

Technická univerzita v Liberci

Hospodářská fakulta

Studijní program: 6209T - Systémové inženýrství a informatika

Studijní obor: M6209 – Manažerská informatika

Datové sklady v informačních systémech

Data warehouses in information systems

Číslo závěrečné práce: DP-MI-KIN-2004-11

PAVEL MĚCHURA

Vedoucí práce: Ing. Vladimíra Zádová (Katedra informatiky)

Konzultant: Ing. Robert Kleiner (truconneXion a.s.)

Počet stran: 73

Počet příloh: 0

Datum odevzdání: 5.1. 2004

UNIVERZITNÍ KNIHOVNA
TECHNICKÉ UNIVERZITY V LIBERCI



3146078676

ZADÁNÍ DIPLOMOVÉ PRÁCE

pro:

Pavel Měchura

Studijní program:

Systémové inženýrství a informatika (6209T)

Studijní obor č. M 6209

Manažerská informatika

Vedoucí katedry Vám ve smyslu zákona č. 1111/1998 Sb. o vysokých školách a navazujících předpisech určuje tuto diplomovou práci :

Název tématu:

Datové sklady v informačních systémech

Zásady pro vypracování:

1. Koncept datového skladu.
2. Architektura a hlavní komponenty datového skladu.
3. Nástroje pro řešení a využití datových skladů.
4. Specifika řešení datových skladů a datových tržišť.

Rozsah diplomové práce: 50-60

(do rozsahu nejsou započítány úvodní listy, přehled literatury a přílohy)

Doporučená literatura :

Rud, O.P.: Data Mining, computer Press, 2001

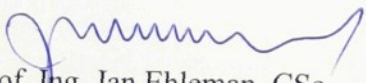
Dohnal, J., Pour, J.: Architektury informačních systémů, Ekopress, 1997

Kimball, R., Mergy, R.: The Data Warehouse Toolkit – Multidimensional Modelling, Willey, 2002

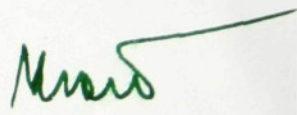
Vedoucí diplomové práce: Ing. Vladimíra Zádová

Odborný konzultant: Ing. Dana Nejedlová

Termín odevzdání diplomové práce: 5.1.2004


Prof. Ing. Jan Ehleman, CSc.
vedoucí katedry




Prof. Ing. Jiří Kraft, CSc.
děkan Hospodářské fakulty

V Liberci dne 31.3.2003

Prohlášení

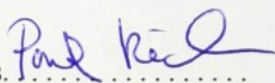
Prohlašuji, že jsem diplomovou práci vypracoval samostatně s použitím uvedené literatury pod vedením vedoucího a konzultanta. Byl jsem seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 o právu autorském, zejména §60 (školní dílo) a §35 (o nevýdělečném užití díla k vnitřní potřebě školy).

Beru na vědomí, že TUL má právo na uzavření licenční smlouvy o užití mé práce a prohlašuji, že souhlasím s případným užitím mé práce (prodej, zapůjčení apod.)

Jsem si vědom toho, že užití své diplomové práce či poskytnutí licence k jejímu užití mohu jen se souhlasem TUL, která má právo ode mne požadovat přiměřený příspěvek na úhradu nákladů, vynaložených univerzitou na vytvoření díla (až do její skutečné výše).

Po pěti letech si mohu tuto práci vyžádat v Univerzitní knihovně TU v Liberci, kde je uložena, a tím výše uvedená omezení vůči mé osobě končí.

V Liberci dne 21. 2004

Podpis. 

Resumé

Cílem této práce je obecně charakterizovat komplexní problematiku v oblasti datových skladů. Zaměřuje se na typické problémy související s procesem získávání, ukládání a zpracování dat, se kterými se podnik střetává ve svém běžném hospodářském životě a které mohou případně ovlivnit jeho konkurenceschopnost. Blíže představuje nejvyužívanější přístupy a postupy k budování datových skladů, se kterými se lze setkat. Je zaměřena také na využívání datového skladu pomocí nástrojů pro koncového uživatele, zvláště pak na oblast OLAP a dolování dat. Stručně popisuje několik produktů, které mají v oblasti datových skladů v České republice značný tržní podíl. V poslední kapitole se blíže věnuje problematice tvorby datových skladů pro dynamické prostředí internetu.

Summary

The goal of this thesis is the description of global problems in process of data warehousing. It discusses typical problems related to gaining, saving and manipulation of data, which can influence the company's competitive advantage. It maps the most used accesses and processes for data warehouse building which we can possibly meet. It also analyzes end user tools of data warehousing, especially the OLAP and data mining. It describes in short several products with great market share in the data warehousing sphere in the Czech Republic. The last chapter specifies problems with building of data warehouse for the internet environment.

Poděkování

Tímto bych chtěl poděkovat paní Ing. Vladimíře Zádové za její podnětné připomínky a cenná doporučení týkající se tématu této diplomové práce.

Dále bych chtěl poděkovat Ing. Robertu Kleinerovi ze společnosti truconneXion a.s. za konzultace a předání praktických zkušeností z oblasti implementace datových skladů.

OBSAH

1	ÚVOD.....	10
1.1	Cíl práce.....	10
2	KONCEPT DATOVÉHO SKLADU	11
2.1	Od dat ke znalostem.....	11
2.1.1	Data	11
2.1.2	Informace	12
2.1.3	Znalosti.....	12
2.1.4	Moudrost	13
2.1.5	Informační zahlcení.....	13
2.2	Business intelligence.....	15
2.2.1	Obecné schéma architektury moderních informačních systémů	16
2.2.2	Provozní systémy vs datové sklady	19
2.2.3	Provozní systém	19
2.2.4	Nevýhody OLTP systémů	20
2.3	Datový sklad - podstata	21
2.3.1	Rozdíly mezi provozními systémy a datovým skladem	23
2.3.2	Datová tržiště – datamarty.....	24
2.3.3	Využití datového skladu.....	26
2.3.4	Účel datového skladu	27
2.3.5	Uživatelé datového skladu	28
2.3.6	Náklady datového skladu	30
3	ARCHITEKTURA A HLAVNÍ KOMPONENTY DATOVÉHO SKLADU	31
3.1	Přístupy k budování datového skladu	31
3.1.1	Centralizovaný datový sklad- Bill Inmon.....	32
3.1.2	Sjednocené data marty – Ralph Kimball	37
3.2	Datové pumpy – ETL procesy.....	38
3.2.1	Proces ETL.....	39
3.2.2	Extrakce vstupních dat	39
3.2.3	Transformace vstupních dat	40
3.2.4	Uložení dat	40
3.3	Databáze datového skladu.....	41
3.4	OLAP	43
3.4.1	Multidimenzionální architektura	45
3.4.2	Základní operace v OLAP systémech	48
3.5	Data Mining.....	49
3.5.1	Zdroje dat	49
3.5.2	Data mining a datové sklady	50
3.5.3	Hlavní charakteristiky a cíle dolování v datech.....	51
3.5.4	Rozdíl mezi OLAP a dolováním dat	51
3.5.5	Využití data miningu.....	51
3.5.6	Metody data miningu	52

3.5.7	Nasazení projektu dolování dat	54
4	NÁSTROJE PRO ŘEŠENÍ A VYUŽITÍ DW	55
4.1	Přehled řešení datových skladů v ČR.....	55
4.1.1	ORACLE.....	56
4.1.2	MICROSOFT.....	57
4.1.3	AIMTEC a.s.	58
4.1.4	SYBASE	59
4.1.5	SAS Institute ČR, s.r.o.	60
5	SPECIFIKA ŘEŠENÍ DW A DM.....	61
5.1	Využití DW v prostředí internetu.....	61
5.1.1	Zákazník.....	61
5.1.2	Clickstreamová data	63
5.1.3	Clickstream analýza	65
5.1.4	Návrh schématu datového skladu pro clickstream analýzu.....	66
6	ZÁVĚR.....	72
7	SEZNAM LITERATURY	73

