Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra informačního inženýrství

Bakalářská práce

Datové sklady

SUSLIK Alen

Čestné prohlášení

Prohlašuji, že svou bakalářskou práci "Datové sklady" jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a pouze s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autor uvedené bakalářské práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne _____

Poděkování

Rád bych touto cestou poděkoval panu Doc. Ing. Vojtěchu Merunkovi, Ph.D. za odborné vedení mé bakalářské práce.

Datové sklady

Data Warehouses

Souhrn

Tato bakalářská práce zpracovává problematiku datových skladů, které vznikají v souvislosti se zvyšujícím se množstvím údajů v provozních systémech a snahou tyto údaje analyzovat. První část práce popisuje vznik těchto údajů, jejich formy uložení a možnosti jejich dalšího využití. Další část, literární rešerše, se zaměřuje na teoretické pojetí datových skladů, jejich definici, využití a způsoby uplatnění. Praktická část se zabývá modelováním procesů datových skladů podle standardu BPMN. Tato část zahrnuje několik diagramů, které popisují konkrétní procesy a byly vytvořeny pomocí programu.

Summary

The aim of this bachelor's thesis is to demonstrate difficulties with data warehouses (DWH) in connection with the emerging number of data arising from operational systems followed by tendency to analyze such data. In the first part, thesis describes creation of this sort of data, forms of their storing and other ways of their eventual usage. Background research is dedicated to theoretical concept of DWH, their definition, utilization and further forms of assertion. Practical part is aimed at process modelling of DWH in accordance with standard form BPMN. This part includes certain examples of diagrams describing concrete processes which were created through the use of Intalio program.

Klíčová slova: Datový sklad, datové tržiště, On-Line Analytical Processing (OLAP), Online Transaction Processing (OLTP), Business intelligence, Extrakce-Transformace-Nahrávání (ETL), databáze, dimenzionální modelování, reporting, dolování dat, Business Process Modeling Notation (BPMN)

Keywords: Data warehouse, data mart, On-Line Analytical Processing (OLAP), Online Transaction Processing (OLTP), Business intelligence, Extraction-Transformation-Loading (ETL), database, dimensional modeling, reporting, data mining, Business Process Modeling Notation (BPMN)

Obsah

1 Úvod 10
2 Cíl práce a metodika 11
2.1 Cíl práce 11
2.2 Metodika 11
3 Analýza procesů správy dat ve firmě. 12
3.1 Úvod 12
3.2 Typy údajů 12
3.3 Systémy správy dat 12
3.3.1 Databáze 13
3.3.2 Sklady provozních dat 15
3.4 Sledování a hlášení 15
3.5 Rozhodovací potřeby 16
4 Literární rešerše 17
4.1 Definice datových skladů, datových trhů a OLAP 17
4.1.1 Datové sklady 17
4.1.2 Datové trhy 18
4.1.3 OLAP 18
4.1.4 Shrnutí 18
4.2 Architektura datových skladů 19
4.2.1 Shrnutí 21
4.3 Databázové technologie 21
4.3.1 Super-relační databázové systémy. 22
4.3.2 Multidimenzionální databázové systémy 22
4.3.3 Shrnutí 23
4.4 Schémata datového skladu 23
4.4.1 Schéma hvězdy (star) 25
4.4.2 Schéma sněhové vločky (snowflake) 25
4.4.3 Shrnutí 26
4.5 Metadata 26
4.5.1 Shrnutí 27
4.6 Příprava a zavedení údajů 27

4.6.1 Extrakce 28
4.6.2 Transformace 28
4.6.3 Nahrávání 29
4.6.4 Shrnutí 29
4.7 Prezentace dat 29
4.7.1 Tvorba výstupních sestav 30
4.7.2 Analytické zpracování (OLAP) 30
4.7.3 Dolování dat (Data mining) 33
4.7.4 Shrnutí 34
4.8 Realizace datových skladů 34
4.8.1 Metody 34
4.8.2 Lidé 36
4.8.3 Shrnutí 37
4.9 Provoz a správa 37
4.9.1 Shrnutí 38
4.10 Náklady 39
4.11 Přínosy 39
4.12 Uplatnění 40
4.13 Úvod do BPMN 42
5. Praktická část 43
5.1 Cíl projektu 43
5.2 Encyklopedie 43
5.2.1 Slovní popis procesů 43
5.2.2 Seznam událostí 44
5.2.3 Seznam entit 46
5.2.4 Modely chování - diagramy 47
6. Zhodnocení a výsledky 51
7. Závěr 52
8. Seznam literatury 53

Seznam obrázků

Obr. 1 Rozdělení datových modelů podle (Vostrovský, 2008).....14
Obr. 2 Sklad provozních dat podle (Humphries a kol., 2002).....15
Obr. 3 Centralizovaný typ datového skladu podle (Jarke a kol., 2003).....20
Obr. 4 Spolkový typ datového skladu podle (Jarke a kol., 2003).....20
Obr. 5 Úrovňový typ datového skladu podle (Jarke a kol., 2003).....21
Obr. 6 Schéma hvězdy (star) podle (Jarke a kol., 2003).....25
Obr. 7 Schéma sněhové vločky (snowflake) podle (Jarke a kol., 2003).....26
Obr. 8 Relační OLAP podle (Humphries a kol., 2002).....32
Obr. 9 Muldidimenzionální OLAP podle (Humphries a kol., 2002).....32
Obr. 10 Hybridní OLAP podle (Humphries a kol., 2002).....33
Obr. 11 Implementační tým datového skladu podle (Humphries a kol., 2002).....36
Obr. 12 Datové sklady v různých typech českých organizací podle (Reml, 2009).....41
Obr. 13 Využití centralizace dat v Česku podle (Reml, 2010).....41
Obr. 14 Diagram procesu požadavku na informace z datového skladu.....47
Obr. 15 Diagram procesu technické údržby datového skladu.....48
Obr. 16 Diagram procesu reportingového servisu.....49
Obr. 17 Diagram procesu ETL.....50

Seznam tabulek

Tab. 1 Klady a zápory relačních databázových systémů.....22
Tab. 2 Klady a zápory multidimenzionálních databázových systémů.....23
Tab. 3 Příklad nenormalizované tabulky dimenze času.....25
Tab. 4 Porovnání metody velkého třesku a metody etapové.....36

1 Úvod

V současné době se zvyšuje zájem společnosti o systémy, které umožňují operace, jejichž cílem je získat informace podporující operativní i strategická rozhodování. Tato problematika je zároveň předmětem autorova zájmu, a proto bylo pro bakalářskou práci vybráno téma, které s ní úzce souvisí. Práce se bude podrobně zabývat datovými sklady a metodami umožňujícími jejich kvalitní využití. Právě ty jsou dnes základním prvkem informačních systémů, které spojují databázové prvky s prvky analytickými a reportovacími. Podpora rozhodování je prováděna formou reportů, přehledových zobrazení či složitých prediktivních analýz. Tyto systémy vznikly v reakci na zvyšující se konkurenci, složitost obchodních procesů a vzrůstající poptávku po efektivních možnostech zpracování velkého množství ukládaných údajů . Množství údajů roste se stále častějším používáním informačních systémů automatizujících běžné provozní procesy ve všech tržních odvětvích. Pro systémy podporující informační rozhodování se používá pojem *business intelligence*.

Datové sklady mají v systémech *business intelligence* roli datových úložišť. Shromažďují se zde údaje z různých provozních i externích zdrojů upravené do integrované podoby, která je vhodná pro následnou analytickou činnost. Analýza takto uspořádaných a centralizovaných údajů může při správné implementaci přinášet velmi cenné a často jinak obtížně zjistitelné informace, jako například údaje o detailním stavu podnikání nebo o chování zákazníků. Z tohoto důvodu se datové sklady stávají neodmyslitelnou součástí větších informačních systémů téměř ve všech odvětvích, především však v oblasti služeb.

2 Cíl práce a metodika

2.1 Cíl práce

Cílem této práce je vytvořit ucelený pohled na problematiku datových skladů. Tento pohled bude tvořit vstupní část popisující okolnosti vzniku a užitečnost datových skladů a literární rešerše, jež bude selekcí dosavadních znalostí o datových skladech napříč různými názory. Výsledkem rešerše by měl být popis technologií, procesů, metod realizace a uplatnění datových skladů.

Po dohodě s vedoucím práce bude v praktické části namísto realizace databázové části vytvořeno několik diagramů pomocí modelovacích nástrojů a podle pravidel BPMN (Business Process Modeling Notation). Ty budou popisovat některé procesy probíhající v podniku, kde je aplikovaný datový sklad. Diagramy budou tedy zachycovat reálný sled kroků používaných při operacích s datovými sklady. Tato názorná ukázka by měla napomoct lepšímu pochopení podstaty datových skladů a s nimi souvisejících úkonů.

2.2 Metodika

Výše uvedených cílů práce bude dosaženo především vyhledáním, shromážděním a studiem odborných publikací na téma datové sklady a *business intelligence*. Důkladným prostudováním veškeré dostupné literatury včetně článků v odborných časopisech je možné vybrat konkrétní tituly, jež se danou problematikou zabývají nejpodrobněji, či takové, které přinášejí nové informace. V této práci se vychází hlavně z následujících publikací: *Data warehousing : návrh a implementace* (Humphries a kol., 2002), *Fundamentals of data warehouses* (Jarke a kol., 2003) a *Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL* (Lacko, 2003). Jako zdroj nejnovějších či doplňujících informací pak poslouží články z periodik IT Systems, Computer World a Connect!.

V praktické části bude klíčovou metodou analýza procesů, jež probíhají v podnicích v souvislosti s užíváním datových skladů. Následné zobrazení těchto procesů pomocí diagramu bude vytvořeno v programu Intalio, který byl doporučen vedoucím práce. K jednotlivým diagramům bude za pomoci terminologie, jež bude vysvětlena v rámci literární rešerše, připojen popis zobrazovaných procesů.

3 Analýza procesů správy dat ve firmě.

3.1 Úvod

V současné době většina společností denně zpracovává velké množství údajů. Jsou to údaje o klientech, obchodních partnerech, zakázkách, produktech výroby a mnoha jiných aspektech, kterými se společnosti zaobírají. Především se jedná o společnosti pracující v oboru telekomunikací, finančnictví, pojišťovnictví, veřejné správy a výrobní sféry, kde jsou tyto údaje klíčovým prvkem jejich zájmu. Počet zpracovávaných údajů však stále roste, a tím se zvyšují požadavky na jejich správu a organizaci. Podle odhadů připadá na jednoho obyvatele planety 45GB dat a toto množství se každý půldruhý rok zdvojnásobí. Pět procent z celkového množství je pak uloženo v podnikových serverech. Pro úspěšné a kvalitní vedení firmy je nezbytné vytvářet rychlá a správná rozhodnutí. Ty se nyní dají dělat efektivně pomocí nástrojů, které využívají informační technologie umožňující správu a analýzu dat. (Šoule, 2010: 17; Zavoral, 2010: 13)

3.2 Typy údajů

Údaje je možno rozlišovat z hlediska jejich původu, formy záznamu či času. Z hlediska původu lze rozlišovat údaje interní a externí. Interní údaje mají původ v podnikových procesech, především v provozních systémech. Externími jsou údaje z vnějších zdrojů – jedná se o poštu, elektronické datové schránky, údaje z aplikací obchodních partnerů nebo údaje od společností, pro které je podnikatelským cílem zisk z prodeje svých dat. Z hlediska formy záznamu lze rozdělit údaje na údaje v tištěné nebo elektronické podobě. Trendy vývoje elektronického zpracování zaznamenaly rozsáhlý vývoj, a elektronický dokument tak díky jasným výhodám z velké části nahrazuje dokumenty v papírové formě. K jejich úplnému nahrazení však nedošlo a lze předpokládat, že ani v brzké době nedojde. Zaměřením této práce budou data elektronická. Časové dělení na údaje aktuální a historické je nutno brát v potaz. (Humphries a kol., 2002: 32)

3.3 Systémy správy dat

Rozvoj a dostupnost informačních technologií vede k masivnímu nasazování informačních systémů (IS) takřka ve všech odvětvích lidské sféry. Práce s IS znamená práci s daty. Pro konkrétní činnosti jsou vytvářeny aplikace automatizující tuto práci. Téměř každá aplikace

obsahuje funkci pro načítání nebo ukládání dat. Prvním procesem správy dat může být jejich vznik. Jak již bylo zmíněno výše, údaje vznikají v rámci provozních potřeb nebo se přejímají z externího prostředí. Provozní potřeby obsahují *„technologie podporující hladké provádění a soustavné zdokonalování každodenních operací, identifikaci a opravu chyb pomocí hlášených výjimek řízení průběhu prací (work flow) a celkového sledování provozu.“* (Humphries a kol., 2002: 7) Do těchto potřeb patří původní systémy, aplikace OLTP, databáze, sklady provozních dat a nástroje pro sledování a hlášení.

Původní systémy (legacy systems) obsahují jakýkoli v současné době používaný informační systém, který byl vytvořen za použití technologií minulých generací. V důsledku zájmu o automatizaci transakčně orientovaných obchodních procesů byla převážná část projektů z oblasti informačních technologií věnována provozním systémům.

