiťují úvodní analýzy kvality datových zdrojů ve společnostech.

29.2.12 Quants

Kvantitativní analytici pracují ***na algoritmizační optimalizaci a kvantitativní analýze***. Takoví zaměstnanci mívají silný matematický základ a objevují se často v bankách, kde pracují v oblasti risk managementu a finančním sektoru obecně. V oblasti Big Data se kvantitativní inženýři zaměřují na optimalizaci manipulace s daty. Stojí za implementací optimalizačních algoritmů jako je HyperLogLog nebo Count-Min Sketches.

29.2.13 Full-Stack pozice

Rozdělení pozic v Data Driven Company dle Harrise není žádné dogma. Mnoho (hlavně menších) společností zaměstnává tzv. *Full-Stack* zaměstnance. Ti ***se zabývají více oblastmi najednou***. Tento termín je hodně populární u webových vývojářů, kteří dokážou samotný web navrhnout, vyvinout mu front-end i back-end a celý tento web následně obhospodařovat na webovém serveru. V datových odděleních to může být datový analytik se schopností obhospodařování datových pump a manipulací s daty v databázovém systému. Takoví zaměstnanci se hlavně vyskytují v decentralizovaných datových týmech.

29.2.14 Ostatní pozice

Do fungování Data Driven Company spadá daleko více profesí než jen ty datově orientované. Pro správné fungování všech procesů spojených s extrakcí důležitých informací pomocí datové analýzy je třeba řada na první pohled ne zcela souvisejících profesí.

29.2.15 Tým infrastruktury

Infrastrukturní tým zaštiťuje ***provoz všech databázových systémů či celého Hadoop Clusteru***. Řeší neočekávané výpadky a požadavky na změny konfigurace jednotlivých serverů. Aktivně monitoruje využití výpočetních strojů v podniku a komunikuje tyto údaje datovým inženýrům. Zároveň se tento tým stará o ***zabezpečení daných databázových systémů a serverů***.

29.2.16 Právní oddělení

Právnické oddělení je důležité u společností sbírajících data spojené s uživateli produktů či návštěvních webových stránek. Taková data spadají, pod již zmiňované ***nařízení o ochraně osobních údajů***. Toto vymezení nejlépe znají zaměstnanci právního oddělení a ti určují, jaká data společnost může sbírat, jaká ne, a jak s nimi mohou zaměstnanci nakládat. Zabývají se ochranou před únikem dat a určují i retenci – dobu, na jak dlouho daná data může společnost uchovávat a využívat k analýze.

29.2.17 Oddělení informační bezpečnosti

Společnosti využívají podněty a vhledy z dat k různým účelům. Jedna z nich může být prezentace investorům, stakeholderům nebo jsou využívány k marketingovým kampaním. Společnost si musí být jistá že data, která publikuje veřejnosti, ***jsou z relevantních zdrojů a nejsou vymyšlená*** či misinterpretována. Proto je zde oddělení informační bezpečnosti, které ***takové výstupy kontroluje a schvaluje***. Toto oddělení je spojovacím bodem mezi datovým oddělením a PR (Public Relations).

29.2.18 Vývojáři

Pokud společnost získává data z klientských aplikací nebo webových stránek, je důležitá úzká ***spolupráce s vývojovým týmem***, který generuje dané ***datové výstupy***. Ty jsou připojeny konektorem k databázovému systému. Vývojáři úzce spolupracují s business analytiky na identifikaci datových zdrojů a jejich spojením s datovým uložištěm.

29.2.19 Produktové týmy

Pokud má společnost produktové týmy zabývající se vývojem a provozem služeb či produktů generujících zisk, je to automaticky jeden ze zdrojů požadavků na ***analýzy jednotlivých částí systému či uživatelské interakce***. Produktové týmy využívají výstupy daných analýz ke ***zlepšení jejich služeb a produktů nebo k zjištění pozice na trhu*** vůči konkurentům i z jiných než finančních pohledů.

29.2.20 Konzultanti

Společnosti mohou některé služby či části infrastruktury ***outsourcovat.*** To znamená, že spolupracuje s konzultanty (v některých případech je může zaměstnávat) daných společností, od kterých služby přebírá. Typické je to pro ***poskytovatele cloudových služeb či Big Data infrastruktury***. Konzultanti

společnosti však mohou být i vysokoškolští výzkumní pracovníci, kteří jednotlivým týmům a oddělením pomáhají s výzkumem.

29.3 Organizace týmů

Datově orientované společnosti mají tři základní možnosti, jak rozdělit zaměstnance zabývající se daty. Mohou tvořit ***centralizovaný datový tým***, pod který spadají všichni zaměstnanci. Ti pak přerozdělují jednotlivé řešené problémy a agendu mezi sebe. Druhou možností jsou ***decentralizované týmy spadající pod jednotlivé projekty či správu datových zdrojů***. Třetí poměrně novou možností je ***hybridní model*** týmů.

29.3.1 Centralizované

Centralizovaný tým ***se skládá ze všech zaměstnanců*** pracujících s daty. ***Výhodou*** takového týmu je bezesporu ***kolektivní znalost***, která se stále rozšiřuje. Další výhodou je ***vysoká míra zastupitelnosti*** jednotlivých členů ***a rychlost učení*** nových. Mezi ***nevýhody*** však patří v některých případech až ***příliš široká agenda*** a s tím spojená ***nižší znalost detailů*** daných datových zdrojů. V týmu se udržuje tolik informací ze všech částí společnosti a projektů, že se ***řešení případných problémů stává méně efektivní*** a delegace práce je složitá.

29.3.2 Decentralizované

Decentralizované týmy jsou ***přímo spojeny s jedním projektem, datovým zdrojem či produktem***. Tým obsahuje zaměstnance pokrývající potřebnou agendu a ***může být složen nejen z datově orientovaných zaměstnanců***. Mezi ***výhody*** patří ***vysoká míra znalosti detailů*** projektu či datového zdroje. Mezi ***nevýhody*** patří zastupitelnost členů. Malé decentralizované týmy ***se skládají většinou z velmi malého počtu lidí a pozice se nepřekrývají***. Proto ***při absenci*** jednotlivých členů týmu ***vznikají problémy***, které se řeší dočasným zastoupením či přechodem jiných zaměstnanců na *Full Stack* pozice.

29.3.3 Hybridní

Hybridní model se užívá spíše ***ve velkých společnostech vyvíjejících více produktů*** pod různými projekty. Ve společnosti jsou určité ***centrální týmy*** (data scientists, datový inženýři) a ***ostatní pozice*** (statistici, analytici) ***patří do konkrétních týmů a projektů***. Tato kombinace ***eliminuje přetížení*** a špatnou delegaci práce. ***Centrální týmy poskytují služby ostatním*** datově-analytickým zaměstnancům a projektům.

29.4 Zdroje

1. BEYER, Mark, LANEY, Dougles. *The Importance of 'Big Data': A Definition* [online] 21. 6. 2012 [cit. 6. března 2013]. Dostupné z: http://www.gartner.com/DisplayDocument?ref=clientFriendlyUrl&id=2057415
2. LANEY, Dougles. *Highlights from GartnerChat on Big Data* [online]. 28.1.2012 [cit. 6. března 2013]. Dostupné z: http://blogs.gartner.com/doug-laney/highlights-from-todays-gartnerchat-on-big-data/
3. SHORT, James, BOHN, Roger a BARU Chaitnaya. *Report on Enterprise Server Information*. University of California, San Diego [online]. duben 2011 [cit. 7. března 2013]. Dostupné z: http://hmi.ucsd.edu/howmuchinfo_research_report_consum_2010.php
4. GANTZ, John, REINSEL, David. *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digi tal Shadows, and Biggest Growth in the Far East* [online]. prosinec 2012 [cit. 7. března 2013]. Dostupné z: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
5. HILBERT, Martin, LÓPEZ, Priscila. *The world's technological capacity to store, communicate, and compute information*. Science magazine [online] 10. 2. 2011 [cit. 10. března 2013]. Dostupné z: http://www.sciencemag.org/content/332/6025/60
6. META GROUP. *3D Data Management: Controlling Data Volume, Velocity, and Variety.* [online]. 6. 2. 2001 [cit. 7. března 2013]. Dostupné z: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
7. ZIKOPOULUS, Paul C, et al. *Understanding Big Data*. New York: McGraw-Hill Osborne Media, 2011. s 5. ISBN 978-0-07-179053-6.

8. MANYIKA, James, et al. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute [online] červen 2011 [cit. 18. března 2013]. s 15. Dostupné z: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

9. ZIKOPOULUS, Paul C, et al. *Understanding Big Data*. New York: McGraw-Hill Osborne Media, 2011. s 6 - 7. ISBN 978-0-07-179053-6.

10. DUMBILL, Edd. *Volume, Velocity, Variety: What You Need to Know About Big Data*. Forbes magazine. [online] 19. 1. 2012 [cit. 23. dubna 2013]. Dostupné z: http://www.forbes.com/sites/oreillymedia/2012/01/19/volume-velocity-variety-what-you-need-to-know-about-big-data/2/

11. SICULAR, Svetlana. *Gartner's Big Data Definition Consists of Three Parts*. Forbes magazine. [online] 27. 3. 2013 [cit. 23. dubna 2013]. Dostupné z: http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/

12. SNOW, Dwaine. *Adding a 4th V to BIG Data - Veracity*. [online] 16. 7. 2012 [cit. 23. dubna 2013]. Dostupné z: http://dsnowondb2.blogspot.cz/2012/07/adding-4th-v-to-big-data-veracity.html

13. IBM Big Data Team. *IBM Big Data Platform* [online] květen 2011 [cit. 29. dubna 2013]. Dostupné z: http://www-01.ibm.com/software/data/bigdata/

14. MANYIKA, James, et al. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute [online] červen 2011 [cit. 18. března 2013]. s 5. Dostupné z: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

15. MCGUIRE, Tim, MANYIKA, James a CHUI, Michael. *Why Big Data is the new competitive advantage.* Ivey Business Journal. [online] srpen 2012 [cit. 10. května 2013]. Dostupné z: http://www.iveybusinessjournal.com/topics/strategy/why-big-data-is-the-new-competitive-advantage#.UoJFF-F-TJJ

16. PROFFITT, Brian. *Turbine company knows which way the wind blows at your house.* IT World. [online] 12. 6. 2012 [cit. 10. května 2013]. Dostupné z: http://www.itworld.com/big-datahadoop/280783/big-data-meets-big-wind?page=0,1

17. KELLY, Jeff. *Big Data: Hadoop, Business Analytics and Beyond.* Wikibon. [online] 16. 3. 2013 [cit. 13. května 2013]. Dostupné z: http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond

18. *CERN Brochure.* European Organization for Nuclear Research. [online] červenec 2010 [cit. 14. května 2013]. Dostupné z: http://cds.cern.ch/record/1278169/files/CERN-Brochure-2010-006-Eng.pdf?version=1

19. EVELSON, Boris. *The Forrester Wave: Enterprise Business Intelligence Platforms.* Forrester Research Inc. [online] 31. 7. 2008 [cit. 14. května 2013]. Dostupné z: http://www.forrester.com/The+Forrester+Wave+Enterprise+Business+Intelligence+Platforms+Q3+2008/fulltext/-/E-RES44724?isTurnHighlighting=false&highlightTerm=enterprise%20Business%20Intelligence%20Platforms

20. ANUGANTI, Venu. *Typical Big Data Architecture.* [online] 30. 11. 2012 [cit. 18. května 2013]. Dostupné z: http://venublog.com/2012/11/30/typical-big-data-architecture/

21. BANGE, Carsten, et al. *Big Data Survey Europe*. BARC Institute. [online] únor 2013 [cit. 21. května 2013]. Dostupné z: http://www.teradata.com/big-data-survey-europe/

22. LOHR, Steve. *For Today's Graduate, Just One Word: Statistics.* The New York Times. [online] 6. 8. 2009 [cit. 21. května 2013]. Dostupné z: http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=1&

23. MANYIKA, James, et al. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute [online] červen 2011 [cit. 18. března 2013]. s 105. Dostupné z: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

24. DOLÁK, Ondřej. *Big data: Nové způsoby zpracování a analýzy velkých objemů dat.* SystemOnLine. [online] 2011 [cit. 3. června 2013]. Dostupné z: http://www.systemonline.cz/clanky/big-data.htm

25. WHITE, Tom. *Hadoop: The Definitive Guide.* O'Reilly Media, California. 2009. s 10. ISBN 978-0-596-52197-4.

26. GHEMAWAT, Sanjay, GOBIOFF, Howard a LEUNG, Shun-Tak. *The Google File System.* [online] říjen 2003 [cit. 16. června 2013]. Dostupné z: http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/gfs-sosp2003.pdf
27. NYBERG, Chris. *Sort Benchmark* [online]2009 [cit. 16. června 2013]. Dostupné z: http://sortbenchmark.org/
28. WHITE, Tom. *Hadoop: The Definitive Guide.* O'Reilly Media, California. 2009. s 56. ISBN 978-0-596-52197-4.
29. BORTHAKUR, Dhruba. *HDFS Architecture Guide*. [online] 2009 [cit. 19. června 2013]. Dostupné z: http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html
30. ZIKOPOULUS, Paul C, et al. *Understanding Big Data*. New York: McGraw-Hill Osborne Media, 2011. s 56 . ISBN 978-0-07-179053-6.
31. ZIKOPOULUS, Paul C, et al. *Understanding Big Data*. New York: McGraw-Hill Osborne Media, 2011. s 57 . ISBN 978-0-07-179053-6.
32. BORTHAKUR, Dhruba. *HDFS Architecture Guide*. [online] 2009 [cit. 20. června 2013]. Dostupné z: http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html
33. HBase Development Team. *HBase Architecture Guide.* [online] 2010 [cit. 26. srpna 2013]. Dostupné z: http://hbase.apache.org/book/architecture.html#arch.overview
34. IBM Big Data Team. *What is HBase?* [online] 2012 [cit. 26. srpna 2013]. Dostupné z: http://www-01.ibm.com/software/data/infosphere/hadoop/hbase/
35. DEAN, Jeffrey, GHEMAWAT, Sanjay. *MapReduce: Simplified Data Processing on Large Clusters*. [online] 2008 [cit. 27. srpna 2013]. Dostupné z: http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/mapreduce-osdi04.pdf
36. ZIKOPOULUS, Paul C, et al. *Understanding Big Data*. New York: McGraw-Hill Osborne Media, 2011. s 58 . ISBN 978-0-07-179053-6.
37. HDFS Development Team. *MapReduce Architecture Guide.* [online] 2012 [cit. 27. srpna 2013]. Dostupné z: http://hadoop.apache.org/docs/r1.0.4/mapred_tutorial.html
38. WEIL, Kevin. *Hadoop, Pig, and Twitter (NoSQL East 2009).* [online] 30. 10. 2009 [cit. 3. září 2013]. Dostupné z: http://www.slideshare.net/kevinweil/hadoop-pig-and-twitter-nosql-east-2009
39. Apache Pig Development Team. *Getting Started With Pig.* [online] 4. 2. 2013 [cit: 3. září 2013]. Dostupné z: http://pig.apache.org/docs/r0.11.1/start.html
40. Apache Hive Development Team, *Apache Hive sponsors*. [online] 2009 [cit. 5. září 2013]. Dostupné z: http://www.apache.org/foundation/thanks.html
41. LAM, Chuck. *Hadoop in Action.* Manning Publications, Connecticut. 2010. s 247. ISBN 978-1-935182-19-1
42. LEVERENZ, Lefty. *Getting Started With Hive.* [online] 19. 5. 2013 [cit. 10. září 2013]. Dostupné z: https://cwiki.apache.org/confluence/display/Hive/Tutorial
43. FEINLEIB, David. *Big Data Landscape.* [online] 2012 [cit. 10. září 2013]. Dostupné z: http://www.bigdatalandscape.com/
44. KELLY, Jeff. *Big Data Vendor Revenue and Market Forecast 2012 - 2017*. [online] 2012 [cit. 18. září 2013]. Dostupné z: http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2012-2017
45. FLOYER, David. Big Data Database Revenue and Market Forecast 2012 - 2017. [online] 2012 [cit. 18. září 2013]. Dostupné z: http://wikibon.org/wiki/v/Big_Data_Database_Revenue_and_Market_Forecast_2012-2017
46. Cloudera [online], poslední aktualizace 6. listopadu 2013 [cit. 10. listopadu 2013], Wikipedie. Dostupné z: http://en.wikipedia.org/wiki/Cloudera
47. Cloudera Team. *Cloudera's Distribution of Hadoop.* [online] 2013 [cit. 20. září 2013]. Dostupné z: http://www.cloudera.com/content/cloudera/en/products/cdh.html
48. MapR Team. *MapR's Distribution of Hadoop.* [online] 2013 [cit. 20. září 2013]. Dostupné z: http://www.mapr.com/products/product-tour
49. Hortonworks Team. *Hortonworks Data Platform*. [online] 2013 [21. září 2013]. Dostupné z: http://hortonworks.com/products/hdp/
50. Amazon. *What is Amazon EMR.* [online] 2011 [cit. 21. září 2013]. Dostupné z: http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-what-is-emr.html

51. Amazon. *Amazon EMR Features* [online] 2011 [cit. 21. září 2013]. Dostupné z: http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-features.html
52. Windows Azure. *Windows Azure HDInsight Service.* [online] 2013 [cit. 4. října 2013]. Dostupné z: http://www.windowsazure.com/en-us/manage/services/hdinsight/introduction-hdinsight/
53. Qubole. *Qubole Data Service.* [online] 2013 [cit. 8. října 2013]. Dostupné z: http://www.qubole.com/qubole-data-service/
54. Goswami, S., Dāśa, A. and Mukherjee, S., 2019. *Big Data Simplified.* 1st ed. Pearson Education India. ISBN 978-93-539-4150-5.
55. Big data: The next frontier for innovation, competition, and productivity. McKinsey. Global management consulting. McKinsey & Company [online]. Dostupné z: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation
56. Definition of Big Data - Gartner Information Technology Glossary. [online]. Copyright ©2021 [cit. 25.02.2021]. Dostupné z: https://www.gartner.com/en/information-technology/glossary/big-data
57. Discover the 4 V's of Big Data | OpenSistemas. OpenSistemas - [online]. Copyright © OpenSistemas 2016 [cit. 25.02.2021]. Dostupné z: https://opensistemas.com/en/the-four-vs-of-big-data/
58. What Is Big Data? | Oracle. [online]. Copyright © 2021 Oracle [cit. 25.02.2021]. Dostupné z: https://www.oracle.com/big-data/what-is-big-data/
59. Big Data: What it is and why it matters | SAS. [online]. Copyright © 2021 SAS Institute Inc. All Rights Reserved. [cit. 25.02.2021]. Dostupné z: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html
60. Sources of big data: Where does it come from? | CloudMoyo. CloudMoyo | Enabling digital transformation with Cloud and AI [online]. Copyright © 2020 [cit. 01.03.2021]. Dostupné z: https://www.cloudmoyo.com/blog/data-architecture/what-is-big-data-and-where-it-comes-from/
61. White, Tom. Hadoop: the definitive guide. 3. vyd.,O'Reilly, 2012. ISBN 978-05-965-2197-4.
62. Apache Hadoop. Apache Hadoop [online]. Copyright © 2006 [cit. 07.03.2021]. Dostupné z: https://hadoop.apache.org/
63. Apache Hive TM. Apache Hive TM [online]. Copyright © 2011 [cit. 07.03.2021]. Dostupné z: https://hive.apache.org/
64. BAKER, Mark. Cluster computing white paper. arXiv preprint cs/0004014, 2000. [cit. 08.03.2021]. Dostupné z: https://arxiv.org/ftp/cs/papers/0004/0004014.pdf
65. What Is a Hadoop Cluster?. Databricks - The Data and AI Company [online]. Copyright © Databricks 2021. All rights reserved. [cit. 08.03.2021]. Dostupné z: https://databricks.com/glossary/hadoop-cluster
66. Apache Hadoop Architecture Explained (In-Depth Overview). phoenixNAP [online].Copyright © 2021 Copyright phoenixNAP [cit. 08.03.2021]. Dostupné z: https://phoenixnap.com/kb/apache-hadoop-architecture-explained
67. Anderson, C. (2015) *Creating a Data-Driven Organization*. 1st edition. Sebastopol: O'Reilly Media, Incorporated. ISBN 978-14-919-1691-9.
68. SAS, Eight Levels of Analytics, (Cary, NC: SAS Institute, Inc., 2008), [cit. 15.03.2021]. Dostupné z: https://www.datasciencecentral.com/profiles/blogs/eight-levels-of-analytics-for-competitive-advantage
69. Dykes Brent, *Reporting vs. Analysis: What's the Difference*?. Adobe Blog [cit. 16.03.2021] 2021 [online] Dostupné z: https://blog.adobe.com/en/publish/2010/10/19/reporting-vs-analysis-whats-the-difference.html#gs.vsp7gl
70. Harris, J. G., E. Craig, and H. Egan, "*How to Organize Your Analytical Talent*," Accenture [online]. 2021 [cit. 17.03.2021]. Dostupné z: https://www.criticaleye.com/inspiring/insights-servfile.cfm?id=1834
71. CDP Private Cloud. *Cloudera Product Documentation* [online]. 2019 [cit. 22.03.2021]. Dostupné z: https://docs.cloudera.com/cdp-private-cloud/latest/overview/topics/cdppvc-overview.html
72. Facebook Engineering. *HyperLogLog in Presto: Faster cardinality estimation* . [online]. 2018 [cit. 26.03.2021]. Dostupné z: https://engineering.fb.com/2018/12/13/data-infrastructure/hyperloglog/

73. Chambers, B. and Zaharia, M., 2018. *Spark: The Definitive Guide*. 1st ed. Sebastopol: O'Reilly Media, Incorporated. ISBN 978-14-919-1221-8.

29.5 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k vymezení a hodnocení konceptu „*Data Driven Company*“.