Seznam použitých zkratk

atd.	a tak dále
apod.	a podobně
BI	business intelligence, obchodní inteligence
CD	conformed dimensions, odpovídající dimenze
CF	conformed facts, odpovídající fakta
CPU	Central Procesor Unit, počítačový procesor
CRM	Customer Relationship Management, řízení vztahů se zákazníky
č.	číslo
DS	datový sklad
DSS	Decision Support Systems, systémy pro podporu rozhodování
DW	data warehouse, datový sklad
DM	data mart, datové tržiště
EIS	Executive Support Systems, systémy pro vrcholové řízení
GB	gigabyte
IS	informační systém
MIS	Management Information Systems, manažerské informační systémy
ODS	Operational Data Store, operativní úložiště dat
OLAP	On-line Analytical Processing, on-line analytické zpracování
OLTP	On-line Transaction Processing, on-line transakční zpracování
resp.	respektive
USD	americký dolar
SQL	Structured Query Language, strukturovaný dotazovací jazyk
TCO	Total Cost of Ownership, celkové náklady na vlastnictví
TPS	Transaction Processing Systems, transakční systémy
tzn.	to znamená

1 Úvod

Ve vývoji počítačového zpracování informací můžeme rozeznat jeden dominantní trend. Množství dat, které je potřebné zpracovat se neustále zvyšuje. Oproti tomu existuje jasná snaha sjednotit metody přístupu k těmto datům. V první fázi vývoje zpracování dat byla potřebná data obsažena v jednom modulu s programovou logikou, která je zpracovávala. Data vznikala a zanikala současně s během programu. Dat bylo poměrně malé množství, ale programů, které je zpracovávaly bylo mnoho - prakticky každý program byl unikát co se týkalo generovaných dat i programové logiky. K zásadnímu předělu došlo v okamžiku oddělení programových a datových oblastí. Data se začala ukládat do speciálních datových struktur nezávislých na programovém kódu, který je zpracovával. Tyto datové struktury byly přímými předchůdci současných databází. Postupem času a s rostoucími potřebami uživatelů se musely databáze transformovat do podoby, kterou dnes někdo nazývá datovým skladem. Datový sklad znamená různé věci pro různé lidi a pro některé neznamena vůbec nic. Co to vlastně je ? Stojí za to ? Co může přinést ?

1.1 Cíl práce

Tato práce by měla poukázat na významnou úlohu existence datových skladů v rámci informačního systému podniku. Aby mohlo být dosaženo vytyčeného cíle, bylo nutné zasadit proces data warehousingu do širšího kontextu. Z tohoto důvodu je v rámci první kapitoly zmíněn význam hodnoty informací pro jedince i pro podnik. Je zde popsán proces business intelligence a zdůrazněn význam datového skladu jako důležitého úložiště různých zdrojových dat. Druhá kapitola se věnuje jednotlivým součástem datového skladu a popisuje dvě různá pojetí přístupu k implementaci a pojetí datových skladů. Ve třetí kapitole shrnuji nejvíce využívané produkty, které danou problematiku řeší. Poslední část práce věnuji odlišnostem a specifikům datových skladů v prostředí internetu, protože se domnívám, že řešení, která pokrývají i tuto oblast, budou do budoucna stále více převládat.

2 Koncept datového skladu

2.1 Od dat ke znalostem

„Znalosti a informace jsou dnes jediným smysluplným zdrojem. Tradiční výrobní faktory – půda, práce a kapitál – nezmizely, ale staly se druhořadými. Hlavním producentem bohatství jsou informace a znalosti.“ Peter Drucker

Často se říká, že data jsou tou nejcennější komoditou, kterou mají firmy k dispozici. S tímto konstatováním nelze zcela souhlasit. Tou cennou komoditou nejsou „surová“ data, ale **informace**. Právě informace, které mohou být v datech ukryty, budou či spíše dnes už jsou základní konkurenční výhodou pro celou řadu firem.

Pojmy data, informace, údaje, znalosti jsou velmi rozšířené, ale používání těchto pojmů (ve smyslu jejich popisu) je často velmi nepřesné. I mnozí odborníci tyto termíny používají velice volně a často jako synonyma.

2.1.1 Data

Data lze charakterizovat jako posloupnost znaků. Pod posloupností se mohou skrývat prakticky libovolné znaky a značky. Je důležité si uvědomit, že daná posloupnost znaků sama o sobě představuje „surová“ data, která bez jejich využití či interpretace nic neznamenaají.

Z hlediska počítačové vědy se pojem data používá pro označení čísel, textu, zvuku, obrazu popř. jiných smyslových vjemů, které jsou reprezentované v podobě vhodné pro počítačové zpracování.

Z hlediska práce s daty je možno rozlišovat [1]:

- **Strukturovaná data**

Tyto data explicitně zachycují fakta, objekty, atributy atd., přičemž výrazným rysem je existence určitých elementů dat. Typickým příkladem je ukládání dat pomocí relačních databázových systémů, ve kterých se obvykle používá hierarchie elementů:

pole → záznam → relace → databáze

Díky tomuto strukturovanému uložení je potom možné snadno vybírat jen ta data, která jsou třeba pro řešení určité informační potřeby, např. zjištění průměrné hodnoty nějakého atributu.

- **Nestrukturovaná data**

Lze je vyjádřit jako neupřesněný „tok bytů“. Může jít např. o videozáznamy, zvukové nahrávky nebo obrázky. Do této skupiny je možné zařadit také textové dokumenty.

- **Semistrukturovaná data**

Semistrukturovaná data jsou definována jako data, která jsou neuspořádaná či neúplná, jejich struktura se může měnit, dokonce nepredikovatelným způsobem.

Mohou to být data ve webových zdrojích, HTML stránky, XML data atd.

Data jsou základním materiálem, ze kterého mohou vyvstávat informace. Například data „3,14“ nebo „škoda“ jistě reprezentují něco z reálného světa, ale bez dalšího popisu nebo kontextu nedávají smysl.

2.1.2 Informace

Pojem informace se používá v mnoha vědních disciplínách a oborech, a proto existuje i mnoho definic tohoto pojmu. Informace jsou svázány až s nějakou konkrétní interpretací, konkrétním sémantickým významem obsaženým v datech.. Data tedy přecházejí v informace v okamžiku, kdy jsme schopni z nich získat nějakou hodnotu.

- **Hodnota informace vs cena dat**

Hodnota informace je součástí procesu transformace dat na informace, proto má subjektivní charakter.

Hodnota informace nemá přímou souvislost s případnou cenou dat.

Hranice mezi pojmem data a informace jsou v mnoha případech nejasné. V určitých situacích je prakticky nemožné korektně rozhodnout, zda se jedná o data, která nějakou informaci nesou, nebo ne. Vztahu informací a dat se mimo jiné věnuje i speciální vědní disciplína – teorie přenosu dat.

2.1.3 Znalosti

Za znalost lze považovat informaci, která byla zorganizována a analyzována tak, aby byla srozumitelná a použitelná pro řešení problémů nebo pro rozhodování a učení. V rámci firmy je tato

informace/znalost implementována do firemních postupů a procesů. Pojem znalosti lze popsat v několika bodech:

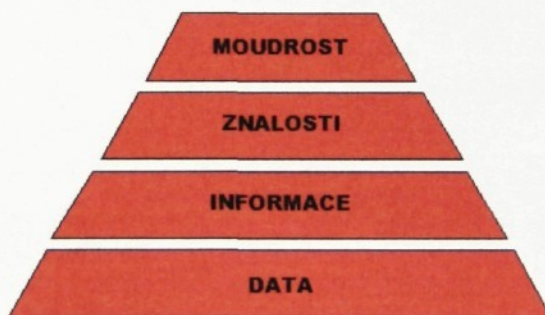
- Jsou základem práce s informacemi
- Mění, rozvíjejí se během procesu učení
- Jsou nutnou veličinou pro proces výběru, rozhodování, interpretace

2.1.4 Moudrost

Tento pojem lze označit jako „schopnost přesného zhodnocení znalostí a jejich následné využití v reálném životě“. [11]

Termín není třeba dále upřesňovat, protože do tématu dotýkajícího se této práce patří především tři předchozí pojmy.