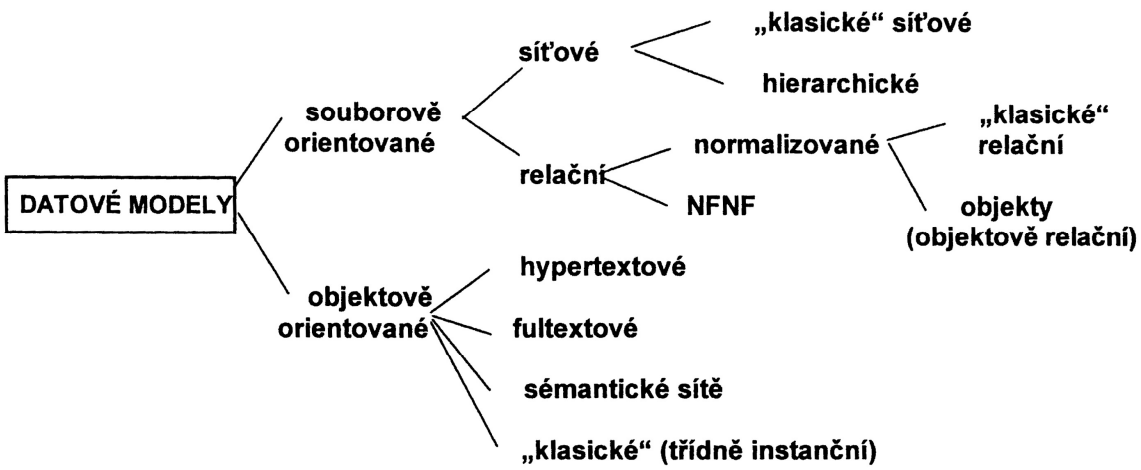
Pojem OLTP (Online Transaction Processing) označuje druh systémů, které jsou postaveny na architektuře klient/server a umožňují vkládání, sběr, automatizaci a správu obchodních transakcí. Systémy okamžitě reagují na požadavky uživatelů, kterých může být ve stejném čase větší množství. Transakční systémy jsou v praxi oblíbené a často používané, a v případě, že obstarávají většinu podnikových aktivit, má takový systém název ERP (Enterprise Ressource Planning). (Humphries a kol., 2002: 9; Lacko, 2003: 20)

Data tedy vznikají v různých podobách daných typem aplikací, předmětem zájmu sběru a stylem tvorby dat. Vzniká tak velké množství různorodých údajů v různých formátech. Jelikož se data vytváří pro budoucí práci s nimi, druhým procesem je ukládání takto získaných dat. Místa, kde se data koncentrují, nazýváme databáze.

3.3.1 Databáze

Databáze je souhrn dat vztahujících se k určitému tématu nebo účelu. Tato množina dat popisující část objektivní reality je využívána a spravována prostřednictvím databázového systému. *„Základní myšlenkou databázové technologie je oddělení uživatelů systému od fyzických souborů s daty a existence komunikačního rozhraní zajišťující nezávislost datové základny na aplikačním software.“* Součástí databáze je tedy komunikační rozhraní, což je program, který umožňuje přístup k datům uložených v databázi. Ten se nazývá systém řízení báze dat (SŘBD) z anglického pojmenování *database management system* (DBMS), nebo jinak také databázový systém.

Systémy řízení báze dat usnadňují práci s daty, umožňují k nim přístup bez znalosti, kde jsou požadovaná data uložena nebo jakým způsobem je získat. SŘBD musí zajistit operace třídění, vytváření součtů, tvorbu vstupních obrazovek a výstup evidovaných dat pro tisk. V této souvislosti však musíme rozlišovat dva pojmy. Druh použitých dat a použitý datový model. Druh dat informuje o kvalitativních vlastnostech uložených dat, přičemž data se dělí na dvě základní kategorie, data statická a data dynamická. Datový model popisuje vlastnosti vazeb mezi údaji v databázi a způsob organizace jejich struktury. Datové modely se dělí na modely souborově orientované a objektově orientované. Souborově orientované modely se dále dělí na síťové a relační. (Lacko, 2003: 20; Merunka, 2008: 113; Vostrovský, 2008: 11-15)



Obr. 1 Rozdělení datových modelů podle (Vostrovský, 2008)

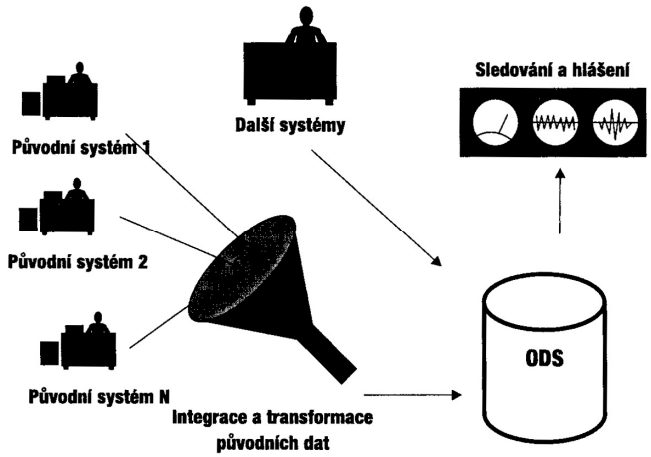
Relační datový model, který patří do první, historicky starší skupiny souborově orientovaných modelů, je blízký souborové technologii stejným způsobem, jakým je navrhnuta architektura organizace paměti von Neumannova počítače. Autorem modelu je matematik laboratoří IBM E. F. Codd, jenž koncem 60. let položil základy prvního prakticky používaného relačního databázového systému. Byl to produkt DB2 firmy IBM. Koncem 80. let došlo k rozšíření modelu, limitovaného tehdejším nízkým výkonem počítačů, firmou ORACLE se stejnojmenným produktem. Ten má dodnes dominantní postavení na světovém trhu. (Vostrovský, 2008: 16)

Objektově orientované modely představují revoluční trend vývoje mimo jiné svým odlišným pohledem na organizaci paměti v počítači. Modely pracují s pojmy jako je

objekt, třída, polymorfismus a dědičnost. Tyto modely lépe korespondují s reálnými daty a jejich reprezentací v databázovém systému, a nachází tak uplatnění v systémech se složitou datovou strukturou. (Švec, 2003: 1, 2, 20)

3.3.2 Sklady provozních dat

Sklady provozních dat (Operational Data Store, ODS) lze definovat jako *„skupinu sjednocených databází navržených na podporu sledování provozu.“* (Humphries a kol., 2002: 10) Podobně je popisuje i W. H. Inmon jako *„architektonickou budovu, kde jsou hromadně uložena sjednocená data.“* (Humphries a kol., 2002: 37) Data z původních či dalších provozních systémů jsou transformována a integrována do jednotného homogenního celku za účelem poskytnutí sjednoceného a aktuálního pohledu na provoz. Aby byl tento pohled možný a stále aktuální, je nutné skladovat data podrobná, předmětově orientovaná a pravidelně aktualizovaná. (Humphries a kol., 2002: 10)



Obr. 2 Sklad provozních dat podle (Humphries a kol., 2002)

3.4 Sledování a hlášení

Pro získání pravidelně celistvého pohledu na provoz slouží nástroje pro sledování a hlášení (reporting, výkaznictví). Pojem reporting znamená vizualizaci informací a je dnes jedním z procesů správy dat a zároveň jednou z klíčových firemních povinností. Na základě získaných informací se vyhodnocuje chod podniku a činní se manažerská rozhodnutí. Z technologického pohledu slouží ke sledování a hlášení stavu pro provozní potřeby sklad provozních dat. (Humphries a kol., 2002: 7,11; Ježek F., 2010: 14)

3.5 Rozhodovací potřeby

Účinné zpracování velkého množství údajů vede k odhalení podstatných informací v nich ukrytých. Na základě takto získaných informací lze vytvářet rozhodnutí a strategie, které mohou být velkou konkurenční výhodou. Pro získání těchto informací je nutné umět údaje využít, analyzovat a zasadit do souvislostí. Je zapotřebí data uložená v provozních informačních systémech, které jsou většinou orientovány funkčně, převést na výkonnostní ukazatele a charakteristiky předmětu sledování (např. zákazníka, produktu, pobočky).

Ačkoli by se mohlo zdát, že pro potřeby rozhodování mohou vyhovovat sklady provozních dat, není tomu tak. Příčinou je především absence statických, souhrnných a historických dat. Tyto sklady dat mohou však sloužit jako zdroje dat pro systémy rozhodovací. K rozhodování nejsou vhodné ani datové struktury transakčních systémů. Jejich komplexně a vysoce strukturovaná forma (většinou ve 3.NF) dosahuje vysokých výkonů spíše při on-line transakcích, než při analýzách, které jsou náročné na výpočetní kapacitu procesorů. V okamžiku, kdy se pro zpracování transakcí, analýzu a podporu používá stejný server, degraduje se výkon použitého hardwaru i operačního systému. Důsledkem je prodloužení odezvy jak uživatelů pracujících s transakčním systémem, tak uživatelů provádějících analýzu. To může být velkým problémem, jelikož transakční systém je systém provozní a v případě jeho přetížení nebo dokonce výpadku, nebude možné s daty pracovat vůbec. Přeruší se tak běžná provozní činnost. Stejně jako sklady provozních dat i transakční systémy postrádají historická data. Zatímco transakční systémy pracují s daty platnými pouze v krátkém časovém horizontu, k analýze a predikci jsou zapotřebí údaje uchovávané i několik let. Největším problémem je však decentralizace dat a nehomogennost jejich struktur. Skutečnost, že není k dispozici integrovaný zdroj údajů ze všech podnikových informačních systémů, vede ke složité a někdy i nemožné analýze.

Integrování dat z většinou heterogenních transakčních systémů je časově náročné a v některých případech i nemožné pro získání globálního obrazu o stavu podnikání. Tyto skutečnosti vedou k nasazování systémů pro podporu rozhodování (*Business Intelligence*, BI) založených na datovém skladu (Data Warehouse, DWH). (Dragolov, 2007: 8; Hroch, Cach, 2007: 2;; Humphries a kol., 2002: 37; Lacko, 2003: 22-24)

4 Literární rešerše

4.1 Definice datových skladů, datových trhů a OLAP

4.1.1 Datové sklady

Zjednodušeně řečeno, datový sklad je centralizované úložiště veškerých podnikových dat, který poskytuje ucelené údaje a zároveň datovou základnu pro detailní analýzu dat. Uznávaný expert v tomto oboru William H. Inmon definuje datový sklad jako *„podnikově strukturovaný depozitář subjektově orientovaných, integrovaných, časově proměnlivých, historických dat použitých na získávání informací a podporu rozhodování. V datovém skladu jsou uložena atomická a sumární data.“* (Lacko, 2003: 48) Jinak ho definuje jako kolekci integrovaných, stálých, předmětově orientovaných databází navržených za účelem podpory systémů pro rozhodování (Decision Support System, DSS). Tyto definice v různých obměnách používá mnoho autorů zabývajících se touto problematikou. Jednotlivé stručné pojmy zmíněných definicí je třeba dále rozvést.(Humphries a kol., 2002: 31, 171; Jarke a kol., 2003: 2,3; Lacko, 2003: 48)

- Subjektová orientace označuje fakt, že se data do skladu zapisují podle předmětu zájmu (zákazník, produkt, pobočka) a ne funkčně podle aplikace, ve které byly vytvořeny (odbyt, fakturace, personalistika).
- Integrace značí sjednocenost dat uložených do skladu. Jelikož jsou data nahrávána z mnoha podnikových provozních systémů, které mohou být nekonzistentní a neintegrované, může dojít k situaci, kdy máme rozdílné údaje o identické položce. V případě výskytu nekonzistentních dat ztrácí datový sklad smysl.
- Časová proměnlivost vymezuje platnost údajů v datovém skladu v určitém časovém intervalu. Údaje se do skladu ukládají v podobě snímků reprezentujících určitý časový úsek.
- Historická data jsou neaktuální údaje uchovávané po delší období, většinou několik let. Často se uvádí průměrná doba uložení dvou let. Data se v datových skladech obvykle nemění ani neodstraňují, pouze se přidávají data nová. Historická data tvoří hlavní obsah datových skladů.

- Atomická data jsou údaje na nejvíce podrobné úrovni sledování. Slouží k odvození sumačních či agregovaných hodnot. Záznamem velmi podrobných dat dochází k nárůstu požadavků na ukládací prostor, jejich absence však znemožňuje provádět rozklady dotazů na podrobnější položky.
- Sumarizační údaje jsou agregovaná data atomická, která jsou nahraná do skladu pro rychlejší odpovědi na dotazy na vysokém stupni granularity. Jsou velmi důležitou součástí zvyšující celkový výkon a použitelnost skladu.

4.1.2 Datové trhy

Datové trhy jsou systémy založené na stejné bázi jako datové sklady. Jedná se také o úložiště dat poskytující informace vhodné pro rozhodování, obsahují však podmnožinu dat uložených v datovém skladu. Tyto datové podmnožiny reprezentují dílčí organizační složky firmy, ať jde o oddělení geografické nebo lokální. Tato možnost se také využívá při budování datových skladů, jakožto velmi náročných projektů. Buď je možné nejprve vybudovat datová tržiště a na nich postavit integrující datový sklad, nebo lze naopak začít datovým skladem, ze kterého se budou čerpat data do konkrétně zaměřených datových trhů. (Humphries a kol., 2002: 35, 36; Lacko, 2002: 51)

4.1.3 OLAP

OLAP, zkratka anglického *Online Analytical Processing*, jsou technologie zahrnující struktury dat a analytické služby, které umožňují přeměnu dat z datových skladů na informace vhodné pro podporu rozhodování. Termín zavedl Dr. E. F. Codd a definuje ho jako *„volně definovanou řadu principů, které poskytují dimenzionální rámec pro podporu rozhodování.“* (Lacko, 2006: 242) Nástroje OLAP jsou součástí systémů na podporu rozhodování (DSS) a systémů pro vrcholové řízení (EIS). (Jarke a kol., 2003: 87) Funkčnost OLAP je charakteristická analýzou konsolidovaných podnikových dat, jež jsou uložena v dimenzionálních datových modelech. Pracují tak s vícerozměrnými poli, tzv. *„hyper-kostka“* nebo *„kostka“*. Více k nástrojům OLAP i dimenzionálnímu modelování je popsáno v kapitolách 4.5 a 4.8.2.

4.1.4 Shrnutí

Přesná definice datových skladů a s nimi souvisejících pojmů je nezbytnou součástí práce zabývající se touto tématikou. Vymezuje klíčové vlastnosti datových skladů a zároveň

upřesňuje použitou terminologii. Datové sklady tvoří datový základ pro navazující datová tržiště a systémy OLAP. Všechny tři popsané technologie se tak nejlépe uplatní při současném nasazení v rámci jednoho informačního systému.

