Analýza a návrh datového skladu pro telekomunikační společnost

Bc. Josef Jurák

Diplomová práce
2006

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky
Ústav aplikované informatiky
akademický rok: 2005/2006

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Josef JURÁK**
Studijní program:	**N 3902 Inženýrská informatika**
Studijní obor:	**Informační technologie**
Téma práce:	**Analýza a návrh datového skladu pro telekomunikační společnost.**

Zásady pro vypracování:

Analýza problematiky datových skladů včetně specifických požadavků firmy působící na telekomunikačním trhu.
Shrnutí požadavků na datové výstupy jednotlivých částí firmy a možnosti dalších analýz nad dostupnými daty.
Návrh řešení samotného DataWareHouse systému včetně napojení na další podnikové systémy, i řešení reportingu a dalších systémů typických pro Business Intelligence.
Řešení bude založeno na systémech MS SQL Server 2005, a bude též využívat služeb MS Analysis, Reporting a Integration Services.

Rozsah práce:

Rozsah příloh:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

Lacko, L.: Databáze: datové sklady, OLAP a dolování dat. Computer Press, Praha, 2003.
Humphries, M., Hawkins M. W. a kol.: Data warehousing, Principy a praxe. Computer Press, Praha, 2002.
Lacko, L.: Business Intelligence v SQL Serveri 2005. Microsoft, Praha, 2005.
Vieira R.: SQL Server 2000 Programujeme profesionálně. Computer Press, Praha, 2001.
Iseminger, D.: Microsoft SQL Server 2000 Reference Library. Microsoft Press, Redmond, 2001.

Vedoucí diplomové práce: **RNDr. Ing. Miloš Krčmář**
Ústav aplikované informatiky

Datum zadání diplomové práce: **14. února 2006**

Termín odevzdání diplomové práce: **26. května 2006**

Ve Zlíně dne 14. února 2006

prof. Ing. Vladimír Vašek, CSc.
pověřený děkan

UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ
Fakulta aplikované informatiky

doc. Ing. Ivan Zelinka, Ph.D.
ředitel ústavu

ABSTRAKT

Business Intelligence a Data warehouse jsou základní prostředky pro podporu rozhodování, které se rychle dostávají do popředí problematiky průmyslu databázových systémů. Rovněž data mining a dolování informací jsou jedny z nejrychleji se rozvíjejících oblastí počítačových věd. Původní dotazovací a reportovací nástroje bývaly užívány k popisu dat v databázi a k jejich získání. Uživatel formuluje hypotézu o vztazích a ověřuje ji prostřednictvím série dotazů. Data mining může být užíván k vytvoření takové hypotézy. Jsou však i jiné možnosti propojení technologií data warehousingu a data miningu. Ve své práci diskutuji několik přístupů takové spolupráce v různých vrstvách a globální trendy v této oblasti, které se staly typickým způsobem sjednocování heterogenních systémů.

Klíčová slova: Business Intelligence, Datový sklad, Dolování dat

ABSTRACT

Business Intelligence and Data Warehouse are essential elements of decision support, which has increasingly become a focus of the database industry. On the other hand, Data Mining and Knowledge Discovery is one of the fast growing computer science fields. Traditional query and report tools have been used to describe and extract what is in a database. The user forms a hypothesis about a relationship and verifies it with a series of queries. Data mining can be used to generate an hypothesis. But there is some other possibilities for collaboration between Data Warehouse and Data Mining technologies. In my thesis I discuss several approaches of such collaboration in different architecture levels and the global trends in this field which has become a significant way of heterogeneous systems integration.

Keywords: Business Intelligence, Data warehouse, Data mining

Velmi děkuji vedoucímu mé diplomové práce Ing. Miloši Krčmářovi, RNDr. za podnětné připomínky a rady, které vedly k její vypracování.

Zvláštní poděkování patří zejména Radce Bauerové za konzultace při analýze požadavků a další podporu při tvorbě této práce.

Ve Zlíně …….

Josef Jurák

OBSAH

ÚVOD

Cílem této diplomové práce je analýza a návrh datového skladu pro telekomunikační společnost.

V dnešní době, kdy je konkurenceschopnost každého podniku podmíněna správným rozhodováním v klíčových situacích, jsou kladeny daleko vyšší nároky na management takové firmy, než tomu bylo dříve. Souvisí to samozřejmě i s vývojem informačních systémů v každé větší společnosti. V 90. letech společnosti začaly masivně investovat do informačních technologií, CRM, ERP a dalších systémů, obsahující i cenné data pro rozhodování. Bohužel data v těchto systémech jsou běžnou cestou nedostupná, jsou umístěna v různých systémech a mohou mít rozličný význam, a tak situace v dnešních společnostech vypadá tak, že řadový manager bez znalosti programování, SQL, případně jiných způsobů získávání dat z databází není schopen z těchto systémů získat data pro své každodenní rozhodování. Je tedy třeba nejen sjednocený přístup k datům pomocí datového skladu, ale i správné pochopení těchto dat, vytváření a ověřování hypotéz, případně jejich pokročilá analýza.

Ve své práci jsem se zaměřil právě na tuto situaci a chtěl bych pomoci vedoucím pracovníkům a manažerům při jejich rozhodovacích procesech. Tzn. je zde potřeba nástroj, který tyto nedostatky nahradí a který v sobě zahrnuje tzv. Business Intelligence. Porovnáním a analýzou těchto prostředků a jejich kombinací na různých úrovních být schopni určit optimální řešení.

Práce tedy sleduje současné trendy v oblasti Business Intelligence a jejích nástrojů, zejména data miningu, online analytical processingu (OLAP) a data warehousingu (DWH) jako prostředků pro správné rozhodování skupiny lidí ve firmě, zaměřené na telekomunikační služby.

Jelikož je datový sklad určen pro telekomunikační, je nejprve třeba analyzovat její specifické potřeby a požadavky na systém.

Prostředky Business Inteligence je tak možno chápat jako sadu nástrojů umožňující transformovat dostupná data na užitečné informace použitelné pro rozhodování. Je však nutné zmínit, že kvalita těchto informací je závislá na kvalitě údajů ve zdrojových systémech, a tak současně s implementací řešení BI do firmy by měl růst tlak na správné zadávání údajů do jednotlivých systémů.

V této práci vycházíme z toho, že podstatnou část software firmy tvoří produkty firmy Microsoft a tudíž analýza prostředků Business Intelligence, analýza možností Business Intelligence i stručný návrh řešení je založen na produktu Microsoft SQL 2005.

I. TEORETICKÁ ČÁST

1 BUSINESS INTELLIGENCE

Business Inteligence (BI) představuje komplex přístupů a aplikací IS/ICT, které téměř výlučně podporují analytické a plánovací činnosti podniků a organizací a jsou postaveny na principu multidimenzionality, kterým zde rozumíme možnost pohlížet na realitu z několika možných úhlů.

1.1 Historie a vývoj

Řešení směřující k podpoře manažerských a analytických úloh v podnikovém řízení se začala objevovat již na konci sedmdesátých let minulého století v souvislosti s rozvojem on-line zpracování dat. Prvotní pokusy a aplikace jsou spojeny s americkou firmou Lockheed. V polovině osmdesátých let byly publikovány první významné práce k tomuto typu aplikací. V druhé polovině osmdesátých let přišly na trh v USA první firmy s komerčními produkty, založenými na multidimenzionálním uložení a zpracování dat, označovanými jako EIS (Executive Information System). Trh s EIS produkty se pak velmi rychle rozvíjel a na začátku devadesátých let (od roku 1993) se tyto produkty začaly prosazovat i na českém IS/ICT trhu.

Koncem osmdesátých a začátkem devadesátých let se v USA začal velmi silně prosazovat i další trend v multidimenzionálních technologiích, a to datové sklady (Data Warehouse) a datová tržiště (Data Marts). Za rozvojem těchto technologií stáli především Ralph Kimball a Bill Inmon. Větší uplatnění datových skladů a tržišť je na českém trhu patrné spíše až v druhé polovině devadesátých let. V souvislosti s datovými sklady a narůstajícím objemem dat v tomto prostředí se v průběhu devadesátých let začaly prosazovat i technologie a nástroje tzv. dolování dat (Data Mining) založené na vysoce sofistikovaných analýzách dat s pomocí nejrůznějších matematických a statistických metod.

1.2 Business Intelligence ve firmě

V prostředí stále tvrdší konkurence musí podnikoví analytici a manažeři rozhodovat pod časovým tlakem a současně s vysokou zodpovědností. To znamená, že pro tato rozhodnutí musí mít dostatek relevantních a objektivních informací, které jsou dostupné rychle, s minimální technickou náročností na manipulaci, a přitom s možností rychle formulovat nové požadavky na další informace odpovídající aktuální obchodní nebo výrobní situaci.

Zpracování a uložení dat v transakčních systémech, především v aplikacích ERP, je založeno vesměs na využití relačních databázových systémů. Toto řešení je v mnohém ohledu velmi výhodné. Data jsou zde přehledně uspořádána, a v případě efektivně navržené datové základny umožňují rychlé provádění jednotlivých transakcí a poskytují odpovídající dobu odezvy na zadané dotazy. Navíc zajišťují integritu dat, bezpečnost přístupu k datům a další potřebné charakteristiky spojené s řízením firmy na taktické nebo operační úrovni. ERP aplikace však mají z hlediska analytických a plánovacích činností podniku některá omezení:

- Neumožňují rychle a pružně měnit kritéria pro analýzy podnikových dat (např. sledovat data o prodeji – v čase, podle zákazníků, produktů, segmentů trhu, obchodních zástupců, podnikových útvarů, a dále i v nejrůznějších kombinacích uvedených kritérií).
- Stejně tak se v obrovských objemech dat současných databází obtížně řeší zajištění okamžitého přístupu pracovníků k agregovaných datům, a to na nejrůznějších úrovních agregace (za podnik, útvar, za všechny zákazníky, skupiny zákazníků, jednotlivé zákazníky atd.)
- ERP a ostatní transakční aplikace jsou primárně určeny pro pořizování a aktualizace dat, přičemž některé z nich pracují neustále téměř na 100% svého možného výkonu; analytické úlohy tyto systémy nadměrně zatěžují a v možných případech nejsou ani díky jejich vytížení možné.
- Dalším problémem je narůstající objem dat v podniku, který se průměru zdvojnásobí každých pět let. Většina firem tak nemá problém s nedostatkem dat, ale naopak firmy jsou jimi zahlceny, a to často redundantními a nekonzistentními daty, která jsou v rozhodovacích procesech obtížně využitelná.

Pokud o těchto omezeních mluvíme v souvislosti se systémy ERP, pak to neznamená, že by aplikace ERP a další transakční úlohy nebyly schopny zmíněné operace realizovat. Jde však o jejich rychlost a pružnost vzhledem k uživatelským požadavkům. Řešení uvedených problémů se tak postupně stalo doménou speciálních technologií a aplikací Business Inteligence.

1.3 Vymezení a definice BI

Termín Business Inteligence zavedl v roce 1989 Howard J. Dresner, analytik společnosti Gartner Group, který jej popsal jako „sadu konceptů a metod určených pro zkvalitnění rozhodnutí firmy“. Vyzdvihuje zde význam datové analýzy, reportingu a dotazovacích nástrojů, které provádějí uživatele množstvím dat a pomáhají mu se syntézou hodnotných a užitečných informací.

Možná právě díky krátké době existence tohoto termínu není pro něj v dnešní době zavedena jednotná definice, podporovaná jakoukoliv organizací, zabývající se standardy (např ANSI). Například definice ze serveru searchCRM.com definuje Business Inteligence jako určitou kategorii aplikací a technologií pro sběr, skladování, analyzování a zpřístupňování dat, jejichž účelem je pomoci podnikovým uživatelům dělat lepší rozhodnutí. BI aplikace podle searchCRM.com zahrnují funkčnost systémů pro podporu rozhodování, dotazování a reportingu, statistických analýz, vytváření prognóz a Data Mining.

Server firmy iolap charakterizuje BI jako sběr a analýzu dat, jejímž celme je lepší porozumění a reakce na změny, kterým organizace čelí. Definici založenou na znalostech a informacích nabízí sever DM Review, který definuje BI jako znalosti o podniku získané za pomoci rozličných hardwarových a softwarových technologií, které umožňují organizaci přeměnit data na informace. Tento popis považuje technologie pouze za prostředek, nikoli za podstatu Business Inteligence.

Česká společnost pro systémovou integraci definuje Business Intelligence jako sadu procesů, aplikací a technologií, jejichž cílem je účinně a účelně podporovat rozhodovací procesy ve firmě. Podporují analytické a plánovací činnosti podniků a organizací a jsou postaveny na principech multidimenzionálních pohledů na podniková data.

Aplikace BI pokrývají analytické a plánovací funkce většiny oblastí podnikového řízení, tj. prodeje, nákupu, marketingu, finančního řízení, controllingu, majetku, řízení lidských zdrojů, výroby, IS/ICT apod.

Do nástrojů a aplikací Business Inteligence se zahrnují:

- Produkční, zdrojové systémy
- Dočasná úložiště dat (DSA – Data Staging Area)
- Operativní úložiště dat (ODS – Operational Data Store)

- Transformační nástroje (ETL – Extraction Transformation Loading)
- Integrační nástroje (EAI – Enterprise Aplication Integration)
- Datové sklady (DWH – Data Warehouses)
- Datová tržiště (DMA – Data Marts)
- OLAP
- Reporting
- Manažerské aplikace (EIS – Executive Information Systems)
- Dolování dat (Data Mining)
- Nástroje pro zajištění kvality dat

K uvedenému přehledu nástrojů, které zahrnujeme do BI, je nutné poznamenat, že i na tuto otázku existují různé názory. Vedle toho, který chápe BI jako široký rámec od manažerských aplikací po reporting, existuje i jiný pohled, který chápe BI pouze jako jeden z nástrojů vedle, resp. nad datovými sklady nebo datovými tržišti. V našem případě se budeme nadále držet první z uvedených variant.

Ze všech předcházejících vymezení a definic však vyplývá, že BI je orientován na vlastní využití informací v řízení a rozhodování, a nikoli na základní zpracování dat a realizaci běžných obchodních, finančních a dalších transakcí. To, jak jsou možnosti BI využity, dnes do značné míry ovlivňuje výkonnost a kvalitu řízení firmy, a v souvislosti s tím nakonec i její celkovou úspěšnost a konkurenceschopnost.

1.4 Principy řešení BI

1.4.1 Multidimenzionální databáze

Informační systémy mohou pracovat se 2 základními typy informací – operativními a analytickými. První typ, operativní informace, slouží pro realizaci obchodních a dalších transakcí v podniku. Jsou uloženy většinou v relačních databázích, zobrazují aktuální stav podniku a v průběhu jednoho dne se mohou i několikrát měnit. Příkladem může být např. účetnictví, data v dokumentech obchodních případů apod. Transakční systémy realizují jejich zpracování v reálném čase a označují se jako OLTP (On Line Transaction Processing)

sing) systémy. Vzhledem k analytickým aplikacím se data OLTP systémů chápou jako primární, zdrojová nebo produkční.

Na druhé straně systémy pracují s analytickými informacemi využívají primární data vytvořená v OLTP systémech.

Pro své uložení a operace s daty se pro tyto systémy vžil v osmdesátých letech minulého století název OLAP – On Line Analytical Processsing. Se zavedením pojmu BI (který ve své podstatě kopíruje výše zmíněný význam výrazu OLAP) a současně s rozvojem nástrojů a technologií pro 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. Dále budeme OLAP chápat pávě v tomto užším technologickém významu.

Na rozdíl od OLTP systémů jsou analytické systémy charakteristické především těmito vlastnostmi:

- Informace poskytují na základě vstupů získaných z primárních dat
- Jejich data jsou uložena multidimenzionálně, resp. v multidimenzionálních databázích
- Obsahují různé úrovně agregace dat, podle hierarchické struktury dimenzí
- Zachycují faktor času a umožňují realizovat časová srovnání, časové řady, predikovat možný vývoj sledovaných ukazatelů apod.

1.4.1.1 Architektura

Pro data analytického typu se nehodí, aby byla ukládána v relačních databázích do podoby třetí normální formy (typické pro transakční systémy). Aby analytické systémy mohly poskytovat různé analýzy a přehledy sloužící pro strategické rozhodování, je nutné, abychom se na jejich data mohli dívat z více hledisek současně. Mělo by tedy být možné vytvářet tzv. multidimenzionální pohledy, což je pro data uložená v třetí normální formě velký problém. Nástroje koncového uživatele musí umožňovat analýzu ve smyslu nacházení souvislostí, které nejsou z primárních dat na první pohled zřejmé. Navíc je nutné procházet

velká množství dat, vypočítávat agregace (které v databázích modelových ve třetí normální formě nejsou automaticky uloženy), rychle měnit pohledy na data, rychle, a co možná automatizovaně, je ukládat do přehledných tabulek a grafů.

Multidimenzionální databáze jsou již optimalizované pro uložení a interaktivní využívání multidimenzionálních dat. Výhodou multidimenzionality, resp. nasazení OLAP technologií (v užším chápání), je rychlost zpracování a efektivní analýzy multidimenzionálních dat.

Základním principem, na němž jsou aplikace Business Inteligence založeny, je několikadimenzionální tabulka umožňující velmi rychle a pružně měnit jednotlivé dimenze, a nabízet tak uživateli různé pohledy na modelovanou ekonomickou realitu. Jde tak v podstatě o princip "n-dimenzionální Rubikovy kostky" naplněné nejdůležitějšími podnikovými daty.

1.4.1.2 Implementace

Databáze produkčních – transakčních systémů bývají modelovány v tzv. třetí normální formě. Jedná se o design databáze, který nejvíce vyhovuje požadavkům na tyto systémy, tedy možnosti data rychle a jednoduše ukládat a současně optimalizovat velikost databáze.

Datové modely produkčních systémů jsou komplexní, obsahují mnoho tabulek a jejich vazby. Neexistuje jedinečný způsob, jak provést dotaz do databáze. Těchto způsobů je větší množství a záleží jen na uživateli, kterou vazbu mezi tabulkami zvolí. Různé dotazy mohou dát stejné výsledky. Tím se ovšem pro běžného uživatele stává databáze velmi nepřehledná. Pro dotazy do více tabulek navíc musí vytvářet jakési propojovací můstky, a jak si dále ukážeme, právě tato propojení jsou největší zátěží systému.

Pro výše uvedené nedostatky se objevily snahy o zjednodušení ERD diagramu, přizpůsobení jej pro tvorbu datových sklad, a především přiblížit data koncovému uživateli. Vznikl tak relační "dimenzionální model" kterému se také běžně říká "Schéma hvězdy (Star schneme), resp. "Schéma sněhové vločky " (Snowflake scheme)

1.4.2 Porovnání OLAP a OLTP databází

Požadavek pohledů na data z více hledisek (dimenzí) tedy s sebou současně přináší i požadavek na optimalizované fyzické ukládání dat, přičemž se většinou jedná o data historická, agregovaná, průběžně rozšiřovaná a ukládaná v jednoduché struktuře vhodné pro analýzu a přizpůsobené potřebám managementu.