- „*Data Driven Company*“ je taková ***firma, která upřednostňuje rozhodnutí prováděná na základě dat*** než na osobních pocitech či intuici.
- Pro vymezení firmy jako „*Data Driven Company*“ jsou podstatné následující komponenty řízení v rámci ***procesu získávání hodnoty*** (v návaznosti na komponenty v rámci anatomie firmy):
  - data, resp. datové zdroje, datové struktury,
  - reporting, jeho obsah a formy,
  - analýzy a odpovídající analytické úlohy,
  - akce, aktivity, úlohy,
  - hodnoty řešení.
- Obdobně jako v jiných kapitolách, i v tomto případě lze doporučit, aby si analytici připravili ***sady analytických otázek*** tak, aby řešení směřovalo k pokrytí hlavních problémů firmy. Náměty takových otázek jsou v dokumentu „*AF II.01: Oblasti řízení*“ v podkapitoly x.8, např. „7.8: Řízení prodeje“.

I) Podniková analytika a cloud computing

[31] Cloud computing pro BI
[32] Využití Cloud řešení a Cloud BI
[33] Možnosti Cloud BI současnosti

Cloud computing a jeho systém jeho služeb je i v oblasti podnikové analytiky a zejména business intelligence uplatňován v praxi ve stále větší míře. Oddíl proto charakterizuje v širší míře tyto možnosti a zahrnuje tyto hlavní body:

- ***Vymezení cloud computingu pro BI*** a podnikovou analytiku obecně, řešení jednotlivých komponent BI v tomto prostředí, architektonické přístupy k implementaci BI v cloudu, různé varianty BI v cloudu a doporučení k jejich výběru.
- ***Skutečné využití BI v cloudu***, analýzy činitelů pro přechod BI do cloudu, analýzu hlavních rozdílů cloudových řešení BI a řešení on premise.
- ***Hodnocení možností využití BI v cloudu*** v současnosti, analýzy důvodů pro taková řešení, architektury BI v cloudu a jejich podstatné charakteristiky včetně dopadů na nejvýznamnější komponenty.

31. BI Cloud computing

Borovec, J. (NDC Group, a.s.)

Účelem kapitoly je:

- vymezit ***možnosti řešení a provozu***, BI aplikací s využitím technologií a služeb cloud computingu,
- definovat ***podstatné efekty a omezení*** BI aplikací pracujících v prostředí cloudu,
- definovat ***architekturu cloudového řešení a jeho rozdíly*** oproti klasickým architekturám,
- charakterizovat služby ***SaaS*** ve vazbě k aplikacím BI.

31.1 Vymezení BI v cloudu

Pod pojem BI v cloudu je možné zahrnout více různých řešení. lze se setkat jak s ryzími ***aplikacemi vytvářenými od počátku pro cloud*** (zejména různé startupy), tak i s ***„transformovanými" klasickými řešeními*** (Evelson, 2012). Tato původně klasická řešení jsou podobně implementována na platformách IaaS/PaaS a jejich distribuční model je tak pro koncového zákazníka stejný jako u ryzích cloud řešení. Existují řešení, která staví na dřívějších postupech tvorby BI, pro které se jeví výhodné využít dříve vytvořeného konceptu a aplikovat filosofii cloudu na původní řešení. V hojném počtu vznikají i ***řešení nová***, která ***kombinují různé přístupy***, ale v zásadě jsou schopny poskytnout zcela nové cloud BI řešení. V kontextu cloudového BI lze identifikovat ***tři hlavní směry***, kterými poskytovatelé jdou:

1. ***Transformace klasických*** BI řešení do cloudu
2. ***Nové*** a komplexní řešení v cloudu
3. Řešení ***pro vizualizaci dat*** v cloudu

Pro pracovníka, respektive koncového uživatele dat pracujícího s BI je vždy výstupem report či jiná analýza, ať už je řešení jakéhokoliv typu, a takové BI obecně splňuje následující ***charakteristické vlastnosti***:

- je určeno ***pro front-line uživatele*** (bez specializovaných IT znalostí),
- funkcionalita je dostupná ***v reálném čase***,
- dostupnost výstupů BI ***z různých zařízení/platforem*** (webová stránka, webová aplikace, mobilní zařízení/aplikace atd.),
- jako ***zdroj*** slouží ERP, CRM, emaily, webové stránky, dokumenty, webové platformy/aplikace/služby a jejich obsah atd.

31.2 Efekty a přínosy cloud BI pro kvalitu řízení podniku a IT

- ***Provozní a investiční úspory***, snížení implementačních nákladů.
- ***Agilita*** řešení.
- Značná míra ***flexibility a škálovatelnosti*** řešení.
- ***Přístupnost*** odkudkoliv prostřednictvím Internetu.
- Kratší doba ***implementace,*** rychlejší a snazší nasazení.
- Přístup ***k nejmodernějším technologiím*** v ekonomicky přijatelné úrovni.
- Není potřeba ***specializovaná infrastruktura***.
- ***Jednoduchost*** využití.
- Tvorba ***analýz a reportů***.

31.3 Otázky, problémy a omezení spojené s cloud BI

- ***Investiční výdaje*** jsou licence na klientský software.

- ***Customizace*** aplikací je relativně obtížnější.
- Vyšší výdaje za internetovou ***konektivitu,*** objem přenášených dat roste.
- ***Bezpečnost*** (bezpečnost identit uživatelů, bezpečnost informací, bezpečnost infrastruktury).
- ***Závislost na poskytovateli*** cloudových služeb.
- ***Správa dat*** třetí stranou.
- ***Obnova dat*** je náročnější pod cizí kontrolou.
- ***Značný objem dat***, který je potřeba přenést přes internet.
- ***Rychlost přístupu*** k datům.
- ***Legislativní aspekty***.

31.4 Architektonický rozdíl klasické a cloudové BI

Tato část se zabývá ***rozdílným (ne)využíváním komponent*** v cloudovém modelu oproti klasickému. Cloud poskytuje nový přístup pro dostupnost výkonu/úložného prostoru, proto již jsou některé komponenty, které dříve napomáhaly provádět náročné operace, nadbytečné. Zásadní změnou oproti klasickému řešení je, že poskytovatelé využívají ***pro datové sklady***, výpočetní výkon a pro další komponenty ***veřejné cloudy*** (Amazon AWS, Microsoft Azure apod.) (Hapl, 2010). Tento vliv cloud computingu pak má ***zásadní význam na výkon*** a cenu BI řešení oproti klasickému.

31.4.1 ETL a cloud

Při zavádění BI jsou jednou z nejdůležitějších částí datové pumpy. Filosofie cloudu aplikovaná na ETL nástroj v mnohém využije jeho výhod a ***tuto komponentu přímo předurčuje ke cloudovému využití***. ETL totiž nemusí pracovat neustále, nýbrž přenáší data kupříkladu každou noc. V tomto ohledu je ***škálovatelný*** (řízený potřebou) výkon ETL nástroje obzvláště výhodou. V době, kdy ETL nepracuje, zdroje provozovatel cloudu využívá u jiných řešení, a naopak v případě zmiňovaného každodenního přenosu je od poskytovatele ***poskytnut maximální výkon*** (je vhodné pro aplikování platebního modelu ***Pay as you go***).

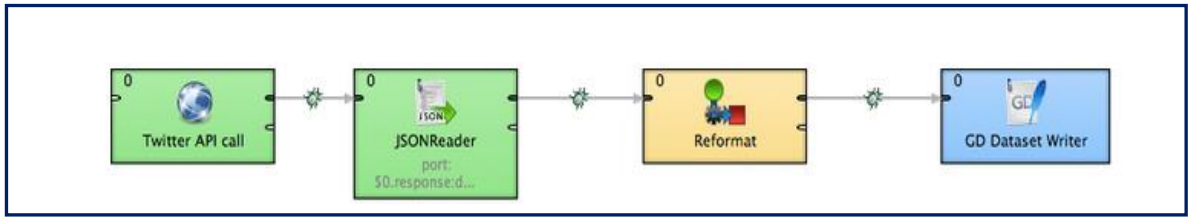
Cloudové ETL by mělo být schopno ***získat data jak z cloudových, tak i z klasických zdrojových systémů***, a to napříč platformami a protokoly. (Thomson, 2013) poukazuje na základní ***funkcionalitu každého ETL*** řešení, které musí nabízet následující ***nástroje, platformy, aplikace a funkce***:

- Podporovat ***standardní protokoly***
  - JDBC, FTP, …
  - XML, HTML, TXT, …
- ***Konektory na běžně používané aplikace a platformy*** (obvykle pomocí API)
  - Amazon, Azure, …
  - Salesforce.com,
  - Dropbox, SkyDrive,
  - Facebook, Twitter,
  - a mnohé další.
- ***Plánovač transformací***.
- ***ETL funkce:***
  - Agregování, řazení, spojování heterogenních a nestrukturovaných dat
  - Logování

Z pohledu cloudu se pro ETL otevírají ***nové možnosti***, za prvé zmiňovaný ***škálovatelný výkon*** a za druhé ***dostupnost API*** u většiny cloudových řešení sloužících jako zdroje pro BI, které velmi usnadňuje vzájemné ***propojení nesourodých aplikací***. Tohoto trendu využívají také další poskytovatelé řešení, tentokrát tzv. konektorů, které se soustředí na propojení různých aplikací obvykle právě pomocí API. ***Konektory*** pak velmi ***ulehčují práci při integraci řešení***, neboť propojení mezi aplikacemi je již připraveno a je nutné ho pouze nastavit. Většina ***ETL*** nástrojů již v základu ***nabízí množství podporovaných***

protokolů, platforem a jiných aplikací, proto použití konektoru není obvykle nutné. Blíže se konektorům věnuje část 31.6 Konektory.

Jako jedna z komponent architektury BI je i ETL nabízeno buďto jako ***celek komplexního řešení***, nebo jako ***samostatné řešení*** pro spojení s datovým skladem zvoleného poskytovatele. Dosud však nejsou k dispozici žádné cloudové ETL poskytujícího webového klienta pro definování transformací. Všechna nalezená řešení nabízejí pouze tenké desktopové klienty, pomocí kterých se transformace v cloudu nastavují. ETL se v těchto aplikacích tvoří obdobně jako v klasických řešeních, což může demonstrovat Obrázek 31-1, který vizualizuje datovou pumpu z řešení GoodData získávající data z Twittrovského účtu, které transformuje do datového souboru.



Obrázek 31-1: GoodData ETL proces z Twittru (Zdroj: GoodData, 2014)

Cloudová ETL se liší od klasických zejména ***většími možnostmi napojení*** na ostatní aplikace. Lze říci, že tomuto faktu nahrává progresivnější vývoj cloudových aplikací a faktu, že uživatelé mají neustále přístup k nejnovější verzi. ***Poskytovatel*** tak v případě popularizace některého jiného řešení ***rychle zareaguje a vytvoří nové propojení*** na tuto aplikaci. Čekat na nasazení nové verze u klasického řešení je mnohonásobně časově náročnější.

Jedním ze ***slabých míst*** cloudového řešení je ***rychlost přenosu mezi aplikacemi***, která může být limitujícím faktorem při průběhu procesu ETL. Těžko si lze představit, že by mohl být problém například při nahrávání do úložiště Amazonu, avšak v případě některých zdrojových aplikací, zejména v případě napojení na on-premise nástroj, může mít rychlost přenosu dat zásadní vliv na rychlost ETL pumpy. V takovém případě hrozí i přes výhody filosofie cloudu, že datové pumpy budou naopak pomalejší než lokální transformace. Lze však očekávat, že cloudové ETL je využíváno zejména pro použití z cloudu do cloudu.

Z textu je patrné, že cloudové ETL se kromě lepšího výkonu (v případě odpovídající konektivity), širší podpory formátů, aplikací a platforem pro připojení, se nikterak zásadně neliší od klasických ETL nástrojů.

31.4.2 Operační datový sklad v cloudu

Operační datové sklady (dále ODS) se hojně vyskytují v architekturách klasických BI systémů. ***V ryze cloudovém BI*** se však s použitím této komponenty ***nelze téměř setkat***. V klasickém BI tato komponenta integruje podobně jako DWH data z různých zdrojů, ale na rozdíl od relativně statického DWH se v ODS data dynamicky mění. ODS slouží především k jednoduchým a rychlým dotazům nad aktuálními daty, což je vhodné pro ad-hoc reporting, případně pro využití výstupů z ODS pro jiné podnikové aplikace, což je podle (Hrabáček, 2006) jeden z nejvýznamnějších přínosů celého BI.

Tak jak je ODS prezentováno, je již dnes pravděpodobně ***integrováno do celkové architektury cloudového BI***. V dané rovině se však zmiňuje o využití ODS v cloudovém prostředí kombinující PaaS a IaaS platformu Microsoft Azure, přičemž zde ODS pracuje stejným způsobem jako u klasického řešení s tím rozdílem, že se tak neděje lokálně.

31.4.3 Datové tržiště v cloudu

Datová tržiště, na rozdíl od datových skladů kombinujících data celé společnosti, jsou menší sklady zaměřené na konkrétní předmět nebo oddělení. Jsou používány k rychlým analýzám jen na základě svého zaměření. ***Jejich použití*** se jeví stejně jako v případě ODS ***již v cloudu nepotřebné*** a lze se s ním setkat jen v klasických a hybridních řešeních. (Goul, 2009) si všímá, že ***někteří poskytovatelé*** cloudového BI prostředí název datového tržiště ***ponechaly, avšak pouze ve významu původní funkcionality analýzy dat pro konkrétní oddělení***, jde však pouze o virtuální rozlišení každého oddělení v daném uživatelském prostředí. Data jsou ***čerpána ze stejného zdroje***, nikoliv z jiného fyzického datového tržiště.

31.5 BI v cloudu bez použití ETL a DWH

Vzhledem k celkovému kontextu a obecně známé teorii se jeví možnost BI řešení bez použití ETL jako nemožná. Avšak v rámci konkurence se ***několik firem*** zaměřilo i na tuto oblast a nabízí ***implementaci vlastního řešení i bez použití ETL***. Jak uvádějí oficiální stránky jednoho takového řešení, příčina tkví v časové náročnosti implementace ETL, která podle (Penichter, 2009) obvykle spotřebuje 70 až 80 procent času při zavádění BI, respektive DWH.

Příkladem je společnost InetSoft (InetSoft, 2014), která takovéto řešení nabízí na základě svého ***patentu Data Block technology***, jež ***nahrazuje ETL a DWH takzvaným Data mashupem***[9]. Tato společnost spatřuje zbytečnou snahu považovat BI za enormní platformu a vytvářet kvůli každému nasazení velké projekty, které zbytečně generují mnoho nepotřebných procesů, tvorbu mnoha ETL pump a datových skladů z čehož plyne, že se stávají mnohem nedostupnější pro široký okruh především menších společností.

Data mashup data ***transformuje a manipuluje s nimi***, pouze pokud jsou vyžadována. To je dle oficiálních materiálů InetSoftu považováno za hlavní výhodu BI bez použití ETL. Na webových stránkách společnosti není snadné najít zmínku o jakékoliv nevýhodě tohoto způsobu, avšak složitěji dostupný dokument[10] přiznává, že u složitějších a datově náročných operací lze očekávat ***o něco pomalejší návratnost výsledků***. ETL je zde přirovnáváno k velké nákladní lodi, zatímco Data mashup k letadlu – „letadlo je rychlejší a lehčí, ale unese méně nákladu".

Další zmiňovanou výhodou je ***možnost rychle zakomponovat nový zdroj dat*** (např. Excelový sešit z externího zdroje) a za pomoci drag-and-drop ho okamžitě přidat do reportingů. Nevýhoda dotazování se do databáze má být zmenšena za pomoci mezipaměti cache. Bližší srovnání výhod a nevýhod zobrazuje Tabulka 31-1.

Tabulka 31-1 – Srovnání výhod a nevýhod DWH s ETL a Data mashupu. Zdroj: (Bord, 2010)

DWH a ETL		**Data mashup**	
Výhody	Nevýhody	Výhody	Nevýhody
Rychlý běh	Složitě se mění	Ad hoc změny a flexibilita	Pomalý běh
Zpracování velkého zdroje dat	Dlouhý vývoj a přizpůsobení	Rychle zpracuje data z různých zdrojů	Lze zpracovat menší množství dat

Rozhodně tak nelze očekávat, že způsob ***získávání dat přímo z produkčních databází je možné aplikovat na všechna řešení***, vždyť nenáročnost na produkční databáze je jednou z hlavních deviz BI, proto je tento způsob více než překvapivý. Stejně jako ETL dokáže technologie Data Mashup ***kombinovat různé datové zdroje***, pracuje s databázemi, tabulkami, soubory atd. Řešení nabízí stejné možnosti jako ostatní BI řešení, obyčejný uživatel BI tak zřejmě rozdíl mezi BI s ETL a bez něho ani nepozná.

Integrace dat pomocí ETL obvykle tvoří nejnáročnější část celého BI projektu, z toho důvodu se jeví vynechání tohoto prvku jako zajímavá alternativa. Je zapotřebí si však uvědomit, že síla BI spočívá právě v nově a efektivně konsolidovaných datech. (Penichter, 2009) v souvislosti s BI řešením, které integruje různorodá data z databází transakčních systémů, textových a tabulkových souborů, webových stránek případně mnohých dalších a které si bere data z různě rozmístěných úložišť eventuálně od odlišných dodavatelů, rozhodně ***nelze uvažovat o nepoužití ETL***. V takovém případě si nelze ani představit, že by se analytický nástroj mohl nezávisle, kdykoliv odkazovat například na data uložená v externích transakčních databázích. V daném případě lze očekávat problémy na obou stranách – nesprávný report na straně BI a na druhé straně například přetížený databázový systém. Na základě výše zmíněných faktů se lze domnívat, že ***BI bez ETL a DWH může existovat, musí však být dobře zanalyzováno***, s kolika daty se bude pracovat a jak často budou generovány reporty, potažmo real-time reporty, aby nedošlo k přetížení produkční databáze.

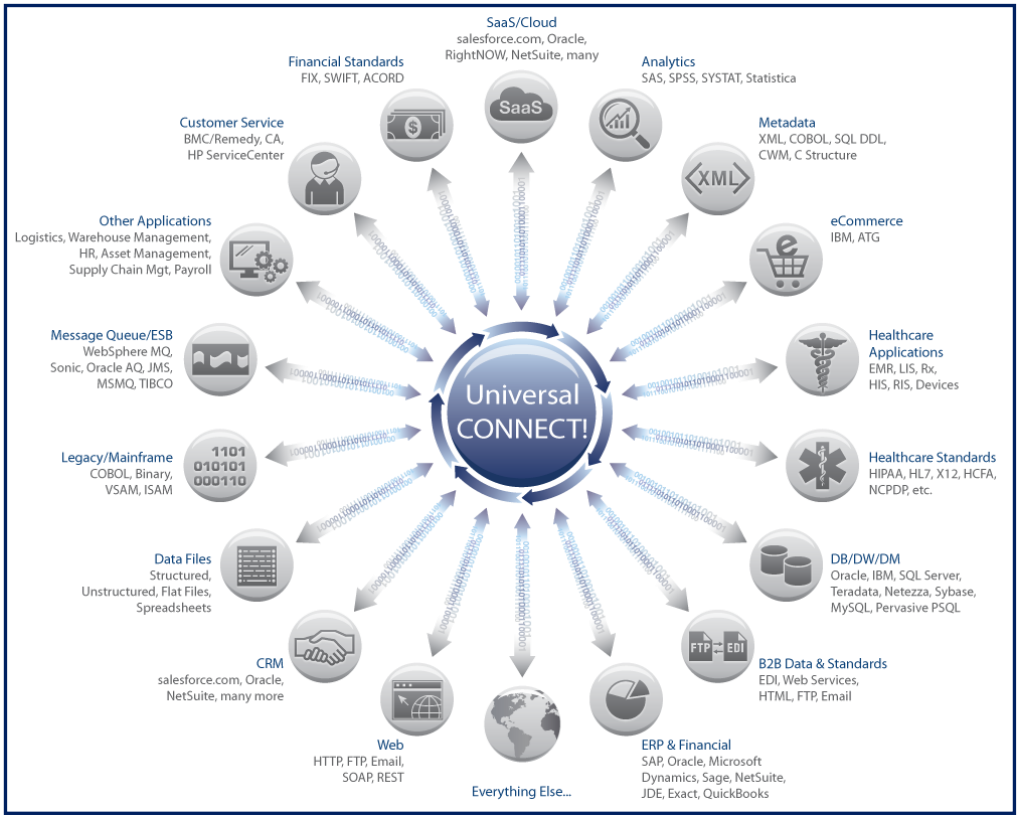
31.6 Konektory

Podnikové systémy nabízí velké množství typových řešení, nicméně tato řešení je obvykle možné téměř neomezeně (záleží na řešení) ***rozšiřovat, resp. propojovat s dalšími s využitím jejich API*** rozhraní,

[9] Více o problematice Data Mashupu lze najít na oficiálních stránkách: http://www.inetsoft.com/info/enterprise_mashup_server/ ,

[10] Studie o Data mashupu v řešení InetSoft: http://www.inetsoft.com/dl/docs/index.jsp?wp=Data_Mashup.pdf

jež je nabízeno v naprosté většině řešení. ***Klasická řešení*** jsou také obvykle ***vytvářena s ohledem na potenciální součinnost*** s dalšími systémy, proto se lze setkat s případy kombinací několika řešení od různých poskytovatelů. Lze očekávat, že pokud poskytovatel nabízí více typů řešení, ***vzájemná integrace je vyřešena přímo poskytovatelem***. V případě různých poskytovatelů však může nastat problém. Existuje však potenciál v přípravě různých ***konektorů (middlewarů)*** propojujících jednotlivá řešení. Zákazníkovi tak nabízí velice snadnou možnost, jak integrovat různorodé systémy rychle a pohodlně. Pro ***příklad*** lze jmenovat například řešení ***Pervasive connector***. Za pomoci tohoto řešení lze relativně pohodlně integrovat různé podnikové aplikace, soubory a jiná data, jak ukazuje Obrázek 31-2.



Obrázek 31-2 – Integrační konektor PERVASIVE. Zdroj: (Pervasive, 2014)

Toto řešení je velmi ***výhodné v případě více vzájemně propojených aplikací***, neboť přímé propojení se v případě více jak 3 aplikací stává velice komplikované (počet vazeb je roven N x (N-1)). V případě využití konektoru je počet vazeb roven počtu aplikací a zejména při přidání/ubrání/změně jedné aplikace je nutné upravit pouze jednu vazbu, a nikoliv vazby ostatní.