4.2 Architektura datových skladů

Mnoho vývojářů i uživatelů sdílí názor, že na architekturu datových skladů lze nahlížet jako na materiální vrstvy stavěné na sebe. Jednotlivé vrstvy obsahují data v různých podobách a vždy jsou data z nižších vrstev nahrávána do vrstev vyšších.

Existují dva hlavní koncepty struktury datových skladů; třívrstvá koncepce W. H. Inmona a dvouvrstvá koncepce Ralpha Kimpballa, která nepracuje se střední, hlavní vrstvou.

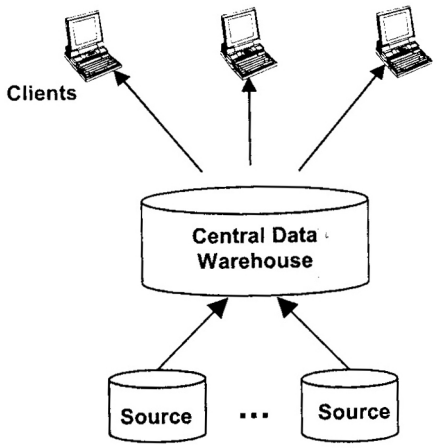
Nejnižší vrstvu v pohledu na datový sklad představují datové zdroje a prostory pro dočasné uložení a zpracování dat extrahovaných z těchto zdrojů. Jak bylo uvedeno ve třetí kapitole, jedná se o původní systémy, produkční databáze a různé externí databáze se strukturovanými, nehomogenními a decentralizovanými daty. Tato vrstva odpovídá první vrstvě v obou konceptech.

Střední vrstvou je *„centrální“* úložiště dat - databáze, ve které se uchovávají konsolidovaná, vyčištěná a historizovaná data. Je to hlavní vrstva datového skladu a R. Kimpball ve svém konceptu vynechává právě tuto vrstvu. V některých případech se používají sklady provozních dat (ODS) jako mezivrstva mezi nejnižší vrstvou datových zdrojů a střední vrstvou centrálního úložiště dat. Sklady provozních dat obsahují již transformovaná, integrovaná a agregovaná data, a mohou se tak využívat jako zdroj aktuálních dat pro centrální úložiště.

Další vrstvou jsou *„místní“* datové sklady, které již obsahují vysoce agregovaná data získaná z centrálních datových úložišť přímo podporujících aktivity jako informační zpracování, manažerské rozhodování, dlouhodobé plánování či historickou analýzu. Mezi místní datové sklady patří datová tržiště nebo databáze OLAP. Tato vrstva je opět pro obě koncepce společná, s tím rozdílem, že v řešení bez centrálního úložiště se datová tržiště plní daty přímo ze základní vrstvy dočasných úložišť. (Brodníček, 2009: 428, Jarke akol., 2003: 2)

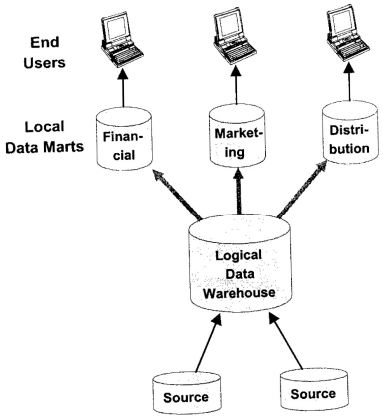
Rozlišujeme tři základní typy využití výše popsaných konceptů (Jarke a kol., 2003: 10,11):

- *Typ centralizovaný*, kde je použit pouze jeden datový sklad, který skladuje všechna data potřebná pro podnikovou analýzu. Přístup k datům je proto nekomplikovaný. Správa takového skladu je jednodušší, ale na druhou stranu je systém méně výkonný než u distribuovaných řešení. Používá se ve společnostech, kde je operační struktura také centralizovaná.



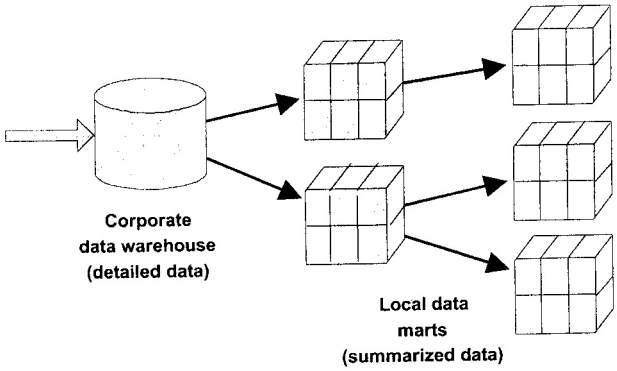
Obr. 3 Centralizovaný typ datového skladu podle (Jarke a kol., 2003)

- *Typ spolkový*, kde jsou data logicky konsolidována, ale uložena v separovaných databázích. Místní datová tržiště obsahují pouze relevantní informace pro dané oddělení. Kvůli redukci množství dat mohou datová tržiště obsahovat podrobnější data. Důležitá je skutečnost, že datový sklad je pouze virtuální.



Obr. 4 Spolkový typ datového skladu podle (Jarke a kol., 2003)

- *Typ úrovňový*, kde fyzicky existuje datový sklad i datové trhy, ovšem na různých úrovních. Ty obsahují kopie nebo sumarizovaná data úrovní předchozích, ne však v tak detailní podobě jako u typu spolkového.



Obr. 5 Úrovňový typ datového skladu podle (Jarke a kol., 2003)

Poslední dva typy jsou decentralizované a nacházejí uplatnění pouze v decentralizovaných operačních prostředích. Jejich výhodou je rychlejší reakce na dotazy a menší množství dat nutných k prohledání. Jsou zároveň odpovědí na škálovatelnost datových skladů, jakožto systémů, které se vyvíjejí a rostou v čase.

4.2.1 Shrnutí

Architektura datového skladu není stejná pro každý projekt, vždy záleží na mnoha faktorech, jak bude implementace skladu v rámci konkrétní společnosti provedena. Třívrstvé řešení W. H. Inmona se jeví jako komplexnější ale také nákladnější řešení.

4.3 Databázové technologie

Systémy řízení dat používané pro datové sklady i datová tržiště musí být velmi výkonné systémy, které splňují požadavky na komplexní dotazování vyžadované jejich uživateli a na podporu aplikací OLAP. V projektech datových skladů se používají dva následující typy systémů (Jarke a kol., 2003: 7):

- super-relační databázové systémy
- multidimenzionální databázové systémy

4.3.1 Super-relační databázové systémy.

V klasických relačních systémech řízení báze dat (RSŘBD) jsou data uložena v dvojrozměrných tabulkách. Pro přístup k datům se nejčastěji používá jazyk SQL (Structured Query Language). Pro použití RSŘBD s aplikacemi OLAP museli vývojáři systém obohatit o nové, tzv. *„super-relační“* prvky. Ty obsahují podporu pro rozšiřování ukládacích formátů, relační operace a speciální indexová schémata. (Jarke a kol., 2003: 94)

Pro rychlou odezvu aplikací OLAP jsou data organizována použitím schémat dimenzionálního modelování. (viz. kapitola 4.4) Porovnání výhod a nevýhod RSŘBD popisuje následující tabulka. (Lacko, 2003: 34)

Tab. 1 Klady a zápory relačních databázových systémů.

výhody	nevýhody
potenciál odborníků ve firmách používajících model	absence komplexních analytických nástrojů
využití v transakčních databázích i datových skladech	potenciální omezení objemu údajů, ke kterým je možné v adekvátním čase přistoupit
potenciál softwaru a vývojových nástrojů	

Mezi významné dodavatele RSŘBD patří například (Humphries a kol., 2003: 157):

- DB2 od společnosti IBM
- SQL Server od společnosti Microsoft
- Oracle 9imod společnosti Oracle
- Red Brick Warehouse od společnosti IBM
- RDBMS Engine – Systém 11 od společnosti Sybase

4.3.2 Multidimenzionální databázové systémy

Multidimezionální databáze (MultiDimensional DataBase, MDDB) překonávají dva hlavní nedostatky RSŘBD organizací údajů do vícerozměrných polí – n-dimenzionálních krychlí. Každá taková dimenze představuje perspektivu uživatele, například data o prodeji ve firmě mohou obsahovat dimenzi produktu, regionu a času (což jsou zároveň nejčastěji používané dimenze, u dimenze času se dá tvrdit, že je přítomná vždy). Tyto krychle jsou výsledkem sumarizovaných a agregovaných údajů z klasických dvourozměrných relačních tabulek. Takto zpracované a uložené údaje nevyžadují žádné operace spojení tabulek a přesně vyhovují požadavkům pro práci a vizualizaci uživatelů aplikací OLAP. Mezi nevýhody MDDB patří jejich větší náklady na restrukturalizaci při změně dimenzí a větší nároky na

úložný prostor. (Jarke a kol., 2003: 8, Lacko, 2003: 31,32) Souhrn výhod a nevýhod obsahuje následující tabulka. (Lacko, 2003: 34)

Tab. 2 Klady a zápory multidimenzionálních databázových systémů.

výhody	nevýhody
rychlý komplexní přístup k velkému objemu dat	problémy při změně dimenzí
možnost komplexních analýz	vyšší nároky na kapacitu úložiště
silné schopnosti pro modelování a prognózy	

Dodavatelé MDDB jsou (Humphries a kol., 2002: 157):

- Essbase od společnosti Arbor
- Enterprise od společnosti BrioQuery
- DI-Driver od společnosti Dimensional Insight
- Express Server od společnosti Oracle

4.3.3 Shrnutí

Ačkoli jsou mezi multidimenzionálními a relačními databázemi velké rozdíly, je častou praxí smíšené využití obou technologií v různých částech architektury datového skladu. Pro centrální datové sklady se běžně používá relační databázová technologie, jelikož snáze podporuje rozrůstání skladů při stále adekvátním výkonu plnění i zpracování dotazů. Naopak pro datová tržiště používaná lokálně (geograficky nebo firemně) a obsahující podmnožinu již sumarizovaných a více agregovaných podnikových dat se může využívat výhod multidimenzionálních databází a poskytovat tak lepší výkon pro vytváření výstupů.

4.4 Schémata datového skladu

Podle požadavků na informace uložené v rámci skladu se nabízejí dvě techniky pro návrh datového modelu (Humphries a kol., 2002: 116, 163):

- *normalizace*, kde je schéma databázového systému navrženo pomocí technik běžně používaných pro aplikace OLTP,
- *dimenzionální modelování* zahrnující množinu modelovacích technik, které získaly značnou popularitu během posledních let, kdy byly úspěšně implementovány v projektech datových skladů v průmyslovém odvětví. *„Dimenzionální modelování poskytuje množství technik a principů pro denormalizaci databázových struktur a*

tedy vytvoření schémat vhodných pro podporu rozhodování.“ (Humphries a kol., 2002: 165)

Normalizované struktury nejsou vhodné pro vyhodnocování dotazů na podporu rozhodování kvůli nutnému spojování velkého množství tabulek, kde jsou data uložena neredundantně. Mají výhodu rychlé manipulace s daty a nižší nároky na kapacitu úložiště. Mohou tak vyhovovat pro datové sklady, které primárně neslouží k rozhodovacím potřebám, ale jako centrální úložiště podnikových dat.

Dimenzionální modelování používá speciální schémata pro ukládání dat, pomocí kterých lze modelovat vícerozměrné struktury nad klasickými dvourozměrnými relačními modely. (Jarke a kol., 2003: 94) Používají se dva základní typy schémat. Oba pracují se dvěma typy tabulek (Humphries a kol., 2002: 165, 166):

- tabulky faktů
- tabulky dimenzí

Tabulky faktů obsahují záznamy obchodních transakcí. Jsou to číselné, měrné jednotky o stavu obchodování, jako např. počet prodaných kusů a ceny statků. Tyto tabulky jsou provázané s tabulkami dimenzí pro doplnění popisu faktů. Je běžné, že tabulky faktů jsou úplně normalizované a také rozdělené do menších útvarů. To zlepšuje dotazovací výkon a usnadňuje obnovu ze zálohy. (Jarke a kol., 2003: 93)

Tabulky dimenzí obsahují logicky nebo hierarchicky uspořádané údaje o záznamech z tabulky faktů. Jsou to popisy faktů poskytující kompletní informace. V některých případech tabulka neobsahuje žádné popisy a dimenze slouží jen jako klíč, index pro vyhledávání hodnot ve vícerozměrném poli.(Jarke a kol., 2003: 89) Příkladem užívaných dimenzí je například v literatuře nejčastěji uváděná dimenze času, produktu, regionu a klienta.

Podle uspořádání a případného použití normalizovaných či denormalizovaných tabulek dimenzí rozlišujeme dva typy schémat.

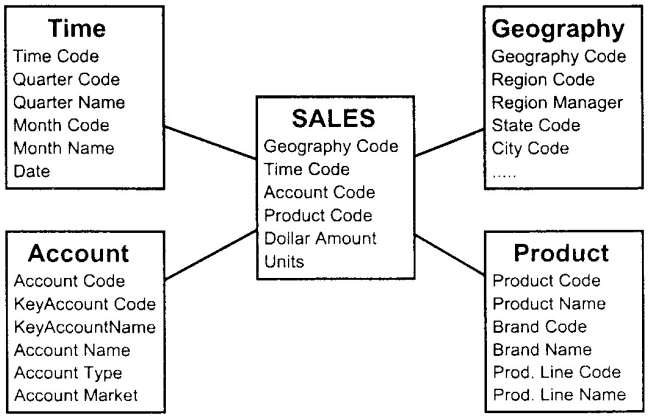
- schéma hvězdy (star)
- schéma sněhové vločky (snowflake)

4.4.1 Schéma hvězdy (star)

Schéma hvězdy obsahuje jednu nebo více tabulek faktů, na kterou jsou napojeny denormalizované tabulky dimenzí. Propojení se provádí pomocí spojení cizích klíčů z tabulky faktů s primárními klíči z tabulek dimenzí. Nevýhodou takového schématu je jeho náročnější vytváření a vyšší nároky na kapacitu databáze, odměnou je však vysoký dotazovací výkon. Ten je zapříčiněn nepotřebností spojovat tabulky dimenzí, protože všechny údaje o dimenzi jsou v jedné tabulce. Tento typ schématu se používá pro datové sklady zaměřené na analýzu dat.