Jsou tedy zřejmé tyto základní rozdílové charakteristiky mezi OLTP a analytickými systémy:

- OLTP systémy jsou primárně určeny pro pořizování dat. Tomu odpovídá jejich celková architektura, a zejména pak databázový model. Ten je charakteristický tabulkami ve třetí normální formě, velkým počtem tabulek a jejich spojení, snahou o nulovou redundanci dat. Současně jsou tabulky indexovány pouze v nejnutnějších případech.
- Data jsou do OLTP systémů pořizována v reálném čase, v běžném provozu probíhají desítky až statisíce transakcí za minutu. Systémy jsou tak zatěžovány kontinuálně.
- Analytické systémy jsou oproti OLTP určeny primárně pro podporu dotazování. Z toho vyplývá jejich architektura (několikavrstvé databáze) a databázové modely, vyznačující se zmenšeným počtem tzv. nenormalizovaných tabulek, vyšší frekvencí indexů, duplicitou uložení dat apod.
- Drtivá většina analytických systémů aktualizuje svá data periodicky (nejběžněji v denních a měsíčních intervalech). Nejnovější trendy sice umožňují i aktualizaci dat v reálném čase, avšak v dnešní době se jedná o případy výjimečné. Zatížení analytických systémů je z tohoto důvodu nárazové – velké zatížení je typické pro období nahrávání dat, a následně lze pozorovat nepravidelnou zátěž podle frekvence a složitosti jednotlivých analytických úloh, probíhajících nad analytickými systémy.
- Zatímco OLTP systémy udržují data na maximální úrovni detailu (tedy na úrovni jedné transakce se všemi jejími detailními atributy), analytické systémy ukládají pouze data relevantní pro analýzy, tedy buď agregovaná na vyšší úroveň než jednotlivá transakce, nebo zahrnující pouze některé její atributy.

Aplikace Business Inteligence vycházejí z několika základních principů.

- Aplikace jsou orientovány výlučně na analytické a plánovací potřeby uživatelů, nikoli transakce,
- Data jsou uložena multidimenzionálně
- Dimenze mají většinou hierarchickou strukturu, které odpovídají agregační funkce v aplikacích

- Jako zdroj dat slouží produkční – transakční systémy
- Data jsou v databázi ukládána s časovým rozlišením
- V další části kapitoly se již budeme věnovat jednotlivým komponentám, resp. součástem řešení BI, a jejich rozdělení do vrstvené architektury.

1.5 Komponenty BI

Konkrétní uspořádání jednotlivých komponent v řešení BI se může výrazně měnit podle situace a potřeb daného zákazníka nebo podniku. To znamená v rozsahu od těch nejjednodušších řešení až po řešení nejkomplexnější a také technologicky, finančně i pracovně nejnáročnější.

Za dobu vývoje oblasti se ustálila obecná koncepce architektury řešení BI. Rozmanitost problémů řešených pomocí nástrojů BI, stejně jako rozmanitost nástrojů samotných, vede však k tomu, že tato obecná architektura má několik vývojových větší a také její konkrétní aplikace v reálných situacích se podstatně liší.

Lze identifikovat několik vrstev s tímto obsahem:

Vrstva pro extrakci, transformaci, čištění a nahrávání dat (komponenty datové transformace), která pokrývá oblast sběru/přenosu dat ze zdrojových systémů do vrstvy pro ukládání dat v řešení BI.

- ETL systémy – neboli systémy pro extrakci, transformaci a přenos dat
- EAI systémy – neboli systémy pro integraci aplikací

Vrstva pro ukládání dat (databázové komponenty), která zajišťuje procesy ukládání, aktualizace a správy dat pro řešení BI

- Datové sklady (Data Warehouse) – základní databázová komponenta řešení BI
- Datová tržiště (Data Marts) subjektově orientované analytické databáze, součást nebo nadstavba datového skladu
- Operativní datová úložiště (Operational Data Store) podpůrné analytické databáze.
- Dočasná úložiště dat (Data Staging Areas) – databáze pro dočasné uložení dat před jejich vlastním zpracováním do databázových komponent řešení BI

Vrstva pro analýzy dat (analytické komponenty), pokrývající činnosti spojené s vlastním zpřístupněním dat a analýzou dat.

- Reporting – analytická vrstva, zaměřená na standardní nebo ad hoc dotazovací proces do databázových komponent řešení BI
- Systémy On-Line Analytical Processing (OLAP) – vrstva zaměřená na pokročilé a dynamické analytické úlohy
- Dolování dat (Data Mining) – systémy zaměřené na sofistikovanou analýzu velkého množství dat.

Prezentační vrstva (nástroje pro koncové uživatele) zajišťující komunikaci koncových uživatelů s ostatními komponentami řešení BI, tedy zejména sběr požadavků na analytické operace a následnou prezentaci výsledků:

- Portálové aplikace založené na technologiích WWW
- Systémy EIS – Executive Information Systems
- Různé analytické aplikace.

Vrstva oborové znalosti (oborová znalost /know how), zahrnující oborovou znalost a tzv. best-pracitces nasazování řešení BI pro konkrétní situaci v organizaci.

Aplikace Business Intelligece kromě toho využívají následující obecné komponenty pro správu a manipulaci s daty:

- Nástroje pro zajištění datové kvality, tedy nástrojí zajišťující, že data přesně odpovídají realitě.
- Nástroje pro správu metadat, zabývající se zjednodušeně řečeno popisem a dokumentací systémů i probíhajících procesů.
- Technickou znalost, zahrnující programovací a technologicky závislé schopnosti implementačního týmu.

1.5.1 Zdrojové (produkční) systémy

Produkční (zdrojové) systémy - také někdy označované jako primární, transakční, OLTP nebo "legacy" jsou takové systému podniku, ze kterých aplikace Business Intellignece získávají data a nepatří do skupiny BI aplikací. Vlastností všech těchto systémů je jejich ar-

chitektura podporující ukládání a modifikaci dat v reálném čase. Oproti BI aplikacím tyto systémy nejsou navrženy pro analytické úlohy. Příkladem mohou být ERP, SCM, CRM systémy, specializované systémy pro podporu personálních oddělení, pro podporu finančních oddělení a další. Zdrojem pro řešní BI nemusí být pouze vnitřní systémy podniku, ale i externí systémy (např. databáze podnikatelských subjektů, telefonní systémy, výstupy statistických úřadů či vládních institucí apod.)

Produkční systémy jsou hlavní, a často i jediným vstupem do BI. V praxi je většinou spektrum zdrojových, resp. produkčních, systémů pro BI velmi různorodé a heterogenní jak obsahově, tak technologicky. Úkolem řešení BI je pak zajistit analýzu těchto zdrojů z pohledu potřeb řízení firmy, výběr relevantních dat pro řízení, a následně jejich vzájemnou integraci. Právě tato část projektů BI je pracovně, časově i finančně nejnáročnější, ale představuje zcela nezbytný předpoklad úspěšných aplikací BI.

1.5.2 Extraction, Transformation and Loading – ETL

ETL je jednou z nejvýznamnějších komponent celého komplexu BI. Běžným označením pro prostředky ETL je rovněž datová pumpa. Jejím úkolem je data ze zdrojových systémů získat a vybrat (Extraction), upravit do požadované formy a vyčistit (Transformation) a nahrát je do specifických datových struktur, resp. datových schémat, datového skladu (Loading). Tyto nástroje lze tedy použít pro přenos dat mezi dvěma (či více) libovolnými systémy. Stejně jako nástroje pro zajištění datové kvality a správu metadat však ETL systémy získaly na důležitosti až s rozvojem analytických systémů, tedy s explicitní potřebou pro zajištění přenosu dat mezi různými aplikačními systémy v rámci různorodého databázového prostředí. Na rozdíl od aplikací EAI (popsaných v následujících částech kapitoly) pracují nástroje ETL v dávkovém (Batch) režimu, data jsou tedy přenášena v určitých časových intervalech. Pro dnešní BI řešení se jedná většinou o denní, týdenní a měsíční intervaly.

1.5.3 Enterprise Application Integration – EAI

Nástroje EAI vznikly, a dnes jsou v naprosté většině případů využívány, ve vrstvě zdrojových systémů. Jejich cílem je integrovat a primární podnikové systémy a razantně redukovat počet jejich vzájemných rozhraní. Tyto nástroje pracují principiálně na 2 úrovních.

- na úrovni datové integrace, kde jsou EAI platformy využity pro integraci a distribuci dat
- Na úrovni aplikační integrace, kde jsou EAI platformy využity nejen pro integraci a distribuci dat, ale především pro sdílení určitých vybraných funkcí informačních systémů.

Na rozdíl od nástrojů ETL pracují EAI platformy v reálném čase.

Své využití v Business Inteligence řešení nachází zejména vrstva datové integrace, kde jsou nástroje EAI využity pro přenos dat do datových úložišť v reálném čase. EAI tak doplňuje dávkový přenos a umožňuje vznik nové generace datových skladů, tzv. Real – Time Data Warehouse.

1.5.4 Dočané úložiště dat – DSA (Data Staging Areas)

Důvodem existence dočasného úložiště dat je dočasné uložení extrahovaných dat z produkčních systémů, a jeho hlavním úkolem je podporovat rychlou a efektivní extrakci (výběr) dat. Data Staging Area (DSA) je tak zde první popsanou komponentou BI řešení, sloužící pro ukládání dat. DSA slouží k prvotnímu ukládání netransformovaných dat ze zdrojových systémů. Jedná se o nepovinnou komponentu řešení BI, která nachází své uplatnění.

- u neustále zatížených produkčních systémů, kde je potřeba transferovat jejich data s minimálním dopadem na jejich výkonnost
- u systémů, jejichž data je třeba před zpracováním konvertovat do databázového formátu (např. systémy pracující s textovými soubory apod.)

Dočasné úložiště dat (DSA) obsahuje data s následujícími charakteristikami:

- detailní – data nejsou integrována,
- nekonzistentní – data nejsou kontrolována proti externími číselníkům či ostatním datům v datovém skladu,
- Neobsahující historii – přenášejí se pouze aktuální data ze zdrojového systému
- Měnící se – při každém snímku se berou pouze data, která ještě nebyla zpracována, pro jejich zpracování a přenosu do dalších komponent BI řešení se tato data z DSA odstraní,

- V přesně stejné struktuře, v jaké jsou uložena ve zdrojových systémech

1.5.5 Operativní úložiště dat- ODS (Operational Data Store)

Operativní úložiště dat (ODS) je další komponentou datové vrstvy, kterou nemusíme nalézt ve všech řešeních Business Inteligence. Existují dva základní přístupy k definici ODS

První přístup definuje ODS jako jednotné místo datové integrace aktuálních dat z primárních systémů. Jedná se o zdroj pro sledování konsolidovaných agregovaných dat s minimální dobou odezvy po zpracování (tedy sledování v téměř reálném čase). V mnoha případech takové ODS slouží jako centrální databáze základních číselníků (zákaznický, produktový) nebo pro podporu interaktivní komunikace se zákazníkem (např. podporu pracovníků call-center, kdy ODS dodává aktuální konsolidovaná data o zákazníkovi, jeho profilu, použitých produktech apod.) Takto definované databáze podporují vkládání a modifikaci dat v reálném čase a jsou typicky napojeny na EAI platformy.

Druhý přístup k definici ODS vymezuje operativní úložiště dat jako databázi navrženou s cílem podporovat relativně jednoduché dotazy nad malým množstvím aktuálních analytických dat. Na rozdíl od prvního přístupu, podle tohoto vymezení vzniká ODS jako derivace již existujícího datového skladu a obsahuje pouze aktuální záznamy vybraného množství dat.

Stejně jako DSA, i ODS obsahuje data:

- Bez historie – pouze aktuální snímky,
- Měnící se po každém nahrání.

Oproti DSA však ODS (díky transformačním operacím) obsahuje data:

- Konsolidovaná
- Konzistentní
- Subjektově orientovaná
- A v určitých případech i doplněná o agregace

Nejmarkantnější rozdíl mezi DSA a ODS je v jejich použití. Zatímco DSA slouží pouze jako dočasné úložiště dat před jejich zpracováním v datovém skladu (přičemž po zpracování jsou data vymazána), ODS slouží jako databáze podporující analytický proces. Jinými slovy do DSA nemají přístup koncoví uživatelé, zatímco ODS je budována právě s cílem

zpřístupnit uživatelům nebo ostatním systémům data pro analýzy či dotazy s minimálním zpožděním oproti jejich pořízení. Typickým příkladem využití operativního úložiště dat je referenční databáze produktů nebo zákazníků. Tato referenční databáze slouží jako jednotný konsolidovaný zdroj příslušných dat pro všechny systémy nebo uživatele podniku.

1.5.6 Datový sklad – DWH (Data Warehouse)

Technologie datových skladů představuje v současné době jeden z nejvýznamnějších trendů v rozvoji podnikových informačních systémů. Datový sklad (DWH) lze definovat mnoha způsoby. Za základ však považuji definici jednoho ze zakladatelů Data Warehousingu, Billa Inmona:

Datový sklad (DWH) je integrovaný, subjektově orientovaný, stálý a časově rozlišený souhrn dat, uspořádaný pro podporu potřeb managementu.

Tyto pojmy lze pak interpretovat takto:

- **Subjektově orientovaný** – data jsou rozdělována podle jejich typu, ne podle aplikací, ve kterých vznikla. Jde tedy o případ, kdy jsou data o zaměstnanci uložena pouze jednou, a to v jednotné databázi datového skladu, kdežto v produkčním systému bývají rozptýlena do různých souborů podle toho, pro kterou aplikaci mají být použita.
- **Integrovaný** – data jsou ukládána v rámci celého podniku, a ne pouze v rámci jednotlivých oddělení.
- **Stálý** – datové sklady jsou koncipovány jako "Read Only", což znamená, že zde žádná data nevznikají ručním pořízením, a nelze je ani žádnými uživatelskými nástroji měnit. Data jsou do DWH načítána z operativních databází či jiných externích zdrojů a existují zde po celou dobu života datového skladu.
- **Časově rozlišený** – aby bylo možné provádět analýzy za určitá období, je nutné, aby byla do DWH uložena i historie dat. Načítaná data sebou tedy musí nést i informaci o dimenzi času.

1.5.7 Datové tržiště – DMA (Data Mart)

Princip datových tržišť je obdobný jako v případě datových skladů. Rozdíl je v tom, že datová tržiště – Data Marts, jsou určena pro omezený okruh uživatelů (oddělení, divize,

pobočka, závod apod.). Podstatou jsou tak decentralizované datové sklady, které se budou postupně integrovat do celopodnikového řešení. V některých případech slouží dále Data Marts, i po vytvoření celopodnikového datového skladu, jako mezistupeň při transformacích dat z produkčních databází.

Data Marts je tak problémově orientovaný datový sklad, určený pro pokrytí konkrétní problematiky daného okruhu uživatelů a umožňující flexibilní "ad hoc" analýzu. Výsledkem vytváření datových tržišť je zkrácení doby návratnosti investic, snížení nákladů a podstatné zmenšení rizika při jejich zavádění.

1.5.8 OLAP databáze

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

1.5.9 Reporting

Reportingem rozumíme činnosti spojené s dotazováním se do databází pomocí standardních rozhraní těchto databází (např. SQL příkazů v rámci relačních databází).

V rámci reportingu lze identifikovat:

- Standardní reporting, kdy jsou v určitých časových periodách spouštěny předpřipravené dotazy;
- Ad hoc reporting, kdy jsou databáze (většinou) jednorázově formulovány specifické dotazy, explicitně vytvořené uživatelem.

1.5.10 Manažerské aplikace – EIS (Executive Information Systéme)

Cílem EIS je podporovat manažerské procesy, jako jsou podnikové analýzy, plánování či rozhodování. Zatímco nástroje EIS dnes podporují vyšší, střední i nižší úroveň řízení, reporting slouží především na nižší úrovni, případně na střední, pro které vytváří pravidelné podpůrné dokumenty (výkazy, přehledy atd.) Technologický rozdíl mezi těmito nástroji je v tom, že zatímco nástroje reportingu přistupují přímo do operačních datových skladů nebo databází produkčních systémů, nástroje EIS vytvářejí vlastní multidimenzionální sémantickou vrstvu, prostřestřednictvím které uživatelé přistupují k analytickým datům.

Manažerské aplikace EIS jsou typem aplikací IS/ICT, které v sobě integrují všechny nejdůležitější datové zdroje systému, významné pro řízení organizace jako celku. S tím jsou spojeny i specifické nároky na prezentace informací a jejich zpřístupnění vedoucím pracovníkům firmy. EIS je tak především analytický a prezentační nástroj.

Aplikace typu EIS se objevily již v osmdesátých letech. Jsou definovány jako systémy, které poskytují manažerům (dnes cca na střední a vyšší úrovní řízení) on-line přístup k relevantním informacím v účinné a přehledné formě. Účinná a přehledná forma zobrazení pak znamená, že tyto systémy jsou již navrhovány s ohledem na jejich uživatele, kterými jsou časově vytížení manažeři, s mnohdy malými znalostmi v oblasti počítačů a ovládání aplikací. Právě tento specifický druh uživatelů odlišuje tyto systémy od jiných informačních systémů podniku.

Pro EIS jsou významná tato specifika:

- Jsou navrhovány speciálně pro poskytování "manažerských " informací, umožňují sledovat firemní procesy, plnění cílů organizace, poskytují přehled o celém podniku apod.
- Integrují širokou škálu interních i externích datových zdrojů a zajišťují výběr dat ze všech podstatných řídících úloh.
- Jsou schopné přistupovat ke konkrétním datům stejně tak, jako vytvářet data agregovaná.
- Poskytují nástroje pro on-line analýzy zahrnující především analýzy trendů, drill-up, drill-down, slice and dice a identifikaci výjimek.
- Jsou jednoduše ovladatelné (standardně myší či pomocí technologie touchscreen) a zajišťují vysokou vypovídající hodnotu výstupů prostřednictvím grafického uživatelského prostředí.
- Mohou být využívány přímo manažery bez nutnosti zprostředkování.

Prvotním cílem EIS byla podpora vrcholového managementu a jeho strategického rozhodování. Současné trendy však směřují k tomu, že se využívání těchto systémů stále více orientuje na střední úrovni řízení, a dále na různé specialisty.

1.5.11 Dolování dat (Data Mining)

Dolování dat umožňuje pomocí speciálních algoritmů automaticky objevovat v datech strategické informace. Je to analytická technika pevně spjatá s datovými sklady, jako s velmi kvalitním datovým zdrojem pro tyto speciální analýzy.

Dolování dat lze charakterizovat jako 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, nikolik předem specifikované uživatelem nebo implementátorem, a jedná se především o odvozování prediktivních informací, nikoliv pouze deskriptivních. Dolování dat slouží manažerům k objevování nových skutečností, čímž pomáhají zaměřit jejich pozornost na podstatné faktory podnikání, umožňují testovat hypotézy, odhalují ve stále se zrychlujícím a složitějším obchodním prostředí skryté korelace mezi ekonomickými proměnnými apod.