31.7 Architektury a implementace cloud BI

V této části jsou analyzovány ***varianty nasazení komponent BI do či mimo cloudové prostředí***. Cílem je posoudit vhodnost jednotlivých řešení a poukázat, ***za jakých okolností se každá z variant uplatní***. Další část ***doporučuje, na které vlivy je nutné brát zřetel*** při rozhodování mezi klasickým a cloud řešením a říká, který vliv nahrává klasickému řešení a který cloudovému. Text však pouze naznačuje, které okolnosti vyhovují více klasickému, a které cloudovému řešení.

31.7.1 Architektury cloudového BI

BI řešení jsou velice komplexní systémy, jejichž ***architektura se skládá ze tří hlavních komponent*** – z ***DWH*** uchovávající data pro analýzu, ***ETL*** nástroje, který získává a transformuje data do podoby potřebné v DWH a ze ***samotného analytického nástroje*** pro analýzu a reporting dat. Spojení klasického a cloudového řešení nabízí různé ***kombinace, které komponenty*** (DWH, ETL, analytické nástroje) budou ***dostupné lokálně***, a které naopak ***cloudově***. Současný trh nabízí ***různé variace řešení***, napomáhají tomu mimo jiné i tzv. ***veřejné cloudy*** na platformách IaaS, či PaaS (Amazon AWS, Microsoft

Azure), na které ***lze přesunout klasická řešení*** a využívat tak výhod cloudu, byť s nástrojem určeným pro klasická řešení.

Především ***datové sklady je vhodné*** umístit ***do veřejného cloudu***, neboť elasticita cloudu poskytuje „neomezený" prostor pro rostoucí DWH. Toto rozdělení je nutné před implementací řešení patřičně zvážit. Níže prezentované ***tabulky ukazují kombinace***, se kterými se lze setkat. Je nutné připomenout, že činnosti při zavádění jednotlivých řešení se liší na základě výchozího řešení, variantě řešení, požadavků a rozsahu cílového řešení.

Podstatnou otázkou pro následující varianty je ***umístění zdrojových systémů*** (ERP, CRM atd.), neboť výsledná varianta je touto skutečností velmi zásadně ovlivněna. Jen pro ilustraci, postrádá smysl mít zdrojové systémy v cloudu, DWH na vlastní infrastruktuře a BI nástroj opět v cloudu. Při volbě z následujících variant tak má zásadní vliv právě umístění zdrojových systémů. Jednotlivé ***varianty se zabývají pouze komponentami BI***, nicméně ***umístění zdrojových IS v cloudu*** lze brát jako ***jeden z motivátorů přesunu*** do cloudu.

31.7.1.1 Varianta 0

Všechny komponenty BI s využitím klasickém modelu.

Tabulka 31-2 – Architektury Varianta 0

Varianta 0	klasické	cloudové
ETL	x	
DWH	x	
BI	x	

Tuto variantu můžeme brát jako ***výchozí*** pro varianty následující. Od toho stavu se také budou odvíjet patřičné ***kroky, vedoucí ke změně*** na jinou z variant.

31.7.1.2 Varianta 1

Tabulka 31-3 – Architektury Varianta 1

Varianta 1	klasické	cloudové
ETL	x	
DWH	x	
BI		x

DWH ponechat jako klasické řešení stejně tak ETL a pouze analytickou část BI přivést do cloudu.

Varianta č. 1 se jeví ***vhodná pro již existující klasické BI*** řešení, kde DWH je správně vytvořen a kde dochází k přenesení BI komponenty do cloudu (třeba z důvodu lepší dostupnosti výsledků – webové prostředí, mobilní telefony). Toto řešení je relativně ***snadné*** na konfiguraci, neboť je nutné pouze napojení nového BI na původní DWH, konfigurace BI a seznámení uživatelů s novým řešením.

31.7.1.3 Varianta 2

Tabulka 31-4 – Architektury Varianta 2

Varianta 2	klasické	cloudové
ETL	x	
DWH		x
BI		x

DWH spolu s BI umístit do cloudu a nástroj ETL využívat jako klasické.

Varianta č. 2, podobně jako č. 1, je vhodná pro relativně ***rychlé přenesení BI i DWH komponent*** do cloudu a s ETL stále běžícím on-premise. Výhodou tohoto řešení je právě ***využití původního konceptu ETL***. Takové řešení poskytuje například společnost Romana Staňka GoodData[11], která nabízí buďto vlastní ETL on-premise nástroj CloudConnect Designer nebo využití vlastního ETL nástroje. Lze očekávat, že po úpravách cílových tabulek bude možné využít původní ETL v téměř nezměněné podobě. Pokud původní datové zdroje zůstávají nezměněné a on-premise, jeví se tato varianta pro

[11] Na oficiálních stránkách společnosti a v představení řešení není snadné poznat, že společně s řešením není nabízeno cloudové ETL. Zájemce se tuto skutečnost dozví až při návštěvě vývojové sekce stránky https://developer.gooddata.com

částečný přechod do cloudu jako výhodná. Tato varianta využívající cloudu je ***vhodná pro rostoucí společnost*** se zvyšujícími se nároky na množství ukládaných a analyzovaných dat.

31.7.1.4 Varianta 3

Tabulka 31-5 – Architektury Varianta 3

Varianta 3	klasické	cloudové
ETL		x
DWH		x
BI		x

Všechny komponenty BI do cloudu.

Varianta č. 3 využívá všech komponent v cloudu a její využití lze obecně doporučit pro ***nově zaváděné BI řešení***, kde není nutno migrovat data z původních DWH, tak i v případě, kdy se v cloudu, již nachází nebo do cloudu migrují i další informační systémy se zdrojovými daty.

Stejně jako v případě modulů u klasických IS, tak ***lze kombinovat i různé cloudové komponenty*** od různých dodavatelů, avšak v tomto případě existuje snad jen argument nižší ceny a lepší funkcionality některé z komponent. V každém případě je nutné do srovnání ***započítat i náklady na integraci***, neboť nižší cena každé komponenty by mohla být snadno vykoupena vyššími náklady na vzájemnou integraci (pokud by integrace vůbec byla možná). Ke vzájemnému propojení mohou sloužit různé konektory. V případě, že jsou zvoleny všechny komponenty od jednoho dodavatele, lze očekávat již plně integrované řešení, a tudíž i snadnější implementaci. Mnozí ***poskytovatelé také nabízejí vlastní konektory*** na řešení od konkurence[12].

31.7.2 Klasické nebo cloud řešení

Rozhodnutí je složité, protože ***oba přístupy nabízí výhody a nevýhody, které je nutno jednotlivě vyhodnocovat*** a na jejich základě určit, jakým způsobem ovlivní práci, jaká rizika přinesou, jak naplní očekávání a jak budou finančně výhodná. Toto jsou pouze některé ***faktory ovlivňující rozhodnutí*** mezi oběma způsoby. Situaci navíc komplikuje fakt, že ony výhody a nevýhody jsou velice často protichůdné, na druhou stranu tak lépe zapadnou do filosofie každé společnosti. (Claybrook, 2012) poukazuje na tyto protichůdné vlastnosti obou řešení (Tabulka 31-6).

Tabulka 31-6 – Vlastnosti cloudového a klasického řešení. Zdroj (Claybrook, 2012)

	Cloudové	Klasické
Cena/ náklady	Platba pouze za to, co je využíváno formou Pay as you go, za uživatele, za měsíc	Náklady na pořízení HW, SW licencí, administrátory, prostory, elektrickou energii, klimatizace atd.
Customizace	Customizace velice omezená	Customizovatelné, záleží však na řešení
Hardware	O HW a SW se stará poskytovatel	Zákazník se stará o vlastní HW infrastrukturu, včetně základní SW vybavení
Bezpečnost	Přístup přes internet zvyšuje riziko	Bezpečnost je řízena lokálně, data jsou uchována lokálně
Přístup	Přístup pomocí internetu – webového prohlížeče, mobilního telefonu	Přístup z klientů, případně vzdálenou plochou či VPN. Složité pro mobilní zařízení.
Integrace	Složitější integrace s lokálními zdroji dat – MS Excel apod.	Připraveno k spolupráci s lokálními aplikacemi
Správa	Poskytovatel je zodpovědný za správu systému, včetně jeho dat	Společnost si sama spravuje systém i data

[12] Poskytovatel cloudového BI řešení Birst nabízí propojení na tyto systémy: Google Analytics, SAP BW, MS Analysis Service Cubes, Hyperion, MS SQL Server, Oracle, DB2, MySQL, Teradata, Sybase IQ, Generic JDBC, Amazon Redshift a další (Zdroj: http://www.birst.com/product/technology/data-sources).

(Claybrook, 2012) pak navrhuje ***přiřadit důležitost k jednotlivým vlastnostem*** a na základě toho provést prvotní hodnocení. Dále pak doporučuje ***uvažovat následující fakta a otázky***:

- Věnovat důkladnou pozornost tomu, co poskytovatelé klasického nebo cloudového řešení nabízí a slibují.
- Nabízí poskytovatel řešení veškerou potřebnou funkcionalitu?
- Analyzovat cenu – (není vždy na prvním místě).
- Je třeba uvažovat komplexně nad bezpečností, customizací, ovládáním, vyhovění legislativě atd.
- Znát potřeby pro kapacitu úložiště, přenos dat, počet uživatelů.
- Pokud možno, nové řešení nejprve vyzkoušet (u cloudových služeb jsou obvykle dostupné plně funkční trial verze na omezenou dobu).

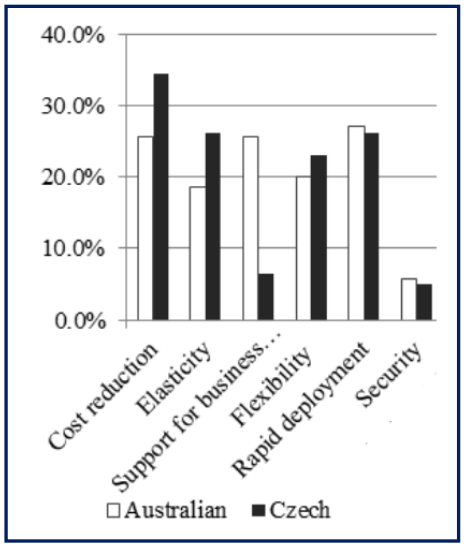
ProfiBricks (Profitbricks, 2014) přidává v úvahu ***následující aspekty***:

- Analyzovat Total cost of ownership (TCO) u zvažovaných řešení. Nutno znát přesné platební podmínky, za co se platí.
- Zhodnotit, jakým způsobem je řešena podpora zákazníka (jak dlouhá je reakční doba).
- Věnovat pozornost Service level agreement (SLA), dostupnost, komunikace, sankce.
- Získat co nejvíce referencí.

31.7.3 Motivace přechodu do cloudu

Motivací k přechodu do cloudu mohou být ***potenciální přínosy***. Pro podporu těchto tvrzení je možné prezentovat výstup ze společného průzkumu australské Technické univerzity, Sydney a české Vysoké školy ekonomické v Praze v českých a australských společnostech (Margaris & Feuerlicht, 2012)13. Výzkum ukázal ***hlavní společné motivátory přechodu do cloudu***: snížení nákladů, elasticitu a flexibilitu. Zatímco v českých společnostech je hlavní motivátor snížení nákladů, australské společnosti pohlížejí na danou oblast komplexněji a cena zde není mandatorním činitelem.

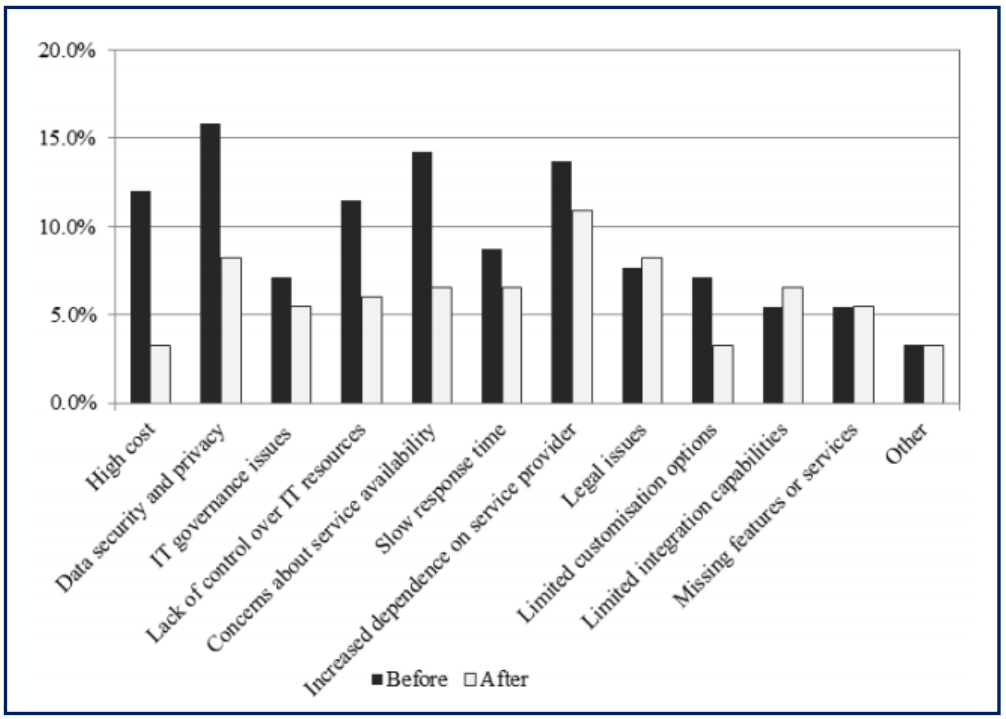
Průzkum nabízí i pohled na ***obavy***, které české společnosti pociťovaly před a jak se tyto obavy změnily po zavedení cloudového řešení. Bariéry australských společností víceméně korespondují s těmi českými, z tohoto důvodu je zde prezentován pouze graf českých společností[14].



Obrázek 31-3 – Motivátory pro přechod do cloudu. Zdroj (Margaris, a další, 2012)

[13] Pozn. k průzkumu: Průzkum byl proveden v obecné rovině cloudových distribučních modelů SaaS, IaaS, PaaS a jakýchkoliv služeb, jež mohou být poskytovány. V každé zemi byl hodnocen vzorek cca 100 společností, přibližně polovina společností v obou zemích měla více než 250 zaměstnanců; v České republice tvořilo 47% procent firem se zaměřením na IT, v Austrálii to bylo procent 17%.

[14] Bariéry australských společností je možné najít ve zdrojovém odkazu u Obrázek 31-4 – Bariéry českých společností pro přechod do cloudu Zdroj: .



Obrázek 31-4 – Bariéry českých společností pro přechod do cloudu Zdroj: (Margaris & Feuerlicht, 2012)

Dalším ***motivátorem*** může být ***migrace zdrojových systémů*** (ERP, CRM, …) do cloudu, v tomto ohledu se jeví jako přínosné přenést/vytvořit i celé BI založené na cloudu. Zobecnění toho, ***proč přecházet do cloudu*** a čeho se případně obávat ukazuje následující ***SWOT analýza***.

Tabulka 31-7 – SWOT analýza motivace přechodu do cloudu (částečně převzato od (Bibi, Katsaros, & Bozanis, 2012))

S **Silné stránky** • Nízká počáteční investice • Snadné zavedení • Jednoduchá údržba • Škálovatelnost (počet uživatelů, množství dat) • Online přístup • Úbytek technických specialistů	**W** **Slabé stránky** • Dostupnost • Spolehlivost • Omezená customizace a konfigurace • Migrace dat • Změna procesů/firemní kultury • Závislost na poskytovateli
O **Příležitosti** • Elasticita, škálovatelnost • Rychlá doba uvedení na trh • Různé typy placení (pay per user/as you go,…) • Přístup pomocí mobilních zařízení	**T** **Hrozby** • Výběr cloudové poskytovatele • Zabezpečení • Důvěrnost, integrita a dostupnost dat • Právní stránka – uložení dat napříč právními rámci • Problémy s kompatibilitou • Žádná garance návratu investice

31.7.4 Migrace mezi cloudy

Významnou otázkou pojící se s využíváním cloudu je ***možnost migrace mezi poskytovateli***, hlavním elementem tohoto procesu je ***zpětné získání všech dat***. Často se uvádí jako významná nevýhoda cloudu uzamknutí se k jednomu poskytovateli. V kontextu BI je toto riziko nutno podstoupit v každém případě, ať už se jedná o klasické či cloud řešení vždy jsme podle využitých komponent nějakým způsobem omezení.

Důležité je mít ***smluvně ošetřeno, kdo je vlastníkem dat***. Tyto informace lze získat z ***Obchodních podmínek*** („*Terms and conditions*“) od poskytovatele služby. ***Zákazník by měl být vždy vlastníkem dat***, souhlasí-li s převedením vlastnického práva na poskytovatele, vystavuje se tak možném komplikacím spojené s právní stránkou věci a data jsou vystavena hrozbě zneužití, neboť s nimi může nakládat i někdo jiný.

31.8 Výběr varianty a implementace BI

V části 31.7.1 byly navrženy jednotlivé varianty řešení s kombinacemi cloudových a klasických komponent BI. V této části jsou specifikovány důležité ***aspekty,*** které by měly být brány v úvahu ***při volbě vhodné varianty a její implementaci***. Každá varianta poskytuje výhodu v určité oblasti a generalizovat, která je ze čtyř zmíněných nejlepší, se jeví velice obtížně. Co se týče výběru varianty ***lze za významné považovat např. tyto aspekty***:

- dostupnost prostředků (HW, resp. financí),
- potřeba konkurenční výhody (cloudové řešení jsou obvykle standardizovaná řešení a velkou konkurenční výhodu nenabízejí),
- citlivosti dat (zvážení do jaké míry je vhodné citlivá data např. o zákaznících a zakázkách přesunout do cloudového úložiště),
- náročnost migrace dat,
- dostupnost dat a propustnost sítě/vzhledem k množství dat

Implementace BI řešení se může ***lišit podle použitých architektur*** (pro komplexní řešení to jsou různé varianty nasazení) a komponent (např. implementace pouze komponenty pro vizualizaci dat SaaS). Vzhledem k širokému rozsahu těchto možností ***implementace BI pro PaaS, IaaS a SaaS*** jsou pro zjednodušení charakterizovány ***pouze dvě možnosti***, a to:

- přechod z klasického na cloudové řešení
- a zavedení nové cloudové komponenty pro vizualizaci dat.

Přechod z existujícího klasického BI na cloudové je ***náročný projekt***. Zejména v případě robustního komplexního řešení se všemi komponentami architektury BI. V tomto případě je nutné mít tým specialistů, který převede databáze a nakonfiguruje nový BI systém. V tomto případě se tak cloudová implementace od klasické liší pouze jinou formou a je ***velmi časově náročná***.

Vizualizační cloudové BI ve formě SaaS může být ***implementováno*** naopak ***velice rychle*** a zde nastupuje výhoda cloudu. Lze očekávat, že i v případě zavádění klasického vizualizačního BI by bylo ***třeba několik specialistů***, cloudové BI však lze ***získat*** v podstatě jednoduchou ***registrací na internetových stránkách provozovatele*** vybraného cloud BI řešení. Začíná-li společnost na zelené louce a nejlépe získává všechny služby od jednoho poskytovatele, nasazení je velice jednoduché. V takovém případě se lze domnívat, že k využívání skutečně postačí pouhá registrace/zaplacení licence. Nezačíná-li se na zelené louce, zavedení i pouhého vizualizačního BI si vyžádá určitý časový fond pro propojení na datové sklady a jiné zdroje a opětovně se nadefinují reporty a dashboardy.

31.9 Zdroje

Bibi, Stamatia, Katsaros, Dimitrios a Bozanis, Panayiotis. 2012. Business Application Acqusition: On-Premise or SaaS-Based Solutions? [Online] 5. 2012. [Citace: 5. 10. 2013.] doi: 10.1109/MS.2011.119. Ročník 29. číslo 3. s. 86-93.

Bord, Mike. 2010. Data Mashups - presentation. *InetSoft Technology Corp. .* [Online] 5. 5 2010. [Citace: 5. 2 2014.] http://www.slideshare.net/MikeBord/inetsoft-self-learning-data-mashups-may10.

Claybrook, Bill. 2012. On-premises vs. SaaS: Making the choice. *SearchCloudApplications.* [Online] 30. 7 2012. [Citace: 29. 1 2014.] http://searchcloudapplications.techtarget.com/tutorial/On-premises-vs-SaaS-Making-the-choice.

Evelson, Boris. 2012. To Be (To Cloud) Or Not To Be (Not To Cloud) BI. *Boris Evelson's Blog.* [Online] 16. 2 2012. [Citace: 3. 2 2014.] http://blogs.forrester.com/boris_evelson/12-02-16-to_be_to_cloud_or_not_to_be_not_to_cloud_bi.

GoodData. 2014. Preparing a Data Loading Process. *GoodData.* [Online] 2014. [Citace: 22. 3 2014.] https://developer.gooddata.com/getting-started/tutorial/preparing-data-loading-process.

Goul, Michael. 2009. Internal Data Cloud Sandboxes: The New Data Marts. *Ebiz.* [Online] 11. 5 2009. [Citace: 10. 3 2014.] http://www.ebizq.net/topics/soa_management/features/1185.html.

Hapl, Martin. 2010. Implementace SaaS BI Platformy v Cloudu. *GoodData, s.r.o.* [Online] 2010. [Citace: 9. 3 2014.] http://si.vse.cz/archive/proceedings/2010/hapl-implementace-saas-bi-platformy-v-cloudu.pdf. SYSTEMS INTEGRATION 2010. s. 190 - 197.

Hrabáček, Jan. 2006. Operační datový sklad nabízí integraci znalostí. *SystemOnline.* [Online] 2006. [Citace: 15. 2 2014.] http://www.systemonline.cz/business-intelligence/operacni-datovy-sklad-nabizi-integraci-znalosti.htm.

InetSoft. 2014. InetSoft Technology: Business Intelligence Without ETL or a Data Warehouse. *InetSoft Technology Corp. .* [Online] 2014. [Citace: 5. 2 2014.] http://www.inetsoft.com/company/biwithoutetl/.

Margaris, Nikos a Feuerlicht, George. 2012. Cloud Computing Adoption: A comparative study. *Published by WSEAS Press.* [Online] 2012. [Citace: 8. 3 2014.] http://www.wseas.us/e-library/conferences/2012/Vienna/COMPUTERS/COMPUTERS-71.pdf. ISBN: 978-1-61804-134-0. s. 441-448..