Tab. 3 Příklad denormalizované tabulky dimenze času.

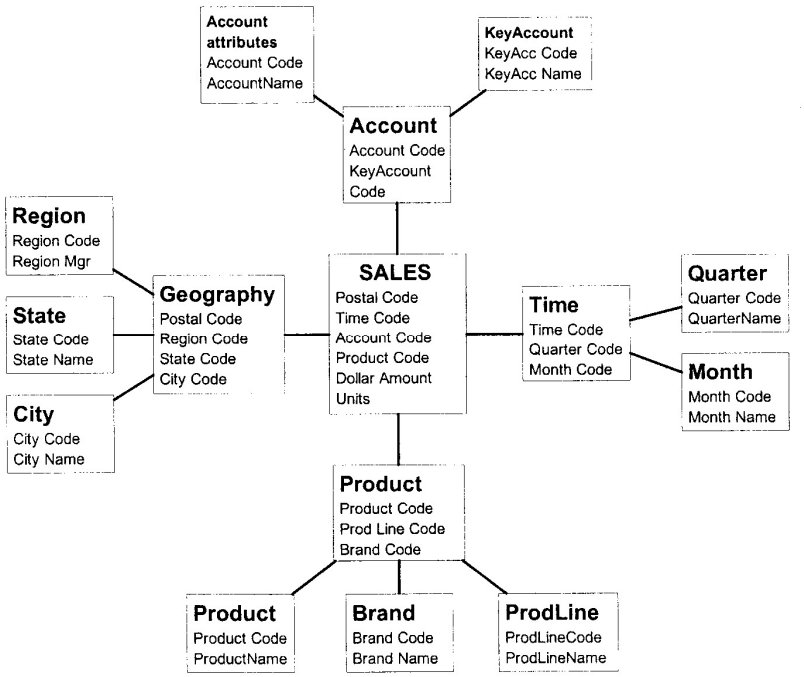
time_id	date	day	month	year	day_of_m	week_of_y	month_of_y	quarter
284	2010-10-11	Monday	October	2010	11	41	10	Q4
285	2010-10-12	Tuesady	October	2010	12	41	10	Q4
286	2010-10-13	Wednesday	October	2010	13	41	10	Q4
287	2010-10-14	Thursday	October	2010	14	42	10	Q4
288	2010-10-15	Frity	October	2010	15	42	10	Q4
289	2010-10-16	Saturday	October	2010	16	42	10	Q4



Obr. 6 Schéma hvězdy (star) podle (Jarke a kol., 2003)

4.4.2 Schéma sněhové vločky (snowflake)

Schéma sněhové vločky je obdobou schématu hvězdy s tím rozdílem, že používá normalizované tabulky dimenzí. To znamená, že na tabulku faktů jsou stejně jako u schématu hvězdy napojeny tabulky dimenzí, jedná se však pouze o jejich první hierarchické stupně. Na ně jsou pak napojeny stupně.další. Výhodou je úspora místa snížením počtu ukládaných identických dat, nevýhodou slabší dotazovací výkon. Tento typ se více užívá u datových skladů používaných hlavně pro centralizaci podnikových dat.



Obr. 7 Schéma sněhové vločky (snowflake) podle (Jarke a kol., 2003)

4.4.3 Shrnutí

Z možných technik návrhu datového modelu je nejčastěji používáno dimenzionální modelování a jeho možnost využít tabulky faktů a dimenzí pro tvorbu schémat. Tento způsob zajišťuje datovým skladům právě ono vhodné uložení dat pro jejich budoucí analýzu a vysoký dotazovací výkon.

4.5 Metadata

„Metadata jsou již tradičně definována jako data o datech.“ (Humphries a kol., 2002: 177) Jsou důležitou součástí architektury datových skladů, která slouží k popisu jednotlivých prvků skladu, způsobů jeho plnění či pravidel použitých při transformaci dat. Metadata pomáhají vysvětlit uživatelům, co jaká data reprezentují, kde je hledat nebo jakým způsobem vznikla. Metadata mohou obsahovat také informace o vztazích mezi jednotlivými datovými položkami, statistiky a bezpečnostní pravidla používání dat. Výše popsané vlastnosti metadat se mohou rozdělit podle určení do tří skupin (Humphries a kol., 2002: 180): *Administrativní metadata* popisující technické komponenty pro správce skladu, *metadata koncových uživatelů*, která pomáhají pracovat s daty skladu při popisu, vyhledávání a vytváření dotazů, a *metadata pro optimalizaci* usnadňující návrh a zlepšující

výkon datových skladů. Práci s metadaty má na starost *administrátor metadat*, který definuje struktury a spravuje úložiště s metadaty. (Humphries a kol., 2002: 65)

4.5.1 Shrnutí

Správa metadat je jedna z hlavních součástí informačních systémů a stejně tomu je i u datových skladů. Metadata slouží ke kontrole a řízení běžných dat. Tento prvek je u datových skladů ještě významnější a měla by mu být věnována patřičná pozornost.

4.6 Příprava a zavedení údajů

Předmětem obsahu datového skladu jsou data. Jedná se o upravená data z mnoha různých provozních systémů, databází a jiných zdrojů, která jsou integrovaná do jednoho celku. Pro plnění skladu se používají nástroje a postupy ETL (Extraction, Transformation, Loading). Jde o časově, finančně i technicky náročnou část budování datového skladu, která může tvořit i více než polovinu celkového projektu. Podle W. H. Inmona je právě integrovaný pohled na uložená data nejdůležitějším aspektem datových skladů. (Jarke a kol., 2003: 27) Jednotlivými etapami ETL jsou (Lacko, 2003: 60):

- extrakce (extraction)
- transformace (transformation)
- nahrávání (loading)

Podle Jarke a kol. (2003: 6) nástroje ETL umožňují nebo automatizují tyto úkony:

- extrakce
- čištění
- transformace
- nahrávání
- replikace
- analýza
- vysokorychlostní přesun dat
- kontrola datové kvality
- analýza metadat

4.6.1 Extrakce

Extrakce je první fází (v závislosti na oblasti vynášení údajů, viz následující kapitola), při které se získávají data z rozličných zdrojů. Rozdílnost může být např. softwarová (rozdílné operační systémy, souborové formáty), hardwarová (PC, iMac, mainframe), databázová (Oracle, IBM DB2, Microsoft SQL Server) nebo časová (operační, archivní systémy).

Nástroje a technologie pro extrakci by měly poskytnout přístup do datového zdroje, popis jeho možností i uložených dat a následnou extrakci – kopírování vybraných údajů. Klíčovým faktem pro tyto funkce je, že nástroje umožňují takový pohled na data ze zdroje, že vypadají jako by měla stejný formát, jenž je použit v datovém skladu. U systémů, kde je struktura zdroje i datového sladu relační, je možné použít nástroje softwarových firem, které se nazývají *middleware* nebo *gateways*. (Jarke a kol., 2003: 55) Humphries a kol. (2002: 151) pro změnu uvádí, že nástroje konektivity – *middleware* poskytují: *„transparentní přístup do databází rozdílných typů provozovaných na různých platformách.“* K této věci Jarke a kol. (2003: 29) dodává, že tento software pro extrakci může být řešen interaktivními programy nebo prostředími pro manuální kódování.

Humphries (2002: 151) popisuje dvě metody extrakce. *Metodu celkové extrakce*, při které je celý datový sklad pravidelně obnovován všemi použitelnými daty z provozních systémů, a *metodu založené na změnách*, která nahrává pouze změněná data. Ta narozdíl od metody celkové není náročná na síťové prostředky, vyžaduje však náročnější programování a schopnost extrakčních nástrojů identifikovat změny v údajích datových zdrojů.

4.6.2 Transformace

Fáze transformace obsahuje procesy zabývající se úpravou přenášených dat do formátů vyžadovaných datovým skladem. Některé údaje je zapotřebí např. rozdělit nebo spojit na více resp. méně atributů. Údaje z datových zdrojů také mohou trpět mnoha problémy, které snižují kvalitu dat, potažmo celého skladu v případě užití takových „špinavých“ dat. Problémem může být například nejednoznačnost údajů v podobě použití různých formátů a názvů pro stejné objekty, použití zkratek, chybějící hodnoty nebo výskyt duplicit. Podle doby a místa se rozlišuje několik způsobů provádění transformace. Může se provádět sériově nebo paralelně se zaváděním údajů do skladu. Sériově se transformuje před zaváděním údajů, paralelně při zavádění. U sériového provádění lze rozlišit místo

provádění. Transformace probíhá na zdrojovém systému (tzv. *model lokálního vynášení*) nebo následuje po extrakci v jiném prostředí (tzv. *model vzdáleného vynášení).* (Humphries a kol., 2002: 154; Lacko, 2003: 60-64)

4.6.3 Nahrávání

Etapa nahrávání je završením procesu ETL. Transformovaná data ze zdrojových systémů nebo z vynášecí oblasti se přesouvají do datového skladu. Součástí nahrávání bývá generování klíčů pro jednoznačnou identifikaci položek a indexování obsahu pro rychlejší přístup. Současné nahrávání a indexování může vést ke snížení rychlosti přenosu, indexování se proto provádí až v druhém kroku. Při nahrávání dat se může jednat o celkové plnění počáteční, plnění celkové obnovovací (metoda celková) nebo plnění přírůstkové (metoda založená na změnách). V prvních dvou typech se jedná o časově a technicky náročné operace, které zpravidla nevyhovují požadavkům na provoz datového skladu. Je běžnou praxí, že se datové sklady plní ve frekvenci jednoho dne a více, záleží to však na zaměření. Celkové plnění by mohlo znamenat opakující se dlouhé období, kdy je sklad nedostupný pro provádění analýz. Přírůstkové plnění je tak vhodnou volbou nezatěžující nadměrně datové zdroje ani datové sklady. Proces by měl být plně automatizován a naplánován na vhodnou chvíli s ohledem na provozní dobu zdrojových systémů a požadavků na výstupy datového skladu. Některým globálním datovým skladům nevyhovuje ani přírůstkové plnění, jelikož jsou na sklad vyvíjeny neustálé požadavky z různých částí světa. (Jarke a kol., 2003: 51,52; Lacko, 2003: 65)

4.6.4 Shrnutí

Proces ETL je spolu s prezentací dat nejdůležitější činností provozu datového skladu. Při této činnosti dochází k pravidelnému nahrávání a úpravě údajů do požadované podoby. Pro následnou kvalitní analytickou a reportovací práci je správně, aktuálními a přesnými údaji plněný datový sklad hlavní podmínkou.

4.7 Prezentace dat

Datové sklady jsou vhodným prostředkem k získávání informací pro rozhodování. Není to však systém tvořící tato rozhodnutí. Pokud je sklad již naplněn konsolidovanými a vyčištěnými údaji z provozních nebo jiných systémů, lze z něho čerpat data. Tímto se

zabývají navazující technologie, nástroje a systémy pracující s datovými sklady. Jejich účelem je efektivní prezentace údajů uživatelům, a to v papírové nebo elektronické, snadno srozumitelné formě. Rozlišuje se několik možností výstupu (nástrojů) pro přístup k datům (Brodníček, 2009: 429; Humphries a kol., 2002: 158; Ježek L., 2010: 16):

- tvorba výstupních sestav
- analytické zpracování (OLAP)
- dolování dat (data mining)

4.7.1 Tvorba výstupních sestav

Tvorba sestav je standardním, grafickým nebo sofistikovaným výstupem založeným na datech z datového skladu. Slouží pro běžné, každodenní sledování ukazatelů výkonnosti společnosti ve formě statických sestav. Ty jsou oprávněným uživatelům distribuovány v tištěné podobě, e-mailem nebo pomocí intranetové sítě. V tomto případě údaje z datového skladu umožňují tvorbu centralizovaných reportů, které jsou mnohem kvalitnější než reporty z decentralizovaných systémů. Výstupní sestavy podle Lacka (2006: 12) využívá většina uživatelů pracujících s datovými sklady. (Humphries a kol., 2002: 158; Ježek L., 2010: 16):

4.7.2 Analytické zpracování (OLAP)

Analytické zpracování používá pro ukládání dat vícerozměrné datové struktury, kde jsou pro sledované faktické hodnoty dopředu vypočítané různé kombinace dimenzí i sumační položky. Tím tato technologie umožňuje provádět velmi rychlé on-line operace s velkými objemy dat. Zrychlení může být až 1000 násobné oproti systémům OLTP. Každá dimenze má hierarchii úrovní a lze na ni nahlížet z různých detailních pohledů. Např. dimenze času má úrovně dne, týdne, měsíce a roku. Pro svojí vícerozměrnou povahu se jim říká „kostky“.

Hlavními uživateli nástrojů OLAP jsou manažeři a analytici. Analytici vybírají vhodná data a jejich analýzou mohou poskytovat manažerům informace pro podporu rozhodování. Grafické prostředí nad vícerozměrnou strukturou jim poskytuje interaktivní operace s daty dle aktuálních potřeb.