Existují různé druhy nástrojů pro dolování dat. Některé z nich jsou určeny specialistům se znalostmi statistiky, některé řídícím pracovníkům. Cílové určení úloh dolování dat je však podobné většině úloh Business Inteligence, mají tedy poskytovat strategické informace Širokém spektru manažerů v organizaci. To, co odlišuje dolování dat od jiných statistických nástrojů, je právě zaměření na odlišné uživatele. Statistické úlohy dolování dat jsou prováděny automaticky podle určených algoritmů, a tak jejich cílovým uživatelem může být i manažer bez speciálních znalostí statistiky, nikoliv pouze specialista, který návazně zhotovuje reporty pro manažera.

Dolování dat je založeno na množství matematických a statistických technik. Uvedeme zde jen příklady některých z nich.

- Rozhodovací stromy – prediktivní model, který zobrazuje data v podobě stromu, kde každý uzel určuje kritérium pro následné rozdělení dat do jednotlivých větví. Strom tak rozděluje veškerá zdrojová data do segmentů, kde každý list odpovídá určitéu segmentu definovanému předešlými uzly. Data, která jsou zařazena do určitého segmentu, se vyznačují shodnými vlastnostmi. Rozhodovací stromy mohou být založeny na množství algoritmů. Příklady některých z nich jsou: ID3, CART (Classifickation and Regresssion Trees) a CHAID (Chi-squared Automatic Interaction Detektor). Rozhodovací stromy jsou velmi častou technikou zejména díky své snadné interpretaci.

- Neuronové sítě – jsou nejčastěji využívány pro tvorbu prediktivních modelů. Jsou založeny na obdobných principech, které napodobují organizaci nebo způsob chování lidského mozku, založeném na systému neuronů. Existuje velké množství variací neuronových sítí, které aplikují různé algoritmy (včetně tzv. samoučících se algoritmů) pro nacházení podobností a vzorů a tvorbu prediktivních modelů z velkých databází.
- Genetické algoritmy – simulující biologickou evoluci pro dedikování, jak by měly být atributy formovány, vyvíjeny, modifikovány atd.
- Clustering a klasifikace – clustering je technika sloužící pro rozdělení dat do skupin s obdobnými charakteristikami, klasifikace definuje podstatné atributy skupin v podobě klasifikačních kritérií. Umožňují identifikovat a charakterizovat různé segmenty v datech.

Kromě pojmu Data Mining se lze setkat s dalšími pojmy označujícími víceméně zde definovaných koncept. Mezi tyto výrazy patří dobývání znalostí z databází (Knowledge Discovery in Databáze), Information Harvesting, Data Archeology nebo Data Distillery, Data Minging se začal komerčně nasazovat zejména v posledním desetiletí, kdy rozvoj vědy a technologií umožnil tento náročný proces.

1.5.12 Nástroje pro zajištění datové kvality

Nástroje pro zajištění datové kvality zažívají svůj prudký rozvoj s růstem nasazení analytických aplikací, zejména díky faktu, že pro úspěch nasazení řešení je, kromě již zmíněné funkcionální a technické znalosti, třeba korektní obsah. Vzhledem k povaze řešení – podpoře analytické práce – je důležité, aby tato práce probíhala nad korektními daty, dokumentující reálnou situaci podniku.

Nástroje pro zajištění datové kvality se proto zabývají zpracováním dat s cílem zajistit jejich:

- Úplnost – jsou identifikována a ošetřena data, která chybí nebo jsou nepoužitelná (z různých důvodů)
- Soulad – jsou identifikována a ošetřena data, která nejsou uložena ve standardním formátu

- Konzistenci – jsou identifikována a ošetřena data, jejichž hodnoty reprezentují konfliktní informace
- Přesnost – jsou identifikována a ošetřena data, která nejsou přesná nebo jsou zastaralá
- Unikátnost – jsou identifikovány a ošetřeny záznamy, které jsou duplicitní
- Integrita – jsou identifikována a ošetřena data, která postrádají důležité vztahy vůči ostatními datům.

Implementace datové kvality je jedním z horkým témat současnosti.

1.5.13 Další komponenty související s BI

Oborová znalost / know-how je součástí každého řešení Business Intelligence. Lze ji definovat jako znalost fungování prostředí, kde se BI řešení implementuje, kombinovanou se znalostí možností technologii BI a znalostí nejvhodnějších řešení, založených na technologii BI pro danou oblast.

Nástroje pro správu metadat odpovídají požadavkům, které získaly na důležitosti až s implementaci řešení BI. Metadata jsou definována jako data o datech, a v této souvislosti slouží pro dokumentaci konkrétních implementací informačních systémů podniku. Metadata jsou tedy popisem veškerých informačních systémů i jejich jednotlivých částí. Z pohledu řešení BI zahrnují zejména datové modely, popisy funkcí, obchodních a transformačních pravidel, reportů či podavků na reporty.

Dále vymezení BI není zcela jednoznačné. Existuje mnoho nástrojů, které s úlohami BI souvisí, mnohdy jsou do architektury BI i zařazeny. Mezi takovéto komponenty patří především Systémy pro podporu rozhodování **(DSS – Decision Support Systems) a Expertní systémy (ES – Expert Systems)**

DSS jsou systémy, které jsou určeny především pro manažery na nižších úrovních řízení, kterým poskytují informace pro jejich řídící práci a navrhují řešení rozpoznaných problémů na základě vytvořených modelů. Nebývají založené na multidimenzionálních datových modelech, a na rozdíl od většiny úloh BI, nevyužívají data pouze pro čtení, ale uživatelé si v nich vytvářejí vlastní rozhodovací modely.

ES jsou pak systémy, které se snaží simulovat řešení vzniklých problémů tak, jak by byly řešeny experty v daném oboru. Obsahují bázi znalostí, (jsou získány od expertů), na které

jsou uplatněna formální logická pravidla, aby je bylo možno využít v počítačových systémech.

2 DATA WAREHOUSING

Když se zamyslíme nad tím, co je to datový sklad, odpověď na tuto otázku je jednoduchá. Je to určitým způsobem strukturované úložiště údajů. Kdybychom ale hledali nějaké analogie mezi klasickým skladem a datovým skladem, tak to není tak jednoduché, jak to na první pohled vypadá.

V klasickém skladu skladujeme buď materiály, součástky a polotovary, které vstupují do výrobního procesu, nebo naopak skladujeme výrobky předtím, než se budou expedovat. Nikdo totiž nemá zájem skladovat dlouhou dobu polotovary a už vůbec ne hotové výrobky. Čím rychleji je dokážeme vyexpedovat a prodat, tím lépe pro ekonomiku firmy. V datovém skladu naproti tomu chceme shromažďovat a uchovávat informační bohatství firmy za co nejdelší období. Spíše než ke klasickým skladům můžeme datové sklady přirovnat k depozitářům muzeí. I v tomto případě se snažíme shromažďovat exponáty, třídit je časově, geograficky, podle druhů a podobně.

2.1 Datový sklad

Nejznámější definice datové skladu pochází od Billa Inmona.

Datový sklad je podnikově strukturovaný depozitář subjektově orientovaných, integrovaných, časově proměnlivých, historických dat použitelných na získávání informací a podporu rozhodování. V datovém skladu jsou uložena atomická a sumární data.

Údaje se získávají a ukládají do produkčních (operačních) databází, které mohou být v různých odděleních firem nebo dokonce v rozličných geografických lokalitách. Tyto údajev pravidelných intervalech sesbíráme, předzpracujeme a zavedeme od datového skladu. Datový sklad je v podstatě též databáze, jen je organizovaná podle trochu jiných pravidel, tabulky například nemusí být normalizovány a podobně. Datový sklad je tedy soubor technologií pro efektivní skladování údajů, tak aby tyto údaje po jejich přeměně na informace sloužily podpoře rozhodování.

Definice Billa Inmona je velmi stručná a výstižná. Pravděpodobně ale bude potřebné si tuto definici přečíst několikrát a zamyslet se nad jednotlivými pojmy, které definici tvoří.

- **subjektová orientace:** Údaje se do datového skladu zapisují spíše podle předmětu zájmu než podle aplikace, ve které byly vytvořeny. Při orientaci na subjekt jsou da-

ta v datovém skaldu kategorizována podle subjektu, kterým může být např. zákazník, dodavatel, zaměstnanec, výrobek a podobně. Orientace na aplikaci naproti tomu znamená, že údaje jsou v systému uloženy podle jednotlivých aplikací, například údaje aplikace pro odbyt, údaje aplikace pro fakturaci, údaje aplikace pro personalistiku.

- **integrovanost:** Datový sklad musí být jednotný a integrovaný. To znamená, že údaje týkající se konkrétního předmětu se do datového skladu ukládají jen jednou. Proto musíme zavést jednotnou terminologii, jednotné a konzistentní jednotky veličin. Není to jednoduchá úloha, protože údaje přicházejí do datového skladu z nekonzistentního a neintegrovaného operačního prostředí. Proto musí být údaje v etapě přípravy a zavedení upraveny, vyčištěny a sjednoceny. Když údaje nejsou konzistentní a důvěryhodné, datový skald ztrácí význam.
- **časová variabilita:** Údaje se ukládají do datového skladu jako série snímků, z nichž každý reprezentuje určitý časový úsek. Na rozdíl od operačního prostředí, kde jsou údaje platné v okamžiku přístupu, v datových skladech jsou údaje platné pro určitý časový moment, časový snímek. Zatímco v operačním databázovém prostředí se údaje ukládají za kratší časové období, nejčastěji dní, maximálně měsíců, v datovém skladě jsou údaje za delší časové období, typicky několik roků. Klíčové atributy v datovém skladě obsahují čas, který v operačních databázích nemusí být uváděn. Jakmile je v datovém skladu zaznamenán konkrétní snímek dat z operativní databáze, nemohou být už tyto údaje v datovém skladě modifikovány.
- **neměnnost:** V operačních transakčních databázích jsou údaje do databáze jednak vkládány, jednak modifikovány, a i mazány. Údaje v datovém skladě se obvykle nemění ani neodstraňují, jen se v pravidelných intervalech přidávají nové údaje. Proto je manipulace s údaji mnohem jednodušší v datových skladech. V zásadě můžeme připustit jen dva typy operací. Zavedení údajů do datového skladu a přístup k těmto údajům. Žádné změny údajů nesou přípustné. Z toho vyplývá, že většina metod pro optimalizaci a normalizaci údajů a transakční přístup k údajům je v datovém skladě nepotřebná.

Údaje se získávají a ukládají do produkčních (operačních) databází, které mohou být v různých odděleních firem, nebo dokonce v rozličných geografických lokalitách. Tyto údaje v pravidelných intervalech sebereme, předzpracujeme a zavedeme do datového skla-

du. Datový sklad je v podstatě též databáze, jen je organizovaná podle vzpomínaných pravidel.

Tabulka 1 Rozdíly mezi daty v produkčních databázích a daty v datovém skladu

	Produkční databáze	**Datový sklad**
Funkce	Zpracování dat, podpora podnikových operací	Podpora rozhodování
Data	Procesně orientovaná, aktuální hodnoty, detailní	Předmětově orientovaná, aktuální i historická, sumarizovaná, zřídka detailní
Užití	Strukturované, opakované	Ad hoc, částečně opakující se reporty a strukturované aplikace
Procesy	Vstup dat, dávky, OLTP	Dotazy koncových uživatelů, OLAP

V datovém skladu můžeme vykonávat různé analýzy pro potřeby rozhodování manažerů, obchodníků. Nástroje pro budování a provoz datových skladů však představují poměrně velkou počáteční investici za hardware a software, takže datové sklady využívají většinou banky, pojišťovny, mobilní a telekomunikační operátoři, velké obchodní řetězce a podobně. K údajům v datovém skladě a výsledkům analýz nad těmito údaji mohou mít prostřednictvím webu přístup manažeři firmy po celém světě a obchodní partneři na základě přidělených přístupových práv. Informace jsou totiž zbožím.

Tím, že údaje z datového skladu můžeme poskytovat i svým obchodním partnerům, se nám může postupně část vynaložených nákladů na datový sklad vrátit. Například dodavatelé velkého obchodního domu určitě ocení informace o tom, jaké zboží jde nejvíce na odbyt, případně o jaké zboží bude pravděpodobně zájem v nejbližších týdnech. Mýtus o tom, že datové skaldy jako oblast informatiky je záležitost složitá a náročná na čas i finance, by se zdánlivě se zdánlivě mohl prolomit jednak po přečtení a pochopení definice a jednak při pohledu na principiální schéma datového skladu. Problém je formulovatelný do jednoho krátkého odstavce. Stačí získat údaje z operačního prostředí a po úpravě je zavést do datového skladu, nad těmito údaji vykonat analýzy a jejich výsledky zpřístupnit uživatelům, tedy manažerům a analytikům.

A ještě dodáme, že datový sklad v malém je možné simulovat i na trochu lépe vybaveném desktopovém počítači. Co je potom na datových skladech tak náročné a složité?

2.1.1 Návrh a koncepce

Asi největší procento z hodnoty datového skladu se skrývá v jeho návrhu a koncepci. Know-how a to nejen z oblasti datových skladů, ale hlavně z oblasti podnikání naší firmy, je zkrátka to nejcennější, co ve firmě máme, když samozřejmě nepočítáme firemní značku zavedené firmy. V tomto případě si kvalitu na rozdíl od hardwaru a softwaru nemůžeme koupit, kvalitu musíme vytvořit, případně ji pro nás vytvoří dobře placený vývojový tým.

2.1.2 Hardware a software

Co se týká nákladů, ani hardware se nedá zahanbit, ale z hlediska filosofie datového skladu se jedná jen o technické prostředky, které jsou nahraditelné. Pozor, technické prostředky, ne údaje, které jsou v nich uloženy. A tak musíme uvážit, kam se v přímé úměře výkonu a spolehlivosti na jedné straně a ceny na druhé straně chceme pohybovat. Samozřejmě pro rychlý přístup k obrovskému množství údajů potřebujeme výkonné servery nebo serverové farmy.

Nástroje pro vytváření datových skaldů a analýzy údajů jsou také velmi drahou záležitostí. Stále více se prosazuje trend integrace analytických sužeb přímo do instalací databázových serverů, přičemž v některých případech se za analytické služby platí licenční poplatky, v některých případech je to zahrnuto v ceně databázového serveru.

2.2 Datové trhy

Datové trhy (anglicky datamart) jsou určité přesně specifikované podmnožiny datového skladu, které jsou určeny pro menší organizační složky firmy. Datový sklad je totiž z hlediska investic i objemu prací velmi náročný projekt. Proto se v některých případech přistupuje k budování datového skladu po částech, což znamená, že pro některé důležité organizační složky se vytvořily jakési podmnožiny datového skladu – datové trhy. Kromě ekonomického efektu má tento postup i psychologický efekt, protože fungující podmnožina datového skladu prohlubuje důvěru v úspěšnost a potřebnost datového skladu jako celku.

Datové trhy mohou vzniknout i opačným postupem, to jest nejdříve se vytvoří centrální integrovaný datový skald a z něho se potom vytvoří několik datových trhů. Toto řešení je flexibilnější a klade menší nároky na provoz a údržbu. Datové trhy tedy mohou existovat jako subsystémy datových skladů nebo i samostatně jako jednoduchý datový sklad.

2.3 Metody budování datového skladu

Pravděpodobně nejdůležitějším krokem při budování datového skladu je výběr nejvhodnější metody. Musíme brát do úvahy nejen organizační strukturu a informační „kulturu“ firmy, ale předvídat i možné potíže, které se během budování datového skladu nevyhnutelně objeví. Nejznámější a nejčastěji používané metody jsou:

- Metoda „velkého třesku“
- Přírůstková metoda

2.3.1 Metoda „velkého třesku“

Mnohé firmy a vývojáři a možná i někteří konzultanti se domnívají, že je možné realizovat implementaci datového skladu pomocí jediného projektu. Ale vývoj datového skladu je náročná záležitost a pravděpodobně se ji nepodaří vyřešit najednou a už vůbec ne v rozumném čase. To je ale největší slabina, protože kdybychom i nakonec projekt datového skladu metodou velkého třesku zrealizovali, mezitím se mohou změnit nejen technologie, ale i požadavky uživatelů. Metoda „velkého třesku“ se skládá z tří etap:

- Analýza požadavků podniku
- Vytvoření podnikového datového skladu
- Vytvoření přístupu buď přímo, nebo přes datové trhy

Jedinou výhodou metody „velkého třesku“ je skutečnost, že můžeme celý projekt kompletně vypracovat ještě před začátkem jeho realizace. Budování datového skladu je dynamický proces, při kterém je skoro jisté, že se změní nejen technologie, ale i požadavky uživatelů, není možné tento fakt považovat za skutečnou výhodu. Takže převažují spíše nevýhody, které se dají vyjmenovat velmi stručně a jsou velmi závažné. Je tu jednak velké riziko změny požadavků a hlavně trvá velmi dlouhý čas, než se projeví první výsledky obrovských investic do datového skladu, jinak řečeno, než se dostaví „hmatatelný“ obchodní zisk.

2.3.2 Přírůstková metoda

Přírůstková metoda, jinak nazvaná i evoluční předpokládá budování datového skladu po jednotlivých etapách, tedy místo vybudování celého datového skladu postupně přibývají přírůstková řešení, která samozřejmě zapadají do celkové architektury datového skladu. Začneme tedy budováním několika málo předmětných oblastí, typicky jednou nebo dvěma. Toto částečné řešení implementujeme například jako škálovatelný datový trh a poskytneme ho koncovým uživatelům. Tím se částečně uspokojí „hlad“ managementu po návratnosti investicí, když první subsystémy začnou fungovat a přinášet výhody v krátkém čase po zahájení projektu. Když se částečné řešení, navíc důsledně otestované skutečnými uživateli osvědčí, můžeme do systému přidat další předmětnou oblast, čímž získáme novou funkcionalitu. A takto bychom mohli pokračovat až do úplného vytvoření datového skladu.

Budování datového skladu přírůstkovou metodou je tedy iterativní proces, který udržuje neustálou spojitost mezi datovým skladem a potřebami uživatelů. U přírůstkové metody na rozdíl od metody „velkého třesku“ převažují výhody nad nevýhodami. K hlavním výhodám přírůstkové metody patří:

- Přírůstkové budování datového skladu zachovává kontinuitu budovaného projektu s požadavky a potřebami uživatelů.
- Umožňuje implementovat škálovatelnou, tedy rozšiřitelnou architekturu.
- Zabezpečí rychlejší zisk a tedy i rychlejší návratnost investic.

Když se rozhodneme pro budování datových skladů přírůstkovou metodou, můžeme si vybrat jednu z dvou variant

- Přírůstková metoda směrem „shora dolů“.
- Přírůstková metoda směrem „zdola nahoru“.