Penichter, Tom. 2009. Biggest BI Mistakes - Mistake #1: Not Using an ETL Platform. *TGP Associates.* [Online] 3. 5 2009. [Citace: 25. 1 2014.] http://blogs.tgpassociates.com/post/2009/06/03/Biggest-BI-Mistakes-Mistake-1-Not-Using-an-ETL-Platform.aspx.

Profitbricks. 2014. How To Choose A Cloud Hosting Provider: Find The Right Provider For You. *Profitbricks.* [Online] 2014. [Citace: 3. 2 2014.] https://www.profitbricks.com/choosing-a-cloud-provider.

31.10 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k vymezení BI řešení na bázi cloud computingu.

- BI s využitím technologií a služeb cloud computingu lze charakterizovat jako zcela ***původní nová řešení***, nebo řešení ***transformovaná*** z předchozích „klasických" řešení na bázi on-premise.
- Analýza spojená s přípravou BI pro cloud se musí zabývat ***rozdílným využíváním nebo nevyužíváním komponent*** v cloudovém modelu a v klasickém modelu.
- V analýze je nezbytné ***posoudit i různé varianty a architektury a možnosti implementace*** cloud BI:
  - Všechny komponenty BI s využitím klasickém modelu.
  - DWH ponechat jako klasické řešení stejně tak ETL a pouze analytickou část BI přivést do cloudu.
  - DWH spolu s BI umístit do cloudu a nástroj ETL využívat jako klasické.
  - Všechny komponenty BI do cloudu.
- Je nezbytné velmi dobře pochopit a vyhodnocovat ***faktory ovlivňující rozhodnutí*** mezi způsoby klasického cloudového řešení BI.

- Je nutné vyhodnocovat ***možnosti migrace mezi poskytovateli***, hlavním elementem tohoto procesu je ***zpětné získání všech dat.*** Proto je nutné mít ***smluvně ošetřeno, kdo je vlastníkem dat***. Tyto informace lze získat z ***Obchodních podmínek*** („*Terms and conditions*“) od poskytovatele služby. ***Zákazník by měl být vždy vlastníkem dat.***

32. Využití Cloud služeb a Cloud BI řešení zákazníky

Borovec, J. (NDC Group, a.s.)

Účelem kapitoly je:

- hodnotit ***možnosti využití služeb BI v cloudu*** pro různé typy zákazníků,
- analyzovat ***problematické případy*** využívání takových služeb,
- analyzovat ***hlavní rozdíly*** v řešení a provozování služeb BI v cloudovém prostředí.

32.1 Analýza důvodů pro využití Cloud řešení

Cloud Computing je významně rozšířený ***fenomén současnosti***. Na uvedený fakt poukazuje i každoroční výzkum americké společnosti *Flexera*. Výzkum z ledna 2019 ukázal, že až 94 % respondentů používá Cloud řešení. (Flexera 2019)

Mezi nejvyužívanější služby v prostředí Cloud patří ***na prvním místě DBaaS neboli Database as a Service*** (databáze poskytovaná formou služby bez nutnosti její údržby zákazníkem (Techopedia Inc. 2019b)), ***na druhém místě Push Notifications neboli služba posílající oznámení klientským zařízením*** (IBM 2019) a ***na třetí příčce*** se umístilo ***používání Data Warehouse*** (*DWH* neboli datový sklad s trvalými, časově rozlišenými a subjektově organizovanými daty (MBI 2019e)) v prostředí Cloud. (Flexera 2019)

Podle uvedeného průzkumu je ve firmách ***nejvíce používáno tzv. hybridní Cloud řešení*** (firma v dané variantě kombinuje využívání vlastní Cloud infrastruktury pro privátní účely a veřejného Cloud řešení (Mell a Grance 2011)). Konkrétně se jedná o 69 procent podniků účastnících se daného výzkumu. (Flexera 2019)

Byla provedena i analýza využití podniky, technologických možností a modelů Cloud řešení včetně způsobů jejich nasazení. Z uvedené analýzy vyplývá minimálně jeden ***benefit, který Cloud Computing přináší: značná pružnost*** v možnostech jeho firemní realizace. Další přínosy, kvůli kterým firmy Cloud služby využívají, jsou:

- ***Snížení nákladů na implementaci, provoz a údržbu IT ve firmě*** – studie uvádějí odhad, že náklady na informační technologie v evropských firmách využívajících Cloud služby se snížily o 20 % až 50 %.
- ***Proměnění tzv. provozních nákladů (Operational Expenditures neboli OPEX) na kapitálové náklady (Capital Expenditures neboli CAPEX)*** – jelikož firmy hradí investice z kapitálových nákladů je využití Cloud služeb výhodné. Cloud služby snižují provozní náklady (například náklady na energie, chlazení a obecný provoz IT oddělení a infrastruktury ve firmě), jelikož provoz IT infrastruktury je v rukou dodavatelů. Snižovány jsou však i kapitálové náklady (například na pořizování serverů, harddisků a IT infrastruktury obecně). Zmíněné úspory na operativních i kapitálových nákladech umožňují podnikům vyčlenit více na strategické investice, tj. na kapitálové náklady.
- ***Přizpůsobitelnost a škálovatelnost Cloud služeb pro potřeby firem*** – kdy aplikace vykazují různorodé zatížení (například sezonní či dokonce zcela nepředvídatelné). Danou charakteristiku Cloud služeb podniky oceňují i v případě dočasně využívaných aplikací.
- ***Zkrácená doba nasazení aplikací (tzv. Time to market)*** –je způsoben předpřipravenou infrastrukturou, kterou může zákazník snadno upravovat dle potřeby.
- ***Více času pro management na řešení strategického směřování podniku*** – manažeři nemusí řešit technologický rozvoj a provoz IT oddělení, mohou se soustředit více na analýzy dat, které jim ulehčí rozhodování o strategickém směřování podniku.

- ***Snazší a rychlejší obnova byznysových služeb zákazníkům a IT infrastruktury po výpadku*** *(tzv. Business Continuity Management neboli řízení kontinuity byznysových procesů a služeb)* – možnost kontinuálního zálohování dat a přesunutí aktivit uživatele na jiné servery či Cloud službu při výpadku.

32.2 Činitelé negativně ovlivňující adopci Cloud služeb a Cloud BI včetně návrhu možných řešení

Faktory omezující adopci cloudových řešení se liší ***podle velikosti společnosti***. ***Velké a střední*** společnosti považují riziko ***prolomení digitální bezpečnosti*** za nejvíce limitující činitel. Důvodem je nejistota klientů o způsobu, jakým jsou cloudové služby na straně dodavatelů zabezpečeny. Zákazníci však rovněž často mají ***pochybnosti i o samotné důvěryhodnosti dodavatelů***, a to i v případě že dodavatel odběratelům jasně sděluje způsoby zabezpečení jejich dat.

Za ***druhý*** hlavní omezující faktor velké a střední společnosti považují ***nejistotu o fyzickém umístění dat***. Společnosti uvádějí obavy o bezpečnost, portabilitu (mezi cloudovými službami a jinými systémy) a přístupnost dat. Jinými slovy, převládá ***strach ze ztráty kontroly*** nad svými daty při využívání cloudových služeb.

Třetím limitujícím faktorem pro velké a střední společnosti jsou jejich ***nejasnosti o legislativě zabývající se cloudovými službami*** a technologiemi. Především se jedná o nejistotu společností v otázce, která legislativa je využitelná, protože dodavatel cloudových služeb může sídlit v jiném státě než odběratel. Rovněž společnosti uvádějí, že jim chybí jednoduchý mechanismus vyřešení právních konfliktů mezi odběratelem a dodavatelem. (Deloitte 2017)

Pro malé firmy jsou pořadí i ***faktory*** limitující adopci cloudových služeb ***odlišné***. Na ***první*** příčce se umisťuje ***nedostatek znalostí*** malých společností o cloudových technologiích a službách, což jim znemožňuje efektivně nasadit komplexní cloudová řešení. Na ***druhém*** místě jsou ***obavy z prolomení bezpečnosti*** cloudových služeb (včetně stejných pochybností o dodavatelích uvedených o odstavec výše). ***Třetí*** místo obsadily ***vysoké náklady***, které by bylo nutné investovat ***do přechodu*** na cloudové služby. Na ***čtvrtém*** místě jsou rovněž jako omezující faktory umístěny ***nejasnosti ohledně legislativy*** a fyzického umístění dat. (Deloitte 2017)

Výše uvedené ***limitující faktory*** se přímo vztahují i ***na Cloud Business Intelligence***. Za hlavní limity adopce Cloud BI lze považovat níže uvedené faktory seřazené podle závažnosti uváděné společnostmi (Indriasari et al. 2018):

- ***Bezpečnost dat*** – Dodavatel by podle společností z průzkumu měl zajistit dostupnost, integritu, a především důvěrnost klientských dat uložených v Cloud BI řešení. Bohužel u některých stále přetrvávají ***obtíže s důvěrou v dodavatele***, což pro ně v současnosti činí adopci cloudových služeb neuskutečnitelnou. Pro získání důvěry mohou dodavatelé projít auditem, že splňují požadavky mezinárodních standardů bezpečnosti. ***Splňovat je mohou například podle normy ISO/IEC 27001*** pro Systémy řízení bezpečnosti informací (Information Security Management Systém neboli ISMS) a později více specificky dle normy ISO/IEC 27018 zabývající se přímo ochranou osobních údajů v cloudovém prostředí.
- ***Kontrola nad daty*** – Pro klienty je důležité, aby firma dodavatele Cloud BI služeb ***měla zavedenou tzv. IT Governance*** (neboli správu IT ve firmě na strategické úrovni) a rovněž i standardy dodávky služeb (například podle rámce ITIL nebo ISO/IEC 20000-1 zabývající se mimo jiné i dostupností a kvalitou pro Cloud Computing). To může být podstatným prostředkem pro získání důvěry klientů, se kterou jsou v současnosti obtíže. Dodavatelé by měli zkrátka zákazníkům prokázat, že jejich data jsou zabezpečena jak po technické stránce (šifrováním, zálohami, fyzickým zabezpečením serverovny), tak po stránce organizační (prostřednictvím zmíněných standardů).
- ***Zralost dodavatelů*** – Dodavatel si musí uvědomit, že potenciální zákazníci potřebují nejprve provést analýzu byznysových požadavků a posléze ***zhodnotit nabídky jednotlivých dodavatelů*** a poté teprve vybrat vhodného kandidáta. Uvedený proces může být díky značnému

množství a rozdílné kvalitě dodavatelů Cloud BI služeb ***složitý, a proto může uplynout dlouhá doba***, než se z potenciálního klienta stane odběratel. Dodavatelé by proto měli splňovat standardy dodávky služeb a bezpečnosti (viz odstavec výše) a rovněž je vhodné, aby se pokusili s ostatními dodavateli vytvořit jednotný cenový model pro Cloud BI služby. ***Nejednotnost v cenách*** je způsobena stanovováním cen podle odbíraného počtu jader, množství paměti, různých funkcionalit, velikosti uložiště, počtu uživatelů atd. Je očividné, že takové množství kritérií, podle kterých dodavatelé stanovují cenu, zesložiťuje zákazníkům výběr vhodné firmy poskytující Cloud BI. Pokud tedy dodavatelé chtějí daný proces na straně zákazníků zkrátit, je vhodné sjednotit navzájem tvorbu ceny pro Cloud BI služby.

- ***Výkon*** – V cloudovém prostředí ***se v současnosti přenáší stále větší objemy dat***, které je třeba analyzovat (Big Data neboli velkoobjemová data). Z uvedeného důvodu rostou i nároky zákazníků na výkon Cloud BI služeb, především datových skladů. Z hlediska dodavatele je proto třeba neustále kontrolovat a popřípadě ***navyšovat množství a kapacitu serverů***, neboť výkon Cloud BI systémů je kapacitou serverů dodavatele přímo ovlivněn. Nedostačující výkon Cloud BI řešení může totiž zapříčinit odliv zákazníků od daného dodavatele.

32.3 Analýza rozdílů ve způsobech provozu podnikových informačních systémů

Pro ***komplexní závěr analýzy cloudových řešení současnosti*** je vhodné rovněž uvést jejich srovnání s ostatními podobami provozu informačních systémů. Jedná se o standartní fyzické tzv. *stand-alone* či *on-premises* řešení provozované na infrastruktuře klientského podniku a o virtualizaci (vytvoření logických či jinak řečeno virtuálních prvků v infrastruktuře nad fyzicky existujícími výpočetními jednotkami).

32.3.1 Základní rozdíly ve způsobech provozu firemních informačních systémů

- ***On-premises řešení*** – Jedná se o řešení tzv. *na místě* (jak vyplývá z překladu názvu) neboli ***realizované lokálně v organizaci*** nebo datovém centru organizace.
- ***Virtualizace*** – Uvedený způsob provozu informačního systému znamená ***rozdělení výkonu fyzických zařízení*** (například serverů v serverovně) ***mezi různé množství imaginárních výpočetních prostředků***. Jinými slovy, klienti jsou vystaveni představě, že disponují plnými fyzickými výpočetními prostředky (například serverem se dvěma jádry procesoru), avšak ve skutečnosti se jedná pouze o obraz vytvořený z části výkonu skutečných fyzických zařízení (například z pěti osmijádrových serverů). Důležitou součástí je tzv. ***hypervizor***, který se stará o oddělení fyzické infrastruktury od operačního systému (vytváří zmíněný imaginární obraz). ***Virtualizace je technologickou součástí cloudových služeb***, avšak využívá se i odděleně.
- ***Cloud Computing*** – Pro shrnutí lze zmínit, že se jedná poskytování podnikových aplikací formou služby prostřednictvím sítě internetu, přičemž výpočetní prostředky nejsou provozovány odběratelem.

32.3.2 Porovnání implementace, údržby a customizace (úprav) v rámci daných způsobů provozu

V případě ***lokálního (stand-alone či on-premises) řešení*** je ***výhodou*** fakt, že organizace ***disponuje značnou kontrolou*** na procesem implementace, údržby a customizace informačního systému. ***Nevýhodou je dlouhá doba***, kterou implementace a úpravy zaberou a rovněž komplikace s aktualizacemi software, jelikož se mohou objevit problémy s integrováním nekompatibilních verzí různých programů.

Virtualizace rovněž nabízí ***vysokou míru kontroly podniku nad provozem*** informačního systému a bohužel je rovněž neméně ***složitá*** v porovnání s předchozí metodou z hlediska nutných technických znalostí IT personálu pro její zprovoznění. ***Výhodou*** virtualizace však je ***oddělení výpočetních prostředků*** (prostřednictvím výše zmíněného hypervizoru), což umožňuje například ***provozovat zastaralý software nad moderním hardware***, či provádět změny ve virtuálně vytvořeném operačním systému, aniž by došlo ke změnám v hostujícím operačním systému. Uvedená výhoda ***odstraňuje***

například výše zmíněné problémy s integrací nekompatibilních výpočetních prostředků. Další výhodou virtualizace je rovněž ***rychlejší nasazení výpočetních prostředků***, jelikož virtuální zařízení a systémy jsou odděleny od fyzických a je tedy snazší provést výměnu na jedné vrstvě, aniž by bylo třeba měnit cokoliv na vrstvě druhé. ***Nevýhodou*** virtualizace je ***omezení fyzickou výpočetní infrastrukturou*** podniku, jelikož není možné přidělit objem výkonu nad limit fyzických zařízení.

V případě ***cloudových služeb odpadá nutnost udržovat fyzickou infrastrukturu*** (tj. i o virtualizaci se stará dodavatel) a je možné rychle nasadit a využít nabízené aplikace bez nutnosti hlubokých technických znalostí odběratelského IT personálu. Navíc je možné odebírat pouze tolik výpočetních prostředků, kolik zákaznický podnik potřebuje. Zmíněná fakta jsou výhodou v především v ***ušetření času a peněz*** na důležitější podnikové aktivity, které by jinak bylo nutno investovat do implementace, údržby, provozu a úprav infrastruktury a aplikací. Problémem však může být případná ***nedostupnost připojení*** k internetu,

32.3.3 Srovnání bezpečnosti uvedených způsobů realizace informačních systémů společnosti

Při provozování informačního systému na ***lokální IT*** infrastruktuře se o bezpečnost stará ***samotná organizace***. Uvedený fakt je však výhodou pouze v případě, že má daná organizace dostatečně vybudované bezpečnostní protokoly a její pracovníci disponují dostatečnými znalostmi v oblasti budování kybernetické bezpečnosti organizace. Slovem dostatečně je myšleno ***dostatečně vzhledem k předmětu podnikání***, jelikož například banky či zdravotnická zařízení vyžadují přísnější a komplexnější úroveň bezpečnosti než například start-upy či firmy prodávající drobné módní nebo jiné doplňky. Každopádně veškerá ***zařízení a systémy je třeba zabezpečit individuálně*** a je třeba dané zabezpečení pravidelně revidovat, s čímž souvisí i nutnost čas od času aktualizovat software, což může posléze způsobit i výše zmíněné integrační problémy uvedené výše.

Při ***virtualizaci*** se o bezpečnost stará rovněž ***samotná firma*** využívající svoji infrastrukturu, avšak je možné veškerou bezpečnost ***vyřešit centrálně prostřednictvím hypervizoru***, který kromě oddělení fyzické a virtuální vrstvy rovněž umožňuje nastavit bezpečnostní procedury a politiku pro všechny spravované servery a systémy.

Při ***využívání cloudových služeb*** starost o udržování bezpečnosti dat ***ze strany klientského podniku odpadá***. Je však potřeba se ujistit, že ***dodavatel data řádně šifruje*** a je třeba zjistit od dodavatele, jak zajišťuje kybernetickou bezpečnost své infrastruktury. Někteří dodavatelé totiž data nešifrují, či jejich bezpečnost je z klientského úhlu pohledu nedostatečná, což v současnosti vzbuzuje nedůvěru některých firem ke cloudovým řešením. Proto je důležité důkladně prostudovat a vyjednat s dodavateli ***smlouvu o úrovni služeb*** (SLA či Service Level Agreement), aby byly zajištěny mechanismy proti ztrátě či poškození dat a rovněž bylo zajištěno pojištění při případné škodě.

32.3.4 Srovnání nákladů každého z uvedených způsobů realizace informačních systémů

V prvním případě se jedná o ***vysoké náklady na vybudování a*** provoz daného informačního systému, jelikož podnik se stará o veškerou infrastrukturu sám.

Druhý způsob je ***méně nákladný na vybudování a údržbu*** infrastruktury a platformy pro virtualizaci, neboť je možné provozovat pouze několik konsolidovaných (sjednocených) či dokonce jediný server na kterém je spuštěno více virtuálních zařízení a operačních systémů.

V případě cloudových služeb odpadají veškeré výdaje zákaznického podniku na budování a údržbu infrastruktury (včetně IT personálu) a ***platí se pouze přesně za takový objem služby***, jaký je klientem ***užíván*** neboli přesně za odebíraný výkon, počet přístupů pro uživatele, funkcionality atd. (tzv. *pay-per-use* či *pay-as-you-go* podoba placení). Je však třeba zmínit, že ceny se mohou u různých dodavatelů lišit (uvedené je problémem především u Cloud BI služeb, a to včetně možného řešení u některých dodavatelů může být ve výsledku pro danou firmu dražší využívat cloudových služeb například oproti virtualizaci.

32.4 Zdroje

DELOITTE, 2017. *Measuring the economic impact of cloud computing in Europe* [online]. Text. B.m.: Evropská komise [vid. 2019-06-26]. Dostupné z: https://ec.europa.eu/digital-single-market/en/news/measuring-economic-impact-cloud-computing-europe

FLEXERA, 2019. Cloud Computing Trends: 2019 State of the Cloud Survey. *Cloud Management* [online] [vid. 2019-06-19]. Dostupné z: https://blogs.flexera.com/cloud/cloud-industry-insights/cloud-computing-trends-2019-state-of-the-cloud-survey/

INDRIASARI, Elisa, Harjanto PRABOWO, Meyliana WIDJAJA a Achmad HIDAYANTO, 2018. Key benefits, challenges, and risk of cloud business intelligence: A systematic literature review. *International Journal of Mechanical Engineering and Technology*. **9**, 819–831. ISSN 0976-6340.

MELL, Peter a Tim GRANCE, 2011. *The NIST Definition of Cloud Computing* [online]. NIST Special Publication (SP) 800-145. B.m.: National Institute of Standards and Technology [vid. 2019-06-19]. Dostupné z: doi:https://doi.org/10.6028/NIST.SP.800-145

TECHOPEDIA INC., 2019a. What is a Domain Controller? - Definition from Techopedia. *Techopedia.com* [online] [vid. 2019-08-28]. Dostupné z: https://www.techopedia.com/definition/4193/domain-controller-dc

TECHOPEDIA INC., 2019b. What is Database as a Service (DBaaS)? - Definition from Techopedia. *Techopedia.com* [online] [vid. 2019-06-19]. Dostupné z: https://www.techopedia.com/definition/29431/database-as-a-service-dbaas

32.5 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k analýze využívání BI řešení na bázi cloud computingu.

- Hlavní důvody ve využívání cloudových technologií a služeb jsou:
  - ***značná pružnost*** v možnostech jeho firemní realizace,
  - ***snížení nákladů*** na implementaci, provoz a údržbu IT ve firmě,
  - ***proměnění tzv. provozních nákladů*** (Operational Expenditures neboli OPEX) ***na kapitálové náklady*** (Capital Expenditures neboli CAPEX),
  - ***přizpůsobitelnost a škálovatelnost*** Cloud služeb pro potřeby firem,
  - ***zkrácená doba nasazení*** aplikací (tzv. Time to market),
  - ***více času pro management*** na řešení strategického směřování podniku,
  - snazší a ***rychlejší obnova byznysových služeb*** zákazníkům a IT infrastruktury po výpadku.
- ***Faktory omezující využití*** cloudových řešení se liší ***podle velikosti:***
  - ***pro velké a střední firmy*** jsou to prolomení digitální bezpečnosti, nejistota o fyzickém umístění dat, nejasnosti o legislativě zabývající se cloudovými službami a technologiemi,
  - ***pro malé firmy*** to jsou nedostatek znalostí, obavy z prolomení bezpečnosti cloudových služeb, vysoké náklady, které by bylo nutné investovat do přechodu na cloudové služby, nejasnosti ohledně legislativy a fyzického umístění dat.
- V rámci ***analýzy cloudových řešení*** je podstatné i jejich ***srovnání s ostatními forma provozu*** informačních systémů, a to standartní fyzické tzv. ***stand-alone*** **či** ***on-premises*** řešení a o ***virtualizaci*** (vytvoření logických či jinak řečeno virtuálních prvků v infrastruktuře nad fyzicky existujícími výpočetními jednotkami).