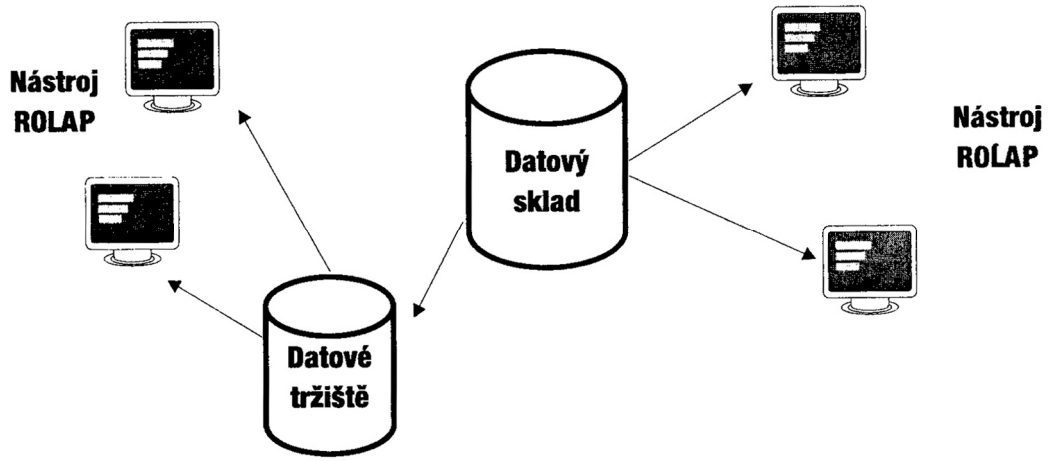
Zároveň se projevuje snaha tvůrců softwarů začlenit analytické nástroje do běžně užívaných aplikací, na které jsou pracovníci zvyklí. Toho je příkladem kancelářský tabulkový procesor MS Excel od společnosti Microsoft nebo správa OLAP krychlí pomocí jeho kontingenčních tabulek. (Balažovič, 2008: 3; Lacko, 2002: 207) Software podporující analytické zpracování využívá následující operace (Jarke a kol., 2003: 90):

- agregace (Roll-up) – možnost pohledu na sumarizovaná data
- zanořování (Drill down) – je opakem agregace, od vyšší úrovně dimenze lze nahlížet na detailní informace
- filtrování (Selection) – zobrazuje data podle zadaných kritérií
- krájení (Slicing) – na základě výběru úrovně z dimenze je zobrazena podmnožina, která neobsahuje nevybrané úrovně
- rotace (Pivot) – je možnost změny orientace dimenzí, např. u dvourozměrného pole prohození řádků za sloupce

Podle povahy systémů, s jakými pracují nástroje OLAP, se zpracování údajů dělí na (Humphries a kol., 2002: 87-89):

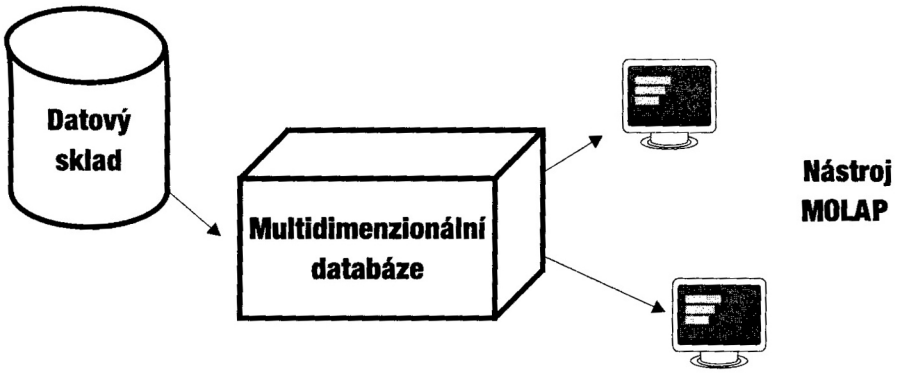
- relační OLAP (ROLAP)
- multidimenzionální OLAP (MOLAP)
- hybridní OLAP (HOLAP)

Relační zpracování OLAP (ROLAP) využívá pro analytické nástroje údaje z datových skladů vybudovaných pomocí relačních databází. Nástroje ROLAP poté poskytují uživatelům multidimenzionální pohled na data. Metadata jsou uložena v úložišti ROLAP a pomocí serveru OLAP je generován kód SQL sloužící pro přístup k datům požadovaných uživateli. Data tak zůstávají v relační databázi a nedochází k redundanci dat. (Lacko, 2003: 116)



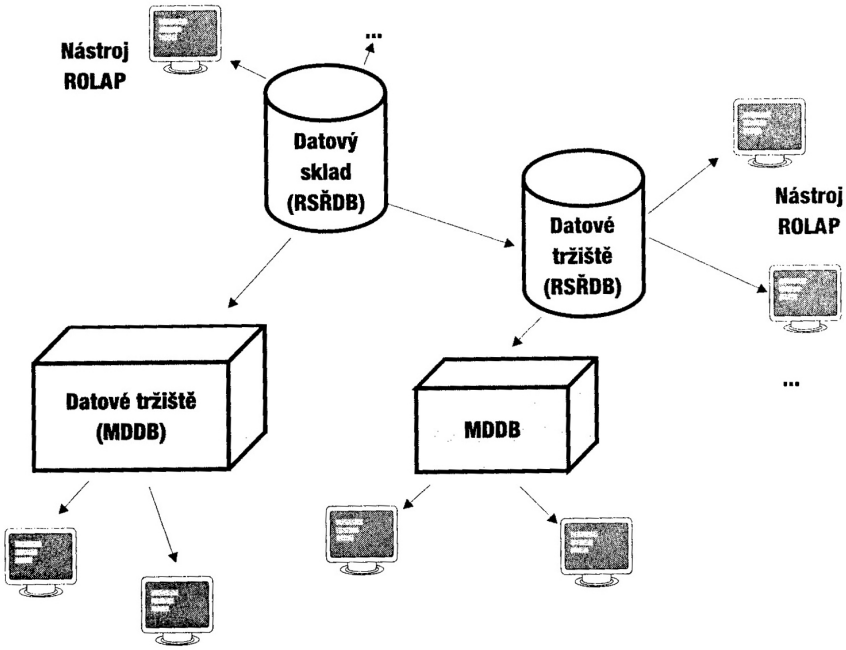
Obr. 8 Relační OLAP podle (Humphries a kol., 2002)

V mechanizmu multidimenzionální OLAP (MOLAP) jsou data nahrána z datového skladu a následně uložena v dalším úložišti, které má multidimenzionální povahu datových struktur. Během procesu ukládání jsou do úložiště nahrány spočítané předběžné výsledky, jež slouží pro velmi rychlé získávání údajů z mnoha dimenzí. Takové údaje z úložiště MOLAP lze nahrávat ke klientům na server a ulehčovat tím zatížení sítě při dotazech. Výhodou je maximální výkon, nevýhodou redundance dat. Při použití více dimenzí mohou vzniknout extrémní nároky na úložný prostor.



Obr. 9 Muldidimenzionální OLAP podle (Humphries a kol., 2002)

Hybridní OLAP (HOLAP) kombinuje obě předchozích technologie, využívajíc výhody a potlačujíc nevýhody. Detailní data zůstávají v relační databázi pro nástroje ROLAP a slouží pro jednoduché dotazy. Agregované údaje jsou ukládány pomocí nástroje MOLAP do multidimenzionálních úložišť. Těmi jsou často datová tržiště.



Obr. 10 Hybridní OLAP podle (Humphries a kol., 2002)

4.7.3 Dolování dat (Data mining)

Dolování dat patří mezi pokročilejší a náročnější způsoby analýzy dat. Tyto způsoby dovedou odhalit skryté souvislosti mezi údaji, určovat vzorce chování zákazníků nebo stanovovat trendy vývoje. Dolování dat je založeno na matematických a statistických metodách, složitější postupy pak využívají metody umělé inteligence a genetické algoritmy. Ze statistických metod se používá korelace, lineární a logická regrese, diskriminantní analýza a metoda předpovídání. Jsou to postupy hledání závislostí mezi na první pohled nesouvisejícími prvky a jejich kvantifikace. V případě trendů vývoje se používají časové řady.

Proces dolování dat je činnost natolik specifická, že se nedají definovat automatizované procesy. Vždy se musí vytvářet modely přizpůsobené konkrétním podmínkám i požadavkům managementu. Lze však vytvořit metodiku, která podle Lacka (2006, 314, 319) proces dolování dat rozděluje do čtyř etap:

1. výběr algoritmů a modelu
2. učící fáze aplikovaná na existujících případech
3. testování modelu
4. analýza a predikce nových případů

Po vybrání vhodného modelu následuje zpřesnění jeho parametrů a testování na vybraných a předzpracovaných datech (např. pomocí nástrojů ETL). Pokud vyhovuje, lze ho použít na množinu vstupních dat pro získání souvislostí. Tyto skutečnosti mohou dále sloužit pro podporu rozhodování. (Lacko, 2006: 319, 320)

Vytěžování dat je mezioborový přístup, v současné době nejrozšířenější ve finančnictví a marketingu. Zaznamenává však rychlý rozvoj v průmyslu a lékařství. (Kubásek, 2009: 424)

4.7.4 Shrnutí

Prezentace dat v jakékoliv podobě, analýza a datové dolování jsou hlavním činnostmi, kvůli kterým společnosti zavádějí datové sklady. Jedná se o postupy, pomocí kterých lze získávat detailní a sjednocené informace napříč celou firmou, díky kterým lze vyhledat původy dění a nebo být připraven na situace v budoucnosti. To vše probíhá v reálném čase umožňujícím operativně jednat.

4.8 Realizace datových skladů

4.8.1 Metody

Projekt datového skladu se skládá z mnoha na sebe navazujících kroků, které musí proběhnout úspěšně, aby se datový sklad stal fungujícím a tak i rentabilním prvkem firemních informačních systémů. Jednotlivé kroky musí brát v potaz nejen organizační strukturu firmy ale i případné potíže, které jsou bohužel téměř nevyhnutelné. K eliminování nebo alespoň minimalizování těchto potíží je nutné určit strategii. Ta by měla vycházet z požadavků na informační podporu určitých podnikových procesů nebo podnikových dat. Součástí strategie je i určení správného postupu realizace tím, že budou zvoleny adekvátní metody budování a použití projektové a technické dokumentace. (Valečková, 2007: 6) V literatuře se uvádějí dvě metody realizace, které odpovídají běžným postupům systémového inženýrství při zavádění nových systémů. V praxi je to nejčastěji (Lacko, 2002: 52):

- metoda velkého třesku
- metoda přírůstková (inkrementální)

Metoda velkého třesku je koncipována jako realizace jediného celkového projektu. Na počátku je vyhotovena analýza požadavků všech uživatelů. Podle té je vytvořen souhrnný návrh řešení sloužící jako podklad pro implementaci. Tento postup je vhodný spíše pro menší projekty, u datových skladů může celkový projekt trvat i více než rok. Během takto dlouhého období je možné, že se změní požadavky na sklad, změní se zdrojové systémy nebo přestane mít projekt podporu vrcholového managementu. To vše zvyšuje rizikovost projektu, a proto se při realizaci datových skladů uplatňuje nejčastěji metoda přírůstková.

Metoda přírůstková je způsob, kdy se rozděluje celkový projekt na menší etapy. Každá etapa má prioritu a určitý cíl, jenž doplňuje architekturu celého projektu. Jednotlivé etapy trvají kratší dobu, a umožňují tak dřívější zapojení dílčích systémů do provozu. To může pozitivně ovlivnit management, který reaguje na rychlejší návratnost vynaložených investic, nebo uživatele usnadněním jejich práce. Celková architektura datového skladu je pak škálovatelná. Dovoluje přidávat další části podle potřeb firmy, a reaguje tak na nové požadavky uživatelů. Příkladem může být vybudování datového skladu v první etapě a datových tržišť v etapách následujících nebo zahrnutí jen některých zdrojových systémů či požadavků v jednotlivých etapách. Každá etapa obsahuje následující kroky (Dragolov, 2007: 9,10):

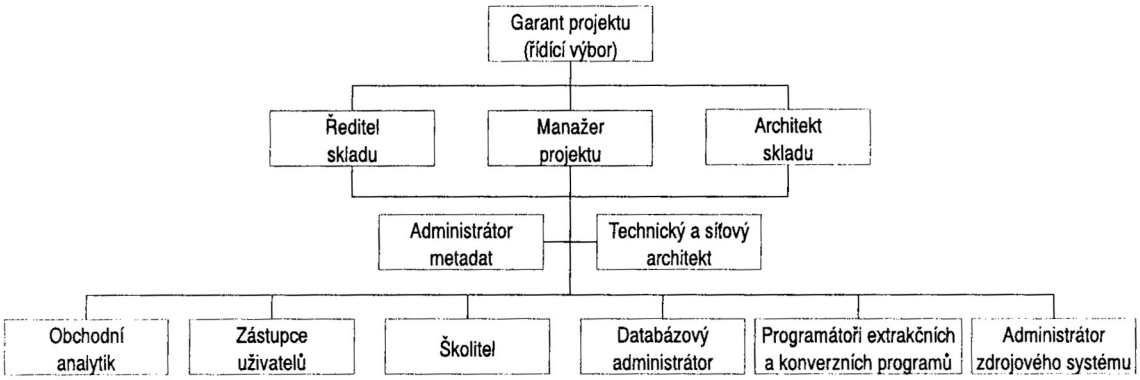
1. analýza požadavků - specifikace uživatelských reportů, jejich periodicita, distribuce
   - analýza zdrojových systémů, metod ETL, OLAP
2. implementace - návrh a vytváření struktury skladu a kostek OLAP
   - realizace funkcí pro plnění daty a jejich čištění
   - realizace výstupních nástrojů
3. testování a ladění
4. ověřovací provoz
5. školení
6. rutinní provoz
7. podpora a údržba

Tab. 4 Porovnání metody velkého třesku a metody etapové.

	výhody	nevýhody
metoda velkého třesku	-celkový projekt před realizací	-delší doba trvání a návratnosti investic -nemožnost reakce na změny -větší riziko neúspěchu projektu
metoda etapová	-škálovatelná architektura -brzké zapojení subsystémů a rychlejší návratnost investic -možnost reagovat na požadavky -možnost změny pořadí etap	-možné komplikace při změně datové struktury -riziko nárůstu objemů prací v jednotlivých etapách

4.8.2 Lidé

Součástí realizace datového skladu je vždy tým lidí s rozdílnými odbornými schopnostmi. Potřebnost různě kvalifikovaných pracovníků je vynucena velkým množstvím částí datového skladu. Složení týmu odpovídá náročnosti a rozsáhlosti právě zpracovávaného projektu, ale obvyklé složení vývojového týmu bývá podle M. Humphries a kol. (2002: 67) v podobě uvedené na obrázku č.11**.**



Obr. 11 Implementační tým datového skladu podle (Humphries a kol., 2002)

V této podobě má každý člen týmu odpovědnost za správné provedení jemu svěřeného úseku. Tato podoba týmu je navrhnuta pro realizaci datového skladu, přičemž budoucí podoba týmu lidí spravujících sklad vyžaduje odlišnou, více trvalou strukturu. Z technických pozic jsou klíčové role administrátorů metadat, databází zdrojových systémů a architektů skladu, síťového a technického vybavení. Projekt datového skladu je pak běžně implementován dodavatelskou firmou, není tedy budován vlastními silami. Zároveň je ale vhodné do týmu začlenit interní zaměstnance mající znalosti s informačními systémy, které jsou nasazeny ve firmě. Především jde o systémové a databázové administrátory spravující provozní systémy, tedy budoucí zdrojové systémy datového skladu.