Obr. č. 1: Pyramida hierarchie informačních úrovní



Pramen: vlastní

2.1.5 Informační zahlcení

Fenoménem dnešní společnosti je úspěch. Uspěť na poli osobním či na poli firemním vyžaduje kromě jiných důležitých vlastností, také vhodný přístup k informacím. Úspěšná firma či jedinec musí mít schopnost nalézt vhodnou informaci, analyzovat ji a následně ji použít ve svůj prospěch. Informace v dnešním propojeném světě však vznikají mnohem rychleji a v daleko větších objemech než kam sahají schopnosti člověka.

Tento problém se obecně nazývá *informační exploze*. Informační zahlcení je způsobeno neschopností vytěžit potřebné znalosti z obrovského množství informací.

Dle [1] je toto několik stavů, kdy lze mluvit o informačním přehlčení z pohledu jedince:

- Člověk nedokáže porozumět dostupným informacím
- Člověk nemá tušení, zda nějaké informace existují
- Člověk je zavalen množstvím informací, které je nucen vstřebat
- Člověk neví, kde má dané informace hledat
- Člověk ví, kde dané informace může nalézt, nezná však způsob jak se k nim dostat

Příčiny informačního zahlcení[6]:

Objem informací

Je nesporné, že rostoucí objem informací je tou základní příčinou informačního přehlčení. V závislosti na tomto rostoucím objemu se zvyšuje obtížnost nalezení té správné a vhodné informace. V případě, že roste objem dat, je uživatel nucen prohledávat více materiálu, střetává se s informacemi užitečnými i zbytečnými. Záleží pak pouze na něm, jak je schopen tyto informace roztrždit a pracovat s nimi.

Verifikace přesnosti informací

Tento problém se týká především internetu jako nejvýznamnějšího informačního zdroje současnosti a především budoucnosti. Internet může být zdrojem chybných nebo zbytečných dat, verifikace je proto velmi obtížná.

Informační gramotnost

Informační gramotnost lze charakterizovat jako „schopnost efektivního přístupu a hodnocení informací vzhledem k určité potřebě“ [1]. Tato gramotnost ovlivňuje schopnost uživatelů efektivní práce s informacemi.

Člověk, který je informačně gramotný je schopen:

- Rozpoznat vznik určitého problému a definovat jej
- Definovat otázky, které problém vyvolává
- Identifikovat informace, které jsou třeba pro řešení problému
- Nalézt potřebné informace
- Vyhodnotit nalezené informace
- Informace analyzovat a zakomponovat do řešení

2.2 Business intelligence

Předchozí kapitola se zabývala pojmy data, informace, znalosti z pohledu jedince. Z pohledu podniku je však porozumění těmto pojmům, následné práce s nimi a jejich využití zásadní konkurenční výhodou či nevýhodou. Pro efektivní řízení podniku je nutné porozumět okolním stavům a situacím, vně či uvnitř. Proto podniky začaly postupně budovat své informační systémy. Tyto informační systémy měli sloužit uživatelům jako podpora v jejich každodenních činnostech. Uživatelé těchto systémů byli jak řadoví zaměstnanci, tak i vrcholní manažeři. Spolu s vývojem informačních technologií došlo k postupnému vytvoření dalších, specializovaných systémů jako DSS, EIS, MIS atd. Tyto systémy a jejich souvislosti budou popsány v následující kapitole.

Proces business intelligence lze chápat jako konzistentní a efektivní přístup k práci s firemními daty, který má vliv na správnost strategických a taktických rozhodnutí, a tím posléze i na obchodní úspěch podniku.

Základem procesu business intelligence je přetváření zdrojových (zpravidla transakčních) dat na znalosti, s pomocí nichž jsou následně přijímána správná rozhodnutí. V rámci tohoto procesu jsou data čištěna, integrována, transformována do využitelné podoby a následně analyzována a dále zpracovávána.

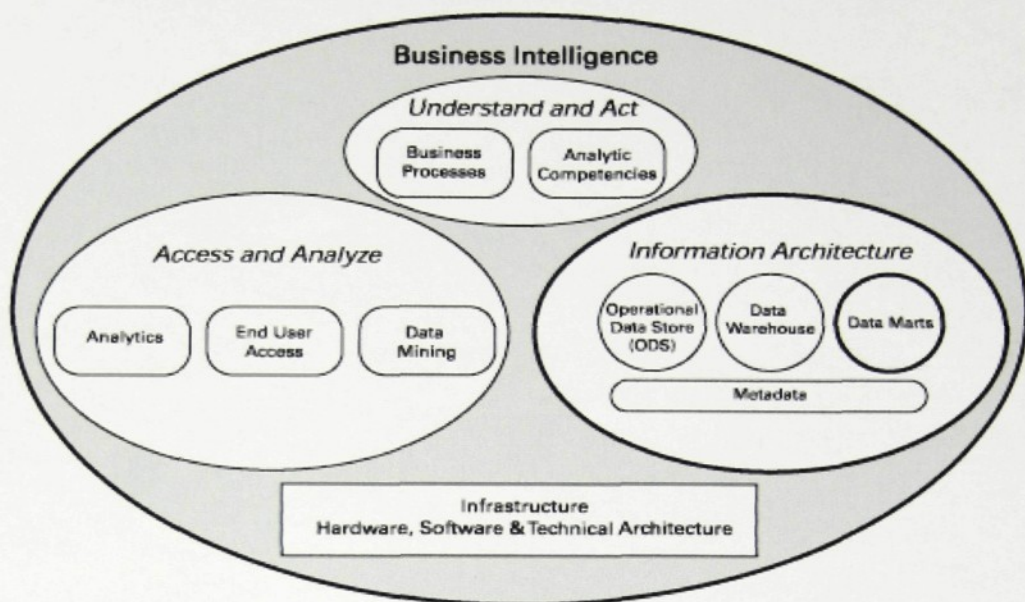
Podle [i2] se BI řešení skládají v podstatě z následujících dílčích procesů:

- sběr dat (business measures gathering)
- transformace a historizace dat (data warehousing, data quality management)
- analýza dat (analysis, datamining, database research)
- přijetí rozhodnutí a akcí (taking actions and decisions, follow up)
- vyhodnocení akcí a měření (feedback)

Pro firmu může BI plnit informační funkci např. v následujících oblastech:

- Analýza zákazníků – součást CRM (Customer Relationship Management)
- Řízení a plánování cash – flow
- Controlling nákladů a výnosů
- Predikce objemů prodeje, ziskovosti, výnosnosti..
- Podpora marketingových kampaní atd.

Obr. č. 2: Znáznornění schématu Business Intelligence



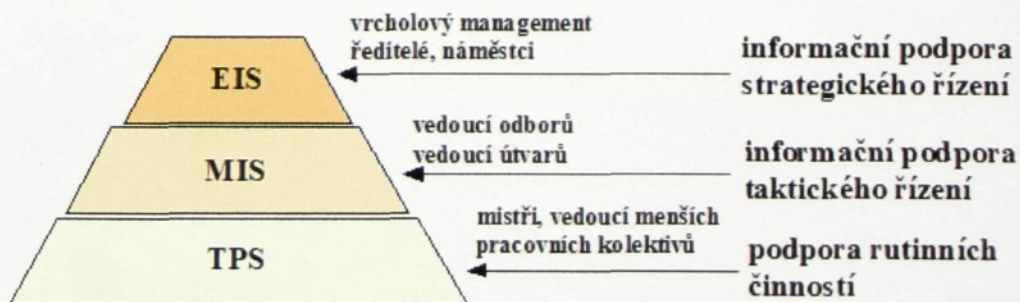
Pramen: <http://www.dc.uba.ar>

Samotná oblast BI dnes pokrývá mnoho oblastí, jako jsou např. datové sklady, dolování dat, OLAP atd.... Tyto pojmy budou v následujících kapitolách samostatně definovány a bude popsáno jakou úlohu tvoří a jak spolu souvisejí.

2.2.1 Obecné schéma architektury moderních informačních systémů

Při tvorbě architektury IS v různých organizacích musí být vždy respektovány konkrétní podmínky dané organizace, a proto architektura IS je vždy do jisté míry jedinečná. Mnoho odborníků vychází ze schématu třístupňového řízení, které lze vyjádřit pomocí pyramidy, která rozděluje jednotlivé úrovně řízení v rámci podniku.