33. Možnosti Cloud Business Intelligence současnosti

Borovec, J. (NDC Group)

Účelem kapitoly je:

- hodnotit ***důvody pro praktická uplatnění služeb BI v cloudu*** v současných podmínkách,
- analyzovat ***architektury*** BI služeb v cloudu a jejich dopady na kvalitu provozu,
- vyhodnotit možná ***řešení jednotlivých komponent*** BI architektury v cloudu.

Základem k pochopení praktického užití Business Intelligence je jeho komplexní vymezení.

33.1 Důvody pro praktické využití standardního Business Intelligence

Dále uvedené oblasti jsou hlavními důvody, proč firmy využívají Business Intelligence:

- ***Finanční hodnota*** – Business Intelligence umožňuje ***zvýšit výdělečnost podniku*** prostřednictvím zvýšení příjmů a snížení nákladů.
- ***Produktivita*** – Pomocí BI lze ***zvýšit efektivitu procesů*** ve firmě (například zvýšit množství kvalitních hotových produktů a obsloužených zákazníků a snížit celkový čas nutný pro činnosti v procesu), rovnoměrně rozložit či přímo snížit pracovní zátěž pracovníků (například prostřednictvím poskytnutí kvalitních informací z BI analýz nutných pro jejich práci).
- ***Důvěryhodnost*** – ***Zvýšení důvěry zákazníků a dodavatelů*** v podnik je další z výhod, kterou je BI schopno nabídnout prostřednictvím zvýšení jejich spokojenosti s produkty a službami firmy. Spokojenost zákazníků je možno zvýšit pomocí na míru připravených byznysových plánů, které lze vypracovat na základě BI analýz v datech (například sledováním, které služby zákazníci rádi odbírají a které se naopak neprodávají a je zbytečné nimi zákazníka obtěžovat).
- ***Snížení rizik*** – Rizika lze pomocí BI řešení snížit již výše zmiňovanými informovanými rozhodnutími a rovněž i ***odhalováním nezjevných problémů*** ve firmě. Dané problémy je možné odhalit prostřednictvím důkladných BI analýz, tj. hledáním vzorců v datech (souvislostí, podobností a rozdílů). Rovněž lze snížit riziko vzniknutí problému pomocí možnosti BI nástrojů ***nastavit limitní hodnoty různých ukazatelů***, při jejichž dosažení nástroj příslušné pracovníky upozorní na vznikající problémy (například na rapidní snížení produkce v některém ze závodů).

33.2 Analýza problémových stránek standardního Business Intelligence včetně návrhu možných řešení

Ačkoliv výhod při využívání Business Intelligence řešení lze vyjmenovat značné množství, existují i určité problémové stránky, které je vhodné analyzovat, např. (Janoschek 2018):

- Hlavními problémy všech společností při implementaci Bussiness Intelligence řešení jsou ***obtíže s nasazením nástrojů,*** obvykle s ***integrací částí architektury BI***, což může být zapříčiněno ***nedostačujícími a nedostatečně komunikovanými návrhy***, ***nejasně definovanými požadavky*** na řešení (což vyplývá obvykle z nedostatečné komunikace se zákazníkem) a ***nedostatkem*** (především finančních) ***zdrojů***
  - ***V případě malých společností*** se navíc jako významný faktor negativně ovlivňující nasazení ukazují ***potíže s proškolením zaměstnanců***, přičemž zde by opět pomohla dostatečná komunikace tentokráte s koncovými uživateli, aby se vyjasnily vhodné termíny školení (aby měli uživatelé znalosti v živé paměti) a případné další.
  - ***V případě středních a velkých společností*** se navíc objevují příliš ***krátké termíny pro dokončení projektu*** nasazení BI infrastruktury, platformy a aplikací a rovněž

obtíže s migrací dat do BI řešení, vyplatí se začátkem provést důkladnou analýzu informačních systémů podniku a veškeré přesuny konzultovat s poskytovateli příslušných systémů.

- Při používání BI řešení jsou největšími komplikacemi všech společností ***problémy s výkonností dotazů do BI databází,*** které mohou být vyřešeny například jejich ***paralelizací***. Aktuálně lze danou paralelizaci realizovat například ***technologií*** **MapReduce**.
- Dalším problémem podniků jsou ***administrativní komplikace během provozu BI*** systémů, které lze ***řešit zavedením následujících procesů***:
  - ***řízení výkonu*** (s důrazem na hledání úzkých a slabých míst systémů),
  - ***řízení kapacit*** (s důrazem na zajištění potřebných kapacit vzhledem k zatížení systémů),
  - ***řízení záloh a obnova systémů*** (s důrazem na ochranu klíčových i jiných podnikových dat),
  - ***konfigurační management*** (s důrazem na zmapování systémů, na plynulý průběh změn ideálně bez ovlivnění byznysových procesů a na přehlednou správu verzí systému)
  - ***kontinuální monitorová***ní (s důrazem na sledování chyb systému a jeho výkonu včetně včasného upozornění na kritické stavy) (Loshin 2013).
- ***Ve středních a velkých společnostech*** se dále objevují ***problémy s nedostatečnou zainteresovaností a zapojením byznysových uživatelů***. Zmíněnou potíž lze ***řešit jednak komunikací*** IT oddělení s manažery (zde včetně včasného upozorňování na problémy s uvedením časových a finančních dopadů a možného řešení) i koncovými uživateli a rovněž také dostatečným proškolováním koncových uživatelů.
- Dalším problémem středních a velkých podniků je ***nedostatečná správa dat a řízení jejich kvality***. Uvedenou komplikaci je možné napravit či ji předejít ***zavedením politik a procesů, které zajistí sběr požadavků na data*** (po stránce obsahové, formální a kvalitativní) a celopodnikový konsensus se zavedenými politikami a standardy pro práci s daty Rovněž je důležité zavést ***řízení kvality dat a metadat***, což úzce souvisí se správou dat jako takových.

33.3 Analýza Cloud Business Intelligence

Cloud BI nabízí možnost ***soustředit se více na předmět podnikání*** a analýzy z byznysového hlediska než na správu IT infrastruktury, avšak potýká se s nedůvěrou zákazníků.

Cloud Business Intelligence lze vymezit jako ***model, ve kterém je alespoň některá komponenta BI řešení umístěna v cloudovém prostředí*** (tj. mimo infrastrukturu podniku). Komponenty, které se obvykle umisťují do cloudového prostředí jsou ETL/ELT pumpy, datový sklad a reportingové, analytické a další vizualizační nástroje.

Rovněž lze na Cloud Business Intelligence pohlížet ***z hlediska modelů dodávky služeb***. V daném případě poskytování Cloud BI v modelu software jako služba (***SaaS***) znamená nabízení ***předpřipravených BI software***, přičemž z firemního prostředí zákazníka je do cloudového řešení ***potřeba migrovat pouze data***. Při dodávání cloudového BI formou platformy jako služby (***PaaS***) se jedná o BI systém, jehož ***funkcionalitu si může zákazník upravit podle potřeby***, tak aby ukazatele z BI analýz mohly přesně měřit výkonost byznysu (tj. sledovat mimo jiné i naplnění takzvaných klíčových ukazatelů výkonnosti). V daném případě pak jde dodavateli o ***nabídnutí co nejširší funkcionality na svých platformách zákazníkům***. Poslední možností je ***poskytnutí pouhé infrastruktury*** jako služby (tj. hardware a software s využitím virtualizace, přičemž veškeré dané infrastrukturní součásti jsou spravované dodavatelem), na kterém ***si zákazník samostatně buduje své BI*** řešení.

Veškerá Cloud BI řešení by měla být podložena smlouvou o dodávce služeb (***SLA***) uzavřenou mezi dodavatelem a odběratelem. Součástí by měla být ***definice způsobu platby za služby, pojištění*** při výpadku služeb včetně ***vymezené hierarchie priorit a kritičnosti*** pro zákazníka (tj. které systémy musí být provozuschopné a jaká je jejich kritičnost pro byznys včetně definic, co přesně pro

zákaznickou firmu znamená provozuschopnost systémů a co má učinit dodavatel, když nebudou provozuschopné a jaké budou ***sankce***, pokud tak neučiní a jak se budou v čase sankce zvyšovat).

33.4 Architektura Cloud Business Intelligence, využití v podnikové praxi

V návaznosti na daná témata je třeba ***vymezit i architekturu Cloud BI***. Jednotlivé veškeré dále zmíněné komponenty Cloud BI jsou v zásadách fungování shodné s částmi tradičního interního BI řešení.

33.4.1 Propojení se zdrojovými systémy, mezi komponentami Cloud BI a mezi různými aplikacemi navzájem

Zdrojové podnikové systémy ***nemusí nutně být umístěny v cloudovém prostředí*** pro zavedení Cloud BI do podniku. Je nutno ***integrovat veškeré komponenty cloudového BI*** (tj. ETL pumpy, datový sklad a reportingové, analytické a další vizualizační aplikace) a případně integrovat BI řešení s dalšími podnikovými aplikacemi (například CRM). K danému účelu ***slouží API*** nebo popřípadě i ***konektor*** v podobě propojovacího software (tzv. middleware) v cloudovém prostředí, pokud spojení pouze pomocí API nestačí k propojení značně různorodých systémů. ***Konektory*** lze považovat de facto za ***cloudové EAI*** obvykle s vlastním cloudovým operativním uložištěm dat (ODS). Dané konektory jsou označovány jako nástroje či platformy datové integrace.

Základem veškerého propojení je, aby veškerá data, která je třeba nahrát do cloudového BI řešení či obecně jakýchkoliv jiných aplikací, byla skutečně úspěšně do cílového prostředí v pořádku doručena. Proto je vhodné z pozice zákazníka ***zjistit, zda v případě využití middleware jsou data během přenosu do cílového prostředí zálohována*** a zda jsou v daném software implementovány mechanismy automatické obnovy například při výpadku spojení. Jinými slovy, při případném přerušení spojení by middleware měl být samostatně schopen odhalit poškozená (nekompletní) data a opravit je (či alespoň automaticky znovu přenést daná data ze zdrojových systémů) a dokončit automaticky proces nahrávání do cílového systému. (Juan-Verdejo et al. 2014)

Většina poskytovatelů cloudového BI v základu však mnohdy ***nabízí možnost napojit jejich systémy prostřednictvím API na mnoho různých aplikací od jiných dodavatelů***. Rovněž pokud si podnik pořídí kompletní integrované řešení (například od společnosti IBM, SAP, Amazon, Microsoft atd.) není třeba z pohledu zákazníka propojení vůbec řešit. Pokud se však zákaznická firma dostane do situace, že nelze propojit vzájemně některé systémy ***lze využít middleware***. Zákaznicky nejoblíbenějšími konektory jsou podle Gartner v současnosti produkty společnosti Informatica. Daná společnost nabízí komplexní platformu umožňující propojit zdrojové systémy s Cloud BI řešeními (například od společností Amazon, Microsoft, SAP, Google a dalších), podnikové systémy s dodavatelskými systémy a rovněž i cloudové systémy s on-premises či s jinými cloudovými systémy.

Platforma nabízí ***možnost vytvoření jednotného místa (HUB či rozbočovač)***, ze kterého je možné ***přistoupit k datům ze všech systémů***. Daná platforma navíc disponuje algoritmem využívajícím ***strojové učení***, který dokáže samostatně identifikovat důležité datové oblasti (jako jsou uživatelé, objednávky, role apod.) a vztahy mezi nimi. Uvedená funkcionalita může značně usnadnit mapování celého systému (například pro potřeby přestavění architektury systémů). Zmíněný algoritmus navíc dokáže ***identifikovat citlivá data v podniku***, rozpoznat možnosti přístupu k daným datům a upozornit příslušné odpovědné osoby na možná rizika týkající se datové dostupnosti. (Gartner, Inc. 2019a; Informatica 2019;

33.4.2 Extrakce, transformace a nahrávání v rámci cloudového prostředí

ETL pumpy se starají o extrakci, transformaci a nahrávání dat do cílových databází. Daný úkon však neprobíhá neustále, ale po dávkách. Uvedený fakt činí ***přenesení ETL do cloudového prostředí výhodným*** jak pro zákazníka, tak pro dodavatele. Zákazníci oceňují množství dat, které je možné přenést prostřednictvím výkonu dodavatelských serverů. ***Dodavatelé zhodnocují možnost naplno využít své infrastruktury*** pro obsloužení značného množství zákazníků.

Jinými slovy, náklady na údržbu infrastruktury dodavateli pokrývá ***možnost rozpadnout její využití mezi mnoho zákazníků*** a nenechat tak servery ani na krátkou dobu nevyužité. Daným způsobem dodavatelé efektivně generují prostředky například i na údržbu a aktualizace ETL nástrojů a zákazník se tedy nemusí starat o nic jiného než o používání datových pump pro své byznysové analýzy. Určitou nevýhodou může být ***omezený výkon ve spojení s některými on-premises*** interními zdrojovými systémy, které mohou být úzkým místem omezujícím výkonné cloudové ETL pumpy.

ETL proces v cloudovém prostředí nemusí nutně probíhat shodným způsobem, jako v interním BI. Ve skutečnosti ***existuje několik možných přístupů***, jak data extrahovat ze zdrojových systémů, transformovat je a nahrát do datového skladu. Jedná se například o ***přístupy využívající sémantického datového modelu s využitím systému popisu zdrojů neboli RDF*** (Resource Description Framework) rozdělující data na subjekt, predikát a objekt. Jak dané rozdělení konkrétně vypadá lze uvést na příkladu: Chris Nerdal (subjekt) sídlí v (predikát) Norsko (objekt). Databáze využívající RDF se v cloudovém ETL procesu využívá pro rozdělení dat ze zdrojových systémů do shodných skupin a jejich následné vyčištění od duplicit, zkontrolování neporušenosti integrity a odfiltrování nerelevantních/chybných dat. Jedná se tedy de facto o cloudové dočasné uložiště dat. Pro extrakci dat z dané databáze do datového skladu ***je posléze využit jazyk SPARQL***, který je specializován na práci s RDF databázemi.

Další přístupy nevyužívají sémantického datového modelu a ***provádějí běžnou extrakci dat s rozlišením různých typů*** (text, obrázek, video atd.) a následnou ***transformaci v dočasném cloudovém uložišti (cloudovém DSA)*** s čištěním, ukládáním dat a metadat (***ve formě XML souborů***) do datového skladu. U druhého příkladu se tedy jedná o velice podobný proces, jako u interního ETL. (Diouf et al. 2018)

ETL v cloudovém prostředí umí extrahovat data ***jak z cloudových, tak z on-premises*** aplikací. Cloudové ETL nástroje v základu ***podporují API pro různé databázové systémy*** (např. Oracle či Microsoft SQL databáze), ***protokoly*** pro přenos souborů (např FTP), ***jazyky XML, HTML*** a nabízejí funkce agregace, řazení, spojování různorodých dat (včetně nestrukturovaných), vytváření záznamů o aktivitách ETL pump (tzv. logy) a časového a logického plánování procesu transformace. Součástí ETL pump bývají i konektory. ETL může být ***integrovanou součástí celkového cloudového BI řešení anebo jsou nabízeny i samostatné*** ETL nástroje, které se posléze propojí s datovým skladem.

Mezi zákazníky nejoblíbenější ***samostatné ETL*** cloudové nástroje patří podle Gartner jednak řešení od společnosti ***Informatica*** a dále pak například od firmy ***Fivetran*** (Gartner, Inc. 2019b). ETL od společnosti Fivetran nabízí například ***automatické upravení zdrojového schématu***, aby při extrakci a následných transformacích nedošlo ke ztrátě jakýchkoliv dat přenášených do cílového schématu datového skladu. Dále rovněž nabízí ***mnoho předpřipravených pump*** pro různé transformace dat, možnost analyzovat historii dat (i data smazaná) a rovněž možnost ukládat některá data interně ve firmě. (Fivetran, Inc. 2019; Panoply Ltd. 2019)

33.5 Datový sklad v cloudovém prostředí

Datový sklad je, jako všechna cloudová řešení, ***udržován a spravován poskytovatelem***, přičemž zákazník je užívá pro své byznysové potřeby a může se soustředit plně na byznysové analýzy a na výběr dat, která je potřeba do datového skladu nahrát. ***Kapacitu datového skladu*** si navíc zákazník může u poskytovatele ***navýšit podle potřeby*** (tj. objemu dat, náročnosti operací nad nimi a množství přistupujících uživatelů). ***Cloudové datové sklady*** lze obvykle prostřednictvím API a často integrovaných konektorů (případně i ETL nástrojů) ***propojit se značným množstvím systémů***. Rovněž k nim lze obvykle ***přistupovat prostřednictvím mnohých zařízení*** (mobilní telefon, tablet, počítač atd.), jelikož datové sklady často bývají dostupné ***zprostředkováním webových aplikací***.

Datový sklad je sestaven obvykle ***ze sady XML souborů***, ve kterých jsou ***uložena data ve formě dimenzionálních a faktových tabulek*** včetně příslušných propojení. Dané XML soubory jsou rovněž také ***distribuovány po serverech poskytovatele***, aby bylo možné z hlediska zákazníka kdykoliv datový sklad škálovat (například zvýšit kapacitu uložiště).

Škálovatelnost je zajištěna ***dynamickou redistribucí*** dat. To znamená, že pokud je jakýkoliv server nadužíván, nebo určitá část dat nabývá příliš na objemu či je značně měněna, tak jsou data znovu rovnoměrně redistribuována po síti. ***Redistribucí*** je zajištěn ***optimální a nenarušený výkon*** po celou dobu užívání datového skladu uživateli. Základem správy daných distribuovaných dokumentů, aby se předešlo ztrátě dat či složitému manuálnímu skládání rozdělených dat, je většinou ***specifikace v podobě Document Type Definition (DTD)*** neboli definice typu dokumentu ***anebo ve formě XML Schema Definition (XSD)***. Specifikace v daných podobách ***určuje*** všem ostatním distribuovaným XML dokumentům, ***jak mají být strukturovány***, aby se předešlo problémům s jejich následným automaticky řízeným (prostřednictvím řídícího serveru) znovusložením pro potřeby datových analýz. Dotazování nad datovým skladem se leckdy uskutečňuje v paralelním provedení pomocí modelu MapReduce (Dkaich et al. 2016)

Cloudový datový sklad ***poskytuje*** například zákaznicky značně oblíbená firma (podle průzkumu Gartner) ***Amazon v rámci platformy Redshift***. Daná platforma disponuje značným výkonem kvůli robustní infrastruktuře Amazonu a paralelnímu zpracování dotazů nad datovými sklady. Platforma je rovněž ***snadno propojitelná s BI řešeními a analytickými nástroji*** od různých dodavatelů (SAP, Oracle atd.) a rovněž nabízí připojení k mnohým dalším systémům (prostřednictvím API, ETL a dalších integračních konektorů). (Kerner 2019; Panoply Ltd. 2019; Gartner, Inc. 2019c)

33.6 OLAP databáze v cloudovém prostředí

Databáze OLAP v cloudovém prostředí ***využívají paralelizace***, která jim pomáhá zpracovat dotazy obvykle stovek až tisíců uživatelů najednou. Rovněž ***provádění dotazů*** je ***rovnoměrně rozděleno mezi servery***, což ještě více zvyšuje výkon cloudového provedení OLAP databází. (Al-Aqrabi et al. 2019)

Multidimenzionální ***OLAP kostky*** jsou v cloudovém prostředí ***realizovány formou XML souborů*** s využitím formátu ***DTD***. Daný formát usnadňuje ***propojení daných souborů*** a umožňuje tak vytvořit vztahy mezi dimenzionálními tabulkami navzájem a mezi dimenzemi a faktovými tabulkami. Vytváření OLAP kostek začíná nejprve extrakcí příslušných dat z datového skladu (DWH je v cloudové podobě obvykle také tvořen sadou XML souborů). Následně je ***vytvořena soustava dimenzí a faktové tabulky fyzicky realizovaná XML*** soubory se strukturou a propojeními definovanými formátem DTD. Pro extrakci dat a vytváření OLAP kostek je v některých případech využíváno modelu ***MapReduce***, který zajišťuje ***značnou paralelizaci procesu extrakce a seskupování dat***. Posledním krokem je vybrání nepotřebných atributů v dané kostce a případné ***propojení kostky na určité webové či analytické aplikace***. (Dkaich et al. 2016; Al-Aqrabi et al. 2019)

Vytváření OLAP multidimenzionálních databází/kostek je většinou součástí analytických nástrojů (například SAP Analytics Cloud, Domo, IBM Cognos Analytics).