2.3.2.1 Přírůstková metoda směrem „shora dolů“

U této metody nejdříve na základě požadavků uživatelů vytvoříme konceptuální model datového skladu, přičemž stanovíme hierarchii předmětných oblastí. Následně sestavíme konceptuální modely jednotlivých předmětných oblastí. Jinými slovy řečeno, postupně vytváříme datové trhy jednotlivých předmětných oblastí v rámci struktury datového skladu.

Tato metoda poskytuje poměrně rychlou implementaci jednotlivých datových trhů a tím i návratnost investic. V porovnání s metodou „velkého třesku“ přírůstková metoda „shora dolů“ je zatížena podstatně menším rizikem, neboť není tak náročná na analýzu. Mezi hlavní nevýhody přírůstkové metody směrem „shora dolů“ patří zvýšené vstupní náklady dříve, než je možné předvídat návratnost investic.

2.3.2.2 Přírůstková metoda směrem „zdola nahoru“

Tato přírůstková metoda je velmi podobná metodě „shora dolů“, ale prioritu mají údaje před obchodním ziskem. Jak je zřejmé ze schématu, budujeme nejdříve datové trhy předmětných oblastí v rámci struktury datového skladu.

U této metody vystupuje do popředí oddělení IT podniku. Co se týká výhod a nevýhod, převažují nevýhody. Nakolik se konceptuální modle odvíjí od zdrojových systémů, je celková rozšiřitelnost v některých případech značně problematická. Oddělení IT se v mnohých podnicích nepovažuje za „lídra“ v oblasti strategie a marketingu, proto se o mnohých připravovaných změnách a strategických záměrech dovídá oddělení IT zpravidla jako poslední. Proto mohou navrhnout a v některých případech i zrealizovat něco, co je vzhledem ke strategickým záměrům podniku už neaktuální. A navíc oddělení IT je zvyklé pracovat spíše s údaji než s informacemi, proto úloha lídra pro oddělení IT není vždy šťastné řešení.

2.3.2.3 Fáze přírůstkové metody

Když si prohlédneme schéma přírůstkové metody, tak zjistíme, že se skládá z následujících kroků:

- Strategie
- Definice
- Analýza
- Návrh
- Sestavení
- Produkce

2.4 Příprava údajů – etapa ETL

S nasazením technologie business intelligence a data warehouse (BI/DW) prakticky nikdy nezačínáme, jak se říká, „na zelené louce“. Obvykle se před zavedením těchto technologií pro procesy rozhodování používají údaje získávané z primárních transakčních systémů OLTP (Online Transaction Processing), v lepším případě zpracované do sestav. Tyto sestavy sou potom (zpravidla ručně nebo pomocí softwaru typu MS Office) zpracovávány do manažerských podkladů pro účely rozhodování.

Údaje pro proces business intelligence data warehouse tedy pocházejí z různých nehomogenních zdrojů. Mohou to být údaje ze souborových databází (Access, dBase...), údaje z databází spravovaných některým databázovým serverem (Oracle, Informix, Microsoft, SQL, Server, Sybase, Interbase, Ingres...). Mohou to být údaje vyexportované nějakou databázovou platformou do tzv. flat file a podobně. Příprava a zavedení údajů je důležitou součástí každého řešení datového skladu vyextrahovat, vyčistit, upravit a až následně ve vhodné formě do datového skladu zavést. *(obrázek – Datový sklad)*

2.4.1 Extrakce, transformace a zavedení

Nástroje a postupy ETL (Extraction, Transformation, Loading), případně vyjádřené jinou terminologií ETT (Extraction, Transformation, Transport), jsou velmi důležitou součástí každého projektu datového skladu. Celý proces ETL je komplexní a ve většině případů časově poměrně náročný. U některých implementací může zabrat i více než polovinu celkového času, úsilí a tedy i velkou část nákladů potřebných na vytvoření datového skladu.

Když se blíže podíváme na jednotlivé etapy procesu ETL:

- Extrakce – výběr dat prostřednictvím různých metod,
- Transformace – ověření, čištění, integrování a časové označení dat,
- Loading – přemístění dat do datového skladu.

zjistíme, že hlavním cílem etapy ETL je centralizace údajů, tzn. jejich shromáždění z mnoha zpravidla nehomogenních a různorodých zdrojů z databází OLTP, a naplnění datového skladu určenými údaji v požadovaném čase. Tyto dva ekvivalentní pojmy (ETT nebo ETL) popisují řadu procesů, jejichž úlohou je extrakce údajů ze zdrojových systémů, jejich transformace a vyčištění a přenos údajů do datového skladu. Údaje se tedy v této etapě nejen přenášejí, ale i zpracovávají, například indexují, sumarizují, zjišťují se případ-

né změny struktur zdrojových údajů potřebných pro datový sklad, podle potřeby se mění struktura klíčů a udržují se metadata, tedy data o datech, v tomto případě předpisů a definicí pro přenos a zpravování údajů.

Počátečním naplněním datového skladu údaji z operačních databází úloha ETL samozřejmě nekončí, datový sklad se přece v pravidelných intervalech plní aktualizovanými daty. Je potřebné si uvědomit, že na úrovni ETL se pracuje s údaji, z kterých se později stanou informace. I tak by ale údaje zaváděné do datového skladu měly být kvalitní, přesné, k věci a aktuální, a tedy užitečné pro uživatele. Kromě kvality údajů je samozřejmě důležitá i jejich dostupnost, aby mohli jednotliví uživatelé datového skladu, například manažeři, obchodníci a analytici, používat datový sklad účinně a efektivně.

2.4.2 Oblast vynášení údajů

Ještě předtím, než se budeme podrobněji věnovat jednotlivým etapám ETL, je potřeba potřebné upozornit na jakési „staveniště“ datového skladu, kterým je oblast pro vynášení údajů. Z hlediska implementace se může jednat například o paměť operačních dat, adresář textových nebo flat souborů, tabulky v relační databázi nebo vlastní struktury údajů, které používají nástroje určené pro vynášení dat. Z hlediska principu můžeme použít dva modely vynášení:

- Model lokálního vynášení.
- Model vzdáleného vynášení.

Výběr modelu závisí na uživatelských požadavcích a samozřejmě na objemu údajů a kvalitě a rychlosti připojení.

Model lokálního vynášení

Procesy úpravy a transformace údajů se v tomto případě vykonávají nejdříve, a to lokálně v operačním prostředí a až potom se přenášejí do vynášecí oblasti. Tento způsob může značně zatěžovat operační paměť počítače, takže je ho možné využívat jen u méně zatížených systémů.

Model vzdáleného vynášení

V tomto případě se „surová“ data nejdříve přenesou operačního prostředí do vynášecí oblasti, nebo dokonce přímo do prostředí datového skladu a až tady se zpracují.

2.4.3 Extrakce

Údaje, které chceme přenést do datového skladu, jsou jednak umístěny v různých nehomogenních operačních prostředích, hardwarových platformách (PC, mainframe, iMac...), operačních systémech (Windows, Unix, Linux, Sun, Solaris...), databázových systémech (MS SQL Server, Oracle, Informix, IBM DB2...), a co je ještě horší, mohou se vyskytovat v rozličných formách. Úlohou extrakce je právě získat údaje právě z takových zdrojů.

Kapitolu samu pro sebe tvoří archivní data, která obsahují historické údaje. Hlavní rozdíl mezi datovým skladem a archivem je v tom, že údaje v datovém skladu se pravidelně obnovují. Údaje z archivů jsou nezastupitelným zdrojem historických dat při prvním naplnění datového skladu. Naproti tomu archivní data nepoužíváme pro obnovu údajů v datovém skladu.

Kromě interních údajů z našeho podnikatelského prostředí je někdy potřebné pracovat i s externími údaji. Tyto údaje můžeme získat analýzou konkurenčního prostředí, zakoupením údajů o zákaznících, nebo i stáhnutím údajů volně přístupných na internetu. Z toho vyplývá, že tady nemůžeme periodicky odebírat vzorky, jako jsme byli zvyklí u interních údajů. Externí údaje proto vyžadují nepřetržité monitorování za účelem určení, kdy jsou dostupné.

Pro extrakci jsou k dispozici různé postupy, nástroje a technologie. Můžeme vytvářet vlastní aplikace ve vyšších procedurálních programovacích jazycích, C++, C# nebo v procedurálních nadstavbách jazyka SQL (T-SQL, PL/SQL...). Pro menší množství údajů je výhodné vytvořit přístupovou bránu (gateway). Tato metoda ale pro větší objemy údajů zatěžuje síť. Někdy můžeme použít výstupy z vlastních podnikových systémů, které umožňují konverzi a vyčištění údajů. Při správně navrhnuté etapě ETL máme k dispozici metadata pro všechny váze této etapy. Tato metadata obsahují informace o místě, typu, přístupových privilegiích a struktuře zdroje údajů.

2.4.4 Transformace

Říká se, že je lepíš nemít žádná data, než mít data nekvalitní. Když jsou data v datovém skladu nekvalitní, snižuje to důvěru v taková řešení a datový sklad se oprávněně nevyužívá. Jádro problému je v tom, že nekvalitní data se dost často vyskytují ve zdrojových systémech. Použití nekvalitních údajů vede k chybným nebo minimálně nepřesným sestavám a následně k chybným obchodním rozhodnutím.

Čištění údajů

Údaje z externích zdrojů mají určitou kvalitu, která je buď postačující, nebo nepostačující pro jejich zavedení do datového skladu. Často dokonce bývá kvalita údajů značně proměnlivá, například, když pracovník vyplňoval údaje po „veselé“ noci a podobně. Pro čištění údajů je v anglické terminologii vžito několik rovnocenných termínů, například cleaning (čištění), scrubing (vyčištění), a cleansing (pročištění).

Čištění údajů může být někdy velmi náročné, a tedy i nákladné. V některých případech dokonce ani nemá smysl čistit údaje s vysokými náklady, když je přínos pro podnikání zanedbatelný. Dokonce když i systémy OLTP obsahují kvalitní údaje, tyto údaje nemusí být zárukou kvalitního datového skladu. Systémy OLTP totiž neobsahují historické údaje.

Transformace

Transformace samotná je soubor úloh a úkonů, které vedou ke zvýšení kvality údajů, hlavně k odstranění anomálií. Anomálie nejsou v systémech OLTP zpravidla na závadu, když to řekneme trochu ironicky, tak se tyto anomálie v systémech OLTP roky budují. Vývoj některých systémů trvá poměrně dlouho, obměňují se verze softwaru, mění se vývojová prostředí a technologické platformy, na kterých se tento software vyvíjí. Mění se operační systémy.

Jako příklad můžeme uvést přechod z operačního systému MS DOS na Windows. V DOSu se zdála být pozice kódové stránky pro češtinu a slovenštinu od bratrů Kamenických neotřesitelná. V ní se ukládaly skoro všechny textové údaje, jména, adresy a podobně. kdo z uživatelů Windows si ale dnes vzpomene, co to ta „stránka Kamenických“ vlastně byla? Snad jen pokud bude postaven před úlohu zpracovávat textové údaje, kde byla tato kódová stránka použita.

Do hry ale kromě technických záležitostí vstupuje i lidský faktor, například pravopisné chyby, pořadí zadávání atd. Během čištění dat, tedy sjednocuje formátování údajů, sjednocuje přiřazení datových typů, jednotek míry a peněžních měn. Údaje v databázích OLTP často obsahují různě kódované údaje nebo i primární klíče, které se skládají z více částí. Například kód výrobku, ze kterého se dá vyextrahovat například datum výroby, číslo závodu, ve kterém byl výrobek vyroben a podobně. Při zavádění těchto údajů do datových skladů je potřebné rozložit tyto údaje na atomické hodnoty.

Nejednoznačnost údajů

Jedním problémů, které musíme při transformaci údajů řešit, je i nejednoznačnost údajů. Například údaje o pohlaví zákazníka mohou být uloženy různým způsobem.

Chybějící hodnoty a duplicitní záznamy

Starosti nám mohou způsobovat i chybějící hodnoty (sloupce relačních databází obsahující hodnotu NULL), případně otevřená nebo skrytá duplicita záznamů. Duplicita údajů je menší problém, když je něco navíc, tak se to dá vždy odstranit. V některých případech to ale může být dost časově náročné. Větší problém představují chybějící údaje. V takovém případě máme více možností. U malého objemu chybějících údajů je můžeme ignorovat. Někdy můžeme chybějící hodnoty doplnit z jiných zdrojů. Když nemáme jinou možnost, můžeme s vhodným příznakem dočasně ponechat v systému OLTP a zapracovat později.

Konvence názvů pojmů objektů

Když slučujeme údaje z různých zdrojů, které v podstatě popisují stejný jev, ale mají jednotlivé entity vedeny pod různými názvy, tak musíme sloučit terminologii a vytvořit jednotnou konvenci názvů.

Různé peněžní měny

Kapitolou samo o sobě jsou hodnoty měny. Suma 200,50 znamená něco úplně jiného v dánských korunách než ve forintech. Tato problematika je velmi aktuální například při přechodu na euro. V přechodném období se uváděly ceny v místních měnách i v EUR.

Formáty čísel a textových řetězců

Údaje jsou v relačních databázích a souborech uloženy v různých druzích formátů. Největší problém je s ukládáním číselných údajů. Nejčastěji se pro tyto údaje používají numerické a řetězcové datové typy. Do numerického datového formátu se ukládá číslo jako numerická hodnota, do řetězcového datového typu se číslo ukládá jako postupnost číslic a jiných znaků, například desetinných čárek a mezer.

V čem je tedy problém? Když máme například rodné číslo ve tvaru 6808214321, můžeme ho v této podobě uložit jako číslo, ale i jako textový řetězec. V častěji uváděné podobě 680821/4321 můžeme toto rodné číslo uložit jen jako textový řetězec. Podobně je to i s poštovními směrovacími čísly. V podobě 12345 ho můžeme uložit i jako číslo, i jako textový řetězec. Ve tvaru 123 45 ho můžeme uložit jen jako textový řetězec. Numerické formáty se pro ukládání PSČ nepoužívají i z jiného důvodu. Mnoho z nich totiž začíná

nulou a tak se pětimístné PSČ03483 změní na čtyřmístné 3483, protože nuly před první platnou číslicí se v numerických formátech vynechávají.

Referenční integrita

Kromě hodnot jsou v údajích skryty i různé vtahy, například master – detail, organizační struktura firmy, hierarchická struktura zaměstnanců a podobně. Ale údaje jsou dynamické, organizační struktura se mění, často bez dokumentace a adekvátních změn v databázích OLTP. Když se zruší nějaké oddělení a zůstanou po něm nějaké záznamy, mohou tito „údajoví sirotkové“ zkreslit údaje a tedy nepříznivě ovlivnit kvalitu údajů.

Chybějící datum

Čas plní v datovém skladě významnou úlohu. Od něj se vše odvíjí a skoro každá analytická databáze má časovou dimenzi. V mnohých transakčních systémech se údaje neoznačují časovými údaji, v jiných je naopak čas důležitou veličinou. Například datum objednávky, transakce a podobně. Časový údaj musí být přítomný v datech před jejich zavedením do datového skaldu, nebo se musí určit a přidat při zavádění dat. Je potřeba dobře uvážit, kdy a kde se transformace uskuteční. Můžeme jí vykonávat sériově nebo paralelně se zaváděním údajů. U sériového způsobu se transformace vyková před zavedením dat do datového skladu. U paralelní metody se tento proces vykonává souběžně se zaváděním.

2.4.5 Přenos

Završením etapy ETL je přenos údajů z paměti zdrojových dat nebo přechodné vynášecí oblasti do datového skladu. Přenos spočívá v přesunu údajů a jejich uložení do databázových tabulek. Přenos spočívá v přesunu údajů a jejich uložení do databázových tabulek. Přenos by měl být plánovaný a automatizovaný. Při prvotním naplnění datového skladu může jít o obrovské množství údajů. Potom se už údaje zavádějí v pravidelných časových obdobích, například každý den, a to v takových objemech, kolik údajů za dané období v databázích OLTP vznikne.

Zavádění dat by mělo být plánované a hierarchizované. Stejně by mělo být automatizované na nejvyšší možnou míru. Po zavedení údajů zpravidla probíhá jejich indexování, aby byl přístup k nim optimalizovaný. Pro jednoznačnou identifikaci údajů se používají i uměle vytvářené klíče, s jejich pomocí zajistíme jednoznačnost každého řádku v tabulce. Data datového skladu jsou totiž často kombinací mnoha transformovaných záznamů, které nemají žádné přirozené klíče, které by se daly použít pro jednoznačnou identifikaci.

2.4.6 Chyby a problémy etapy ETL

Proces ETL vždy neproběhne úspěšně. Problémy mohou být se spolehlivostí úložiště údajů (disky jsou mechanická zařízení, která se opotřebovávají), může dojí k výpadkům spojení, zdroje údajů se mohou měnit, například při upgradu systému OLTP, které se nedokumentují v metadatech. Důležité je ověření údajů, protože když údaje nejsou ověřeny, tak může dojít k problémům při extrakci a transformaci. Podle závažnosti selhání je potřebné začít nanovo nebo můžeme pokračovat od místa selhání. Nepřesné nebo neúplné údaje mohou být příčinou nepřesnosti výsledků analýzy, což následně může vést k nesprávným obchodním, anebo ještě hůře k strategickým rozhodnutím.

2.4.7 Testování etapy ETL

U etapy ETL se asi nejvíce projeví pravdivost trochu ironické věty: „nikdy není čas udělat něco pořádně, ale vždy je dost času udělat to potom ještě jednou“. Proto je třeba etapu ETL důkladně otestovat nejdříve na simulovaných a později i na ostrých údajích. V etapě testování se už poprvé projeví i to, zda jsme etapu ETL dobře a podrobně zdokumentovali.

Ani po otestování a plném nasazení ETL nemáme jistotu, že vše bude stále fungovat k naší spokojenosti. Entropie (zjednodušeně řečeno, že věci ponechané samy sobě spějí od deseti k pěti) je totiž neúprosná. Objem datového skladu roste rychle a metrika zavádění a granularity dat vyžaduje pravidelnou revizi. Na vykonání procesů ETL je možné použít jedna specializovaná nástroje ať už od externích dodavatelů, nebo vyvinuté pro konkrétní projekt, různé brány a datové pumpy mezi databázovými systémy a interně vyvinuté nebo dodavatelské nástroje.

3 ANALÝZA OLAP

V předcházejících kapitolách jsme už stručně naznačili základní rozdíly mezi systémy typu OLTP a OLAP. Pochopení teoretického rozdílu není pro aplikace typu BI/DW až tak důležité. O mnoho důležitější je správně navrhnout strukturu databáze pro danou oblast použití. Analýza OLAP slouží ke zpracování údajů uložených v datovém skladu do podoby pro koncové uživatele, tedy manažery a analytiky.

K této kapitole je možné přistupovat dvěma způsoby, buď si nejdříve prostudovat (nebo zopakovat) teoretický úvod do problematiky OLAP, nebo si nejdříve pročíst část o OLAPu pro začátečníky.