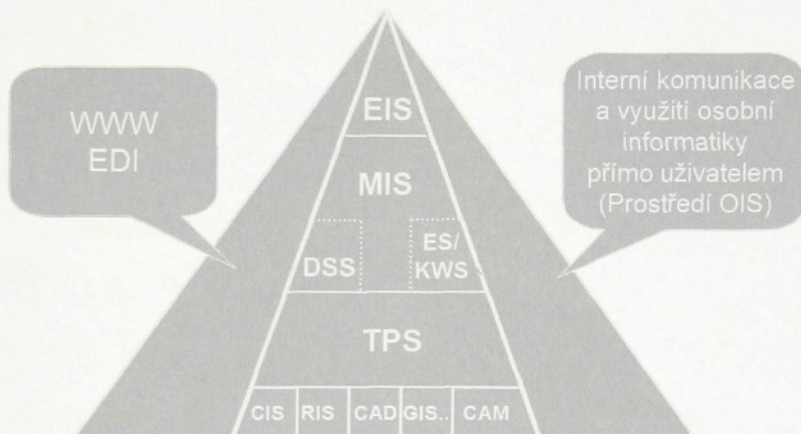
Obr. č. 3: Úrovně řízení z informačního pohledu



Pramen: Orientace ve světě informatiky, Norbert Žid a kol., str. 46

Tuto tradiční architekturu lze postupně doplnit dílčími stavebními bloky a jednotlivé úrovně lze rozdělit do více specifických automatizačních úloh. Proto se v současné době můžeme setkat s následujícím rozdělením. (obrázek č. 4)

Obr. č. 4: Typy úloh v informačních systémech



Pramen: *Orientace ve světě informatiky*, Norbert Žid a kol., str. 47

Stručná charakteristika jednotlivých bloků IS

TPS - Transaction Processing Systems

Cílem těchto systémů je automatizace běžných úloh v podniku. Jedná se např. o různé evidenční systémy, skladové hospodaření, vedení účetnictví apod. Především tyto systémy jsou zdrojem prvotních dat. Typickou vlastností je fakt, že velká část celkového zpracování v těchto systémech je vykonávána už při vkládání dat, resp. těsně poté. Další charakteristickým rysem je fakt, že na této úrovni existuje mezi podniky nejvíce rozdílů a specifičnosti. Na úrovni TPS existuje celá řada podoblastí, které mají specifickou programovou podporu. Jedná se např. o CAD (Computer Aided Design), CAM (Computer Aided Manufacturing), GIS (Geographical Information System) atd. V současné době jsou tyto systémy asi nejvíce rozšířené. Je to dáno jejich modulárním charakterem a historickým vývojem.

MIS – Management Information Systems

Tyto systémy jsou určeny zejména pro taktickou úroveň řízení. Měli by poskytovat vedoucím pracovníkům různé přehledy a agregované sestavy podle různých hledisek jako je čas, místo apod. Tyto systémy vycházejí z datové základny vytvořené TPS.

DSS – Decision Support Systems

Jedná se o skupinu úloh, které podporují obvykle taktické rozhodování a opírající se o optimalizační a simulační algoritmy. [2]

KWS – Knowledge Work Systems

Jsou to systémy, které jsou orientované na aplikaci expertních systémů a využití znalostních bází. [2]

EIS – Executive Information Systems

Informační systémy pro vrcholové řízení zastřešují všechny předešlé systémy. Z jejich názvu je patrné, že jsou určeny pro vrcholové manažery podniku. Jsou podkladem pro strategické rozhodování v různých oblastech. EIS využívají všech dostupných informačních zdrojů vytvořených nižšími úrovněmi IS. Jsou koncipovány tak, aby vycházely vstříc způsobu myšlení řídicích pracovníků a respektovaly to, co tyto pracovníky zajímá nejvíce (a co je v angličtině označováno jako: Critical Success Factors neboli kritické faktory úspěchu) [i3]. Ke své realizaci přitom využívají nejmodernější grafiku, nejvýkonnější prostředky počítačových komunikací a nejvyspělejší techniky uchovávání, zpracování a vyhledávání dat.

OIS – Office Information Systems

Tento typ systémů zastupuje sadu aplikací zaměřených na podporu kancelářských prací. Typicky se jedná o různé textové či tabulkové procesory, aplikace pro podporu týmové práce, programy pro správu dokumentů či klienta elektronické pošty apod.

Tab. č. 1: Klasifikace IS podle úrovně řízení

Úroveň řízení	Rozhodování Plánování	Typ IS	Informační úroveň
Vrcholový management	Strategické	EIS	Moudrost
Střední management	Taktické	DSS	Znalost
Operační management	Operativní	MIS	Informace
Běžní zaměstnanci	-----	TPS	Data

Pramen: <http://www.ipesoft.sk/index.html>

2.2.2 Provozní systémy vs datové sklady

2.2.3 Provozní systém

Většina firem dnes provozuje nějaký svůj informační systém. Tyto systémy především zajišťují automatizaci různých provozních činností a jednotlivých obchodních procesů. Často se tak označují jako transakční či provozní systémy – tzv. OLTP systémy. (On-Line Transaction Processing). První systémy tohoto typu začaly vznikat na počátku 80. let a hlavního rozmachu vývoje dosáhly koncem 80. a začátkem 90. let.

Příklady použití OLTP systémů

- Rezervace letenek
- Záznamy telefonních hovorů
- Evidence zákazníků
- Vystavování objednávek
- Záznam obchodních případů
- Skladová evidence
- Účetní aplikace atd.

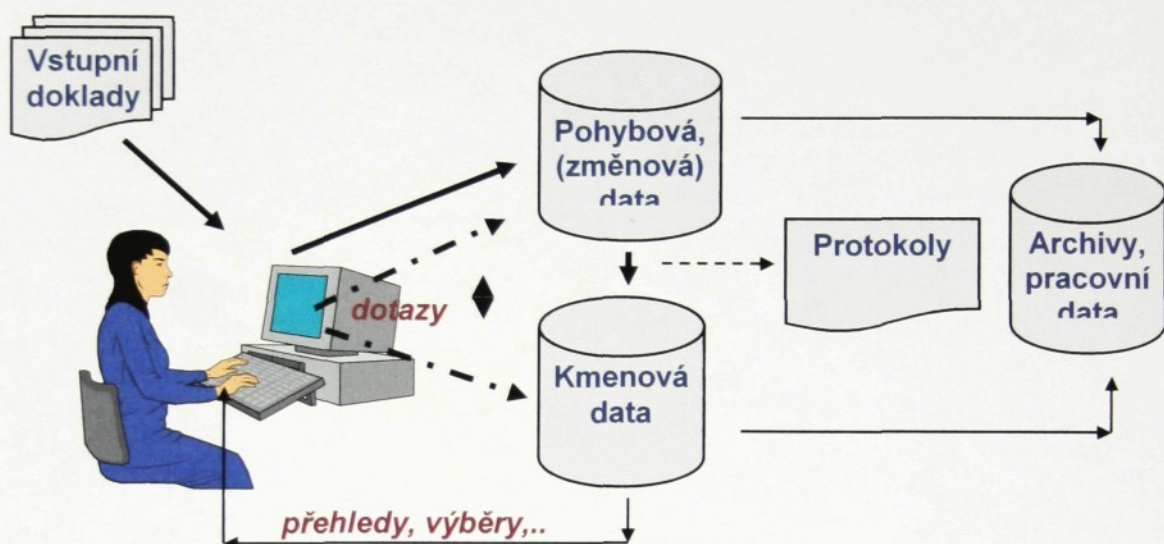
Principy OLTP

Používáním OLTP systémů postupně vzniká velké množství dat, které může být pro firmu velice cenným informačním zdrojem příp. konkurenční výhodou. Kromě toho provozní systémy musí být schopny nad těmito elementárními daty rychle a efektivně provádět relativně jednoduché transakce jako vyhledání záznamu(ů) podle zadaných kritérií, vložení nových záznamů, aktualizace již existujících nebo jejich vymazání. Operace tohoto typu zpracovává provozní informační systém obvykle pro velké množství uživatelů, typicky zaměstnanců, kteří jej používají pro svou každodenní práci.