33.7 Cloudové reportovací a analytické aplikace

Jedná se ***vizualizační komponentu cloudového BI***, která umožňuje nad daty vytvářet manažerské reporty a dashboardy, ***předpřipravovat a ukládat databázové dotazy či vytvářet scénáře vývoje v datech*** při změně hodnot různých proměnných. (Al-Aqrabi et al. 2019)

Analytické nástroje jsou dostupné obvykle ***prostřednictvím webových aplikací*** a nabízí analytikům a manažerům možnost si samostatně sestavit analýzy podle jejich potřeb (což je dáno rovněž i značným výkonem dodavatelské infrastruktury, která umožňuje rychle vytvořit například OLAP kostky a nad nimi dashboardy). Nástroje rovněž nabízejí ***možnost vytvářet kolaborativní analýzy*** (prostřednictvím internetu) ve spolupráci s týmem v reálném čase. Instrumenty je rovněž možné propojit s ostatními Cloud BI komponentami a jinými aplikacemi prostřednictvím mnohých aplikačních rozhraní (API) a konektorů. (Columbus 2019; Muntean 2015; MBI 2019c)

Analytické a reportovací aplikace ve spojení s ostatními cloud BI komponentami navíc nabízejí klíčové funkcionality důležité pro automatizaci operací v podniku (což je jedno ze základních kritérií pro tzv. čtvrtou průmyslovou revoluci či jinak také průmysl 4.0). Jedná se například o ***automatizovanou predikci budoucího vývoje v datech*** (na základě strojového učení) či o ***monitorování vývoje v datech***

v reálném čase včetně automatického upozorňování na dosažení kritické úrovně (například na zpožďování výroby v určité výrobně). Uvedenými funkcemi disponuje například Microsoft Power BI, Qlik Sense Cloud či SAP Analytics Cloud. (Techedge S.p.A. 2018)

33.8 Řízení kvality v cloudovém prostředí

Musí se analyzovat, jak je v tradičním BI realizována a posuzována kvalita dat a jak jsou spravována metadata. Dané ***principy jsou shodné i pro cloudovou Business Intelligence*** s rozdílem, že ***o kvalitu a správu dat i metadat se stará především dodavatel***. (Unguraş a Mirabela 2018)

Zda je produkt ***kvalitní posuzují zákazníci***, a proto musí dodavatelé při budování svého produktu a cloudových služeb ***zjistit nejprve očekávání a požadavky*** potenciálních zákazníků. Jedině daným způsobem je možné z hlediska dodavatele budovat produkt, který bude skutečně kvalitní. Kvalitu cloudových řešení (a obecně jakýchkoliv řešení) ***nelze však definovat pouze spokojeností*** zákazníka. Produkt například může fungovat v souladu s očekáváními zákazníka, avšak nemusí být již dostatečně zabezpečen. Proto lze považovat za vhodné ***definovat kvalitu cloudových služeb komplexně***. Komplexní zajištění kvality cloudových služeb zahrnuje splnění zákaznických očekávání a požadavků a rovněž i ***soulad cloudového řešení s příslušnými standardy kvality***. (Unguraş a Mirabela 2018; Certification & Information Security Services GmbH 2019)

Za aplikovatelné lze považovat především níže uvedené ***standardy v oblasti norem aplikovatelných pro cloudové služby.***

- Za prvé se jedná o standard pro ***řízení bezpečnosti informačních systémů ISO/IEC 27001 či konkrétnější ISO/IEC 27018*** zabývající se ochranou osobních údajů v cloudovém prostředí a ***ISO/IEC 20000-1*** pokrývající rovněž i dostupnost cloudových služeb dodavatele (včetně tzv. Business Continuity Management neboli řízení kontinuity/nepřerušenosti průběhu procesů organizace i během výpadku některého ze systémů).
- Rovněž je možné se držet ***obecného standardu ISO 9001*** pro zlepšení ***kvality procesu implementace*** softwarového produktu a rovněž konkrétních ***pokynů k bezpečnosti v oblasti cloudových služeb obsažených v ISO/IEC/IEEE 90003***.
- Jako další aplikovatelný standard je ***ISO/IEC 25010*** poskytující kroky k zajištění kvality softwarového produktu i při jeho ***užití různými uživateli***. Dále také daný standard nabízí pokyny k ***zajištění datové kvality***, kterou lze považovat za velice důležitou pro cloudová řešení. Standard ISO/IEC 25010 však může být pro potřeby některých podniků až příliš komplexně rozčleněn a množství kategorií pro jednotlivé zmíněné oblasti kvality může být obtížné zavést v praxi. Jmenovaný standard by měl proto sloužit spíše ***jako základ pro zavedení datové kvality a jakosti produktu*** než jako sled kroků, které je bezpodmínečně nutné v každém podniku poskytujícím cloudové služby naplnit. (Bautista Villalpando et al. 2014; SysAid Technologies Ltd. 2019; Certification & Information Security Services GmbH 2019; ISO 2018a; 2019; 2019; 2018b; 2018c; 2013; 2011; 2015)

Při výběru vhodných plánů, firemních politik, systémů řízení a jednotlivých činností ***záleží především na cloudových službách***, které dodavatel nabízí. Jinými slovy, ***ne všechny standardy může být pro dodavatele výhodné zavádět*** v kompletní podobě či vůbec nemusí být třeba se některými zabývat. Například menší firmy mohou shledat značně nákladným a ve výsledku nepřínosným zavádět rigorózně definované ISO/IEC 25010 pro zajištění kvality produktu a datové kvality. Může se jim však například vyplatit ***zavést postupy ISO 9001 anebo jen některé procesy z ISO/IEC 20000 například tzv. Incident management*** starající se o vyřešení dopadů incidentů (např. výpadek serveru ovlivňující chod byznysových činností). Nicméně u velkých společností, kde je třeba udržet vysokou úroveň kvality komplexního produktu může být již naopak vhodné se zmíněného ISO/IEC 25010 držet.

Úplným ***základem pro udržení jakosti*** (a klientské spokojenosti) při vytváření a poskytování cloudových služeb je však v každém případě ***zajistit dostupnost, přesnost, úplnost a konzistenci*** (stejně jako je již definováno výše u tradičního BI). Uvedený fakt je navíc obohacen o nutnost ***poskytovat zákazníkům podporu*** prostřednictvím internetu určitým komunikačním prostředkem na dálku (například

možností kontaktovat zákaznickou podporu prostřednictvím webového chatu či příslušného webového formuláře v jakýkoliv čas) (Unguraş a Mirabela 2018).

Rovněž ***význam metadat zůstává stejný jako v případě tradičního BI***, avšak jejich ukládání je v cloudovém prostředí případné realizovat odlišným způsobem. Metadata pro cloudové BI je vhodné ***uchovávat v distribuované podobě*** na rozdíl od on-premises BI, kde jsou metadata obvykle uložena v centralizovaném firemním repositáři. Daný způsob pomáhá provádět rozsáhlé datové analýzy v reálném (nebo téměř reálném) čase. V současné době dané provedení metadat umožňuje například Apache Hadoop. (Deshpande 2017; The Apache Software Foundation 2019; Wikibon 2011; Harcharan a Seema 2018)

33.9 Komponenty určené především pro standardní BI a jejich role v cloudovém prostředí

Některé ***komponenty v cloudovém prostředí jsou integrovány*** v rámci jiných komponent, případně nejsou vůbec zmiňovány anebo ***postrádají využití*** pro Cloud Computing. Jedná se především o ***operativní uložiště dat (ODS) a datová tržiště. ODS*** (včetně Enterprise Application Integration) bývá integrováno v rámci cloudových ***platforem datové integrace***. Datová tržiště nejsou v rámci cloudových řešení používána. Pouze v určitých případech se lze setkat s tzv. zdánlivými datovými tržišti rozdělující virtuálně data například podle oddělení, avšak data jsou fyzicky nahrávána z jediného datového skladu společnosti. (Náhlovský 2015; MBI 2019c)

33.10 Analýza odlišností standardního a cloudového BI, včetně návrhu hledisek pro rozhodování mezi oběma provedeními

Jsou zde analyzovány dvě základní pojetí k budování Business Intelligence. Jako prvním se tato část zabývá standardním interním (on-premises) pojetím (včetně jeho výhod a problémů i s návrhy řešení). Druhá část se věnuje důkladné analýze cloudového BI. V uvedených analýzách jsou zmíněny ***podstatné rozdíly mezi cloudovým a on-premises BI***, avšak je vhodné sumarizovat dané analýzy standardního a cloudového BI ve shrnujícím porovnání obou přístupů.

Za ***základní rozdíl*** lze považovat ***odpovědnost za správu BI***. V případě cloudového BI je za veškeré cloudové komponenty a data v cloudovém prostředí odpovědný dodavatel na rozdíl od on-premises BI, kde se o svá data a systémy stará společnost interně.

S uvedeným rozdílem souvisí i ***oblast integrace***. Pro ***cloudové BI*** obvykle ***dodavatelé nabízejí plán integrace*** a celý proces je možné automatizovat. V případě interního BI je obvykle třeba opakovaně manuálně integrovat systémy a data pro potřeby jednotlivých analýz. Interní BI je však obvykle snazší integrovat s dalšími podnikovými on-premises systémy. Taktéž ***úpravy (customizace) BI v podniku jsou obvykle snáze proveditelné*** v interní variantě ve smyslu, že firma má nad změnami v systému kontrolu a je důkladně obeznámena se svým BI řešením. Uvedený fakt však záleží na komplexnosti IT infrastruktury v podniku, jelikož firma musí sama řešit přizpůsobování i všech ovlivněných modulů, aplikací, systémů, a to včetně následného testování. ***V cloudových řešeních*** bývají možnosti úprav ***obvykle omezenější***. Nicméně jakékoliv případné změny dohodnuté se zákazníkem implementuje dodavatel.

Dodavatel rovněž v cloudové variantě BI ***spravuje veškerou infrastrukturu*** (hardware a software) a zákazníkovi se tak otevírá příležitost soustředit veškeré úsilí do předmětu podnikání a byznysových analýz. Interní on-premises BI si naopak spravuje zákaznický podnik samostatně, a to včetně veškeré hardware a software infrastruktury s čímž souvisí i obvykle nepříliš široké možnosti přístupu k BI aplikacím. K internímu BI řešení je totiž obvykle přistupováno prostřednictvím klientských aplikací či přes vzdálenou plochu na rozdíl ***od cloudového BI***, které obvykle nabízí ***přístup pomocí webové aplikace*** prostřednictvím internetového prohlížeče ***na jakémkoliv zařízení*** (například na mobilním telefonu, tabletu, počítači atd.).

U standardního ***on-premises BI*** má však podnik značnou ***kontrolu nad řízení bezpečnosti*** svých dat a systémů. Uvedený fakt je však výhodou jen v případě, že podnik má dostatečné zkušenosti s řízením bezpečnosti informačních systémů.

Při využívání ***cloudového BI*** podnik ***platí jen za to, co využívá*** obvykle formou pay-as-you-go. V případě interního BI musí podnik pokrýt pořizovací, udržovací a aktualizační náklady na hardware i software, mzdy správců IT infrastruktury, nájemné za potřebné prostory, náklady na energie a další s daným BI řešením související náklady. (Cristescu 2016)

Každý podnik rozhodující se mezi cloudovým a interním BI by měl podle svých potřeb nejprve zvážit zmíněné rozdíly a přednosti a nedostatky obou provedení. Dané výhody a nevýhody je poté vhodné postupně jednotlivě posoudit z následujících hledisek navržených do formy analytických otázek (Cristescu 2016; Náhlovský 2015):

- ***Jak změní daná výhoda či nevýhoda*** průběh činností v podniku. Jaké systémy, pracovní role, procesy, činnosti, strategie atd. může ovlivnit?
- ***Jaká rizika*** jsou s danou předností či nedostatkem v daném BI provedení spojena a jak danými riziky může být podnik ovlivněn důležitého uvedení pravděpodobnosti nastání rizika a jeho závažnosti například na stupnici od 1 do 4).
- ***Jaká je pravděpodobnost***, že se daná výhoda při konkrétním využití v podniku skutečně realizuje.
- Jak je daná ***výhoda konkrétně realizována*** v daných BI řešeních a v nabídkách dodavatelů.
- Je vhodné si vyjasnit následující oblasti z pozice zákazníka, tj. ***nechat si dodavateli zodpovědět*** následující dotazy:
    - Pokrývá nabízené řešení kompletně požadovanou funkcionalitu?
    - Jak je v daném řešení konkrétně zajištěna bezpečnost systémů a dat proti prolomení a odcizení?
    - Jaké jsou možnosti customizace?
    - Jaké jsou s daným řešením spojeny legislativní požadavky a jak jsou konkrétně nabízeným produktem a službami pokryty?
    - Umožňuje dané provedení BI pokrýt nároky podniku (i z dlouhodobého hlediska) na počet a přístup uživatelů k systému, kapacitu uložiště a rychlost a kvalitu přenosu dat?
    - Poskytuje dodavatel možnost vyzkoušet si dané řešení?
    - Jsou reference na dodavatele především kladné, podrobné a je jich značné množství?
    - Jakým způsobem poskytuje dodavatel zákaznickou podporu?
    - Jak dlouhá je doba reakce a jak dlouhá je doba, než dodavatel začne řešit nahlášené incidenty? Uvedené ***doby je vhodné sjednat ve smlouvě SLA*** včetně definice postihů za jejich nedodržení s progresivním navyšováním sankcí pro dodavatele za každou další prodlevu. V SLA by rovněž měla být pokryta konkrétně u daného řešení dostupnost, tj. v jakých časech je bezpodmínečně nutné, aby bylo řešení dostupné a je prioritní znovu zprovoznit při případném výpadku a jak dlouho mohou být maximálně nedostupné (včetně v čase progresivně se navyšujících sankcí za prodlevu oproti dané době). Rovněž by smlouva měla pokrývat způsoby komunikace (včetně možností eskalace při momentální nedostupnosti příslušných osob) a obsahovat přesně vymezený cenový model.
- Jak jsou obě varianty v konkrétním provedení pro podnik ***nákladné*** a která z nich je pro realizování v daném podniku skutečně ***finančně příznivější***. Důležité je rovněž provést analýzu celkových nákladů na vlastnictví (Total Cost of Ownership neboli ***TCO*)** u konkrétních nabídek jednotlivých dodavatelů. Je třeba od dodavatelů zjistit přesný cenový model pro konkrétní provedení v podniku.

33.10.1 Cloud Business Intelligence jako vizualizační nástroj

Variantou cloud BI důležitou pro potřeby analýzy a návrhů je varianta, při které je ***Cloud BI využito pouze jako prostředek pro datovou vizualizaci***. Podnik tedy v této variantě přenáší do cloudového prostředí pouze ***analytickou, vizualizační a reportingovou komponentu***. Daný přístup je tedy užitečný především z hlediska rozšíření možností přístupu prostřednictvím internetu k datům uloženým interně ve firmě a souvisejícím základním BI analýzám na různá zařízení (mobilní telefon, tablet, notebook) kdykoliv odkudkoliv. (MBI 2019c; Náhlovský 2015)

Určitou ***nevýhodou*** v daném provedení je potřeba ***udržovat vlastní infrastrukturu*** (datový sklad, ETL nástroj a související hardware) i omezení možností datových analýz. ***Pro využití pokročilejších funkcí***, jako například strojového učení například k automatickému upozorňování na důležité vývoje v datech, je ***obvykle potřeba značného výkonu serverů dodavatele***. Aby však mohl být zmíněný výkon využit, je třeba uložit data do cloudového datového skladu, a nikoliv přenést pouze vizualizační komponentu do cloudového prostředí. Přesunutí pouze vizualizační komponenty do cloudu rovněž obvykle znamená ***omezení množství agregací v dané komponentě***, jelikož je třeba data neustále nahrávat z interního uložiště do cloudového prostředí pro provedení kalkulací. Je tedy obvykle potřeba provést ***některé propočty předem***, uložit je a teprve poté je možné provést finální výpočet. (MBI 2019c; Ephrati 2017).

33.10.2 Full-Stack Cloud Business Intelligence

V této variantě jsou veškeré komponenty Business Intelligence v cloudovém prostředí (tj. ETL, datový sklad a analytický, reportovací a vizualizační nástroj). Daná realizace bývá rovněž označována jako tzv. ***Full-Stack Cloud Business Intelligence*** neboli ve volném překladu jako BI s plným zásobníkem kvůli obsažení všech součástí v integrovaném cloudovém provedení. (TrustRadius 2019; Náhlovský 2015)

Jako nejvýznamnější ***výhody*** lze zmínit obvykle mnohem ***snadnější nasazení*** než v případě on-premises řešení a rovněž ***přenesení starostí o provoz*** řešení ze zákazníka ***na dodavatele***, což v praxi znamená obvykle ***nižší náklady na údržbu*** řešení oproti internímu BI a oproti předchozí variantě (pouze vizualizační nástroj v cloudovém prostředí). V daném provedení jsou rovněž k dispozici široké ***možnosti hloubkových analýz nad daty*** včetně automatizovaného vyhledávání vzorců v datech a následných upozornění pro management o důležitém vývoji. To je možné především kvůli uložení dat zákazníků u dodavatele (datový sklad v cloudovém prostředí), kde je následně ***možné využít značného výkonu*** poskytovatelských serverů. Daný výkon umožňuje využívat například strojového učení, což nabízí klientům automatizované analýzy s hlubším pochopením významu dat. (Techedge S.p.A. 2018)

Uvedená varianta je ***nejvhodnější při budování nového BI*** v podniku. Ideální se daná varianta jeví v případě, kdy není potřeba žádná data přesouvat z původně používaného interního datového skladu. Rovněž ***výhodou je***, když již podnik využívá nebo ***migruje některé další systémy se zdrojovými daty do cloudového prostředí***. Dané řešení je rovněž nejvíce ***výhodné pro malé a střední podniky***, jelikož oproti velkým podnikům nepotřebují koupit značné množství výkonu a přístupů pro mnoho uživatelů. Školení jsou méně nákladná, jelikož je třeba proškolit méně uživatelů než u velkých společností. ***Velké společnosti*** uvedená fakta (současně s obavami o bezpečnost svých dat) bohužel v některých případech odradí od realizace v daném provedení, ačkoliv daná varianta přináší významné výhody pro všechny společnosti.

Daná varianta je ***nejsnáze realizovatelná***, pokud zákazník odebírá ***řešení od jediného dodavatele***, jelikož veškeré komponenty jsou v daném provedení již integrovány. V praxi se obvykle jedná o řešení realizované v modelu software jako služba (Software as a Service) s obsloužením více zákazníků najednou (tzv. multi-tenant přístup). Realizace ***s využitím komponent od různých dodavatelů*** se může projevit jako obtížná, nicméně lze ***využít konektorů*** ke vzájemné integraci. Důvodem k implementaci jmenované varianty Cloud BI s použitím částí od různých dodavatelů může být například u některých komponent výhodnější cenová nabídka či vhodnější nebo širší spektrum funkcí pro potřeby podniku.

33.11 Analýza bezpečnostních aspektů cloudového Business Intelligence

Bezpečnost Cloud BI aplikací lze rozdělit na ***dvě hlavní části***:

- Za prvé se jedná o ***bezpečnost z hlediska procesů v podniku*** (tj. jak je nastaven management bezpečnosti informací v podniku), která může být naplněna například postupováním podle standardů.
- Za druhé lze hovořit ***o digitální bezpečnosti***, kterou lze dělit dále na oblasti autentizace, autorizace a zabezpečení dat.

Základem ***digitální bezpečnosti*** jsou ***autentizační mechanismy*** zajišťující ideálně pouze jediné přihlášení v jakémkoliv ze systémů integrovaných na Cloud BI řešení. Daný způsob se nazývá ***Single Sign-On*** (ve zkratce označovaný jako SSO) a zajišťuje, že uživatel se přihlásí pouze jednou a údaje nutné pro autentizaci v ostatních systémech jsou dále k dispozici všem propojeným systémům automaticky (bez nutnosti opětovného přihlášení). ***Single Sign-On je v Cloud BI řešeních možné zařídit prostřednictvím protokolů SAML 2.0*** (Security Assertion Markup Language neboli značkovacího jazyka pro vynucení zabezpečení) ***a OAuth 2.0*** (Open Authentication neboli otevřené autentizace). Oba protokoly využívají ***centrálního autentizačního serveru*** (jinak také označovaného jako poskytovatel identity neboli Identity Provider či Corporate Identity Provider). (Indu et al. 2018; Pierre 2019; Ali et al. 2015) ***Scénář průchodu uživatele systémem*** je u obou protokolů shodný v následující posloupnosti (Indu et al. 2018; Karaki 2019):

- Uživatel se chce přihlásit do jakékoliv z propojených aplikací.
- Uživatel je odkázán na centrální autentizační server (poskytovatele identity), kde se prostřednictvím šifrované komunikace autentizuje.
- Uživateli je poslán tzv. ***session cookie*** (tj. data o vytvořené relaci dále uložená v prohlížeči uživatele označovaná zkratkovitě také pouze jako ***session*** (AllAboutCookies.org 2019)) v zašifrované podobě o tom, že má nyní přístup do systémů.
- Uživatel nyní může přistupovat ke všem propojeným systémům. Každá další aplikace, ke které uživatel přistupuje pouze ověří, že uživatel má platné session cookies.
- Pokud relace vyprší (relace jsou zpravidla časově omezené) nebo aplikace nedokáže nalézt platnou session cookies, pak je celý proces opakován.

Druhou částí digitální bezpečnosti v Cloud BI je autorizace. ***Autorizace*** je obvykle řešena prostřednictvím několika pojetí:

- První možností je ***přidělovat práva přístupu*** konkrétním uživatelům.
- Daný způsob je však značně obtížný na administraci, pokud je aplikován jako jediný způsob autorizace v Cloud BI řešení. Druhé pojetí je založeno na ***přidělování tzv. rolí uživatelům***. Role definuje možnosti přístupu v aplikaci. Jedná se například o role administrátora, běžného uživatele, manažera atd. Uvedené pojetí je vhodnější pro přehlednější administraci přístupů v Cloud BI řešení.
- Pojetí, kde jsou přidělována práva uživatelům ***na základě politik***. Za politiku lze považovat takřka jakékoliv byznysové pravidlo, které je třeba v systému aplikovat. Například, že všichni uživatelé z vývojového oddělení mají přidělena administrátorská práva. Jiným příkladem může být, že uživatelé nemají právo přistupovat k systému po půlnoci (například kvůli pravidelným zálohám, aktualizacím apod.). Uvedená práva lze definovat jak na úrovni Cloud BI aplikace, tak na úrovni serveru. V případě definice na serverové úrovni k serveru následně aplikace přistupuje, aby přidělila přistupujícím uživatelům příslušná oprávnění. (Indu et al. 2018; Ali et al. 2015)

Bezpečnost dat je v Cloud BI zajišťována jednak ***šifrováním dat*** samotných a rovněž i ***zabezpečením komunikace po síti*** internetu. Zabezpečení komunikace bývá realizováno prostřednictvím vrstvy ***VPN*** (Virtual Private Network neboli virtuální privátní sítě) nad autorizací a autentizací, která umožňuje vytvořit zašifrovanou síť pro zákaznický podnik prostřednictvím ***maskování klientských IP adres***. Veškerá komunikace rovněž obvykle probíhá prostřednictvím šifrovaného ***HTTPS*** (Hypertext Transfer

Protocol Secure neboli zabezpečeného hypertextového přenosového protokolu). (Indu et al. 2018; Ali et al. 2015; Pierre 2019; Juan-Verdejo et al. 2014).