3.1 Teoretický úvod do problematiky OLAP

Termín OLAP zavedl Dr. E. F. Codd na popsání technologie, která by pomohla překlenout mezery mezi využitím osobních počítačů a řízením podnikových dat. Pro OLAP existuje více různých definic, například: OLAP je volně definovaný řád principů, které poskytují dimenzionální rámec pro podporu rozhodování.

Pojem OLAP se poměrně často zaměňuje s jiným pojmem DSS (Desicion Support Systems) – systémy na podporu rozhodování. Tyto systémy umožňují pracovníkům přijímajícím rozhodnutí přístup k údajům potřebným na „tvorbu“ takových rozhodnutí.

3.1.1 Fakta a dimenze

Každá krychle OLAP byla vytvořena na základě dvou druhů údajů: faktů a dimenzí.

Fakta jsou numerické měrné jednotky obchodování. Tabulka faktů je pochopitelně největší tabulka v databázi a obsahuje velký objem dat. Jak se dozvíme později, tabulky faktů a dimenzí mohou vytvářet určitá schémata, například hvězdicové schéma (star schema) nebo schéma sněhové vločky (snowflake schema). Hvězdicové schéma obyčejně obsahuje jen jednu tabulku faktů, jiné, hlavně schémata DSS, mohou obsahovat i více tabulek faktů. Prvotní fakta, například objem prodeje, se mohou kombinovat nebo vypočítat pomocí jiných faktů a vytvořit tak měrné jednotky. Měrné jednotky se mohou uložit v tabulce faktů, případně vyvolat, když je to nevyhnutelné, na účely vykazování.

Dimenze obsahují logicky nebo organizačně hierarchicky uspořádané údaje. Jsou to vlastně textové popisy obchodování. Tabulky dimenzí jsou obyčejně menší než tabulky faktů a

data v nich se nemění tak často. Tabulky dimenzí vysvětlují všechna „proč“ a „jak“, pokud jde o obchodování a transakce prvků. Dimenze obecně obsahují relativně stabilní data, dimenze zákazníků se aktualizují častěji. Jak už bylo vzpomenuto, velmi často se používají časové, produktové a geografické dimenze. Tabulky dimenzí obvykle obsahují stromovou strukturu. Například dimenze vytvořená na základě geografických informací, tedy regionální dimenze se člení na jednotlivé úrovně podle konkrétního územně-správní členění dané geografické oblasti.

3.1.2 Úložiště multidimenzionálních údajů MOLAP, ROLAP, HOLAP, DOLAP

Pojednání o jednotlivých druzích multidimenzionálních modelů databází začneme zdůrazněním rozdílů mezi relačními a multidimenzionálními (OLAP) databázovými modely.

Relačně databázový model. Údaje jsou uloženy v dvoudimezionálních tabulkách. Každý řádek v tabulce obsahuje data, která jsou zpravidla obrazem reálného světa, tedy data, která se vztahují k nějaké věci nebo k její části. Sloupce dvoudimenzionálních databázových tabulek obsahují údaje týkající se atributů.

Multidimenzionální databázový model. Datový model multidimenzionální databáze je možné zobrazit jako vícerozměrnou krychli. Tato krychle je vlastně ekvivalent tabulky v relační databázi. Každá krychle má několik dimenzí (ekvivalent indexových polí v relačních tabulkách). Nejlépe si dokážeme představit klasickou trojrozměrnou krychli, ale počet dimenzí v reálných multidimenzionálních databázích je zpravidla větší. Znalci programování ve vyšších programovacích jazycích určitě namítnou, že multidimenzionální krychle není nic jiného než vícerozměrné pole. Prostor pro celou krychli je už dopředu rozvržen. Jednotlivé záznamy se v multidimenzionálních krychlích nacházejí na průsečících dimenzí. Například objem prodeje chladniček (produktová dimenze) za první kvartál roku 2001 (časová dimenze) v Severomoravském kraji (geografická nebo zákaznická dimenze).

S rostoucím počtem rozměrů multidimenzionální databáze velmi rychle rostou i požadavky na úložnou kapacitu. Ale ne na všech průsečících dimenzí se vždy nacházejí údaje. Takovou krychli nazýváme i řídkou krychlí. V praxi se u multidimenzionálních databází používají různé technologie na kompresi objemu použitého diskového prostoru.

3.1.2.1 Multidimenzionální OLAP (MOLAP)

Pro multidimenzionální online analytické zpracování (MOLAP) se získávají data buď z datového skladu, nebo z operačních zdrojů. Mechanismus MOLAP potom uloží analytická data ve vlastních datových strukturách a sumářích. Během tohoto procesu se spočítá tolik předběžných výsledků, kolik je z technického a časového hlediska možné. Údaje v úložišti typu MOLAP se tedy budou ukládat jako dopředu vypočítaná pole. Hodnoty dat i indexů se uchovávají v jednotlivých polích multidimenzionální databáze. Databáze je organizována tak, aby umožnila rychlé získávání příslušných údajů z více dimenzí. Část údajů se může zavést ze serveru ke klientovi, což umožňuje rychlé analýzy bez velkého zatížení sítě. Hlavní výhodou MOLAP je maximální výkon vzhledem k dotazům uživatelů, nevýhodou je redudance údajů, neboť tyto údaje jsou uloženy jednak v relační databázi, jednak v multidimenzionální databázi. Požadavky na úložnou kapacitu mohou v případě použití více dimenzí extrémně narůstat.

3.1.2.2 Relační databázový OLAP (ROLAP)

Relační online analytické zpracování údajů (ROLAP) získává údaje pro analýzy z relačního datového skladu. Tyto údaje z relačních databází se po zpracování předkládají uživateli jako multidimenzionální pohled. Data a metadata se v úložišti ROLAP ukládají jako záznamy v relační databázi. Server OLAP dynamicky používá tato metadata na generování příkazů SQL, které jsou potřebné na získávání dat požadovaných uživateli. U tohoto způsobu zůstávají data uložena v relačních databázích, takže nevzniká problém s redudancí.

3.1.2.3 Hybridní OLAP (HOLAP)

Hybridní OLAP je kombinací úložišť MOLAP a ROLAP, přičemž se využívají výhody jednotlivých typů úložišť a do značné míry se eliminují nevýhody. Údaje zůstávají v relačních databázích a spočítané agregace se ukládají do multidimenzionálních struktur. Při dotazování se údaje vybírají do multidimenzionální paměti cache. U hybridního řešení relační databáze ukládá množství detailních dat a multidimenzionální model ukládá sumární data.

3.1.2.4 Desktop OLAP (DOLAP)

DOLAP (Desktop OLAP) je nejmladší architektura OLAP databází, která se objevila koncem devadesátých let. DOLAP umožňuje připojit se k centrálnímu úložišti OLAP dat a stáhnout si potřebnou podmnožinu kostky na lokální počítač. Veškeré analytické operace jsou pak prováděny nad touto lokální kostkou, takže uživatel nemusí být připojen k serveru. Toto je výhodné zejména pro mobilní aplikace a podporu mobilních uživatelů obecně.

4 DOLOVÁNÍ DAT – DATA MINING

Dolování dat lze charakterizovat jako proces extrakce relevantních přede neznámých nebo nedefinovaných informací z velmi rozsáhlých databází.

Mnoho podniků v současné době spravuje rozsáhlé informační databáze a datové sklady. Reálná data v nich uložená představují obrovský potenciál použitelný pro řízení podniku ve všech jeho oblastech. Cílem dolování dat je tato automaticky či poloautomaticky analyzovat a nalézt v nich ("vytěžit" z nich) důležité informace o vzájemných závislostech mezi vývojem hodnot určitých ukazatelů nebo o strukturách chování (např. nákupní preference zákazníků). Ty lze potom použít např. jako podklad pro změnu marketingové strategie (např. pro stanovení skupin výrobků pro křížový prodej – cross selling).

Pokud bychom porovnali způsob provádění analýz dolování dat s analýzami na bázi OLAP, je zde významný rozdíl. Uživatel při použití analýzy na bázi OLAP používá deduktivní způsob práce – vytváří ve své mysli určitou množinu otázek – hypotéz, které potom za použití OLAP aplikace buď potvrzuje, nebo vyvrací. Analýzy na bázi dolování dat jsou naopak induktivním procese, který pomáhá takovéto hypotézy na základě analýzy skutečných dat vytvářet. Aplikace dolování dat proto zpravidla nepracují s agregovanými OLAP kostkami předem nadefinovaných struktur dat, ale přímo s daty uloženými v datovém skladu podniku nebo v provozních systémech.

V podnikových aplikacích BI se obvykle provádějí jak analýzy dolování dat, tak analýzy na bázi OLAP.

Dolování dat se začalo komerčně prosazovat zejména v posledním desetiletí, kdy rozvoj vědy a technologií vůbec umožnil tento náročný proces.

4.1 Úlohy dolování dat

V praxi existuje mnoho členění typů úloh dolování dat. Běžné dělení úloh v dolování dat se člení na:

- **Explorační analýzy dat –** podstatou je prozkoumat data bez předcházející znalosti, která by určitým způsobem naše hledání usměrňovala. Využívají se zde různé metody či speciální techniky

- **Deskriptivní úlohy** – podstatou je určitým způsobem popsat celou datovou množinu. Z hlediska dolování dat je například takovou metodou shlukování, při kterém dochází k vytvoření skupin, do kterých se dají projevy v datech rozdělit.
- **Prediktivní úlohy** – cílem je předpovědět hodnotu určité veličiny na základě znalosti hodnost ostatních veličin. Z hlediska statistiky je takovou metodou regresní analýza. Predikci v dolování dat provádíme zejména klasifikací příkladů do tříd.
- **Hledání vzorů a pravidel (hledání nuggetů)** – podstatou je hledání určitých vztahů a vzorů chování v datech. Klasickou úlohou je zde analýza nákupního košíku, která má rozkrýt, které druhy zboží jsou zákazníky kupovány současně. Dalším takovým příkladem může být úloha z oblasti bankovnictví, spočívající v detekci vzorů implikujících provádění operací praní špinavých peněz.
- **Hledání podle vzorů –** před prováděním hledání znalostí podle vzorů má analytik k dispozici určitý vzor a cíle je nalézt v datech vzory, shodující se nebo podobné s touto předlohou. Jedná se tedy o rozpoznávání vzorů v datech na základě předem definované šablony. Tyto typy úloh se realizují v oblasti rozpoznávání obrázků a textů. Například při rozpoznávání textů máme k dispozici vzorový informační vektor vyjadřující daný text. Při aplikaci tohoto typu úloh potom porovnáváme ostatní informační vektory s reprezentantem a vyhodnocujeme jejich podobnost, například na základě metod podobnosti vektorů.

4.2 Techniky dolování dat

Úlohy dolování dat je možno řešit s použitím celé řady technik. Mezi nejdůležitější techniky dolování dat patří:

- **Analýza nákupního košíku (Market Basket Analysis)** – je speciální formou clusteringu (detekce shluků) používanou k vyhledávání skupin a prvků, které mají tendenci se pospolu (v jedné transakci). Analýza nákupního košíku hledá opakující se nákupní košíky a popisuje je prostřednictvím implikačních pravidel.
- **Dedukce (Memory Based Reasoning)** – technika, která využívá známé skutečnosti jako model k predikci neznámých skutečností. Dedukce sleduje nejbližší okolí známých instancí a kombinuje jejich hodnoty za účelem odhadu předikovaných hodnot.

- **Detekce shluků (Cluster Detection)** - vytváří modely identifikující datové záznamy, které jsou si navzájem podobné. Detekce shluků nevychází z předem definovaných skupin charakteristiky shluků, i jejich počet vyhledává na základě podobnosti zkoumaných dat.
- **Analýza závislostí (Link Analysis)** – oproti výše uvedeným technikám analýza závislostí nezkoumá prvky na základě jejich vlastností, ale zaměřuje se na vztahy mezi prvky. Jedná se o aplikaci teorie grafů.
- **Rozhodovací stromy a indukce (Decision Trees and Rule Induction),** viz. Obr. 8.1 . – představují výkonné modely, které jsou výstupem statistických a nestatistických metod, např. klasifikační a regresní stromy (CART), chi-kvadrát automatická indukce (CHAID), kritéria informační entropie (C 5.0) apod. Rozdělují záznamy v tréninkových sadách do disjunktních skupin, kde každá skupina může být popsána pomocí jednoduché množiny pravide.
- **Neuronové sítě (Artifical Neural Network)** – jsou v podstatě zjednodušeným modelem neuronových propojení v lidském mozku modelovatelným výpočetní technikou. Jejich principem je nastavení parametrů jednotlivých "neuronů" v procesu učení se z tréninkových vzorků dat, aby výsledná konfigurace co nejlépe vyhovovala následné kvalifikaci a predikci. Neuronové sítě jsou příkladem aplikace jedné z vývojových linií dolování dat – umělé inteligence.
- **Genetické algoritmy (Generic Alogorithms)** – aplikují mechaniku genetiky a přirozeného výběru pro vyhledávání množiny parametrů, například pro použití v predikci. Jinými slovy, genetické algoritmy neslouží k predikci určitých hodnot zkoumaných prvků (jako jsou výše popsané techniky), ale slouží k vývoji, resp. k parametrizaci dalších modelů pro predikci hodnot těchto prvků.

4.3 Proces dolování dat

4.3.1 Definice problému

Prvním krokem v procesu je definice obchodního problému nebo příležitosti, na kterou se máme zaměřit. Úspěšná Data Mining iniciativa je vždy zahájena dobře definovaným projektem. Abychom ověřili, že bude vytvořena určitá nová hodnota, mělo by být zahrnuto vyhodnocení status quo v dané oblasti.

4.3.2 Výběr dat

Poté, co je definován problém, musí být definovány zdroje dat. Ne každý zjištěný datový zdroj je vždy vyžadován pro řešení. Data jsou obvykle extrahována ze zdrojových systémů nebo datových skladů na zvláštní server, kde je realizován Data Mining.

4.3.3 Příprava dat

Příprava dat je časově nejnáročnější částí každého projektu dolování dat, vyžaduje až 80% celkových zdrojů. Data Mining vyžaduje, aby data, která budou analyzována, byla připravena do podoby jednoduché tabulky (každý záznam, který bude modelován, obsahuje mnoho sloupců). Tato metodologie umožňuje vytvoření stovek a občas i tisíců proměnných, které budou vstupovat do modelování.

Tato projektová fáze je nejkritičtější – výsledné modely jsou tak dobré, jak dobrá jsou data, která jsou použitá pro jejich vytvoření. Expertiza v oblasti Data Mining spočívá nejvíce v tom, aby reprezentace podrobných dat měla formu odpovídající všem aspektům řešeného obchodního problému.

Významné zlepšení výsledků může být dosaženo zlepšení metodologie přípravy dat.

4.3.4 Data Mining

Tato fáze zahrnuje využití statistických a nestatistických nástrojů pro vytvoření matematických modelů. Tato fáze je typicky nejkratší a nejjednodušší částí jakéhokoli Data Mining projektu. Většina organizací, která zaměstnává analytiky, je schopna si v tomto směru postupně vystačit i sama.

Data Mining se typicky realizuje na serveru, který je oddělený od datového skladu nebo jiných informačních systémů společnosti. Některé společnosti dokonce vytváří modely na počítačích PC s využitím vzorkování dat.

4.3.5 Zprovoznění modelu (Deployment)

Zprovoznění modelu je proces, kdy se matematické modely implementují do operačního systému, aby mohly být využity ke zlepšení obchodních výsledků.

4.3.6 Obchodní akce

Tato fáze zahrnuje využití zprovoznění modelů pro zajištění zlepšených výsledků v rámci identifikovaného obchodního problému nebo příležitosti.

4.4 Technologie dolování dat

Úlohy dolování dat mohou být realizovány rozmanitými technologiemi, často i kombinací různých technologií. Část procesu Data Mining je typicky realizována nástroji využívajícími statistickou a nestatistickou analýzu dat. Dělení nástrojů na statistické a nástroje Data Mining není samoúčelné. Statistické produkty poskytují většinou potřebné funkcionality se zaměřením na "klasičtější" statistické metody a mají příznivější cenu, ovšem je náročnější s nimi pracovat. Produkty Data Mining mají velmi intuitivní uživatelské rozhraní, díky kterému je práce s nimi velmi efektivní, obsahují některé statistické metody, které nevyžadují hluboké statistické znalosti pro svou parametrizaci, a umožňují automatizaci výsledných modelů oběma výše popsanými postupy.

II. PRAKTICKÁ ČÁST

5 PROČ POTŘEBUJEME BI

V úvodu této práce jsem naznačil možné problémy se zavedením IS do podniku a následnou interpretací uložených dat. V centrálním datovém skladu, kde budeme skladovat veškerá data, nám bude jejich objem narůstat vysokou rychlostí a tím se bude snižovat jejich přehlednost. Jelikož nám jde o informaci, která z dat plyne složitým analytickým procesem, nemůžeme pojmy data (údaje) a informace ztotožnit.

Proces transformace údajů na informace a převod těchto informací na poznatky prostřednictvím objevování nazýváme Business Intelligence. Jinými slovy účelem Business Intelligence je konvertovat velké objemy údajů na poznatky, které jsou potřebné pro koncové uživatele např. v procesu rozhodování.

Často potřebujeme sledovat trend nějaké veličiny, například při obchodování s cennými papíry, nebo potřebujeme najít mezi údaji určité závislosti. Proto moderní databázové servery obsahují rozsáhlou podporu pro budování datových skladů, analýzy OLAP a data mining.

Správně interpretovaná data s pomocí nástrojů Business Inteligence lze využít v podstatě ve všech oblastech lidské činnosti, kde je třeba sledovat a analyticky vyhodnocovat hodnoty určitých ukazatelů. V současné době proto také existuje celá řada aplikací nebo řešení BI specializovaných pro určitou oblast, které se svou funkcionalitou v noha případech překrývají. Cílem této kapitoly je rovněž popsat typické aplikační oblasti, ve kterých jsou v současné době aplikace BI používány, a to např. i u telekomunikační firmy.

5.1 Specifika TELCO firmy

5.1.1 Z pohledu firemních systémů

ISP systémy a telefonní ústředny – jsou pochopitelně zcela specifické pro telekomunikační společnosti. Generují především data o provozu. BI systémy obvykle standardní cestou tyto data nevyužívá. Ale DWH databáze mohou být nepřímo využity pro dlouhodobé uchování těchto údajů, protože provozní systémy obvykle pracují s kratším časovým obdobím. Data z těchto systémů mohou být důležité především pro účely data miningu či adhoc analýz.

Provizioning systémy – jsou též typické pro společnosti, kde je potřeba instalací služby komunikovat s velmi rozličnými technologickými systémy, a to jak automatickým procesem, tak i manuálním zřizováním. Což je typické právě pro telekomunikační společnosti. Data z těchto systému mají ovšem pro BI menší význam.