Dobře navržený provozní informační systém uchovává svoje data obvykle v relační databázi, jejíž datové schéma splňuje všechny nebo některé normální formy relačního uložení dat. [i2] Provozní systém ve většině případů uchovává pouze aktuální data za poslední období, se kterými přímo pracuje a nikoliv data historická. Důvodem je především to, že historická data většinou nepotřebuje, takže by pouze zabírala místo a plýtvala výkonem. Výše zmíněné uložení dat je velice vhodné právě pro provádění jednoduchých elementárních operací, ale je značně nevhodné pro

kladení složitých dotazů pro vytváření různých přehledových a statistických sestav, což je jedna ze značných nevýhod. Další nevýhody budou podrobněji popsány v následující podkapitole.

Obr. č. 5: OLTP systém



Pramen: Data warehousing, návrh a implementace, HUMPRIES, Mark a kol., str. 75

Aplikace OLTP

- Automatizují a sbírají obchodní transakce prostřednictvím výpočetních systémů
- Poskytují výstupní sestavy, které umožňují sledovat stav transakce
- Postaveny především na architektuře klient/server [3]

2.2.4 Nevýhody OLTP systémů

V předchozím odstavci je popsáno, jakým způsobem OLTP systémy fungují a sbírají data. Tyto systémy tak správně vykonávají činnost pro kterou byly navrženy, vytvořeny a uvedeny do provozu. Jestliže se však firma rozhodne určitým způsobem využít tato data pro analýzu či pro rozhodování v nějakém konkrétním případě, nastávají problémy.

Např. firma potřebuje vyřešit následující otázky:

- Jaký počet zakázek můžeme očekávat pro následující půlrok?
- Kolik našich současných zákazníků zvažuje přejít ke konkurenci?
- Jak se vyvíjejí prodeje jednotlivých výrobků v různých oblastech za posledních 12 měsíců?
- Jaká je struktura našich zákazníků v jednotlivých segmentech?

OLTP systémy nedokáží tyto zásadní otázky řešit z několika následujících důvodů [i3]:

Nedostatečná historie dat.

Provozní systémy udržují z kapacitních důvodů data stará pouze několik měsíců, starší data se přehrávají do archívu, kde mají velice omezenou využitelnost.

Nedostatečné techniky a nástroje pro zpracování dat.

Provozní systémy většinou disponují sadou připravených reportů, které jsou navíc orientovány spíše na sledování procesů a jednotlivých transakcí než na globální pohled. Další využívanou možností jsou tzv. ad-hoc dotazy, což jsou dotazy kladené přímo do databáze pomocí standardního jazyka SQL.

Nemožnost zpracovat data z jiných aplikací.

V žádné společnosti není používán pouze jeden systém ačkoli existují velmi mohutná řešení jako je např. SAP R/3. Řada dat navíc vzniká mimo provozní systémy třeba i jako původně pouze osobní aplikace. Důležitá data rovněž pocházejí z externích zdrojů - ať se jedná např. o údaje z finančních trhů anebo data o průzkumu trhu.

Nepříznivý dopad na výkonnost provozních systémů.

Zpracování dat v provozních systémech k souhrnným přehledům a analýzám představuje další zátěž a vede k prodlužování doby odezvy transakčních systémů pro běžné uživatele.

Zde uváděné nevýhody, které transakční systémy omezují v různých oblastech, vedly k nutnosti vytvořit novou datovou základnu, která by lépe vyhovovala specifickým potřebám podniku. Právě proto potřeba vyhodnocování složitých přehledových a statistických dotazů nad daty vedla ke vzniku nového typu informačních systémů, **datových skladů**.

2.3 Datový sklad - podstata

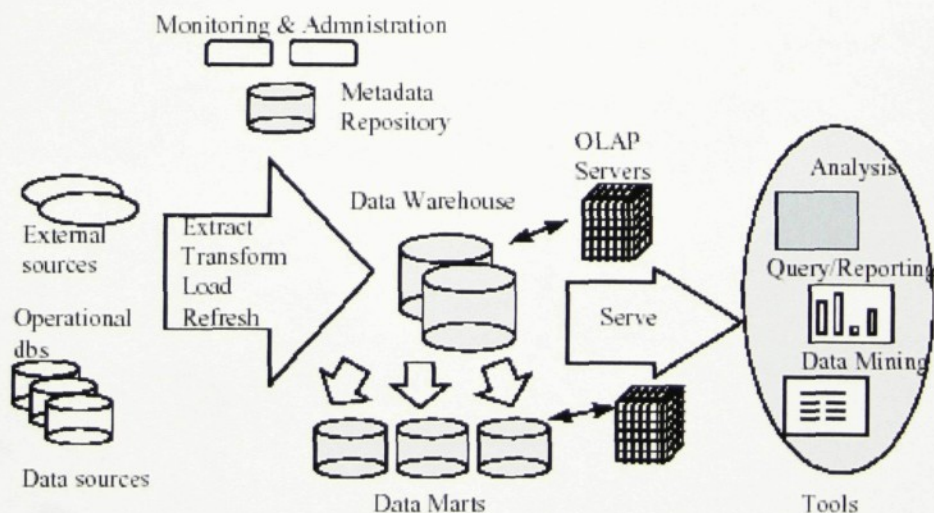
Znalostní báze tvoří intelektuální kapitál podniku, který organizace nepřetržitě transformuje do své vědomostní báze v kontextu své firemní strategie. Informační systém hraje významnou roli při tvorbě a distribuci znalostí. Právě proto je přikládána zásadní důležitost roli, kterou hraje **datový sklad** v podniku. Ten můžeme chápat jako centrální archív předmětově orientovaných originálních dat z podnikových transakčních systémů (OLTP) a externích datových zdrojů.

Technologie datových skladů zdaleka není jen módou jednoho/dvou desetiletí, ale tato technologie využívá metody a postupy, které lidstvo zná již pěknou řádku desetiletí. Nicméně v rámci informačních technologií tvoří datové sklady moderní součást. Poprvé byl pojem **datový sklad** použit na konci 80. let minulého století *Williamem Inmonem*. Ten jej definoval jako strategii přístupu k datům určeným pro rozsáhlé analýzy dat. Současně považuje za datový sklad kolekci sjednocených, předmětově orientovaných databází vyhovujících této strategii. (více II.kapitola).

Výše uvedená definice datových skladů předpokládá, že tato technologie nabídne jak prostředky pro samotné uložení dat, tak pro jejich analýzu. Hlavním cílem nasazení datových skladů by při tom mělo být vnesení pořádku do chaosu a získání dříve obtížně dosažitelných informací. Jedna z mnoha definic uvádí, že datové sklady jsou prostředkem pro poskytnutí lepších informací, na základě kterých je možné přikročit ke kvalifikovanějším a kvalitnějším rozhodnutím. Datový sklad lze také chápat jako „srdce“ manažerských informačních systémů, které jsou označovány jako EIS a DSS. Právě z důvodu, aby aplikace zmíněných typů mohly plnit roli pro kterou jsou určeny, je nutné vytvoření integrované a validní datové základny, tou je právě data warehouse.

Na první pohled se může zdát, že datový sklad se příliš neliší od běžného informačního systému, snad jen co se týká možnosti ukládat velké objemy dat. Mezi klasickým informačním systémem a datovým skladem však existují zcela zásadní rozdíly.

Obr. č. 6: Architektura datového skladu



Pramen: Overview of Data Warehousing and OLAP Technology, Surajit Chaudhuri, str. 15

Je ještě nutno zmínit rozdíl mezi pojmy **data warehouse** (datový sklad) a pojmem **data warehousing**. První pojem lze chápat ve smyslu jeho definice.

Datový sklad představuje uložení dat které má následující charakteristiky [i3]:

- Integruje data z různých zdrojů do jednoho systému
- Obsahuje historii – jsou k dispozici data i za několik minulých let
- Data jsou uložena na různých úrovních sumarizace
- Data se periodicky načítají z provozních systémů (většinou v noci a o víkendech)
- Uživatelé data pouze čtou, tj. neprovádí jejich zadávání ani je nemění
- Data z datového skladu se využívají pomocí širokého spektra metod pro prezentace a analýzy dat

Pojem data warehousing pak představuje v tomto pojetí PROCES, který na základě určité metodiky z dat uložených v datovém skladu, dělá smysluplné informace, nebo umožňuje vytvářet pohledy na informace potřebné v různých procesech.