4.8.3 Shrnutí

Při budování datových skladů vždy záleží na postupu, jak se bude stavět, a týmu lidí, který se o to postará. Správně zvolená metoda a vytvořený tým pracovníků, kde má každý příslušnou odpovědnost, je u velkých projektů informačních systémů, kterými datové sklady jsou, nezbytná. V případě, že je jedna z těchto dvou částí podceněna, nemusí se implementace zdařit a vytvořením nefunkčního datového skladu vzniká výrazná finanční ztráta.

4.9 Provoz a správa

Vybudováním datového skladu dostává podnik do rukou mocný nástroj, který je však třeba vhodně spravovat, aby poskytoval přesná, aktuální a dostupná data. Nejedná se tedy o informační systém, který by bylo možné po nasazení bez údržby využívat. Mezi hlavní provozní činnosti datových skladů patří *pravidelná plnění* skladu novými údaji ze zdrojových systémů a s tím související *kvalita dat, řízení nárůstu dat* a *optimalizace výkonu.* Neméně důležitá je i *bezpečnost* v podobě *správy uživatelských rolí, profilů přístupu* nebo *záloh* v případě poruch. (Humphries a kol., 2002: 193-200)

Pravidelné plnění zajišťuje, že pro uživatele datového skladu jsou vždy dostupná aktuální data. Aby byla data připravena k analýzám v běžné pracovní době, probíhá pravidelné plnění většinou po uzávěrce a při nevyužívání provozních systémů, tedy například přes noc. Během procesu plnění probíhají všechny etapy ETL a po nich následuje přepočítání agregačních hodnot s novými hodnotami. *„Pravidelné plnění skladu bezchybnými daty má na starost tým podpory, který se téměř vždy zodpovídá přímo či nepřímo vedoucímu informatiky (CIO, Chief Information Office).“* (Humphries a kol., 2002: 61) To z důvodu, že oblast kvality dat je vysoce důležitá a větší výskyt tzv. *„špinavých“* dat může vést k chybným rozhodnutím. V tomto ohledu se používají dvě metody nahrávání dat. Metoda, kdy se používají pouze bezchybná, předem vyčištěná data nebo způsob, kdy se nahrávají všechna data a až poté jsou na sklad spuštěny mechanismy nalezení a čištění dat. *„Hlavním účelem datového skladu ale není automaticky opravovat nalezené chyby, nýbrž celý proces oprav vhodnými prostředky a postupy podporovat. Vhodným místem pro opravu chybných dat je především primární informační systém.“* (Šprungl, 2010: 20)

Nárůst dat uložených v datovém skladu je úměrný frekvenci plnění skladu, množství sledovaných obchodních procesů a tím i množství údajů ze zdrojových systémů a v neposlední řadě požadované úrovni detailnosti ukládaných dat. Mnoha set gigabytové až terabytové datové sklady jsou již běžné. Existuje však několik způsobů jak redukovat nárůst dat (Humphries a kol., 2002: 196):

- využití agregací a odebrání detailních dat takto sumarizovaných
- zkrácení sledovaného období pro historická data
- odmazávání nepoužívaných dat na základě statistiky využívání dat při dotazování

Ačkoli tyto techniky dovedou snížit požadavky na úložný prostor, je nutné zvážit, zda se více vyplatí odmazávat údaje, nebo vynaložit prostředky do nového technického vybavení.

S narůstajícím množstvím dat, uživatelů a prováděných dotazů v rámci datového skladu je nutné sledovat výkon a přizpůsobovat ho měnícím se podmínkám – optimalizovat. K tomuto účelu slouží statistiky používání datového skladu a ladění výkonu pomocí lepších možností indexace nebo použití paralelních dotazů. Tyto optimalizace má na starosti databázový administrátor.

Datový sklad obsahuje velmi cenné a mnohdy i citlivé údaje, a proto je v zájmu každého podniku mít tento systém zabezpečený. Proti neoprávněnému přístupu slouží nadefinované uživatelské profily, které umožňují pracovat pouze s vybranými údaji nebo vybranými nástroji. Zabezpečení proti ztrátě dat nabývá na důležitosti s rostoucí závislostí podniku na datovém skladu. Z tohoto důvodu by měly být vytvořeny postupy zotavení a zálohování, aby byla zajištěna kontinuální dostupnost. Pro případ, že je datový sklad jediným úložištěm historických dat, může být alarmujícím varováním studie společnosti Gartner Group, která odhalila, že téměř polovina firem postižených ztrátou dat zkrachuje do pěti let a ostatní mají vážné existenční problémy. (Jankovský, 2009: 17)

4.9.1 Shrnutí

Datový sklad není účetní aplikací, kterou lze po nainstalování nekonečně dlouho a bezstarostně používat. Je to systém, který, aby správně fungoval, vyžaduje určitou údržbu. U datových skladů může být sice údržba časem stejně nákladná jako samotné zavedení, výsledkem však může být systém, který efektivně slouží velké části podniku.

4.10 Náklady

Budování datového skladu je nákladná záležitost a to z několika hledisek. Jde o středně až dlouhodobý projekt, který nekončí okamžikem uvedení skladu do podniku. S narůstajícím množstvím dat i uživatelů se sklad neustále rozšiřuje. Další náklady tvoří technické vybavení v podobě hardwaru a softwaru. Tato technická položka může tvořit až 76% všech prvotních nákladů (60% HW, 16% SW). (Jarke a kol., 2003: 13) Kvalita hardwaru rozhoduje o celkovém výkonu datového skladu, a proto zde platí, že výkon je dán nejpomalejším prvkem systému. (Černý, 2009: 18) Nedílnou součástí projektu jsou služby, tým lidí a náklady s tím spojené.

4.11 Přínosy

Datový sklad je z definice integrované úložiště podnikových dat, které uchovává ověřená, konsolidovaná, správná, aktuální i historická data v předmětově orientovaném datovém modelu. Tato podoba úložiště je vhodná především jako zdroj kvalitních informací pro rozhodování, ale přináší i mnoho dalších výhod. Některé výhody se projeví již samotným implementováním tohoto systému v běžné činnosti pracovníků. Patří mezi ně:

- zvýšená produktivita analytiků při vyhodnocování dat a s tím spojená úspora času
- nezávislost manažerů na IT oddělení při získávání reportů
- pochopení provozu podniku uživateli a možná optimalizace klíčových procesů
- racionalizace pracovních míst určených pro rozhodování
- zpřístupnění dat obchodním uživatelům ve správném čase

Už výše uvedené výhody mohou snižovat provozní náklady až o pětinu (Řechtáčková, 2007: 23), zlepšovat průběhy procesů v podniku a být tak určitou konkurenční výhodou na trhu. Největší konkurenční výhodou jsou však v dnešní době správné informace. *„Víme-li přesně co zákazníci chtějí, jakou kvalitu, za jakou cenu, jaké skupiny zákazníků preferují kvalitu za vyšší cenu, jaké skupiny naopak cenu za nižší kvalitu, čím se tyto zákaznické skupiny vyznačují, kde je můžeme oslovit či kdy je to nejvhodnější.“* (Dragolov, 2007: 8) Hlavní silou datového skladu je tak poskytování údajů z různých úhlů pohledu a na různých úrovních podrobností pro analýzu a následné vytváření manažerských rozhodnutí.

Tímto způsobem lze:

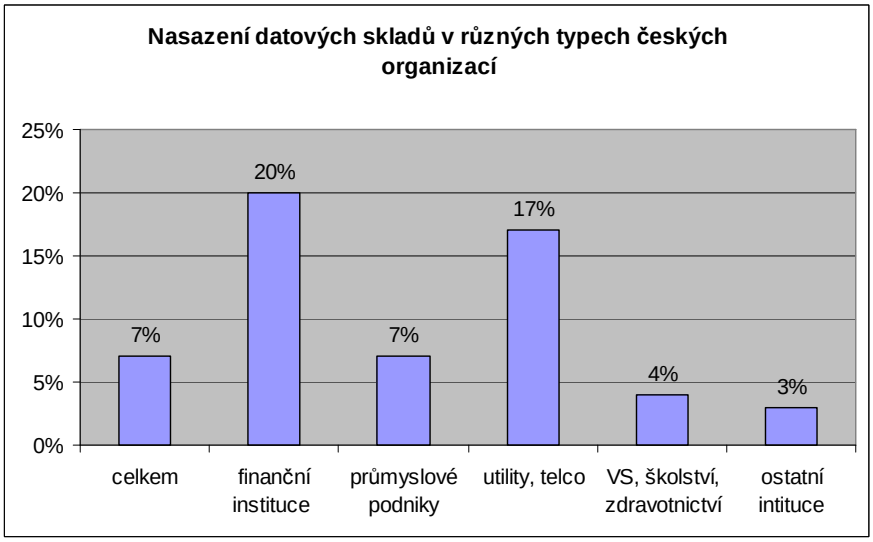
- sledovat různé provozní a finanční ukazatele
- analyzovat chování zákazníků
- identifikovat ihned příčiny
- hledat souvislosti napříč různými odděleními podniku, například mezi finančním, personálním a marketingovým oddělením, nejlépe však napříč celou firmou
- poznat nenápadné změny na trhu a s předstihem reagovat
- vytvářet různé predikce událostí nebo budoucí stavy reality

4.12 Uplatnění

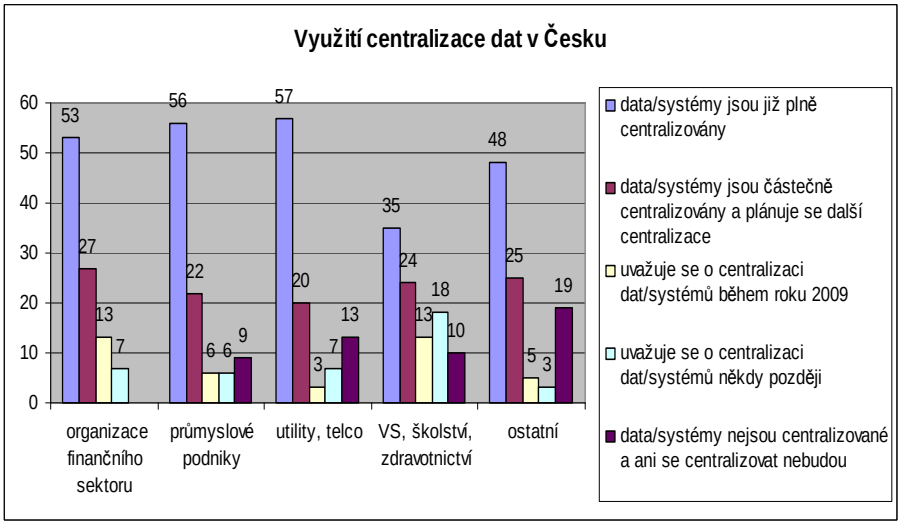
Z globálního hlediska je nasazení systémů datových skladů v největší míře aplikováno velkými korporacemi v sektoru bankovnictví, telekomunikací a veřejné správy. Je to zapříčiněno nejen velkým množstvím klientů a vysokou mírou konkurence, ale i velkým počtem zaměstnanců, různých služeb a procesů v těchto odvětvích. (Vávra, 2007: 27)

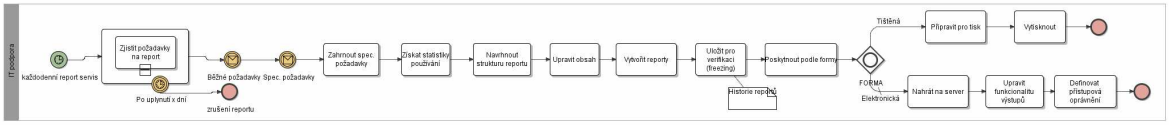
Datové sklady jsou pro ně proto ideálním řešením. Díky jejich univerzálností, kdy mohou pracovat téměř se všemi informačními systémy, je tak lze použít takřka ve všech odvětvích a podnikových oblastech. (Řechtáčková, 2007: 23) Pronikají do oblastí pojišťovnictví, zdravotnictví, farmaceutiky, logistiky a průmyslové výroby.

V ČR odpovídá trend globálnímu vývoji viz obr. 12, tedy nasazení datových skladů hlavně v oblastech telekomunikací a finančnictví. Podle průzkumu společnosti Markent (Reml, 2009: 14) je větší míra použití datových skladů u firem se zahraniční majetkovou účastí a u firem s větším počtem zaměstnanců. Zároveň jsou náklady vynaložené na datové sklady v porovnání zahraničních a českých subjektů asi desetkrát větší u těch zahraničních. Podle jiného průzkumu společnosti Markent (Reml, 2010: 14), který zkoumá využití centralizace dat v Česku, plánuje přibližně 15% subjektů realizovat projekt posilující centralizaci dat – nemusí se však jednat zrovna o formu v podobě datových skladů.