Retail Billing – Billing systém specifický v telekomunikační společnosti je obvykle spjat s možností automatického oceňování provozu (volání, přenesené data), automatického účtování na základě údajů z technologických systémů, a uplatňování různých tarifů a slev. Nese sebou většinou taky informace o účtovaných službách jejich nastavení a taky údajů ze smlouvy. Z pohledu BI je tento systém vedle CRM systémů jeden z nejdůležitějších.

Wholesale Billing – Telekomunikační firma též obvykle musí mít velkoobchodní billing systém, ten však z hlediska BI aktivit není většinou podstatný.

5.1.2 Z pohledu obchodních a marketingových aktivit

Asi nejpatrnější rozdíl je dán přítomností dominantního telekomunikačního operátora na českém trhu a tím i zásadní ovlivnění možného počtu potencionálních zákazníků v jednotlivých segmentech, a určitou závislost na technologiích konkurence. Možnosti prodejních kanálů (partnerský prodej, telesales, přímý prodej, …). A také omezením dostupností telekomunikační technologie, kde je při vhodných zákazníků nutné brát v úvahu lokalitu zákazníka. Dalším významným rozdílem je fakt, že to co zákazníkovi prodáváme je většinou služba, za kterou nám pravidelně platí, a nikoliv jednorázově prodaný produkt. S tím úzce souvisí i péče o zákazníka a budování vztahu se zákazníkem.

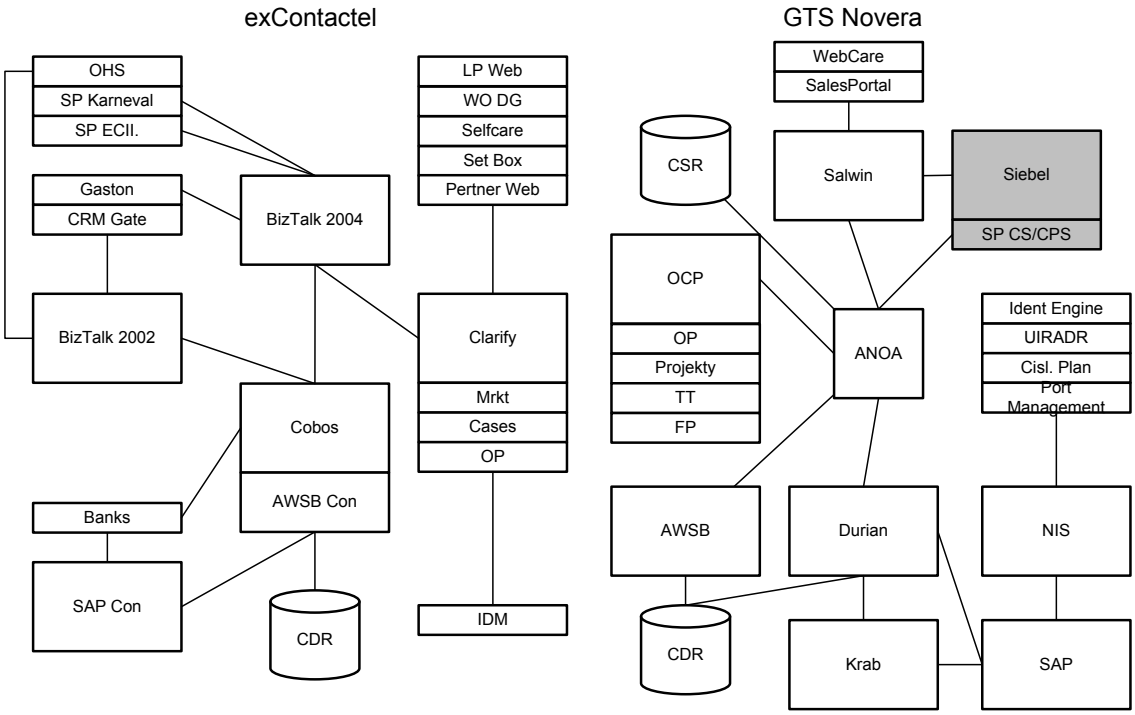
6 POPIS ZDROJOVÝCH SYSTÉMŮ

6.1 Přehled systémů

Tato práce analyzuje BI prostředí společnosti GTS Novera, patřící mezi největší firmy na telekomunikačním trhu v České Republice. Společnost GTS Novera se v době příprav této práce sloučila s několika dalšími společnostmi, Contactel, Nextra a Telenor Networks. V minulosti společnost GTS Novera vznikla sloučením několika dalších společností, GTS Czech, KPNQuest a Aliatel.

Z těchto důvodů jsou systémy v této společnosti značně komplikované. Většina systémů je ve společnosti duplikovaná, a data v systémech jsou nekonzistentní. To je asi největší překážkou při implementaci BI řešení.

V jednotlivých společnostech již na nějaké úrovni je implementován datový sklad, ale ani v jednom případě nevyhovuje požadavkům.



Obrázek 1 Znázornění systému ve společnostech exContactel a GTS Novera

6.2 CRM systém

CRM (Customer Relationship Management) systém je obecně IS systém určen především pro řízení vztahů se zákazníky. V širším pojetí lze CRM chápat jako proces nebo strategii

firmy pro řízení a pochopení vztahů se zákazníkem. Možnosti CRM systémů jsou poměrně široké, a ve většině případů zdaleka nejsou využívány. Dodejme že CRM systém, přesněni databáze CRM systému, je jen z nejpodstatnějších datových zdrojů pro BI řešení.

V našem případě se používá CRM systém společnosti Nortel Networks, Clarify eFrontOffice verze 9, který byl původně implementován ve firmě Contactel.

CRM systém se využívá především pro následující aktivity:

- Správa zákaznické databáze, včetně adresních a kontaktních údajů
- Řízení obchodních případů
- Řízení kampaní
- Řízení CustomerCare aktivit
- Řízení vztahů a komunikace se zákazníkem
- Správa marketingových informaci

Systém Clarify používá OLTP databáze Oracle. Mezi základní databázové entity patří:

- **BUS_ORG** – úroveň zákazníka
- **SITE** – pobočka zákazníka, fakturační místo, místo instalace služby
- **ADDRESS** – adresy zákazníka
- **CONTACT** – kontaktní údaje
- **OPPORTUNITY** – obchodní případy
- **CONTRACT** – uzavřené smlouvy
- **LEAD** – kampaně
- **CASE** – řešení problému se zákazníkem (support)

Ve firmách GTS Novera a Nextra nebyl žádny CRM systém implementován. Pro správu firemních kontaktů se zde používá systém Salwin (vlastní vývoj) a pro obchodní případy OCP (vlastní vývoj), v původní společnosti Nextra je to systém K2.

6.3 Billing

Důvody potřeby vlastního billingového systému byli již popsány, charakterizujme tedy jen situaci a stručný popis systémů.

V původní firmě Contactel se požívá systém Cobos (vlastní vývoj) postavený na databázi MS SQL Server 2000. Ve firmě GTS Novera se používá systém Durian postavený na databázi Oracle.

Oba systémy jsou svou datovou architekturou velmi podobné, mezi základní datové entity patří:

- **Account** – zákazník, fakturační místo
- **Subscriber** – konkrétní instance produktu zákazníka
- **ProductCategory** – produkt
- **Item** – fakturační položka
- **InvoiceHeader, InvoiceDetail** – faktura

V původní firmě Nextra se pro účely billingu používá kombinace produktů K2 a Helios a část vlastního systému. Datová struktura je hodně odlišná od stávajících systémů, a nadále se s tímto systémem nepočítá.

7 POŽADAVKY NA BUSINESS INTELLIGENCE

Technologii BI lze využít v podstatě ve všech oblastech lidské činnosti, kde je třeba sledovat a analyticky vyhodnocovat hodnoty určitých ukazatelů. V současné době proto také existuje celá řada aplikací nebo řešení BI specializovaných pro určitou oblast, které se svou funkcionalitou v noha případech překrývají. Cílem této kapitoly je popsat základní aplikační oblasti, ve kterých sou v současné době aplikace BI používány. Měla by sloužit jako inspirace pro hledání možností, jak využít BI ve svém podniku nebo organizaci.

7.1 Finance

Aplikace BI v této oblasti umožňují dostat pod kontrolu finanční hospodaření podniku. Díky datům z provedených finančních (účetních) operací uloženým v datovém skladu, umožňují BI aplikace získat hodnoty ukazatelů finanční výkonnosti za celou organizaci, za jednotlivé závody, nákladová střediska, projekty, skupiny produktů apod. Výsledkem jsou pracovní panely nebo výstupy, které dovolují v případě, že se ukazatele finančního hospodaření odchylují od plánovaných hodnot, okamžitě zajistit místo, kde dochází k problémům, a přijmout odpovídající nápravná opatření. Nasazení aplikací BI v oblasti finančního řízení a plánování sebou (vzhledem k povinnosti přiřazovat určité typy nákladů na projekty, nákladová střediska apod.) obvykle přinese zavedení vysoké finanční transparentnosti zejména v oblastech řízení nákladů.

Ve finančním řízení podniku jsou aplikace BI používány především v oblastech:

- Finanční plánování a prognózování – zde se užívají zejména analytické nástroje pro tvorbu prognóz a simulaci finančního vývoje organizace. Pomáhají automatizovat procesy finančního plánování (např. vytváření struktury finančních plánů na různých úrovních). Jednotlivé finanční plány mohou být na základě skutečného vývoje situace zachyceného v datovém skladu pravidelně sledovány, vyhodnocovány a případně operativně modifikovány. Díky tou mohou být odhady příjmů a výdajů vytvářeny s mnohem větší mírou jistoty než tomu bylo doposud.
- Finanční výkaznictví a konsolidace – na základě aktuálních informací uložených v datové skladu pomáhají aplikace BI konsolidovat multidimenzionální informace z různých zdrojů (např. přes jednotlivé dceřiné společnosti holdingové struktury) a rychle vytvářet výstupy, které by v klasických transakčních systémech bylo možné

jen s obtížemi provést. Jde Zejména o finanční výstupy prováděné simultánně přes jednotlivé definované dimenze (podniky, pobočky, nákladová střediska, projekty apod.), které umožňují okamžitě zjistit nerovnováhu finančního hospodaření u prvků jakékoliv z těchto dimenzí. V poslední době se do popředí zájmu top managementu firem dostalo standardní výkaznictví vyžadované regulátory trhu,burzovními či jinými institucemi. Jedná se např. o nově zavedené reporty Sarbanes-Oxley, Basel II, reporty telekomunikačních operátorů vůči národním regulátorům či již určitou dobu využívané konsolidované účetní výkazy podle standardů GAAP nebo IAS apod. Tyto požadavky jsou ve značné míře řešitelné (a současně i řešené) právě a pouze aplikacemi BI.

- Analýza nákladů a ziskovosti – díky multidimenzionální struktuře uložených dat umožňují aplikace BI zjistit skutečné náklady a ziskovost spojenou s produkty, dodavateli, prodejními kanály, partnery zákazníky apod. Díky tomu je možné vytvářet lepší předpovědi budoucího vývoje, tvořit konkrétní plány nákladů a zisku na různých úrovních – zejména identifikovat nejvíce a nejméně ziskové zákazníky, produkty apod. a také porozumět důvodům tohoto rozdělení.
- Řízení rizika – aplikace tohoto typu umožňují sledovat a řídit riziko spojené s finančními operacemi, zejména s úvěrovým zatížením, situací na trhu a vlastním provozem organizace. Díky aplikaci pravidel pro řízení rizika na jednotlivé dimenze dat uložených v datovém skladu a jejich spojení s informacemi o aktuálním stavu všech faktorů spojených se vznikem rizika nebo jeho zamezení, je možné vytvářet specifické výstupy týkající se rizik v podniku jako celku, regionu, typu finanční operace, typu výroby atd. Následně mohou zodpovědní pracovníci provádět efektivní protiopatření (přeskupení výroby, zvýšení kapacit, pojištění apod.)
- Finanční optimalizace – umožňují simulovat, plánovat, sledovat a zejména porozumět dopadům fůzí, akvizic, restrukturalizace daňové politiky organizace atd. V této oblasti jsou podporovány analýzy týkající se finančních dopadů spojených zejména s plánováním výroby, produktového mixu a lidských zdrojů ve výše uvedených situacích.

7.2 Obchod

Aplikace BI určené pro obchod jsou velmi podobné těm co jsou pro Marketing, případně aplikace CI. Navíc je potřeba detailnějšího reportování zákazníků a jejich výnosu pro jednotlivé obchodníky, sledování obchodních případu, sledování očekávaných výnosů (order entry) a skutečných výnosu.

7.3 Marketing a Produkt Managment

Oblast marketingu se stále více stává jednou z integračních součástí systémů CRM (Customer Relationship Management) a je podporována v aplikacích typu Customer Intelligece.

Přesto se setkáváme s aplikacemi BI určenými pouze pro podporu marketingu podniku. Jejich cílem je zejména pomáhat při analýze a plánování marketingových kampaní a při analytickém vyhodnocení jejich dopadu. Aplikace BI zde proto můžeme nalézt v následujících oblastech:

- Analýza portfolia produktů a služeb – aplikace BI zajišťují analýzy profitability a nákladovosti jednotlivých produktových řad či konkrétních produktů, analýzy jejich potenciálu vzhledem ke skupinám zákazníků (ať již geograficky či jinak segmentovaným) zajišťují také analýzy jejich vztahu ke konkurenčním výrobkům (tržní podíl oproti konkurenci)
- Klasifikace a segmentace zákazníků – aplikace BI slouží k simulaci a stanovení kritérií pro klasifikaci a segmentaci zákazníků (geograficky, příjmová skupina, věk apod.)
- Proces kontroly interních dat vůči externím databázím, a případné doplňování relevantních informaci ke stávajícím zákazníkům
- Proces správy marketingových kampaní (Campaign Management) a jeho části:
    - Plánování a analýza dopadu marketingových kampaní – aplikace BI provádějí výběry skupin zákazníků a základě kritérií kampaně, stanovují kritéria úspěšnosti kampaně, její vyhodnocení a analýzu jejich celkového dopadu.
    - Analýza marketingových zdrojů – aplikace BI zajišťují analýzu stavu a plánování rozvoje jednotlivých marketingových zdrojů (marketingové materiá-

ly, reklamní čas v médiích, týmy pro přímý marketing atd.) ve vztahu k jednotlivých výrobkovým řadám, lokalitám apod.

- Analýza marketingových nákladů – aplikace se starají o podrobnou analýzu vynaložených marketingových nákladů a jednotlivé kampaně a jejich porovnání s dosaženými efekty,
- Vyhodnocení kampaní – aplikace BI pomáhají sledovat průběh kampaně a podporují proces vyhodnocení úspěšnosti jednotlivých kampaní.

7.4 Customer Inteligence

Customer Inteligence (CI) představuje komplex aplikací IS/ICT, zaměřených na poznání zákazníka, jeho hodnoty, preferencí, rizikovosti nebo pravděpodobnosti odchodu ke konkurenci. Za účele splnění tohoto cíle využívají řešení CI spojení systémů BI a CRM.

Většina zákaznických dat je v dnešní době uložena v systémech patřících do oblasti CRM, které jsou dnes typicky rozřazovány do tří okruhů:

Operativní CRM – je orientované na zefektivnění klíčových procesů "kolem" zákazníka, zejména front office úloh, zahrnuje:

- Automatizace obchodních procesů (SFA – Sales Force Automation)
- Podporu marketingu (MFA – Marketing Force Automation)

Kooperativní (kontaktní) CRM – jedná se o systémy zaměřené na zachycení komunikace mezi společností a jejími zákazníky. Kooperativní CRM zajišťují zejména:

- Podporu zákaznických (kontaktních) center,
- Podporu servisu.

Analytické CRM – zahrnuje již agregace a aplikace znalostí o zákazníkovi, aplikace business a customer inteligence a rovněž speciální CRM analytické aplikace na bázi datových skladů a dolování dat. V rámci analytických CRM se provádějí především následující aktivity:

- Segmentace zákazníků
- Analýza marketingových kampaní
- Predikce chování zákazníků

- Personalizace

Úroveň CRM je vysoce závislá na kvalitě zákaznických dat. Přitom podle průzkumu Meta Group (USA) 67% firem nesbírá data o zákaznících efektivně, pouze 4% na odpovídající úrovni. Z toho vyplývá, že úspěšnost projektů CRM závisí, kromě jiného, i na přístupu k řešení tohoto problému.

Je zcela zřejmé, že možnosti získávání a analýz dat o zákaznících se velmi rychle vyvíjejí v souvislosti s technologiemi a aplikacemi, které zajišťují bezprostřední vtahy se zákazníky. Uplatnění těchto nových zdrojů dat o zákaznících a jejich potřebách na jedné straně, a možnosti poskytování nových informačních služeb zákazníkovi s využitím technologií BI na straně druhé, dalo základ Customer Inteligence. Pro tuto novou kategorii informatických aplikací se nejčastěji uvádějí tyto jejich charakteristiky:

- CI se chápe jako komplexí využití technologických možností BI a datových zdrojů v řízení vztahů podniku se zákazníky, tj. s realizací integrace s CRM
- CI se primárně orientuje na shromažďování a analýzy dat s interakcí se zákazníkem, a využívá přitom procesy a zdroje, které jsou součástí CRM (zázamy z obchodních kontaktů, dokumentace kontaktních center apod.)
- Obdobně jako v případě CRM, musí být i součástí řešení CI nastavení řídících a obchodních procesů respektujících principy CI a záměry jeho uplatnění v řízení podniku.
- CI rozšiřuje běžně aplikované zákaznické analýzy o analýzy opírající se o nové zdroje dat – data z e_business aplikací, podnikových portálů, zachycení tzv. "click stream", tj. přístupy a postupy využití funkcí web aplikací zákazníkem apod.
- Vytváří se tak prostor pro nové analytické aplikace nad novými daty o zákaznících.

Analytické schopnosti CI rovněž umožnily podnikovým analytikům zaměřit se na hodnotu zákazníka, a na základě této hodnoty ovlivňovat kvalitu podpory pro jednotlivé segmenty zákazníků. Koncept hodnoty zákazníka (CV – customer value) spočívá v kvantifikaci jeho minulých i budoucích přínosů a nákladů, a pracuje tedy s následujícími metrikami:

- Ziskovost zákazníka – kalkulovaná jako rozdíl mezi výnosy ze zákazníka sumou přímých a nepřímých nákladů za uplynulé období.

- Riziko ztráty zákazníka – jedná se o riziko dobrovolného odchodu zákazníka ke konkurenci, ale i jeho nedobrovolného odchodu iniciovaného společností. Tento ukazatel se počítá za použití technologií dolování dat a je třeba tyto technologie kombinovat s prediktivní modelováním chování klientů pro určení optimální strategie pro udržení klienta.
- Celoživotní hodnota zákazníka (CLV – Customer Life-Time Value) – je pak součet současné hodnoty zákazníka a predikované hodnoty zákazníka na základě odhadované doby, po kterou bude využívat služby/produkty společnosti.