33.12 Zdroje

ALI, Mazhar, Samee U. KHAN a Athanasios V. VASILAKOS, 2015. Security in cloud computing: Opportunities and challenges. *Information Sciences* [online]. **305**, 357–383. ISSN 0020-0255. Dostupné z: doi:10.1016/j.ins.2015.01.025

BAUTISTA VILLALPANDO, Luis, Anderson RAVANELLO, Jean-Marc DESHARNAIS, Alain APRIL a Abdelouahed GHERBI, 2014. Performance Measurement for Cloud Computing Applications Using ISO 25010 Standard Characteristics. In: *IWSM-MENSURA 2014* [online]. Rotterdam, Nizozemsko: University of Quebec [vid. 2019-10-10]. Dostupné z: doi:10.1109/IWSM.Mensura.2014.33

CERTIFICATION & INFORMATION SECURITY SERVICES GMBH, 2019. *CIS | System Certification | Cloud Computing* [online] [vid. 2019-10-08]. Dostupné z: https://www.cis-cert.com/Pages/com/System-Zertifizierung/Cloud-Computing/ISO-27018-and-ISO-20000-9.aspx

CRISTESCU, Marian Pompiliu, 2016. Traditional Enterprise Business Intelligence Software Compared to Software as a Service Business Intelligence. *Informatică economică* [online]. **20**(1), 39–47. ISSN 1453-1305, 1842-8088. Dostupné z: doi:10.12948/issn14531305/20.1.2016.04

DELOITTE, 2017. *Measuring the economic impact of cloud computing in Europe* [online]. Text. B.m.: Evropská komise [vid. 2019-06-26]. Dostupné z: https://ec.europa.eu/digital-single-market/en/news/measuring-economic-impact-cloud-computing-europe

DESHPANDE, Piyush P., 2017. Hadoop Distributed FileSystem : Metadata Management [online]. **4**(10) [vid. 2019-10-10]. ISSN 2395-0056. Dostupné z: https://pdfs.semanticscholar.org/2406/8d3dcc310be172fccd05c92b66c7a5912020.pdf

EPHRATI, Aya, 2017. Data visualization is not data analytics | Business Intelligence. *Sisense* [online] [vid. 2019-08-22]. Dostupné z: https://www.sisense.com/blog/buyers-beware-data-visualization-not-data-analytics/

FLEXERA, 2019. Cloud Computing Trends: 2019 State of the Cloud Survey. *Cloud Management* [online] [vid. 2019-06-19]. Dostupné z: https://blogs.flexera.com/cloud/cloud-industry-insights/cloud-computing-trends-2019-state-of-the-cloud-survey/

HARCHARAN, Jit Singh a Bawa SEEMA, 2018. Scalable Metadata Management Techniques for Ultra-Large Distributed Storage Systems – A Systematic Review. *ACM Comput. Surv.* [online]. **51**(4), 82:1–82:37. ISSN 0360-0300. Dostupné z: doi:10.1145/3212686

HEADLEY, Megan a Alan COOKE, 2016. *Types of Business Intelligence Tools* [online]. B.m.: TrustRadius Inc. [vid. 2019-10-15]. Dostupné z: https://www.trustradius.com/business-intelligence-bi

INDRIASARI, Elisa, Harjanto PRABOWO, Meyliana WIDJAJA a Achmad HIDAYANTO, 2018. Key benefits, challenges, and risk of cloud business intelligence: A systematic literature review. *International Journal of Mechanical Engineering and Technology*. **9**, 819–831. ISSN 0976-6340.

INDU, I., P. M. Rubesh ANAND a Vidhyacharan BHASKAR, 2018. Identity and access management in cloud environment: Mechanisms and challenges. *Engineering Science and Technology, an International Journal* [online]. **21**(4), 574–588. ISSN 2215-0986. Dostupné z: doi:10.1016/j.jestch.2018.05.010

INFORMATICA, 2019. *Data Integration Cloud | Informatica* [online] [vid. 2019-10-02]. Dostupné z: https://www.informatica.com/products/cloud-integration/integration-cloud.html#fbid=YUc-mWEbYZ9

JANOSCHEK, Nikolai, 2018. *Most Common Business Intelligence Problems | 2,500 User Responses* [online]. B.m.: Business Application Research Center [vid. 2019-09-23]. Dostupné z: https://bi-survey.com/business-intelligence-problems

JUAN-VERDEJO, A., B. SURAJBALI, H. BAARS a H. KEMPER, 2014. Moving Business Intelligence to Cloud Environments. In: *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* [online]. s. 43–48 [vid. 2019-02-26]. Dostupné z: doi:10.1109/INFCOMW.2014.6849166

KARAKI, Jad, 2019. Identity Management: SAML vs. OAuth2 vs. OpenID Connect. *Medium* [online] [vid. 2019-10-19]. Dostupné z: https://medium.com/@jad.karaki/identity-management-saml-vs-oauth2-vs-openid-connect-c9a06548b4c5

LOSHIN, David, 2013. *Business Intelligence : The Savvy Manager's Guide* [online]. 2. vydání. Waltham, MA: Morgan Kaufmann [vid. 2019-02-26]. ISBN 978-0-12-385889-4. Dostupné

z: http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=486587&lang=cs&site=ehost-live

MELL, Peter a Tim GRANCE, 2011. *The NIST Definition of Cloud Computing* [online]. NIST Special Publication (SP) 800-145. B.m.: National Institute of Standards and Technology [vid. 2019-06-19]. Dostupné z: doi:https://doi.org/10.6028/NIST.SP.800-145

NÁHLOVSKÝ, Tomáš, 2015. *Business Intelligence v prostředí Cloudu* [online]. B.m. [vid. 2019-06-16]. Vysoká škola ekonomická v Praze, nám. W. Churchilla 4, 130 67 Praha 3, http://www.vse.cz. Dostupné z: http://www.vse.cz/vskp/eid/49424

PIERRE, Thierry, 2019. *SAP Analytics Cloud - Live Connection and Security - Best Practices - SAP Analytics Cloud - SCN Wiki* [online] [vid. 2019-10-19]. Dostupné z: https://wiki.scn.sap.com/wiki/display/BOC/SAP+Analytics+Cloud+-+Live+Connection+and+Security+-+Best+Practices

SYSAID TECHNOLOGIES LTD., 2019. Hello ISO 20000 (2018), We've Been Expecting You. *Joe The IT Guy* [online]. [vid. 2019-10-10]. Dostupné z: https://www.joetheitguy.com/2019/02/20/hello-iso-20000-2018-weve-been-expecting-you/

TECHOPEDIA INC., 2019a. What is a Domain Controller? - Definition from Techopedia. *Techopedia.com* [online] [vid. 2019-08-28]. Dostupné z: https://www.techopedia.com/definition/4193/domain-controller-dc

TECHOPEDIA INC., 2019b. What is Database as a Service (DBaaS)? - Definition from Techopedia. *Techopedia.com* [online] [vid. 2019-06-19]. Dostupné z: https://www.techopedia.com/definition/29431/database-as-a-service-dbaas

THE APACHE SOFTWARE FOUNDATION, 2019. *Apache Hadoop* [online] [vid. 2019-10-10]. Dostupné z: https://hadoop.apache.org/

TRUSTRADIUS, 2019. *Full-Stack Business Intelligence* [online] [vid. 2019-09-30]. Dostupné z: https://www.trustradius.com/full-stack-business-intelligence

WIKIBON, 2011. *Cloud Meta Data: The Future of Business Intelligence - Wikibon* [online] [vid. 2019-10-10]. Dostupné z: http://wikibon.org/wiki/v/Cloud_Meta_Data:_Driving_New_Business_Model#Metadata_for_Big_Data_and_the_Cloud

33.13 Závěry, doporučení

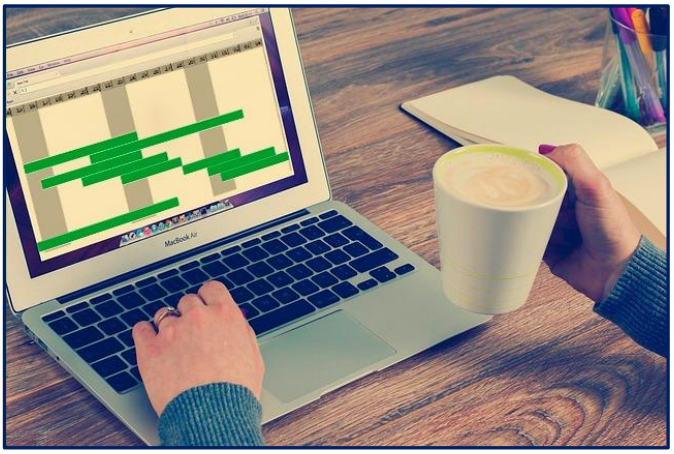
Kapitola představuje ***pracovní závěry*** k analýze variant architektur a využívání BI řešení na bázi cloud computingu.

- Za ***základní rozdíl*** architektur lze považovat ***komponenty*** obsažené v daných variantách Cloud BI. ***Full-Stack Cloud BI*** je kompletním řešením obsahujícím v cloudovém prostředí datový sklad, ETL pumpy a analytický, reportingový a vizualizační nástroj. Cloud BI ve variantě ***pouze pro vizualizaci*** nabízí v cloudovém provedení pouze analytický, reportovací a vizualizační nástroj.
- Další rozdíl spočívá ***v nasazení daných variant***. V kompletním (***full-stack***) Cloud BI je třeba ***veškerá data ze starého interního datového skladu migrovat*** do skladu v cloudovém prostředí a je rovněž třeba propojit nové řešení (především ETL) se zdrojovými i jinými systémy. V případě Cloud BI výlučně pro ***vizualizační potřeby*** je třeba ***pouze napojit nový analytický, reportingový a vizualizační nástroj*** k internímu datovému skladu. Migrace dat v daném případě není třeba, jelikož je využíváno interního datového skladu, avšak pro potřeby výpočtů je beztak nutné menší části dat nahrávat do cloudového prostředí. V obou řešeních je třeba proškolit zaměstnance v práci s novými nástroji.
- Full-Stack Cloud BI a Cloud BI pouze pro vizualizaci se rovněž ***liší v náročnosti provozu***. V případě prvním se o veškerý provoz infrastruktury stará dodavatel, a zákazník se tak může soustředit pouze na vykonávání analýz pro byznysové potřeby. V druhé variantě se dodavatel zabývá pouze provozem nástroje analytického,

reportingového a vizualizačního. O veškerou další infrastrukturu (datový sklad, ETL pumpy) se musí starat zákazník, což znamená rovněž nutnost zaměstnávat potřebný IT personál a disponovat odpovídajícími znalostmi v oblasti BI infrastruktury.

- ***Integrace*** je rozdílná v obou variantách. V případě přesunutí ***pouze vizualizační***, reportingové a analytické komponenty do cloudového prostředí je třeba ji ***propojit s datovým skladem a případně dalšími systémy*** (kam například budou posílány výsledky analýz či reporty). V případě ***Full-Stack*** cloudového BI je třeba ***integrovat celé BI řešení se zdrojovými produkčními systémy*** a následně i s dalšími aplikacemi, které požadují výstupy z analýz a reporty. Pokud však využívá podnik komponent od více dodavatelů, je třeba rovněž ***integrovat komponenty*** mezi sebou (například prostřednictvím konektorů), proto bývá obvykle z pohledu integrace snazší využívat již obvykle integrovaných komponent od jediného dodavatele.
- ***Možnosti analýz*** se taktéž u obou variant odlišují. V případě ***Full-Stack*** Cloud BI je ***možné obvykle realizovat hloubkové analýzy*** i s využitím strojového učení, jelikož veškerá data jsou uložena v datovém skladu na serverech dodavatele. Je tedy ***možné využít výkonu daných serverů*** pro provedení náročných analýz s využitím i automatizovaného upozorňování na trendy v datech. Ve variantě přesunutí pouze analytické, reportingové a ***vizualizační*** komponenty do cloudového prostředí jsou ***možnosti analýzy obvykle skrovnější***, jelikož data ukládá klient v interním podnikovém datovém skladu a poskytovatel cloudových služeb k nim proto obvykle má omezený přístup.
- ***Cena*** se u daných variant ***liší*** rovněž poměrně ***významně***. V případě využití pouze analytického, reportingového a ***vizualizačního*** nástroje je třeba jednak platit obvykle formou ***pay-per-use dodavateli za danou komponentu***. O údržbu dané komponenty se však již podnik dále nestará, nicméně musí udržovat veškerou zbylou BI infrastrukturu, což znamená finanční náklady na provoz hardware, software, energie, nájemné, mzdy IT personálu zodpovědného za BI infrastrukturu atd. U ***Full-Stack*** je třeba pouze platit zpravidla zmíněnou formou ***pay-per-use za užívání kompletního BI řešení*** dodavateli, náklady na údržbu BI infrastruktury odpadají.
- Posledním významným rozdílem je ***rozdíl v bezpečnosti***. V obou variantách je potřeba ze strany zákazníka vložit určitou důvěru v dodavatele, že zabezpečí (pomocí šifrování) veškerá klientská data a že bude ideálně postupovat podle systému řízení bezpečnosti informací (například podle normy ***ISO/IEC 27001***). V případě pouze ***vizualizační***, reportingové a analytické komponenty v cloudovém prostředí dodavatel nakládá pouze s daty nutnými pro propočty v cloudovém nástroji, tj. po menších dávkách. V případě ***Full-Stack*** Cloud BI jsou veškerá data na serverech dodavatele, tj. z pohledu klienta je třeba vynaložit více důvěry v poskytovatele. Z uvedeného důvodu jsou analyzovány základní bezpečnostní aspekty Cloud BI, které by měly vyjasnit způsob zajištění bezpečnosti daného řešení a tím snížit s ním spojenou nedůvěru.

J) Řízení podnikové výkonnosti a jeho formy



[35] Řízení podnikové výkonnosti
[36] Řízení výkonnosti podle uplatnění

Řízení podnikové výkonnosti, Corporate Performance Management, CPM je disciplína, která kombinuje principy a přístupy managementu, metodiky a metriky podporovaná aplikacemi, nástroji a infrastrukturou BI, která umožňuje uživatelům definovat, monitorovat a optimalizovat výsledky a výstupy. Právě úzká vazba na CPM na BI a podnikovou analytiku opravňuje jeho zařazení do tohoto kontextu. Oddíl představuje vymezení uvedené disciplíny v následujících bodech:

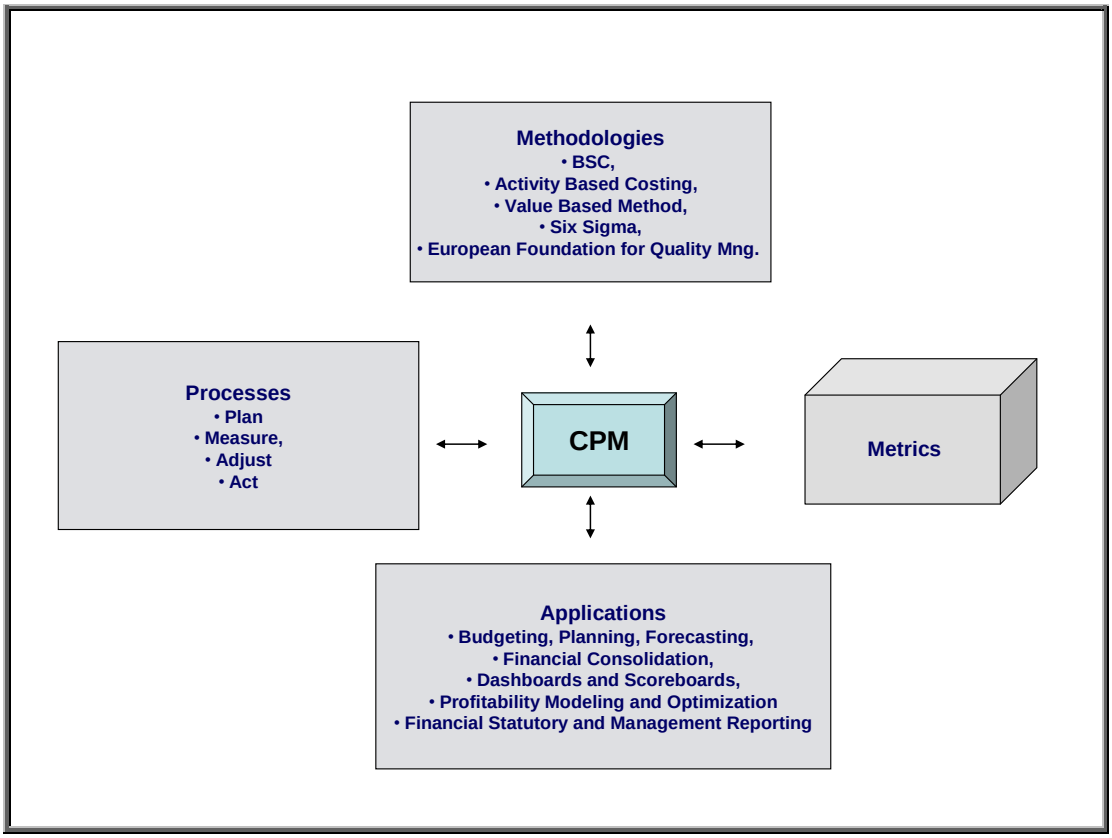
- ***Vymezení řízení podnikové výkonnosti, CPM***, tj. jednotlivých komponent a jejich vazeb, jeho efekty a problémy,
- Uplatnění CPM, resp. PM (Performance Management) ***v různých oblastech řízení firmy nebo odvětví***, např.:
  - Řízení výkonnosti IT (***IT Performance Management***, ***IT PM***).
  - Řízení výkonnosti eGovernmentu (***EGovernment Performance Management***, ***eGPM***).
  - Řízení výkonnosti marketingu (***Marketing Performance Management, MPM***).
  - Řízení výkonnosti prodeje (***Sales Performance Management, SPM***).

35. Řízení podnikové výkonnosti, CPM

Účelem kapitoly je:

- definovat ***obsah a struktury systémů řízení podnikové výkonnosti, CPM*** tvořících často součást nebo nadstavbu BI aplikací,
- analyzovat očekávané, resp. potenciální ***přínosy*** pro řízení firmy a na druhé straně obvyklé problémy spojené s jejich nastavením.

Řízení podnikové výkonnosti, Corporate Performance Management, CPM je kombinace managementu, metodik a metrik podporovaná aplikacemi, nástroji a infrastrukturou, která umožňuje uživatelům definovat, monitorovat a optimalizovat výsledky a výstupy tak, aby bylo dosaženo cílů osobních či cílů organizační jednotky v souladu se strategickými cíli podniku. Corporate Performance Management (CPM) je hlavním představitelem systémů řízení výkonnosti. Základní vymezení CPM vytvořila společnost Gartner, a to je v souladu s obecnou definicí systémů řízení výkonnosti. CPM tvoří čtyři základní segmenty, resp. komponenty řešení, které jsou vzájemně provázané (viz další obrázek).



Obrázek 35-1: Vymezení CPM

Základní ***segmenty CPM*** tvoří:

- Komplex ***manažerských metod***, které tvoří metodologický logický základ podnikového řízení a jejichž principy se respektují v ostatních segmentech CPM. Do těchto metod obvykle patří Balanced Scorecard (BSC), Activity Based Costing (ABC), Value Based Management (VBM), Six Sigma a další.
- ***Podnikové procesy***, tj. plánovací, analytické, monitorovací vytvářející ve svém komplexu procedurální logiku podnikového řízení a navazující na uvedené manažerské metody.

- ***Metriky*** pro podnikové řízení postavené ***na principech business intelligence***, tj. představované klíčovými ukazateli ve vztahu k odpovídajícím dimenzím (zákaznické, komoditní atd.). Podstatnou charakteristikou metrik je jejich přiřazení k podnikovým procesům a vazba na specifikované manažerské metody. Metriky se v rámci CPM člení na tyto ***skupiny***:
  - ***KRI, Klíčové indikátory výsledků***, tj. metriky týkající se nákladů, výnosů a profitability podniku a sleduje se i spokojenost zákazníků,
  - ***KPI, Klíčové indikátory výkonnosti***, tj. metriky týkající se měření výkonnosti lidských zdrojů, výkonnosti interních procesů a další,
  - ***PI,*** ostatní indikátory výkonnosti, tj. specifické indikátory výkonnosti vybraných procesů, zdrojů či pracovních týmů podniku.
- Metody, procesy a metriky jsou základem pro ***plánovací a analytické aplikace*** postavené na technologiích a přístupech business intelligence. Do této skupiny patří především:
  - aplikace pro ***prognózování, plánování*** a tvorbu rozpočtů založené na multidimenzionálním vyjádření a sledování všech vybraných ukazatelů a na silné provázanosti na podnikovou strategii, kde formulace strategie zahrnuje modelování efektů strategických cílů a operací s promítáním do prognóz, plánů a rozpočtů,
  - ***scoreboardy*** analyzující vazby cílových a reálně dosahovaných klíčových indikátorů výkonnosti (KPI),
  - ***dashboardy*** (též digital dashboard, executive dashboard nebo enterprise dashboard),
  - aplikace pro ***analýzy a modelování ziskovosti***, které umožňují analýzy z pohledu různých dimenzí, tj. nákladových objektů, např. zákazníků, segmentů zákazníků, produktů, služeb atd. s cílem sledovat jejich profitabilitu,
  - ***finanční reporting***.

CPM tak představuje ***komplex metod, procesů, dat a aplikací***, který je nezbytné pro konkrétní podnikové potřeby vždy přiměřeně přizpůsobit. Na druhé straně tento koncept je aplikovatelný jak pro podniky různé odvětvové orientace, tak velikosti.

CPM je založeno na principech a technologiích business intelligence, ale není jeho synonymem. CPM je ***rozšiřuje o koncept „řízení“,*** který zahrnuje procesy jako plánování, prognózování a základní východiska podnikové strategie a úzce se váže na metody a metodiky pro řízení výkonnosti (např. BSC, ABC a další). Uplatnění BI aplikací podporují uvedené metodiky a je proto jádrem současného konceptu CPM).
K ***hlavní funkcionalitě*** úloh a nástrojů CPM tak patří komplexní podnikové analýzy na všech úrovních řízení, plánování, rozpočtování a prognózování na bázi klíčových firemních metrik. Lze pracovně vymezit ***tři hlavní pojetí řízení výkonnosti***:

- ***analytické*** pojetí řízení výkonnosti – zaměřené na vymezení a provázanost podstatných komponent řízení podniku,
- ***ekonomické*** pojetí řízení výkonnosti – obvykle chápané jako součást manažerského účetnictví,
- ***manažerské*** pojetí řízení výkonnosti – definující zejména manažerské postupy pro nastavení a hodnocení výkonnosti dílčích útvarů a pracovníků podniku.