2.3.1 Rozdíly mezi provozními systémy a datovým skladem

Zásadní rozdíly mezi OLTP systémy a datovým skladem shrnuje následující tabulka.

Tab. č. 2: Porovnání OLTP systémů s datovými sklady

Hlavní účel	
PROVOZNÍ, OLTP SYSTÉMY	DATOVÝ SKLAD
• Automatizace operací nebo procesů	• Poskytování optimálních informací pro rozhodování
Konceptní rozdíly	
• Dostat data do systému • Uživatelé mají možnost zadávat,měnit,rušit, číst data • Zajišťují automatizaci rutinních činností • Aplikace jsou v podstatě statické (požadavky na funkčnost jsou poměrně stálé) • Podporují každodenní aktivity firmy • Orientované na výkonnost • Proces implementace a využívání je poháněn technologií (impulsem k inovaci systému je nové systémové prostředí, nová verze DB..)	• Dostat informace ze systému • Uživatelé mají možnost pouze číst data • Umožňují kreativitu uživatelů při práci s daty (analýzy, prezentace) • Aplikace jsou dynamické (požadavky na funkčnost aplikace se mění) • Podporují dlouhodobé strategie firmy • Poskytují konkurenční výhodu • Proces implementace a využívání je poháněn potřebami organizace (impulsem k inovaci systému jsou nové potřeby uživatelů)

Technologické rozdíly

- | | |
|---|---|
| <ul style="list-style-type: none">• Zpracovávají velké objemy malých transakcí• Transakce neustále přidávají a aktualizují data• Důležitým hlediskem je omezení redundance dat• Integrita dat se zajišťuje datovým modelem a aplikacemi• Datové modely jsou optimalizované pro online aktualizace a rychlé zpracování transakcí• Používají se převážně normalizované relační datové modely | <ul style="list-style-type: none">• Zpracovává malý počet komplexních dotazů• Data se načítají dávkově• Důležitým hlediskem je rychlý přístup k datům pro účely analýz a prezentací• Integrita dat se zajišťuje při dávkových načítacích procesech (transformace dat)• Datové modely jsou optimalizované pro rychlé zpracování výstupů• Používá se kombinace datových modelů (normalizované a nenormalizované relační modely, sumarizované tabulky, star schéma datové modely, snowflake schéma datové modely, multidimenzionální datové modely) |
|---|---|

Pramen: www.adastra.cz

Z výše uvedeného vyplývá, že klasické informační systémy slouží k momentálnímu zpracování a vyhodnocení jednotlivých transakcí a k základnímu sběru dat a tím pádem i k vytváření momentálního obrazu sledované reality. Datový sklad je naopak dlouhodobým úložištěm, kam data shromážděná klasickými informačními systémy přibývají periodicky po jednotlivých dávkách. Odezvy na dotazy nemusí být okamžité, což se naopak vyžaduje u klasického informačního systému.

2.3.2 Datová tržiště – datamarty

Datová tržiště plní podobně jako datové sklady roli zprostředkovatele důležitých informací získatelných z podnikových dat, nicméně na rozdíl od nich jsou zaměřeny pouze na jednu konkrétní cílovou skupinu uživatelů. Tyto cílové skupiny přitom vyžadují informace o konkrétní oblasti, konkrétním předmětu zájmu – tržiště tedy ve svých databázích obsahují přesně zacílenou podmnožinu dat vyhovující konkrétním analytickým požadavkům skupiny uživatelů. Zjednodušeně by se také dalo říci, že datová tržiště jsou tedy jakousi zjednodušenou verzí datového skladu zaměřenou na užší oblasti.

Data uložená v datových tržištích je možné používat především jako podklad pro cílené analýzy (např. ve kterých regionech došlo k poklesu prodeje o více než deset procent, jak se mění uživatelé služeb apod.). Na základě těchto cílených analýz je pak pro odpovědné pracovníky mnohem

jednodušší přijímat informovaná rozhodnutí, v návaznosti na zaměření tržiště jak taktického, tak strategického charakteru.

Přínos datovým tržišť je ovšem nejen v tom, že umožňují pomocí odpovídajících nástrojů získávat dříve nedostupné informace, ale také v tom, že informace získatelné z běžných provozních systémů jsou s jejich pomocí dostupné jednodušším a rychlejším způsobem.

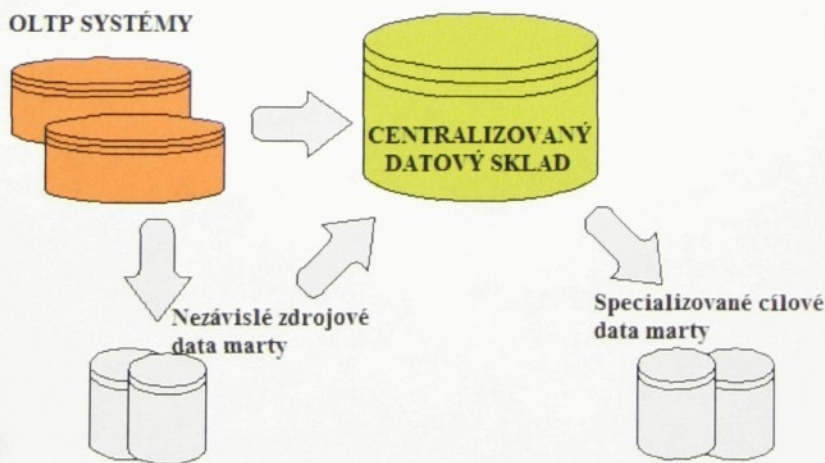
Obdobně jako datové sklady (tyto dvě technologie jsou ale ve skutečnosti stejnou technologií) zažívají v posledních letech v důsledku stále rostoucího uvědomování si hodnoty informací značný boom. Na trhu tak existuje velké množství významných dodavatelů poskytujících především univerzální řešení.

Dle používané metodologie [i4] lze rozlišit dva typy datových tržišť:

1. **Závislý data mart** – toto datové tržiště je budováno nad datovým skladem (je závislé na datovém skladu), data jsou tak určitou předem definovanou podmnožinou dat datového skladu
2. **Nezávislý data mart** – datová tržiště se tvoří místo datového skladu, každý útvar podniku si může tvořit své specializované datové tržiště

Je velmi důležité si uvědomit, že v rámci jednoho podniku sice může existovat více datových tržišť, nemělo by se ovšem jednat o tržiště vzájemně naprosto izolovaná – jinak může dojít přesně k opačnému efektu a místo správných rozhodnutí lze začít přijímat sice rozhodnutí informovaná, ale odtržená od svého okolí. V konečném důsledku tedy rozhodnutí chybí.

Obr. č. 7: Různá pojetí data martů



Pramen: www.sybase.cz

Celkově lze datová tržiště charakterizovat:

- Mají omezený rozsah
- Navržená pro specifické dotazy
- Jsou často podmnožinou datového skladu
- Obvykle používají zdrojová data z datového skladu
- Jsou vhodným nástrojem pro agregaci dat

2.3.3 Využití datového skladu

Datový sklad nepředstavuje řešení samo o sobě. Pro jeho využití musí podnik využívat různé aplikace a nástroje, které uložená data správně zanalyzují a interpretují. Mezi nejpoužívanější techniky využití datového skladu patří [4]:

- **Operativní dotazy**
 - předem nepřipravené dotazy na určité hodnoty
- **Sestavy**
 - standardní generované dávkově
 - operativní vytvářené podle potřeby
- **Multidimenzionální analýza (OLAP)**
 - rychlé prohlížení dat sumarizovaných na různých úrovních z různých pohledů neboli dimenzí
- **Statistické analýzy**
 - např. zjišťování závislosti veličin, identifikace důležitých proměnných, vytváření segmentů
- **Finanční analýzy** (např. ekonometrické modelování, termínové modely)
- **Analýzy časových řad a tvorbu předpovědí**
- **Vizualizace dat**
 - prohlížení dat v dynamicky provázaných grafech pro např. identifikaci neobvyklých a extrémních hodnot a závislostí mezi daty
- **Data mining**
 - specializované techniky pro zpracování velkých objemů dat a hledání skrytých vzorů a souvislostí

- **EIS**
 - připravené aplikace pro vedoucí pracovníky se snadným ovládáním kombinující OLAP, reporting, přehledné zobrazení kritických veličin, jednoduché předpovědi
- **Aplikace vytvořené podle konkrétních potřeb**

2.3.4 Účel datového skladu

Je logické si položit otázku, proč podniky přistupují k budování datových skladů. Proces vybudování datového skladu, jeho údržba a podpora stojí kromě alokovaných lidských zdrojů i nemalé finanční prostředky. Z tohoto důvodu je logické vyjmenovat několik důvodů, které implementaci datových skladů vysvětlují. [5]

Jedna verze pravdy

Data obsažená v datových skladech jsou konzistentní a kvalitní dříve, než k nim mají uživatelé přístup. Tento fakt vede k tomu, že datový sklad se stává primárním zdrojem informací pro celou organizaci, ať už pro rozhodovací procesy strategického či taktického charakteru. Díky tomu už nemůže docházet k různým informačním šumům v rámci podniku, které měli negativní dopad na jeho fungování.