Hlavní závěry:

- Aplikace CI se primárně zaměřuje na dokonalou znalost zákazníka a nastavení relevantních procesů společnosti podle této znalosti
- Strategie a postup řešení CI musí respektovat stav a perspektivy aplikací CRM a BI v daném informačním systému podniku a zahrnovat i návrhy jejich rozvoje, především vzhledem k potřebám řízení vztahů se zákazníky.
- Pro většinu dosavadních BI aplikací byl typický problém "zvládnutí" vztahu jednoduchosti aplikací a jejich ovládání – oproti komplexnosti jejich řešení. Totéž, v ještě větší míře platí i pro aplikace Customer Intelligence, neboť do nich vstupují další datové zdroje a další funkcionalita. Navíc nejsou již orientovány pouze na vlastní zaměstnance, ale některé funkce jsou poskytovány i externím partnerů – bez speciální přípravy
- Úlohy CRM i BI/CI nejsou svébytné ani izolované v informačním systému a musí tvořit jeho logickou součást. Koncepce jejich řešení a plán jejich realizace musí být proto součástí celkové koncepce IS/ICT podniku, resp. jeho informační strategie.

7.5 Aplikace dolování dat

Reálné aplikace, v nichž se Data Mining uplatňuje, je možné rozdělit do několika skupin. Jedná se zejména o kreditní skóring klientů, o prodejně marketingové aplikace, a dále o specializované aplikace pro detekci podvodů. V telekomunikačních firmách se používá behaviorální kreditní skóring, který pro všechny klienty na základě údajů o jich chování předpovídá, kteří z nich nebudou platit za služby. Je často využíván k rozhodování, kterému klientovi bude zaslána marketingová nabídka.

Marketingové aplikace jsou v zásadě rozděleny na 3 typy. První a nejpřínosnější je cílení produktových marketingových kampaní na klienty, kteří mají zájem si daný produkt pořídit. Jedná se o tzv. "propensity to buy" nebo také afinitní modely, které předpovídají budoucí nákup tohoto produktu.l Tyto modely typicky vznikají pro každý významný produkt. Druhou typickou aplikací je předpověď odchodu zákazníků, která umožňuje tomuto nepříznivému vývoji včas předejít. Třetí aplikací je segmentace zákazníků, která rozdělí zákazníky do homogenních skupin podle jejich hodnoty nebo podle jejich chování. Hodnotová segmentace se používá pro strategické rozhodování, jelikož umožňuje sledovat a předpovídat dopad tržních změn.

Behaviorální segmentace se používá pro návrh produktů, volbu komunikačního kanálu, způsob komunikace i vlastní sdělení. Behaviorální segmentace spolu s propensity to buy modely se úspěšně používá pro efektivní cílení marketingových kampaní.

Perspektivní aplikací Data Mining je odhalování podvodů. Tato aplikace má uplatnění především v pojišťovnách a bankách, ale své uplatnění nachází i u telekomunikačních společností.

7.5.1 Segmentace

Segmentaci zákazníků dnes využívá každá významnější společnost pro roztřídění zákazníků do podskupin, pro které se sjednocují obchodní a marketingové postupy.

Příklakem jednoduchého členění zákazníků je jejich rozdělení na korporátní zákazníky, realizující nejvyšší obraty a zisky, a masový trh, členěný dále na malé/střední podniky a fyzické osoby. K rozhodnutí, že nejcennější přivstání klientela bude obsluhována primárně individuálními obchodními manažery a masový trh bude cenovou politikou veden k využití levnějších elektronických kanálů s podporou obchodních balíčků, nejsou jistě nutné pokročilé datové analýzy.

Jestliže například zvažujeme právě návrh jednotlivých balíčků, může být užitečné rozdělení zákazníků podle více než jedné charakteristiky týkající se způsobu obsluhy.

Pro úvahy o možnosti nabídky produktů by mohlo být zajímavé například porovnat zákazníky podle 2 charakteristik.

- Průměrné revenue
- Průměrné revenue konkrétního produktu.

Pokud chceme do proměnných zakomponovat ještě další segmentační proměnnou a segmentovat zákazníky podle tří kategorií:

- Průměrné revenue
- Průměrné revenue konkrétního produktu.
- Trend revenue na účtech za posledních 12 měsíců

Ručně vizuální postup, který je pro jednu nebo dvě segmentační proměnné odůvodnitelný, je zde již prakticky nepoužitelný. Právě v této situaci se uplatňuje Data Mining s nabídkou technik shlukování (Clustering). Shlukovací techniky umožňují po zadání, i většího počtu segmentačních proměnných, najít shluky (clusters), které odpovídají "nejlepším možným" segmentům.

Některé ze segmentačních algoritmů umožňují navrhnout optimální počet segmentů v určitém rozsahu, jiné vyžadují pevně zadat počet segmentů předem. Automatizované nalezení počtu segmentů může být sice vhodné pro první přiblížení, v praxi však často dochází k ruční korekci počtu segmentů. Dostáváme se k hlavnímu kritériu úspěšnosti segmentace. Dobrá segmentace je obchodně užitečná. Pokud rozdělením původně jednoho segmentu Porfitabilní vzniknou dva nové segmenty Profitabilní s malým rizikem a Profitabilní s vyšším rizikem, bude se nejspíše jednat o užitečné rozčlenění. Budeme-li počet segmentů dále zvyšovat, může se ukázat, že při příliš velkém počtu segmentů již další rozčleňování neukáže žádné obchodně zajímavé odlišnosti, nebo je počet zákazníků v některém shluku již natolik malý, že využití takovéhoto segmentu rovněž postrádá marketingově-obchodní smysl.

Obecnou motivací k segmentaci zákazníků je jejich rozčlenění na vnitřně homogenní a navzájem heterogenní podskupiny pro cílení obchodního a marketingového přístupu.

Hodnotová segmentace zákazníka

Segmentačními proměnnými jsou ukazatele hodnoty zákazníka. Hodnota zákazníka může být měřena například kritérii tržba, profit, riziko, loajalita nebo obsluha. V případě některých kritérií (tržba) hodnoty lze nalézt proměnné, které budou vystupovat přímo jako segmentační, v případě jiných kritérií (loajalita) se hledají takové proměnné nebo jejich kombinace, které v dostatečné míře dané kritérium vyjadřují.

Hodnotová segmentace je účelným prvním krokem pro iniciální rozčlenění portfolia zákazníků na hlavní skupiny (popřípadě pro ověření takového již existujícího rozdělení) a bývá doplňována dalšími analýzami, například níže uvedenou behaviorální segmentací.

Behaviorální segmentace zákazníků

V tomto přístupu k segmentaci se pokoušíme primárně odhlédnout od hodnoty zákazníka a zaměřujeme se na jeho "chování", čímž míníme především počet, druh a způsob použití jednotlivých produktů. Například u telekomunikační společnosti můžeme jako segmentační proměnné pro behaviorální segmentaci použít:

- Procento lokálních volání
- Procento vnitrostátních volání
- Procento mezinárodních volání
- Procento volání na mobilní telefony

Behaviorální segmentaci není vždy účelné oddělovat od segmentace hodnotové, výhodná je jejich kombinace. Například po celkové hodnotové segmentaci zákazníků lze na vybrané podskupiny uplatnit segmentaci behaviorální.

Segmentace zákazníků pro odhalení nestandardního chování

Tento typ segmentace se využívá pro analýzu a detekci podvodného chování nebo pro detekci rizika určitého druhu (viz. Obr. 8.2). Zatímco pro dříve uvedené dva druhy segmentací je typický relativně malý počet segmentů (nejčastěji 6 až 15), aby pro obchodně marketingové účely byl jejich počet zvládnutelný, zde se často pracuje s větším množstvím segmentů a hledají se spíše menší, neobvykle se chovající skupiny zákazníků.

Výše uvedená kritéria jsou nutnou, ale nepostačující podmínkou "dobré" segmentace. Hlavním kritériem je jich zmíněná "obchodní užitečnost". Výsledky segmentace se profilují, tj. popíší se všechny zajímavé odchylky segmentačních proměnných od průměrného rozložení v celé množině případů. V rámci popisu segmentů se zahrnou i zajímavé odchylky proměnných, které nebyly zvoleny jako segmentační. Výsledky profilování vedou k označení segmentů krátkým výstižným názvem a výsledná vyprofilovaná segmentace musí projít obchodní oponenturou.

Není-li shledání "dobrou", iterativně se optimalizuje.

Některé základní příklady obchodního použití výsledků segmentace:

- Úvodní nalezení typologií zákazníků
- Rozdělení zákazníků podle hodnoty pro stanovení obecného obchodního přístupu
- Hledání potenciálu cross-sell.
- Sledování změn chování zákazníků. Segmentační model se periodicky skóruje a sledují se migrační proudy mezi jednotlivými segmenty.

7.5.2 Automatizovaný skóring zákazníků

Společnost ve které pracuji má poměrně velké množství zákazníků, které dokáže identifikovat a disponuje databází s jejich obchodní nebo i marketingovou historií. Zákazníků může být několik set, ale také několik set tisíc nebo ještě více. Je pravděpodobné, že několik procent všech největších zákazníků obsluhují individuálně odpovědní obchodní manažeři, kteří podrobně znají jejich situaci a umějí jim nabídnout odpovídající produkty nebo služby. Jak však pracovat se zbývající "masou"? Jak neztratit v její šedi budoucí nejlepší zákazníky? Jak v nepřehledném množství najít rizikové zákazníky? A jak přeci jen ověřit případná opomenutí obchodních manažerů i u těch největších zákazníků? Tedy – jak odpovědět na otázky typu:

- Kteří zákazníci od nás pravděpodobně v nejbližších dvou měsících odejdou?
- Kteří zákazníci nejpravděpodobněji koupí náš produkt ABC, a je jim ho tedy možné nejefektivněji nabídnout?
- Kterým zákazníkům lze s malým rizikem nabídnout lepší cenové podmínky?
- Kteří zákazníci se pravděpodobně chovají nebo v budoucnu začnou chovat nestandardně – např. objednání telekomunikačních služeb s úmyslem nezaplatit atd.
- Kteří neaktivní zákazníci mohou být pravděpodobně aktivováni nabídkou věrnostního programu ?

Myšlenkou skórování je přiřadit a periodicky – např. měsíčně – aktualizovat individuálně pro každého zákazníka jedno nebo více skóre jako jsou Pravděpodobnost odchodu v nejbližším období nebo Marketingový segment zákazníka. Jinými příklady skóre zákazníka jsou například vyčíslení indikativní nebo dlouhodobé očekávané hodnoty zákazníka (Customer Value, Lifetime Value)

Koncovým výstupem skórování pak může být výpis zákazníků s největší pravděpodobnosti odchodu či souborný pokyn call centru nabídnout určitý produkt zákazníkům z určitého marketingového segmentu, nebo mohou mít obchodní manažeři či pracovníci call centra skóre k dispozici on-line, např. v průběhu každého telefonického kontaktu se zákazníkem.

Má-li být skórování automatizovaným procesem, musí se provádět na základě datových údajů uložených v podnikových nebo i v externích databázích.

Pro skórování se typicky využívají následující datové okruhy:

- Behaviorální data – jde o informace, jakým způsobem se v definovaném časovém období zákazníci chovali, v případě zákazníků banky může jíž např. o počet transakcí, počet výběrů z bankomatu, průměrnou velikost transakce, informace o nárůstu nebo poklesu počtu transakcí v posledním období oproti předcházejícímu atd. Do oblasti behaviorálních dat se mohou počítat údaje o výnosech, příp. profitabilitě zákazníků.
- Produktová data – zde se využívají údaje o tom, které produkty zákazník ve sledovaném období využíval, jak často atd.
- Demografická data – u fyzických osob jde o údaje týkající se např. věku, pohlaví, místa bydliště a podobně. U organizací se může jednat o obrat, segment obchodu, druh vlastnictví atd.
- Kontaktní data – zaznamenávají se údaje o reakcích na určité marketingové kampaně, o počtu a druhu dotazů zákazníka na "horkou linku" atd.
- Externí data – v případě, že jsou k dispozici, např. velikost sídla zákazníka atd.

Zdaleka ne vždy jsou k dispozici všechny uvedené okruhy údajů. Nezřídka se pro vytvoření skórovacího mechanismu využívají např. pouze behaviorální data.

Přestože použité modelování v této oblasti využívá velmi sofistikované postupy, není a nesmí být jen jakousi matematickou úlohou. Praxe ukazuje, že úspěch realizace skórovacího mechanismu těsně souvisí se zapojením obchodně odpovědných pracovníků. Především na nich totiž spočívá rozhodnutí, zda nalezené modely mají dobré obchodní využití. Příklad: Predikční model pro indikaci pravděpodobnosti odchodu zákazníků lze "vyladit? Tak, aby označoval buď větší množství potenciálně odcházejících zákazníků (objevuje se pak větší množství "falešných poplachů", které mohou znamenat větší režiii při ověřování),

nebo naopak může označovat jen menší množství zákazníků, u kterých je pravděpodobnost odchodu největší (určité množství zákazníků, kteří odejdou, pak může být opominuto)

Pro koncové uživatele, např. pracovníky call centra, jsou pak klíčová, kromě vlastního skóre, obchodní pravidla, která jsou sestavena v závěrečné fázi modelování. Příkladem obchodního pravidla může být: Je-li pravděpodobnost, že zákazník odejde, větší než 70%, a současně je dlužná částka je menší než průměrný měsíční obrat, a současně jeho průměrný měsíční obrat je větší než 100 000 Kč, pak proveď ověřovací dotaz na spokojenost s poskytovanými službami.

Zatímco modely se vytvářejí na základě historických dat, při aktualizaci skóre se využívají co nejčerstvější údaje.

Skórovaní procedura je v rámci zprovoznění (ang. Deployment) včleněna do informačního systému organizace, typicky jako součást datového skladu nebo tzv. analytického CRM.

Dolování dat je proces extrakce relevantních předem neznámých nebo nedefinovaných informací z velmi rozsáhlých databází. Aplikace dolování dat zpravidla nepracují s agregovanými OLAP kostkami předem nadefinovaných struktur dat, ale přímo s daty uloženými v datovém skladu podniku nebo v provozních systémech.

Úlohy dolování dat se dělí na:

- Explorační analýzy dat
- Deskriptivní úlohy
- Prediktivní úlohy
- Hledání vzorů a pravidel (hledání nuggetů)
- Hledání podle vzorů

Mezi techniky používané při řešení úloh dolování dat patří:

- Analýza nákupního košíku
- Dedukce
- Detekce shluků
- Analýza závislostí
- Rozhodovací stromy a indukce

- Neuronové sítě
- Genetické algoritmy

Dolování dat je v současné době využíváno v komerční sféře zejména v bankovnictví (např. kreditní skóring či detekce podvodů), telekomunikacích, v oblasti řízení jakosti a v marketingu (např. segmentace zákazníků). Mimo tyto čistě komerční sféry se Data Mining hojně využívá také v lékařství, meteorologii, dopravě a v jiných oblastech.

7.6 Přístup k datům

7.6.1 Přístup k databázím OLAP

7.6.2 Programy balíku Microsoft Office jako klienti analytických služeb

Z pohledu uživatele je nejlepší taková klientská aplikace, kterou uživatel důvěrně zná, tedy ideální případ je integrace přístupu k analytickým službám do aplikací, s nimiž klienti doposud pracovali. Mohou to být různé speciální jednoúčelové aplikace, ale když se zamyslíme nad tím, jaký typ softwaru se na klientských počítačích, tedy na běžných pracovních stanicích, které máme ve svých kancelářích, nejčastěji používá, bezpochyby zvítězí kancelářské balíky. Tuto filosofii si uvědomila i firma Microsoft a v poměrně velkém rozsahu integrovala podporu analytických služeb do svého kancelářského balíku Microsoft Office. Když se připojíme k serveru OLAP pomocí klientské aplikace, například pomocí programu Microsoft Excel, sumarizované hodnoty vypočítává server OLAP, klientská aplikace tyto údaje jen zobrazuje, případně je dále zpracovává nebo například ukládá jako pohledy. Při zobrazení nebo změně sestavy se proto mezi serverem OLAP a klientskou aplikací posílá poměrně málo údajů.

Nejen tabulkový procesor MS Excel a kancelářský databázový program MS Access ale mohou pracovat s analytickými údaji. Vynikající myšlenka je možnost nainstalovat a používat Smart Tags (chytré značky). Potom když například píšeme nějakou zprávu nebo analýzu v textovém editoru Microsoft Word a napíšeme slovo, které se vyskytuje v analytické databázi, například píšeme o svých produktech nebo filiálkách firmy, případně geografických lokalitách, toto slovo se zvýrazní, když na něj aktivujeme Smart Tags, tak můžeme do textu vložit údaj, tabulku, nebo dokonce graf, který je výsledkem analýzy pro danou

entitu. Nebo si jen otevřeme výsledky analýzy v programu MS Excel a na základě těchto podkladů napíšeme do textu zprávy vlastní vyhodnocení analyzované situace.

8 ŘEŠENÍ NA MS SQL 2005 SERVER

8.1 Implementace BI v SQL Server 2005

Když porovnáme analytické služby SQL nového Serveru 2005 s momentálně komerčně dostupnou verzí SQL Server 2000, najdeme mnohá vylepšení návrhového prostředí, ale nejvýznamnější rozdíl je v filozofii. Podle různých průzkumů 5 až 10 procent uživatelů používá výsledky analýz, 15 až 25 procent uživatelů tyto informace zkoumá a hledá v nich souvislosti. Největší skupina uživatelů informace používá ve formě různých výpisů a reportů.

Na podnikové úrovni se generují různé druhy reportů například pro obchodní oddělení, finanční oddělení, oddělení lidských zdrojů, ve sféře CRM a podobně. Údaje jsou buď v podnikových databázích anebo v datových skladech. Výhodou je, že údaje jsou už předzpracované a přetransformované v etapě ETL, a přenesené z produkčních systémů do datových skladů (data warehouse), případně datových trhů (data mart). Reporty z reportovacích služeb potom vhodně doplňují údaje z analytických business intelligence aplikací. Nasazení reportovacích služeb je v tomto případě převážně na úrovni podnikových portálů, takže koncoví uživatelé k nim přistupují v rámci podnikového intranetu.

8.2 Srovnání Microsoft SQL Server 2005 s jinými řešeními pro BI

8.2.1 Možnosti

SQL Server 2005 poskytuje škálovatelné nástroje pro business intelligence zpřístupněním důležitých informací uživatelům na všech úrovních organizace, díky kterým se mohou lépe a rychleji rozhodovat.

- Komplexní funkce pro analýzu, integraci a migraci dat napomáhají zákazníkům zvyšovat hodnotu jejich stávajících aplikací.
- Služba Analysis Services umožňuje organizovat data do intuitivních struktur pro podporu předdefinovaných i jednorázových dotazů, které dokáží identifikovat pravidla, vztahy a trendy.
- Komplexní platforma pro generování sestav umožňuje vytvářet sestavy v reálném čase i podle definovaných časových plánů. Tyto sestavy jsou přístupné z webového

prohlížeče, známých kancelářských aplikací i specializovaných obchodních nástrojů.

- Služba Reporting Services je k dispozici ve všech verzích.