Obr. 12 Datové sklady v různých typech českých organizací podle (Reml, 2009)





ku podle (Reml, 2010)

4.13 Úvod do BPMN

BPMN (Business Process Modeling Notation) je standard pro modelování business procesů, který sjednocuje pohled na procesy uvnitř společnosti tak, aby byl srozumitelný všem jejich účastníkům. Snaží se zmenšit komunikační rozdíly mezi analytiky navrhujícími procesy a vývojáři, kteří následně procesy implementují. Notace byla původně vytvořena iniciativou BPMI (Business Process Management Iniciative), nyní je však pod správou neziskové společnosti OMG (Object Managemen Group), která má v portfoliu např. již úspěšnou modelovací specifikaci UML (Unified Modeling Language) (OMG, 2008: 1)

Specifikace BPMN primárně definuje pravidla pro vytváření a používání diagramů (BPD - Business Process Diagram), které slouží k popisu a řízení business procesů. Diagramy jsou ve formátu *flowchart*, který je u analytiků oblíbený pro popis vnitropodnikových procesů. I přes schopnost zachytit komplexní procesy se diagramy skládají z malé množiny grafických objektů, což má usnadnit pochopení procesních dějů analytiky, vývojáři i managementem. (OMG, 2008: 11)

Hlavní čtyři kategorie objektů v BPD jsou . (OMG, 2008: 17):

- flow objects - Events, Activities, Gateways
- connecting objects - Sequence Flow, Message Flow, Association
- swimlanes - Pool, Lane
- artifacts - Data Object, Group, Annotation

Sekundárním cílem BPMN je schopnost převodu grafických procesních diagramů do vykonávacích programovacích jazyků (BPEL – Business Process Execution Language) Touto metodou lze modelovat a zároveň implementovat některé procesy. (OMG, 2008: 1)

5. Praktická část

5.1 Cíl projektu

Cílem je vytvoření několika *business process* diagramů (BPD) pomocí modelovacích nástrojů a podle pravidel BPMN (Business Process Modeling Notation). Ty budou popisovat některé hlavní procesy probíhající v podniku, kde je aplikovaný datový sklad. Diagramy budou tedy zachycovat reálný sled kroků používaných při operacích s datovými sklady. Tato názorná ukázka by měla napomoct lepšímu pochopení podstaty datových skladů a s nimi souvisejících úkonů. K modelování BPD bude použit freeware *Intalio Designer 6.0.3.015.*

Vybrané procesy k modelování jsou :

1. proces ETL
2. proces technické údržby
3. proces reportingového servisu
4. proces požadavku na informace

5.2 Encyklopedie

5.2.1 Slovní popis procesů

Proces ETL popisuje všechny etapy prováděné při získávání „surových“ dat z provozních systémů do datového skladu. Je to primární proces, který je v podobných obměnách součástí každého datového skladu. Proces je pravidelný, většinou je naplánován s denní nebo týdenní frekvencí. Výsledkem procesu je aktualizovaný datový sklad zkontrolovanými platnými daty. Procesu se účastní lidé v podobě oddělení IT a komponenty IT infrastruktury – datové sklady a datové zdroje.

Proces technické údržby je součástí každodenního provozu datového skladu a s tím spojené správy a údržby. Proces zahrnuje několik hlavních činností, které je nutno pravidelně provádět pro udržení datového skladu v dobrém stavu. Proces je pravidelný a je plánovaný podle velikosti a náročnosti datového skladu. Hlavní entitou tohoto procesu je oddělení IT, které komunikuje s operátorem datového skladu a podle zjištěných informací provádí technickou údržbu.

Proces reportingového servisu popisuje činnosti pravidelné tvorby reportů z údajů z datového skladu podle požadavků managementu a ostatních zaměstnanců, kteří běžně pracují s datovým skladem. Reaguje na měnící se aktuálně požadované informace a přizpůsobuje tomu obsah reportů. Jde o v pořadí druhý klíčový proces kolem datových skladů, totiž o výstup informací. V procesu probíhá komunikace mezi uživateli, kteří specifikují požadavky na reporty, a IT podporou, která pomocí statistik využití datového skladu tyto reporty vytváří a distribuuje. Výsledkem procesu je každodenní přístup k aktuálním firemním informacím v tištěné nebo elektronické podobě.

Proces požadavku na informace popisuje komunikaci v případě náhlého požadavku na určité informace, jež mají být získány z datového skladu. Běžné informace mají uživatelé poskytnuty pomocí klasických reportingových sestav, v případě složitějších dotazů nebo specifických požadavků na informace se mohou pokusit informace získat vlastními silami nebo požádat business analytiky, kteří se pokusí informace získat a případně je poskytnout. V procesu komunikují uživatelé a analytici s výstupním modulem datového skladu. Výsledkem je poskytnutí, respektive získání požadovaných informací ve vhodnou chvíli.

5.2.2 Seznam událostí

V jednotlivých procesech se dějí následující události:

1. Proces ETL
    - připojení k datovým zdrojům
    - kontrola připravenosti dat. zdrojů
    - zjišťování rozdílů od posledního nahrávání,
    - nahrání požadovaných dat do pracovního úložiště
    - čištění dat
    - integrování
    - denormalizace
    - sumarizace
    - kontrola dat připravených pro nahrání do dat. skladu
    - nahrávání tabulek dimenzí
    - nahrávání tabulek faktů
    - indexování dat
    - identifikování

- kontrola nahraných dat
- řešení problémů

2. Proces technické údržby
    - zjištění aktuálního stavu
    - zálohování
    - kontrola hardware
    - výměna komponent
    - přidání komponent
    - optimalizování výkonu
    - kontrola dat
    - oprava dat

3. Proces reportingového servisu
    - zjišťování požadavků na report
    - zahrnutí speciálních požadavků
    - získání statistik používání datového skladu
    - návrh struktury reportu
    - úprava obsahu
    - vytvoření reportu
    - uložení reportu pro budoucí verifikace
    - poskytnutí reportu v různých formátech
    - příprava pro tištěnou verzi
    - nahrání reportu na firemní intranet
    - úprava funkcionality intranetového rozhraní
    - definování přístupových práv

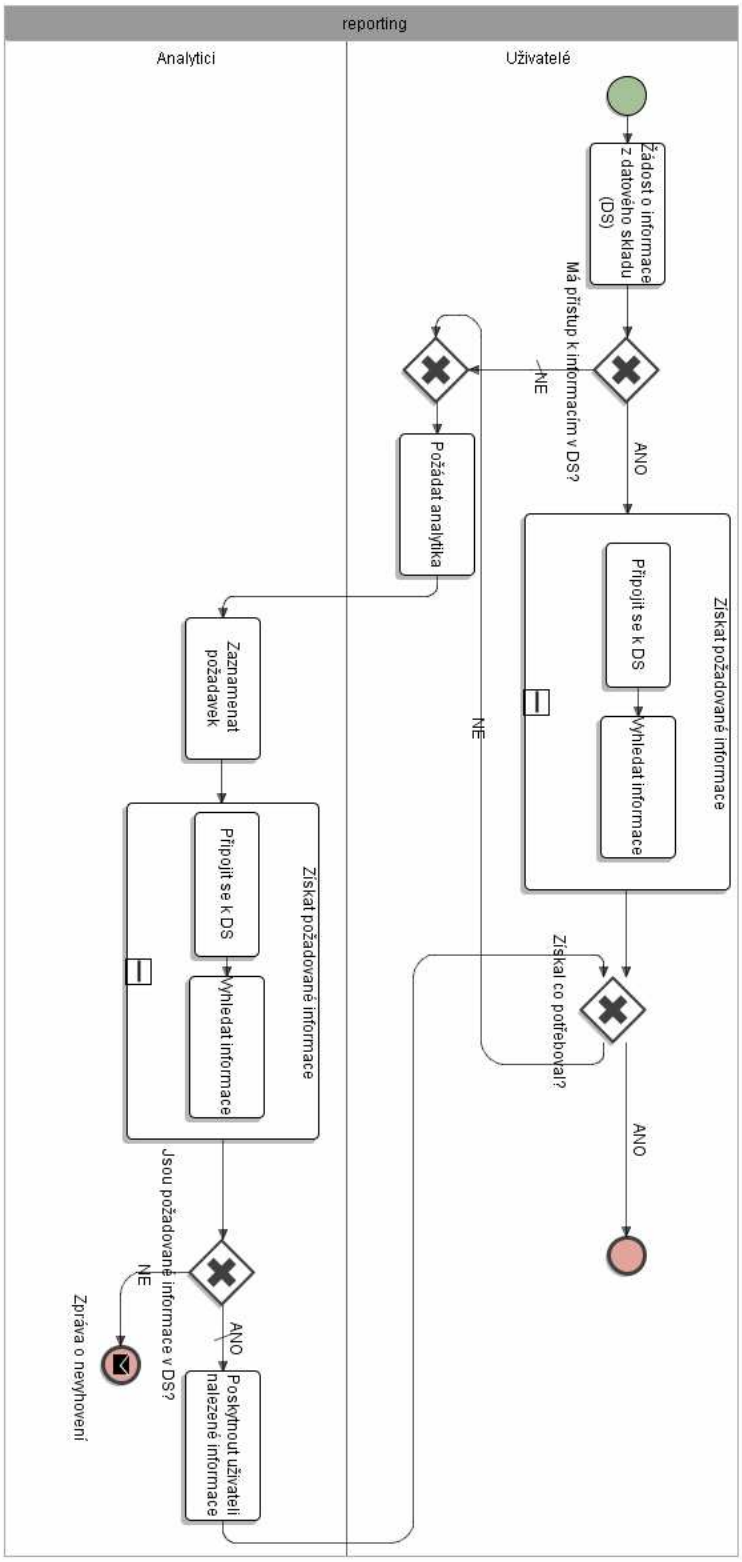
4. Proces požadavku na informace
    - žádost o informace
    - získání požadovaných informací
    - požádání analytika v případě problémů
    - zaznamenání požadavků analytikem pro technickou podporu
    - poskytnutí požadovaných informací uživateli

5.2.3 Seznam entit

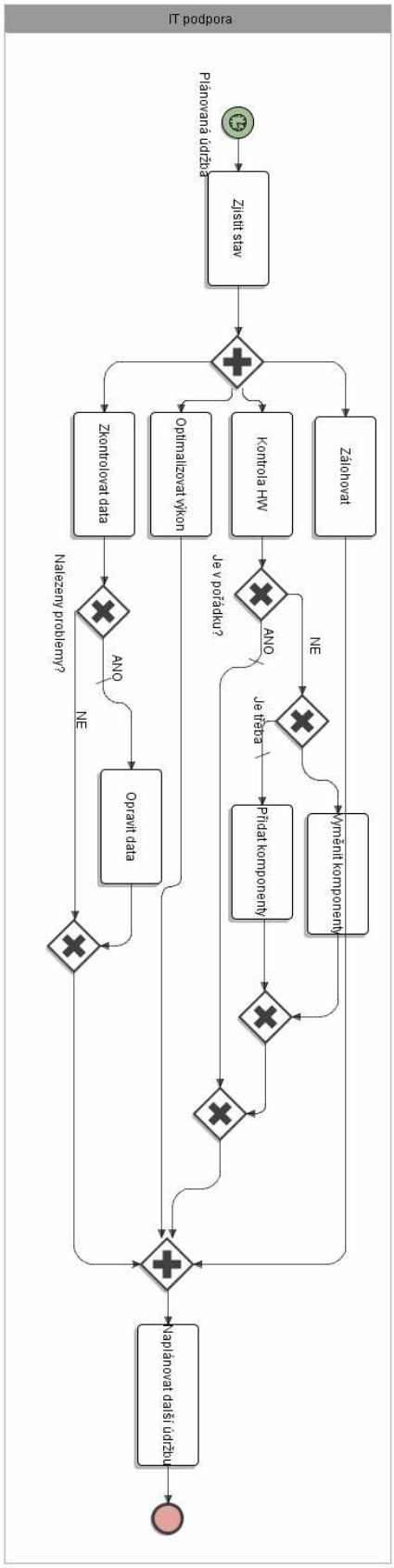
V modelovaných procesech jsou použity následující entity:

- *Uživatelé* - Blíže nespecifikovaní zaměstnanci firmy, kteří pracují s datovým skladem ve formě získávání informací
- *Analytici* Skupina specializovaných firemních odborníků, primárně využívající datový sklad pro zisk údajů a informací pro další rozbory.
- *Podpora IT* Oddělení starající se o veškeré informační technologie použitém firmě, včetně datového skladu.
- *Datový sklad (DWH)* Úložiště veškerých údajů, se kterými se pracuje v souvislosti s vyhledáváním informací a různých souvztažností.
- *Datové zdroje* Primární provozní systémy určené pro automatizování klíčových firemních procesů, generující data vhodná k informačnímu rozhodování.
- *Zálohovací médium* Páskové nebo jiné zálohovací úložiště určené pro bezpečnostní ukládání citlivých firemních dat.
- *Operátor datového skladu* Obsluha datového skladu.