Přesně zaznamenaná minulost

Čas je důležitou veličinou, jestliže se porovnávají různá data. Bez toho, aby manažeři měli k dispozici historická data týkající se jednotlivých roků, měsíců, týdnů či dní, se velmi těžko budou porovnávat aktuální data a údaje. Korektně zaznamenaná minulost samozřejmě také umožňuje vyhledání trendů, negativních či pozitivních. Zaznamenáním minulosti do datových skladů tak dochází k funkčně správnému oddělení historických dat od dat aktuálních. Aktuální data jsou tak zaznamenána v transakčních systémech.

Zpřístupnění dat obchodním uživatelům

Bez existence datového skladu byli obchodní uživatelé získávat podniková data z různých nepříliš uživatelsky přívětivých systémů. Těchto systémů může být samozřejmě několik, a o to více bylo získání požadovaných dat obtížnější. Proto díky datovému skladu mohou obchodní uživatelé přistupovat k jednotným podnikovým datům bez velkých překážek, a to buď prostřednictvím klasické pracovní stanice nebo např. vzdáleným přístupem přes internet. V tomto případě je samozřejmě důležité zabezpečení transferu dat. Nespornou výhodou je také fakt, že existence

datového skladu usnadňuje uživatelům chápání, učení a porozumění složitosti dat a datových struktur.

Oddělení analytického a transakčního zpracování

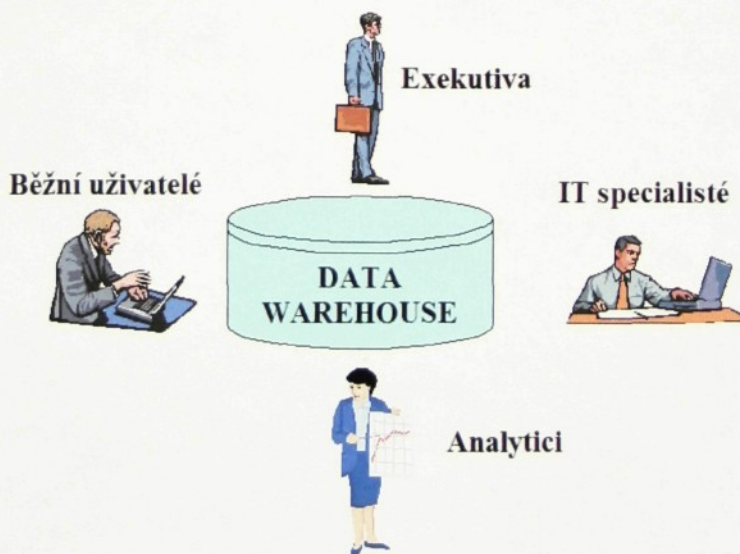
V případě porovnání systému pro rozhodování na jedné straně a systému pro transakční zpracování na straně druhé lze dojít k závěru, že jejich požadavky na architekturu jsou zcela odlišné. Návrh systému, který by tyto požadavky sjednotil, je velmi obtížný a to především z hlediska pozdější správy systému.

2.3.5 Uživatelé datového skladu

V rámci podniku se liší typy uživatelů, kteří datový sklad využívají jako podporu pro své pracovní činnosti, postupy a procesy. Je důležité si uvědomit, že veškeré informační systémy typu EIS, DSS apod. využívají pro svou činnost právě datový sklad jako jakýsi „back end“.

Tyto uživatele lze rozdělit do několika skupin, při návrhu datového skladu se musí s odlišnostmi uživatelů počítat a zakomponovat je do řešení finálního datového skladu. Budoucí uživatelé se mohou lišit podle svého postavení v hierarchii organizace, zkušenostmi s výpočetní technikou, pravomocemi v rozhodování apod. [11]

Obr. č. 8: Uživatelé datového skladu



Pramen: vlastní

Exekutiva

Tato skupina zahrnuje uživatele - manažery, kteří jsou primárně určeni pro rozhodování a řízení společnosti. Nelze u nich předpokládat, že budou schopni a ochotni zaškolit se pro používání sofistikovaných nástrojů, pro svou činnost nepotřebují vědět nic o datových strukturách a podobných technických záležitostech. Často jsou jejich požadavky velice nestandardní, na druhou stranu některé jejich informační potřeby mohou být řešeny standardizovanými a pravidelnými výstupy. Často jim tyto informace dodávají podpůrní pracovníci. Pro tuto skupinu uživatelů datového skladu je důležitou veličinou čas. Pro svá rozhodnutí potřebují informace, které jsou aktuální. Pozdě dodané informace už pro ně většinou nemají žádnou hodnotu.

Analytici

Analytici jsou specialisté, kteří jsou primárně určeni pro různé obchodní analýzy (vyhodnocování různých rizik, vyhodnocování dopadů různých rozhodnutí, analýza poptávky atd.). Pro svou práci využívají metod statistiky, pravděpodobnosti, zabývají se různými modely. Pro jejich podporu jsou určeny různé speciální softwarové nástroje. Tato skupina uživatelů nejlépe rozumí datům obsaženým v datovém skladě a data dokáže korektně interpretovat a vyhledat souvislosti mezi nimi.

„Běžní“ uživatelé

Tímto označením jsou myšleni pracovníci, kteří pracují v různých podnikových odděleních (marketing, logistika, odbyt..). Datový sklad jim slouží jako podpora pro jejich operativní činnosti. Z informací, které jim může datový sklad spolu s navazujícími nástroji poskytnout, je zajímaví pouze informace a souvislosti, které se týkají bezprostředně jejich oddělení. Pro sofistikovanější dotazy využívají podporu IT specialistů.

IT specialisté

Tato skupina není uživateli datového skladu v pravém slova smyslu. Informace, které oni získají, neslouží jim, ale poskytují je ostatním uživatelům. Tato skupina je spíše technologicky orientovaná – stará se o vybudování infrastruktury datového skladu, o jeho běžnou údržbu. Mezi tyto pracovníky patří např. specialisté na databáze, programátoři atd. Pro ně je důležité, aby vývoj, správa a struktura dat byly přehledné a flexibilní. Samotný obsah dat pro ně není důležitý. Na druhou stranu má tato skupina důležitou roli v průběhu návrhu, budování a provozu datového skladu. IT specialisté mohou ovlivnit jak obsah datového skladu tak i způsob komunikace ostatních – problémově orientovaných - uživatelů.

2.3.6 Náklady datového skladu

Stejně jako u klasických IS bude zajímat vedoucí pracovníky i otázka porovnání nákladů a výnosů vybudování a provozování datového skladu. Podnik si samozřejmě musí porovnat přínos vybudování datového skladu oproti nákladům. Tato otázka není vůbec jednoduchá, pro její zodpovězení musí podnik vycházet ze složitých výpočtů a analýz. Jedním z nástrojů, který umožňuje alespoň přibližně určit efektivnost řešení datového skladu je tzv. metoda TCO.