35.1 Efekty a přínosy

- Řízení podnikové výkonnosti v analytickém pojetí směřuje k ***určení a vymezení hlavních podnikových procesů***, resp. úloh, ***metrik, metod a aplikací***, tj. komponent, které významně ***ovlivňují úspěšnost podniku***,
- Definování vazeb mezi uvedenými komponentami ***podporuje systematický charakter*** řízení,
- Vymezení postupů a ***zodpovědností pracovníků za plnění cílů*** ve výkonnosti podniku (manažerské pojetí) zvyšuje úspěšnosti jejich dosažení,
- ***Ekonomické pojetí*** vede k výběru nejdůležitějších ekonomických, převážně finančních ukazatelů a soustřeďuje na ně hlavní pozornost,

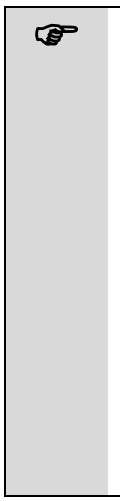
- Hlavním efektem je ***provázanost všech tří uvedených pojetí*** a vytvoření předpokladů pro kvalitní a systematické řízení podnikové výkonnosti,
- Uplatňuje se ***princip uzavřené smyčky*** (*closed-looped CPM*), tj. výsledky monitorování a analýz výkonnosti se okamžitě promítají do adekvátních opatření v byznysu řízení firmy a případně i do redefinování byznys cílů.
- Princip uzavřené smyčky ***posiluje i vazby a komunikaci v řízení*** na strategické, taktické i operativní úrovni. To znamená i stejné pochopení strategických cílů firmy u pracovníků na všech úrovních řízení a jejich podporování.

35.2 Problémy a omezení

- Analytické pojetí výkonnosti se orientuje na uvedené 4 komponenty (procesy, metriky, metody a aplikace). Je otázkou, ***zda neprovázat do tohoto řešení některé další***, zejména role, nebo data.
- Předpokladem pro kvalitní realizaci řízení podnikové výkonnosti je ***pochopení a zájem managementu***, ten není často splněn.
- Pro reálné uplatnění celého konceptu je ***nezbytná kvalifikační příprava*** pracovníků a pro ni se často obtížně hledá časový prostor,
- Princip uzavřené smyčky v CPM vyžaduje nasazení ***analytických aplikací provozovaných v reálném čase***, včetně funkcionality varování (*allerts*),
- Předpokladem uplatnění uzavřené smyčky je i ***vytvoření systému KPI*** s jasně vymezenými vazbami zejména mezi finanční a provozními, zejména obchodními oblastmi řízení.

35.3 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řízení výkonnosti podniků.



- ***Řízení výkonnosti podniků*** (CPM, Corporate Performance Management) představuje koncept řízení podniků založený na ***provázanosti čtyř základních komponent:***
  - metodik a metod podnikového řízení,
  - podnikových procesů,
  - metrik, tj. ukazatelů a jejich analytických dimenzí
  - analytických a plánovacích aplikací.
- Koncept je ***založen na principu aplikací a technologií business intelligence***.
- Z pohledu ***používaných metod*** je vždy nutné posoudit, ***jaké efekty*** ve spojení s aplikacemi informatiky mohou přinést, jaké jsou pro podnik významné a podle toho i nasazení metod i investice do aplikací plánovat.
- Na druhé straně je nezbytné posuzovat i ***problémy, nebo faktory***, které nasazení CPM ve firmě budou ovlivňovat

36. Řízení výkonnosti podle uplatnění

Účelem kapitoly je:

- definovat ***obsah systémů řízení podnikové výkonnosti,*** v rozlišení podle jednotlivých oblastí, či sektorů uplatnění, včetně jejich efektů a případných problémů řešení,
- analyzovat ***možnosti*** řešení systému řízení výkonnosti IT, řízení výkonnosti eGvernmentu v rámci veřejné správy, řízení výkonnosti marketingu a prodeje.

36.1 Řízení výkonnosti IT

Řízení výkonnosti informatiky podniku (***IT Performance Management, IT PM***) představuje jednu z nejvýznamnějších oblastí uplatnění principů řízení informatiky a řízení výkonnosti. Musíme ale ***rozlišovat řízení výkonnosti podniku poskytujícího IT služby***, což je konkretizace konceptu CPM pro podniky tohoto typu, a ***řízení výkonnosti podnikové informatiky***, tj. informatiky jakéhokoli typu organizace. Je založeno na stejném schématu, jako v konceptu CPM, to znamená *na 4 provázaných segmentech řízení informatiky*:

- metodikách a metodách řízení informatiky,
- procesech řízení informatiky,
- metrikách, tj. ukazatelích a s nimi spojených analytických dimenzích,
- aplikacích řízení podnikové informatiky.

36.1.1 Efekty a přínosy

- Řízení výkonnosti IT vede ***k systematickému a řízení ve všech oblastech IT*** při provázanosti nejpodstatnějších komponent řízení.
- Koncept vytváří lepší předpoklady pro ***propojení IT s potřebami firmy***.
- Vytváření se podstatně lepší ***předpoklady pro přípravu analytických a plánovacích aplikací*** na principech podnikové analytiky, resp. Business Intelligence.

36.1.2 Problémy a omezení

- Koncept řízení výkonnosti IT ***není často respektován v podnikových metodikách*** a metodách řízení IT.
- Problémem je i ***nedostatečná motivace na straně IT pracovníků*** i uživatelů pro nasazení takového konceptu.
- Zejména pro systém metrik řízení IT ***nejsou k dispozici relevantní data***.

36.2 Řízení výkonnosti eGovernmentu

Řízení výkonnosti eGovernmentu musí ***respektovat odlišnosti od komerční sféry i odlišnosti jednotlivých typů a úrovní Orgánů veřejné moci (OVM)***. Přístupy a postupy řízení výkonnosti je praktické a nutné i v prostředí orgánů samosprávy uplatňovat, na druhé straně je přitom třeba brát v úvahu činnosti a služby, které OVM musí zajistit ze zákona, nebo vyplynou z provozních priorit.

Návrhy řešení systémů řízení výkonnosti musí reflektovat ***soustavu faktorů***, které budou taková řešení výrazně ovlivňovat. Problémy a rezervy v řízení výkonnosti eGovernmentu mají zásadní ***dopad na řízení výkonnosti a kvality celé veřejné správy*.**

36.2.1 Efekty a přínosy

- ***Snížení celkových nákladů*** na funkce veřejné správy (VS), zkvalitnění služeb VS,
- ***Zkvalitnění služeb*** daného OVM, snížení pracnosti na realizaci služeb, snížení provozních nákladů OVM,
- ***Zkvalitnění organizace poskytování IT*** služeb pro podniky a občany, optimalizace nákladů a pracnosti na poskytované IT služby,
- ***Zkvalitnění služeb podporujících byznys***, vyšší dostupnost informací relevantních pro byznys,
- ***Snížení časové náročnosti*** na využívání služeb VS, zvýšení dostupnosti informací VS nezbytných pro občanské aktivity a funkce.

36.2.2 Problémy a omezení

- Stávající legislativa a ***těžkopádný legislativní proces*** neodpovídají a nevyhovují možnostem a nárokům IT a IT aplikacím. Příkladem je nekonzistence přijímaných legislativních úprav, určení jejich platnosti bez ohledu na nezbytný čas pro jejich promítání do stávajících nebo nových aplikací atd.,
- I přes masivní investice do IT a infrastruktury jsou ***získávané efekty na nich postavených aplikací problematické.*** Důvodem je zejména méně kvalitní analýza a způsob řízení IT projektů. Např. existuje cca 7500 aplikací pro podporu veřejné správy, ale při plánování a řešení nových se často nebere v úvahu již dostupná funkcionalita. Očekávané efekty nových aplikací se nespecifikují, a tedy ani neposuzují,
- Jeden z klíčových problémů je ***úroveň znalostí a motivace na rozvoji eGovernmentu***, jak na úrovni uživatelské sféry (včetně managementu), tak ve veřejnosti.
- ***Není zatím dobře zvládnuto řízení*** (včetně registru) služeb eGovernmentu a s tím do značné míry související ekonomika eGovernmentu

36.3 Řízení výkonnosti marketingu

Řízení výkonnosti marketingu (***Marketing Performance Management, MPM***) je založeno na principech řízení výkonnosti podniku.

MPM zahrnuje služby, řešení, analytické nástroje a technologie, které ***podporují manažera marketingu*** (Chief Marketing Officer, CMO) a další manažery a specialisty v oblasti marketingu při analýzách marketingových dat a metrik (ukazatelů i jejich dimenzí), ***při formulování predikcí na trhu***, při optimalizaci marketingových programů a zdrojů.

36.3.1 Efekty a přínosy

- MPM nabízí ***možnosti přesnějších predikcí*** pokud jde o zákazníky, konkurenci, vývoj na trzích a simulace marketingových scénářů,
- Využití MPM přináší obvykle ***snižování nákladů na marketingové akce,*** efektivnější alokaci personálních i finančních zdrojů na akce, posilování zákaznické loajality a optimalizaci marketingového mixu.
- Na základě marketingových analýz v rámci MPM se vytvářejí ***kvalitnější podklady pro obchodní rozhodnutí*** na strategické úrovni.

36.3.2 Problémy a omezení

- Při ekonomických tlacích na snižování nákladů firmy je právě v oblasti marketingu a MPM třeba klást ***důraz na vyhodnocování ROI***, resp. návratnosti investic do jednotlivých projektů a programů,

- Je třeba dát důraz na ***komplexní a přesně měření nákladů na marketingové kampaně*** a jejich přínosů,
- Je účelné ***potvrdit investice do MPM*** finančním manažerem (CFO).

36.4 Řízení výkonnosti prodeje

Řízení výkonnosti prodeje (***Sales Performance Management, SPM***) je založeno na principech řízení výkonnosti podniku. SPM představuje komplex zejména analytických a plánovacích kapacit a úloh, které společně zahrnují ***funkcionalitu pro přípravu a optimalizaci prodejních plánů, alokování prodejních aktivit podle teritorií***, nastavování prodejních kvót, optimalizaci provizních pravidel a programů.

36.4.1 Efekty a přínosy

- SPM svou analytickou a plánovací funkcionalitou primárně přispívá ke ***zlepšování ekonomických a obchodních výsledků firmy***, např. objektivizací při nastavování obchodních marží,
- Podporuje odhadování a lepší pochopení ***vývojových trendů na trzích***,
- Využití SPM podporuje ***kvalitu a motivaci prodejních týmů*** poskytováním komplexních analytických informací o prodeji a posilování komunikace mezi všemi úrovněmi řízení,
- Kvalifikované nastavení SPM ***podporuje rychlé a flexibilní reakce firmy*** a prodejních týmů na aktuální změny v příslušných tržních segmentech.

36.4.2 Problémy a omezení

- Problémem může být ***integrace procesů (úloh) řízení prodeje a odpovídajících metrik a aplikací na procesy a úlohy ostatních oblastí*** řízení, zejména řízení marketingu, nákupu, financí a výroby,
- Hodnocení ***rozvoje nabídky nových produktů a jejich úspěšnosti*** je třeba promítat do dlouhodobých strategií prodeje a do náplně konceptu Řízení výkonnosti prodeje.

36.5 Závěry, doporučení

Kapitola představuje ***pracovní závěry*** k řízení výkonnosti podniků podle oblastí uplatnění.

- Řízení výkonnosti ***se aplikuje na všech úrovních či v oblastech řízení podniku***. Cíle, stanovené na úrovni celého podniku musí pak být reflektovány a zapracovány do systémů řízení výkonnosti na detailnějších úrovních řízení. Z tohoto hlediska lze tedy systémy řízení výkonnosti ***rozdělit na celopodnikové a dílčí***.
- Řízení ***výkonnosti marketingu***, Marketing Performance Managment, MPM marketingu je aplikované při analýzách marketingových dat a metrik, při formulování predikcí na trhu, při optimalizaci marketingových programů a zdrojů
- Řízení ***výkonnosti prodeje***, Sales Performance Management, SPM je určené pro přípravu a optimalizaci prodejních plánů, alokování prodejních aktivit podle teritorií, nastavování prodejních kvót, optimalizaci provizních pravidel a programů,
- ***Integrované podnikové plánování***, Integrated Business Planning, IBP nabízející zejména efektivní propojení plánování a modelování ekonomických aktivit na strategické a operativní úrovni, především v oblastech finančního řízení, řízení dodavatelských řetězců, řízení prodeje a marketingu,
- ***Řízení produktové výkonnosti***, Product Performance Management, PPM zahrnující analýzy a dopady dodávaných, vyvíjených i plánovaných produktů z pohledu jejich přijetí trhem a ekonomických aspektů jejich úspěšnosti. S touto oblasti souvisí i

Optimalizace produktového portfolia, Product Portfolio Optimization, PPO orientovaná na řízení rozvoje produktů v kontextu celé výroby, resp. celého komplexu obchodních a ekonomických aktivit firmy,

- ***Řízení výkonnosti kontaktních center***, Contact Center Performance Management představuje zejména systém metrika jejich hodnocení ve vztahu k výkonu a kvalitě služeb poskytovaných operátory na kontaktním centru firmy,

37. Zdroje

AALST, WIL van der. Process Mining Data Science v akci. Berlni : Springer, 2016. 978-3-662-49850-7.

ANGEL,I.O., SMITHSON,S.: Information Systems Management - Opportunities and Risks, Macmillan, 1991.

BACAL, R.: *Manager´s Guide to Performance Management*. New York, McGraw-Hill 2012. ISBN 978-0-07-177225-9.

BERKA, P. 2003. Dobývání znalostí z databází. Praha: Academia, 2003. str. 366. ISBN 80-200-1062-9,

BOTHE, O., KUBERA, O., BEDNÁŘ, D., POTANČOK, M., NOVOTNÝ, O.: Managing Analytics for Success, CRC Press, 2022. ISBN 978-1-032-20851-0

BRUCKNER, T. VOŘÍŠEK, J., BUCHALCEVOVÁ, A. a kolektiv: Tvorba informačních systémů: Principy, metodiky, architektury, Grada, 2012, ISBN 978-802477-9027.

BUCHALCEVOVÁ, A.: Metodiky vývoje a údržby informačních systémů. Praha, Grada 2004. ISBN 80-247-1075-7.

BUCHALCEVOVÁ, A., PAVLÍČKOVÁ, J., PAVLÍČEK, L.: Základy softwarového inženýrství, materiály ke cvičením. Praha, VŠE 2008.

CAO, L.: Data Science Thinking. Springer, 2018. ISBN 978-3-319-95091-4.

CIMLER, P., ZADRAŽILOVÁ, D. a kol.: Retail management. Praha, Management Press, 2007. ISBN: 978-80-7261-167-6.

DOHNAL, J., POUR, J.: IT v řízení podniku, Praha, Professional publishing 2016. ISBN 978-80-7431-160-4.

DOHNAL, J., PŘÍKLENK, O.: CIO a podpora byznysu. Praha, Grada, 2011. ISBN 978-80-247-4050-8.

DOHNAL, J.: Řízení vztahů se zákazníky – procesy, pracovníci, technologie. Praha, Grada 2002.

DRESNER, H.: Profiles in Performance. New York, John Wiley and Sons, 2010. ISBN: 978-0-470-40886-5.

ECKERSON, W., W.: Performance Dashboards. New Jersey, Johh Wiley & Sons 2006.

ENGLISH, L. P.: *Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing profits*. New York, John Wiley & Sons 2003.ISBN 0-471-25383-9.

FIBÍROVÁ, J., ŠOLJAKOVÁ, L., WAGNER, J., PETERA, P.: Manažerské účetnictví. Nástroje a metody. Praha, Wolters Kluwer, 2015. ISBN: 978-80-7478-743-0.

GÉRON, A.: Hands-On Machine Leraning with Scikit-Learn, Keras and TensotFlow. O´Reilly, 2023. ISBN: 978-1-098-12597-4.

HALAMA, J.: Řízení datové kvality v Hadoop Ecosystem, DP, VŠE, Praha, 2021.

HOLTSNIDER, B., JAFFE, B.D.: IT Managener´s Handbook. Amsterdam, Elsevier 2012. ISBN 978-0-12-415949-5.

CHANDLER, N.: The CPM Scenario. Gartner BI Summit 2008.

CHRAMOSTOVÁ, V., POTANČOK, M., POUR, J.: Byznys analytika pro manažery, Oeconomia, Praha, 2020.

ISO certifikace – (MANages, a.s., [Online] @2009. [Citace: 26. září 2014.] 2009),

KIMBALL, R., ROSS, M.: Relentlessly Practical Tools for Data Warehousing and Business Intelligence. Indianopolis, John Wiley Publishing 2010. ISBN 978-0-470-56310-6.

KOŽUŠNÍK, M.: Návrh dashboardů pro MBI metriky, DP, VŠE, 2014.

Kratochvíl, T.: Dashboardy - jejich analýza a implementace v prostředí SAP Business Objects 4.0/4.1. DP, VŠE, 2014.

KOTTER, J.P. – Vedení procesu změny: osm kroků úspěšné transformace – (Management Press 2000) – ISBN8072610155,

KRÁL, B. a kol.: Manažerské účetnictví. Praha, Management Press 2006. ISBN 80-7261-141-0.

KRÁL, Bohumil, Jana FIBÍROVÁ, Ondřej MATYÁŠ, Michal MENŠÍK, Jakub STRÁNSKÝ, Libuše ŠOLJAKOVÁ, Jaroslav WAGNER, Martin ZRALÝ a Zbyněk HALÍŘ, 2012. Manažerské účetnictví. 3. doplněné a aktualizované vydání. Praha: Management Press. ISBN 978-80-7261-217-8.

KULHAVÝ, L. - Praktické uplatnění technologií Data Mining v pojišťovnictví - (VŠE, Praha 2011),

KUNSTOVÁ, R.: Efektivní správa dokumentů. Praha, Grada, 2009. ISBN 978-80-247-3257-2.

LABERGER, R.: The Datawarehouse Mentor. New York, McGraw Hill, 2011. ISBN: 978-0-07-174532-1.

LABERGER, R.: Datové sklady. Agilní metod y business intelligence. Praha, Computer Press, McGraaw Hill, 2012. ISBN 978-80-251-3729-1.

LANEY, D., B.: Infonomics, Bibliomotion, Inc., New York, 2018. ISBN 978-1-138-09038-5.

MÁŠA, Petr. 2008. 4IZ560 - Data mining - praktické aplikace. 2008.

NOVOTNÝ, O., POUR, J., BASL, J., MARYŠKA, M.: Řízení výkonnosti podnikové informatiky. Professional Publishing, Praha, 2010. ISBN 978-80-7431-040-9.

NOVOTNÝ, O., VOŘÍŠEK, J. a kol.: Digitální cesta k prosperitě. Praha: Professional Publishing 2011. ISBN 978-80-7431-047-8.

PALADINO, B.: Innovative Corporate Performance Management: Five Key Principles to Accelerate Results. Indianopolis, Wiley Publishing, 2011. ISBN: 978-0-470-62773-0.

PARMENTER, D.: Key Performance Indicators (KPI): Developing, Implementing,and Using Winning KPIs,

POUR, J., MARYŠKA, M., STANOVSKÁ, I., ŠEDIVÁ, Z.: Self Service Business Intelligence, Praha. Grada, 2018. ISBN 978-80-271-0616-5.

PROVOST, F., FAWCETT, T.: Data Science for Business. What You Need to Know About Data Mining and Data-Analytic Thinking. O´Reilly Media. Sebastopol. 2013. ISBN: 978-1-449-36132-7.

ŘEPA, V.: Podnikové procesy. Praha, Grada 2007.

SCHIESSER, R.: IT Systems Management. New York, Prentice Hall 2010. ISBN 978-0-13-702506-0.

SIEGEL, E: Predictive Analytics. New York, John Wiley & Sons, 2016. ISBN 978-1-119-14567-7.

SLÁNSKÝ, D.: Data and Analytics for the 21st Century: Architecture and Governance, Professional Publishing, 2018. ISBN 978-80-88260-16-5.

ŠOLJAKOVÁ, L. FIBÍROVÁ, J.: Reporting. Praha, Grada, 2010. ISBN 978-80-247-2759-2.

SYNEK, M. a kol.: Manažerská ekonomika. Praha, Grada 2011. ISBN 978-80-247-3494-1.

SYNEK, M., KISLINGEROVÁ, E. a kol.: Podniková ekonomika. Praha, C H Beck 2015. ISBN 978-80-7400-274-8.

TOGAF: Architecture Development Method – (The_Open_Group).

TOMEK, G., VÁVROVÁ, V.: Průmysl 4.0 aneb nikdo sám nevyhraje. Praha, Professional Publishing, 2017. ISBN 978-80-906594-4-5.

UČEŇ, P.: Zvyšování výkonnosti firmy na bázi potenciálu zlepšení. Praha, Grada 2008. ISBN: 978-80-247-2472-0.

VAN DECKER, J, CHANDLER, N.: Top Processes for Corporate Performance Management, Gartner, 2011.

VEBER, J. a kol.: Management – základy, prosperita, globalizace. Praha, Management Press 2000.

VOŘÍŠEK, J., POUR, J. a kol.: Management podnikové informatiky, Professional Publishing, 2012, ISBN 978-80-7431-102-4.

VOŘÍŠEK, J. a kol.: Principy a modely řízení podnikové informatiky. Praha, Oeconomia 2008. ISBN: 978-80-245-1440-6.

VOZHAEVA, E.: Vývoj a implementace šablony Executive Cockpit Dashboardu pro monitorování KPI. BP, VŠE, 2024.

ŽŮRKOVÁ, Hana, 2007. Plánování a kontrola – klíč k úspěchu. 1. vydání. Praha: Grada Publishing. ISBN 80-247-1844-8.