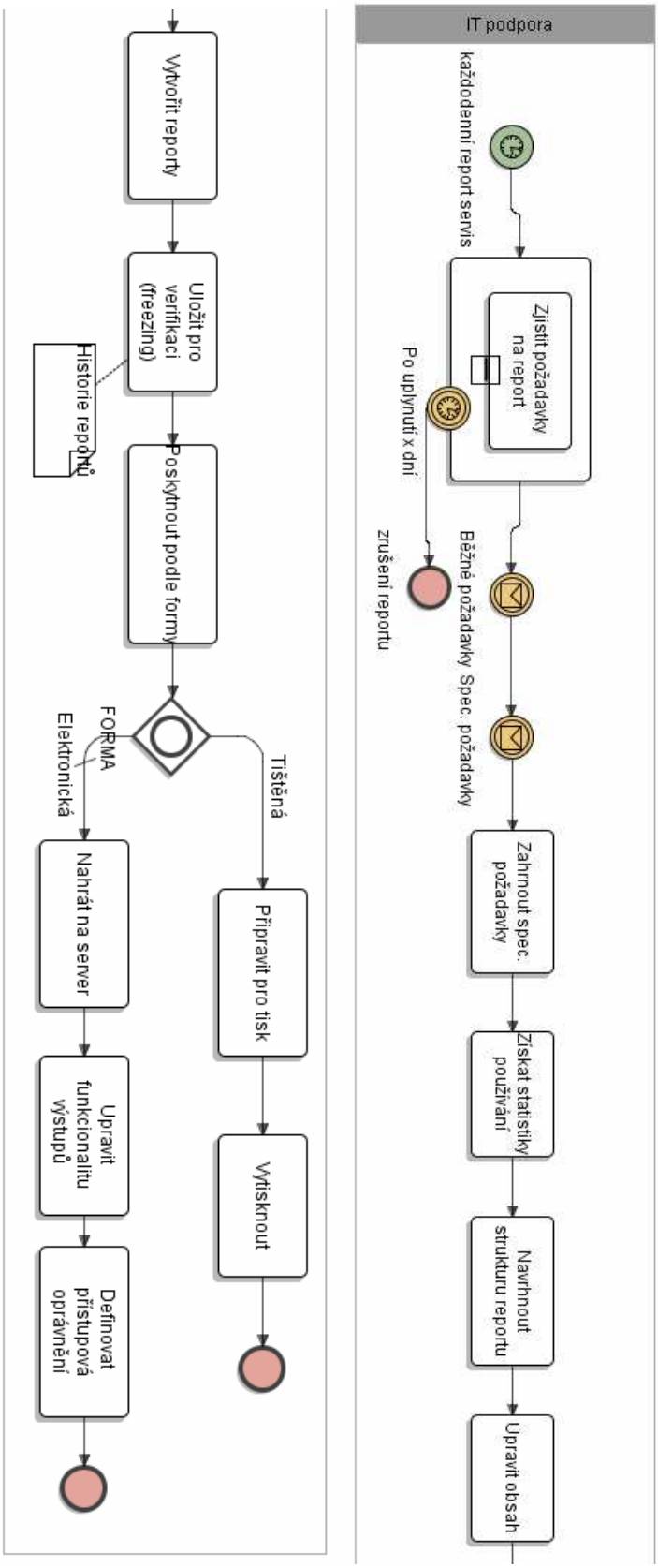
5.2.4 Modely chování - diagramy



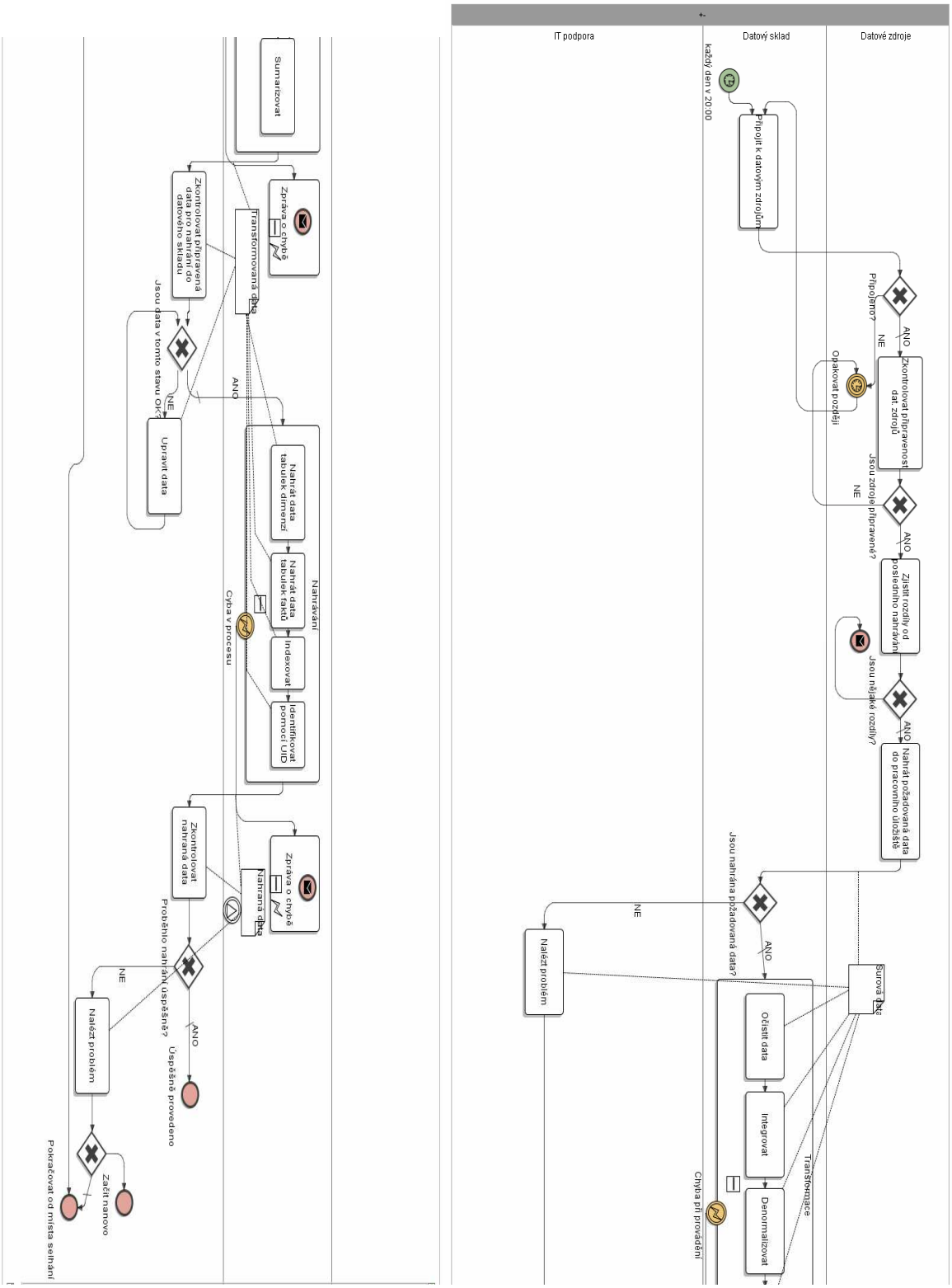
Obr. 14 Diagram procesu požadavku na informace z datového skladu



Obr. 15 Diagram procesu technické údržby datového skladu



Obr. 16 Diagram procesu reportingového servisu



Obr. 17 Diagram procesu ETL

6. Zhodnocení a výsledky

Procesní diagramy slouží k popisu a pochopení dějů, které vznikají v souvislosti s vybudováním datového skladu v podniku. Modelování diagramů je jeden ze stupňů informačního inženýrství. Vytvořením adekvátních modelů se lépe navrhují další navazující části systému a poskytuje se snazší údržba a levnější provoz.

Konkrétní použitá forma *business process* modelů podle notace BPMN je standardem, který umožňuje následný převod modelů do programovacích jazyků a přímou implementaci takto namodelovaných procesů.

Proto by se vytvořené modely v této práci daly po úpravě a začlenění do širšího kontextu použít pro zautomatizování těchto dějů.

7. Závěr

Hlavním cílem této bakalářské práce bylo shrnout dosavadní poznatky o datových skladech a vytvoření procesních diagramů modelujících děje datového skladu. Východiskem pro toto teoretické shrnutí a navazující praktickou práci byla část popisující procesy správy dat ve firmě, jíž se zabývá třetí kapitola. V této části se podařilo uvést všechny důležité informace o tom, kde a jaké údaje ve firmách vznikají a jakým způsobem je s nimi nakládáno. Hlavním závěrem této kapitoly je požadavek na oddělení systémů pro transakční a analytické zpracování ve snaze nasadit nástroje *Business Intelligence* do firemních informačních systémů.

V literární rešerši se konfrontováním různé odborné literatury povedlo shromáždit informace pojednávající o všech důležitých aspektech datových skladů. Zároveň se ukázalo, že poznatky jednotlivých autorů jsou postaveny na podobných základech. Dále lze říci, že jejich prognózy ohledně některých trendů v oblasti *business intelligence* se v současné době potvrzují. Například vzhledem k pokračujícím technickým pokrokům, které vedou ke snižování nákladů na paměťové úložiště, se častěji uplatňují multidimenzionální prostředí. Celkově je kvůli výhodám, které systémy podporující provozní a strategická rozhodování přinášejí, stále se zvyšující poptávka po datových skladech. Jejich nasazení v budoucnu plánuje výrazné množství společností.

Hlavním cílem praktické části bylo vymodelování procesních diagramů datových skladů. To by nebylo možné bez seznámení se s procesy, které probíhají při činnosti datových skladů. K tomuto účelů výborně sloužilo zpracování teoretické literární rešerše. Následné vytváření diagramů prokázalo, že pomocí programů podporující specifikaci BPMN lze modelovat jednoduché i komplexní situace, které jsou schopny zaznamenat každý detail v procesu. Programové prostředí však umožňuje i následné využití v podobě převodu grafických schémat do programovacích jazyků a tím i automatizování procesů vhodných k implementaci. Tato možnost nebyla součástí práce, mohla by na ni však navazovat po určitých úpravách namodelovaných procesů.

8. Seznam literatury

BALAŽOVIČ, Igor. *Business intelligence – Komplexní řešení pro široké spektrum zákazníků.* Speciální vydání časopisu IT Systems, Data warehousing Business intelligence. 2008. s. 2-3. ISSN 1212-4567.

BRODNÍČEK, David. *Úvod do problematiky datových skladů*. Automatizace. 2009, roč. 52, č. 7-8, s. 428-429. ISSN 0005-125X.

ČERNÝ, Jiří. *Data nad zlato*. Connect!. 2009, roč 14, č. 2, s. 16. ISSN 1211-3085.

DRAGOLOV, Daniel. *Správné informace jsou konkurenční výhodou.* Speciální vydání časopisu IT Systems, Data warehousing Business intelligence. 2007. s. 8-10. ISSN 1212-4567.

HROCH, Michal, CACH, Pavel. *Business intelligence ruku v ruce s datovým skladem.* Speciální vydání časopisu IT Systems, Data warehousing Business intelligence. 2007. s. 2-4. ISSN 1212-4567.

HUMPHRIES, Mark a kol.. *Data warehousing : návrh a implementace.* 1. vyd. Praha: Computer Press, 2002. 257 s. ISBN 80-7226-560-1.

JANKOVSKÝ, Zbyněk. *Databázové myšlení: Historická data – má smysl zálohovat?* Connect!. 2009, roč. 14, č. 6, s. 17. ISSN 1211-3085.

JARKE, Matthias a kol.. *Fundamentals of data warehouses.* 2nd ed. Berlin: Sprinter, 2003. 219 s. ISBN 3-540-42089-4.

JEŽEK, František. *Reporting a manažerské výstupy z ERP systému*. IT Systems. 2010, roč. 12, č. 4, s. 14-16. ISSN 1802-002X.

JEŽEK, Lubomír. *Business Intelligence v praxi.* Hospodářské noviny, příloha „ICT Revue“. 2010, roč. 5, č. 2, s. 16. ISSN 0862-9587.

KUBÁSEK, Jiří. *Data mining - móda nebo pragmatismus?*. Automatizace. 2009, roč. 52, č. 7-8, s. 421-424. ISSN 0005-125X.

LACKO, Luboslav. *Business Intelligence v SQL Serveru 2005: Reportovací, analytické a další datové služby.* 1. vyd. Brno: Computer Press, 2006. 391 s. ISBN 80-251-1110-5.

LACKO, Luboslav. *Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL Serveru Oracle.* 1. vyd. Brno: Computer Press, 2003. 485 s. ISBN 80-7226-969-0.

MERUNKA, Vojtěch. *Objektové modelování.* 1. vyd. Praha: Alfa Nakladatelství, 2008. 197 s. ISBN 978-80-87197-04-2.

OMG. *Business Process Model and Notation, V1.1* [online]. 2008. 294 s. Dostupné z WWW: < http://www.omg.org/spec/BPMN/1.1/PDF>

REML, Jiří. *Pokročilá práce s daty – stále vzrůstající trend.* Computerworld. 2009, roč. 20, č. 13, s. 14. ISSN 1210-9924.

REML, Jiří. *Správa dat v českých organizacích.* Computerworld. 2010, roč. 21, č. 1, s. 14. ISSN 1210-9924.

ŘECHTÁČKOVÁ, Jana. *Systémy jediné pravdy, které mění data v peníze. Business intelligence.* Speciální vydání časopisu IT Systems, Data warehousing Business intelligence. 2007. s. 22-24. ISSN 1212-4567.

ŠOULE, Marek. *Správa dokumentu versus ERP systém*. IT Systems. 2010, roč. 12, č. 4, s. 17-18. ISSN 1802-002X.

ŠPRUNGL, Petr. *Jak se vyhnout datové džungli*. CIO business word : IT strategie pro manažery. 2010, s. 20. ISSN 1803-7321. Dostupné též z WWW: <http://businessworld.cz/podnikove-is/jak-se-vyhnout-datove-dzungli-6922>

ŠVEC, Martin. *Objektové databáze* [online]. 2003. 20 s. Závěrečná práce z předmětu. Vysoké učení technické v Brně. Dostupné z WWW: <http://www.fit.vutbr.cz/study/courses/VPD/public/0203VPD-Svec.pdf>

VALEČKOVÁ, Iva. *Na co nezapomenout v projektu budování datového skladu.* Speciální vydání časopisu IT Systems, Data warehousing Business intelligence. 2007. s. 5-7. ISSN 1212-4567.

VAVRA, Tom. *Business intelligence v otázkách a odpovědích.* Speciální vydání časopisu IT Systems, Data warehousing Business intelligence. 2007. s. 26-27. ISSN 1212-4567.

VOSTROVSKÝ, Václav. *Vytváření databází v oracle.* 1. vyd. Praha: Česká zemědělská univerzita, 2008. 134 s. ISBN 978-80-213-1191-6.

ZAVORAL, Petr. *Manažerské nástroje pro rozhodování.* Hospodářské noviny, příloha „ICT Revue“. 2010, roč. 5, č. 2, s. 12-15. ISSN 0862-9587.