8.2.2 Výkon

Srovnání výkonu ve světově uznávaných benchmarcích TPC-H (Transaction Processing Performance) simulujících provoz rozsáhlých informačních systémů se MS SQL Server 2005 dostal ve většině hlavních kategorií na čelní místa, to znamená že jeho poměr cena / výkon vychází lépe ve srovnání s ostatními konkurenty.

Oproti verzi SQL 2000 a i oproti konkurenčním moderním databázovým strojům, které obsahují nástroje pro BI (např. Oracle 10g nebo IBM DB2 Stinger) přichází MS SQL 2005 s novinkou, která právě v oblasti analytických a reportovacích nástrojů znamená dfalší krok dopředu co se týýče zvýšení výkonu. Je to zásadní vylepšení stávajícího mechanizmu MOLAP (multi-dimenzionální on-line analytické zpracování) - nasazení vyrovnávací MOLAP paměti, kam se ukládají nejčastěji pozužívané anebo předpokládané výsledky agregací. Toto proaktivní cachování a predikce výsledků je u SQL 2005 plně automatizované. Udaje tedy na rozdíl od konkurenčních verzí zústávají v relačních databázích a agregované údaje se ukládají do multi-dimenzionálních struktur. Při dotazování se pak údaje ukládají do multi-dimenzionální paměti cache.

8.2.3 Bezpečnost

Microsoft na rozdíl od Oracle má ve všech svých verzích kladem maximální důraz na bezpečnost (Express, Workgroup, Standard, a Enterprise mají všechny bezpečnostní funkce shodné), Oracle má maximální úroveň zabezpečení pouze v nadstandartní verzi Oracle 10g Advanced Security Option.

8.2.4 Cena

8.2.5 Shrnutí

SQL Server 2005 je vysoce produktivní databázová platforma, která vývojářům umožňuje rychleji vytvářet a nasazovat důležité obchodní aplikace. Díky úzké integraci se sadou Microsoft Visual Studio 2005 poskytuje SQL Server 2005 vývojářům novou úroveň mož-

ností. Oproti konkurenci ve všech verzích nabízí maximální úroveň zabezpečení, dále pak obsahuje vynikající nástroje pro analytické zpracování dat a reporting.

Ve výkonových testech a nákladech předčil své největší konkurenty jakými jsou Oracle 10g a IBM DB2. Další nespornou výhodou je těsná integrace s operačním systémem MS Windows Server a jednoduchá autorizace uživatelů (předpokládali jsme síť s Active Directory)

8.3 Novinky

Následující tabulka popisuje nejužívanější komponenty Business Intelligence a jim odpovídající integrované nástroje v SQL Server 2000 a v SQL Server 2005.

Tabulka 2 Srovnání komponent MS SQL Server 2000 a 2005

Komponent	**SQL Server 2000**	**SQL Server 2005**
Integrace, transformace a přesun dat	Data Transformation Services (DTS)	SQL Server 2005 Integration Services
Relační data warehouse	SQL Server 2000 relační databáze	SQL Server 2005 relační databáze
Multidimenzionální databáze	SQL Server 2000 Analysis Services	SQL Server 2005 Analysis Services
Data mining	SQL Server 2000 Analysis Services	SQL Server 2005 Analysis Services
Řízený reporting	SQL Server 2000 Reporting Services	SQL Server 2005 Reporting Services
Ad hoc reporting	Not applicable	SQL Server 2005 Reporting Services
Ad hoc dotazy a analýzy	Microsoft Office produkty (Excel, Office Web Components, Data Analyzer, SharePoint Portal Server)	Microsoft Office produkty (Excel, Office Web Components, Data Analyzer, SharePoint Portal Server)
Vývojové nástroje pro	SQL Server 2000 Enterprise	SQL Server 2005 Business

Komponent	SQL Server 2000	SQL Server 2005
databázový server	Manager, Analysis Manager, Query Analyzer, různé další nástroje	Intelligence Development Studio (Novinka)
Nástroje pro správu databázového serveru	Enterprise Manager, Analysis Manager	SQL Server Management Studio (Novinka)

8.4 Integration Services

Nová součást produktu SQL Server 2005, služba SQL Server Integration Services (SSIS), nahrazuje službu transformace dat Data Transformation Services pro SQL Server 2000 a poskytuje funkce a výkon potřebné k vytváření aplikací pro integraci dat na úrovni podniku.

Vytvoření sledu procesů pro data v celé organizaci může být jedním z nejobtížnějších úkolů. Služba Integration Services usnadňuje vytváření výkonných aplikací pro integraci dat na úrovni podniku.

8.5 Reporting Services

Služba SQL Server Reporting Services je všestranné serverové řešení pro vytváření sestav navržené pro usnadnění vytváření, správy a zasílání papírových i interaktivních webových sestav.

8.6 Analysis Serivices

Díky kombinaci nejlepších aspektů tradiční analýzy OLAP (Online Analytical Processing) a vytváření relačních sestav poskytuje služba SQL Server 2005 Analysis Services (SSAS) model metadat, který vyhovuje veškerým datovým požadavkům.

Služba SQL Server 2005 Analysis Services poskytuje integrované zobrazení obchodních dat pro účely vytváření sestav, analýzy OLAP, přehledů klíčových ukazatelů výkonu (KPI) a dolování dat.

8.7 Notificcation Services

Služba SQL Server 2005 Notification Services (SSNS) je platforma pro vývoj a nasazení aplikací generujících a odesílajících uživatelům přizpůsobená upozornění – včasné zprávy, které lze odeslat na různá zařízení.

8.8 MS Business Intelligence Development studio

Novou částí SQL Serveru 2005 je vývojový nástroj umožňující např. tvorbu reportů, tvorbu ETL procesů, analytické operace, ale i vývoj komplikovanějších procesů pro BI. Je založeno na MS Visual Studio 2005, ale uzpůsobeno právě pro potřeby BI.

9 NÁVRH ŘEŠENÍ

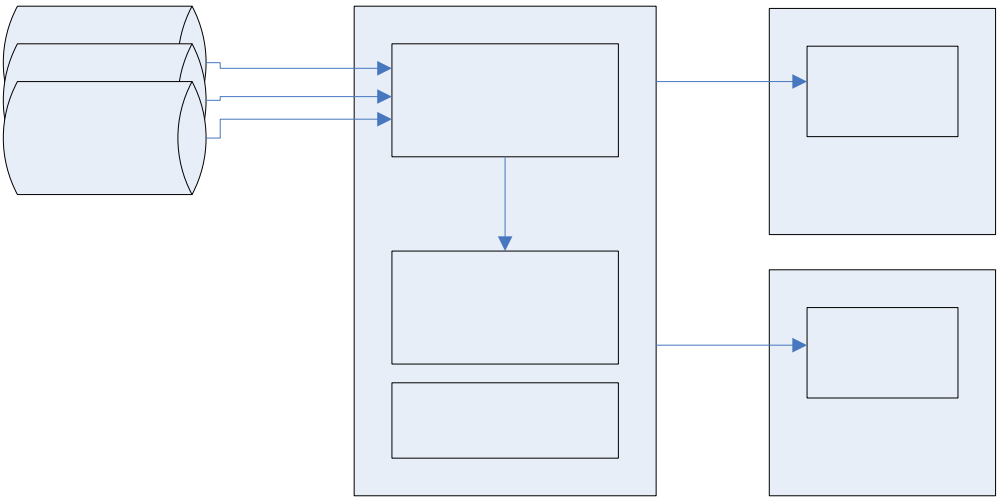
Vybudování datového skladu ve společnosti je rozsáhlý a dlouhodobý projekt, který by měl směřovat ke zlepšení schopnosti pružně reagovat na požadavky obchodníků, produktového managementu i vyhovět požadavkům top managementu na získání informací potřebných pro rozhodování. Datový sklad by měl sjednotit heterogenní prostředí a umožnit lépe sledovat vývoj společnosti (více sloučených společností s různými informačními systémy a zdroji dat).

Vzhledem k rozsahu projektu je nutné použít postupného (přírůstkového) budování datového skladu, na základě priorit obchodních požadavků. Každý z přírůstků lze specifikovat jako samostatný projekt, který bude obsahovat požadavky ze strany obchodu, analýzu řešení BI a specifikaci implementace BI řešení. Popis jednotlivých přírůstků jsou na rámec této práce, mají však společný základ v podobě jednotné koncepce řešení, a společné jednotlivé prvky řešení BI.

9.1 Architektura a koncepce datového skladu

V současné době existuje ve společnosti řada informačních systémů, které obsahují informace o zákaznících, produktech, partnerech a další. Charakter těchto informací se systém od systému výrazně liší a proto existují i různé pohledy na jednotlivé segmenty informací.

Datový sklad GTS Novera bude obsahovat historické informace o chování společnosti a jejích zákazníků. Bude se obnovovat pomocí denních snímků a bude postaven jako jednotný konsolidovaný datový sklad. Prostřednictvím předmětně orientovaných datových tržišť bude poskytovat informace pro zpracovávání analýz.



Obrázek 2 Znázornění architektury DWH

9.2 Doporučený hardware a software

Vzhledem k velikosti společnosti (sloučení více telekomunikačních společností do jedné z největších firmy v rámci ČR působící v tomto segmentu, obsahující velké množství zdrojových systému a velké množství samotných dat) je nutno počítat s náročnými nároky na výpočetní výkon a celkový HW i SW. Vzhledem k tomu a i s ohledem na stabilitu a dostupnost systému bylo zvoleno řešení na 64-bitových serverech moderní architektury IA64 a x64, a softwarové platformě Microsoft v 64-bitové distribuci.

Primární systémy

Datový sklad bude tvořit soustava serverů:

DWH01 - primární datová sklad (nultá, první a uživatelská vrstva, ETL)

- HW: HP Integrity Superdome, CPU 4 x Intel Itanium 1,6 GHz, RAM 64 GB
- SW: MS Windows Server 2003 Datacenter Edition, MS SQL Server 2005 64-bit Enterprise Edition

DWH02 - Reportovací server

- HW: HP ProLiant Server, CPU 4 x Intel Xeon MP 2,8 GHz, RAM 32 GB
- SW: MS Windows Server 2003 Enterprise x64 Edition, MS SQL 2005 Enterprise x64 Edition

DWH03 - OLAP server (zpracování MOLAP, analytické služby)

- HW: HP ProLiant Server, CPU 4 x Intel Xeon MP 2,8 GHz, RAM 32 GB
- SW: MS Windows Server 2003 Enterprise x64 Edition, MS SQL 2005 Enterprise x64 Edition

Datové úložiště - servery budou využívat datové úložiště HP XP12000 které má rozšiřitelnost až na 332 TB, plánováno je 64 disků s celkovou kapacitou cca 15 TB. Předpokládaná velikost databáze samotného datového skladu je 3 až 4 TB, další kapacita je potřeba pro OLAP databáze, a pro zálohu některých historických neagregovaných detailních dat.

9.3 ETL a integrační procesy

ETL (Extrakce, Transformace, Loading) – plnění datového datového skladu z primárních zdrojů bude probíhat s využitím nové architektury MS SQL Serveru 2005 (integrační služby)

Prostředí ve společnosti je tvořeno soustavou heterogenních zdrojů dat, primárně se jedná o OLTP systémy:

- telekomunikační ústředny a ISP technologické systémy
- CRM systémy
- retail billingové systémy
- wholesale billingové systémy

Heterogennost prostředí je taktéž dána různorodostí platforem, na kterých jsou výše uvedený systémy provozovány. Ve firmě se tedy vyskytují systémy Oracle 8 a 9, MS SQL 2000 a 2005 a MySQL, běžících na různých platformách. Zdroje dat z telekomunikačních ústředen mohou být loadovány z externích textových souborů a další zdroje mohou být poskytnuty ve formátu XML

Všechny zdroje dat jsou dostupné v rámci rozsáhlé lokální sítě, resp. soustavou geograficky oddělených lokálních sítí spojených do jedné globální sítě přes WAN sítě zabezpečenými tunelovými spoji (VPN). S rychlostí přístupu ke zdrojům dat tedy není problém.

Plnění datového skladu bude probíhat periodicky v nočních hodinách (jednou denně), zároveň se budou i generovat datové kostky pro sekundární OLAP server. Proces bude spolu

s integrací dat provádět i auditování jednotlivých činností pro případnou kontrolu správnosti ETL procesu a notifikovat administrátory v případě nekorektního průběhu.

9.4 DHW a OLAP systémy

Datový sklad bude rozdělen do několika vrstev:

nultá vrstva – relační databáze DWZ bude obsahovat kopii dat z primárních systémů, předpona jednotlivých tabulek bude identifikovat zdrojový systém.

první vrstva – relační databáze DW bude obsahovat faktové tabulky (DW_F_…), dimenzionální tabulky (DW_D_...), řídící tabulky (DW_S_...) a pracovní tabulky a pohledy

druhá vrstva – OLAP (Online Analytical Processing) běžící na sekundárním serveru s MS SQL 2005 a Analysis Services, tvoří jednotlivé OLAP kostky na sekundárním serveru

databáze DWU – zdroje nezahrnuté do integračních služeb, tzn. neimportují se data z primárních systému ale udržují se ručně

9.4.1 Dimenze a fakta

Detailní popis tabulek dimenzí a faktů je součástí až jednotlivých specifikací implementace přírůstku datové skladu. Uveďme alespoň ty nejzákladnější použité dimenze:

- DW_D_CUSTOMER (dimenze zákazníka)
- DW_D_ACCOUNT (dimenze billingového zákazníka - fakturace)
- DW_D_DATE (dimenze času)
- DW_D_PRODUCT (dimenze produktu)
- DW_D_SERVICE (dimenze konkrétní instance produktu na zákazníkovi)
- DW_D_DEALER (dimenze obchodníka)

9.4.2 Zdrojové systémy

Konkrétní zdrojové systémy jsou součástí jednotlivých specifikací přírůstků implementace. Zdrojové systémy budou voleny podle priorit projektu, a podle budoucí existence samotného zdrojového systému. V první fázi předpokládáme systémy Clarify, Durian, OC a Salwin.

9.5 Reporting

Podle průzkumů 65 až 80 % běžných uživatelů BI služeb požaduje výstupy ve formě reportů. K podobnému výsledku jsem dospěly i při analýze požadavků na BI. Vyhradíme tedy zvláštní server na službu reportovacího serveru (MS SQL 2005) Reporting Services, převážně formou Enterprise reportingu (výhodou je, že údaje budou předzpracované a přetransformované v etapě ETL – denním loadu do datového skladu) a budou tak vhodně doplňovat údaje z analytických BI aplikací.

Přesné definice reportů budou ještě upřesněny na základě dohody a požadavků konkrétních oddělení, stejně tak jako jejich platnost či oprávnění uživatelů a skupin pro přístup k jednotlivým reportům.

Dále se počítá i s možností tvorby ad-hoc reportů pomocí nástroje Report Builder, ale toto bude jen v ojedinělých případech.

9.6 Další BI nástroje

Předpokládá se využití nástrojů které jsou součástí MS SQL Server 2005, zejména tedy nástrojů pro datamining. Zůstává ale i možnost použití nástrojů třetích stran (Enterprise Miner, Clementine, ...).

9.7 Přístup k datům

Přístup k datům v datovém skladu bude zajištěn na úrovni oprávnění definovaných administrátory datového skladu pro jednotlivé uživatele, kteří pak budou moci přistupovat např. přes ODBC nástroji třetích stran anebo přímo z aplikací MS k jednotlivým tabulkám v DW či kostkám na OLAP serveru.

ZÁVĚR

V této diplomové práci jsem provedl návrh datového skladu pro konkrétní telekomunikační společnost, který by mohl být použit i obecněji. Návrh je realizován s pomocí Microsoft SQL 2005.

V kapitole 6. a 7. jsem popsal výchozí stav a situaci firmy, ve které bude implementován centrální datový sklad a nástroje pro business inteligence. Dále jsem zanalyzoval potřeby telekomunikační firmy na nástroj business inteligence a data warehousing. Z této analýzy plyne následující závěr: Vzhledem ke složitosti zdrojových systémů, a tomu že v tuhle chvíli není jasné, které systémy zaniknou a které zůstanou, je nutné budovat BI postupnou přírůstkovou metodou.

V kapitole 8. a 9. jsou analyzovány možnosti nástrojů business inteligence, je vybrán konkrétní produkt a jsou uvedeny jeho hlavní přednosti pro výsledné řešení.

V samotném závěru práce je obsažen návrh řešení systému jako celku, které zajišťuje základní koncept řešení BI.

Popsal jsem (viz kapitola 1.5. a 7.) základní nástroje business intelligence, které nad datovým skladem mohou operovat. Řešení nabízí snadnou tvorbu analýz nad relačně i multidimenzionálně uloženými daty a jejich publikování v rámci organizace i mimo ni. Přináší též možnost vývoje vlastních analytických aplikací, správu metadat celého řešení, návrh, tvorbu a monitorování datových úložišť.

V práci jsem popsal základní techniky nástrojů business inteligence, jimiž jsou data mining a jeho základní principy a techniky, online analytical processing (OLAP) a jeho případné začlenění do systémů a propojení s ostatními částmi celku.

Práce tedy navrhuje komplexní řešení pro telekomunikační firmu, které po dokončení implementace a citlivém nasazení a proškolení umožní podporu rozhodování vedoucích pracovníků a manažerů firmy na základě dat, která budou zpracována technikami data miningu a online analytical processingu (OLAP) do podoby vhodné k publikování a vnitropodnikovým i externím prezentacím a pomohou zvýšit konkurenceschopnost firmy na trhu.

SEZNAM POUŽITÉ LITERATURY

[1] Lacko, Ľ. *Databáze: datové sklady, OLAP a dolování dat.* Praha: Computer Press, 2003. ISBN 80-7226-969-0.

[2] Humphries, M. *Michael W. Hawkins a kol.: Data warehousing, Principy a praxe.* Praha: Computer Press, 2002. ISBN 8072265601.

[3] Lacko, Ľ. *Business Intelligence v SQL Serveri 2005*. Praha: Microsoft, 2005.

[4] Vieira, R. *SQL Server 2000 Programujeme profesionálně.* Praha: Computer Press, 2001. ISBN 8072265067.

[5] Iseminger, D. *Microsoft SQL Server 2000 Reference Library*. Redmond: Microsoft Press, 2001.

[6] Novotný, O., Pour, J., Slánský, D. *Business Intelligence Jak využít bohatství ve vašich datech.* Praha: Grada Publishing, 2005. ISBN 80-247-1094-3.

[7] Arlow, J., Neustat, I.: *UML a unifikovaný proces vývoje aplikací.* Praha: Computer Press, 2003. ISBN 80-7226-947-X.

[8] Kotler, P., Armstrong, G. *Marketing.* Praha: Grada Publishing, 2004. ISBN 8024705133.

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

BI	Business Intelligence.
DWH	Data Warehouse, datový sklad.
SQL	Význam třetí zkratky.
OLAP	OnLine Analytical Processing

SEZNAM OBRÁZKŮ

SEZNAM TABULEK