TCO – Total Cost of Ownership

TCO vychází z principu celkového vyjádření nákladů, kolik bude pořízení a provozování systému stát v průběhu let, kdy se bude využívat. Z analýzy společnosti META GROUP (1998), vyplývá, že u dosavadních projektů datových skladů typické rozložení nákladů vypadá následovně:

- Pořizovací náklady
Tyto náklady lze vypočítat poměrně lehce. Patří sem náklady na požadovaný software, hardware, vybudování infrastruktury, školení, analýza atd.. K přesným číslům se lze dostat např. na základě poptávky po dodávce projektu datového skladu.
- Průběžné náklady na obsluhu systému
- Průběžné náklady na úpravy, přizpůsobování a funkční rozšiřování systému
- Náklady na rozšiřování (velikosti) systému

Táž studie udává, že náklady na provoz typického datového skladu za pět let představují trojnásobek původního rozpočtu. Důvodem je především skutečnost, že nárůst velikosti datového skladu představuje reálně v průměru kolem 40 % ročně. Při konstruování výpočtu na vlastnictví je doporučováno dle [i2] vzít v úvahu především tyto charakteristiky:

- Počáteční velikost vstupních dat v GB
 - Náklady na diskovou paměť pro uložení 1 GB dat v USD
 - Faktor datové exploze v procentech
 - Prvotní náklady na softwarové licence
 - Roční podpora softwarových produktů v procentech
 - Náklady na další hardware – CPU atd..
- + nároky na školení databázových administrátorů a náročnost administrace systému (ladění..)

Na základě těchto informací lze vytvořit kalkulaci pro jednotlivé varianty a spočítat celkové náklady na vlastnictví systému. Výsledky mohou být velice překvapivé, zvláště v horizontu několika let.

3 Architektura a hlavní komponenty datového skladu

3.1 Přístupy k budování datového skladu

Všichni dodavatelé, kteří nabízejí svá řešení pro budování datového skladu, vycházejí z nějaké koncepce. Při podrobnějším zkoumání lze zjistit, že se svým přístupem liší, vždy je však výsledným produktem datový sklad. Tato rozdílnost je dána odlišným přístupem k implementaci datového skladu. Rozdílnost se také promítá do vlastní metodiky.

Existuje mnoho přístupů k budování datového skladu, nepoužívanější jsou však dva odlišné směry. Jejich představiteli jsou William H. Inmon, často nazývaný „otcem warehousingu“, a Ralph Kimball. Tito dva představitelé jsou považováni za nejvýznamnější osoby v oblasti datových skladů resp. data warehousingu.

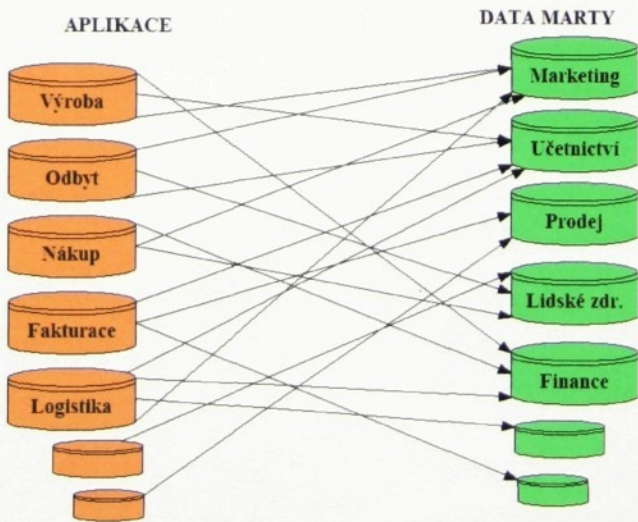
Základní rozdělení dle přístupu k budování DW:

- I. centralizovaný datový sklad - Bill Inmon
- II. data warehouse jako množina data martů - Ralph Kimball

Výchozí stav

Na následujícím obrázku je zachycen stav, kdy podnikové informační systémy jsou zdrojem dat pro data marty. Takto vzniklé data marty jsou označovány jako nezávislé data marty. Je to z důvodu, že zdrojem dat jsou přímo jednotlivé produkční aplikace podniku.

Obr. č. 9: Architektura s nezávislými data marty



Pramen: ComputerWorld 8/2000, str. 35

Data marty jsou budovány pro podporu činností jednotlivých podnikových útvarů. Praxe ukazuje, že málokdy existuje v podniku jednotná definice pro pojmy používané v podniku. Jednotlivé data marty jsou vždy budovány na základě požadavků jednotlivých útvarů společnosti. Z toho vyplývá:

- potřeba vlastních dat
- používání vlastních definice pojmů
- vlastní historie dat
- vlastní periodicitu aktualizace dat.

Zmiňovaný přístup tak vede k tomu, že jednotlivé data marty obsahují:

- redundantní data
- odlišné dimenze (atributy prostřednictvím kterých se "díváme" na data)
- odlišná fakta (veličiny, které sledujeme, např. zisk, ztráta, počet hovorů apod.)

Tímto způsobem vybudované prostředí pro podporu rozhodování však neposkytuje celopodnikový pohled na informace. Podíváme-li se na schématické znázornění architektury, odpovídající zmiňovanému přístupu, pak tato nám může připomínat "pavoučí síť" ("spider net" - obvykle používaný pojem pro architekturu data warehouse založenou na jednotlivých data martech), viz.obrázek č. 9.

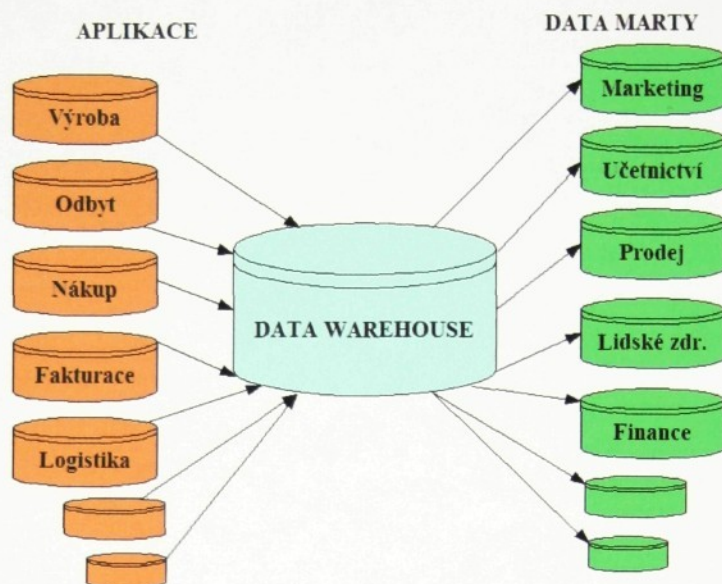
3.1.1 Centralizovaný datový sklad- Bill Inmon

„Podnikový data warehouse obsahuje detailní, atomicky integrovaná historická data“

Bill Inmon

Bill Inmon, přišel s myšlenkou vytvořit architekturu, která by byly minimalizovala redundantní data, a zároveň minimalizovala počet interface mezi produkčními systémy a datovým skladem. Mezi jednotlivé, nezávislé data marty vložil databázi, kterou nazval datovým skladem. Tím došlo ke snížení počtu rozhraní a k definici pojmů tvořících znalostní podnikovou bázi. (viz následující obrázek)

Obr. č. 10: Architektura datového skladu dle B. Inmona



Pramen: ComputerWorld 9/2000, str. 42

Data warehouse je poté místem, kam jsou ukládána podniková data, která prošla procesy ETL (Export $\Rightarrow$ Transformation $\Rightarrow$ Load) – (viz část ETL procesy). Tím dojde k vytvoření jednoho místa podniku, které uchovává validní podniková data a poskytuje je pomocí jednotlivých data martů oddělením podniku.

- V jeho pojetí představuje datový sklad oblast, kde jsou soustředěná data jak z podnikových informačních zdrojů, tak z externích datových zdrojů.
- Granularita dat je na nejnižší možné úrovni, tj. jedná se o tzv. **atomární data**, což vychází z definice data warehouse podle B. Inmona

B. Inmon dále charakterizuje DW jako předmětově orientovanou, integrovanou, stálou a časově rozdílnou kolekci dat, která podporuje rozhodovací procesy.

Vymezení jednotlivých termínů:

Předmětově orientovaný

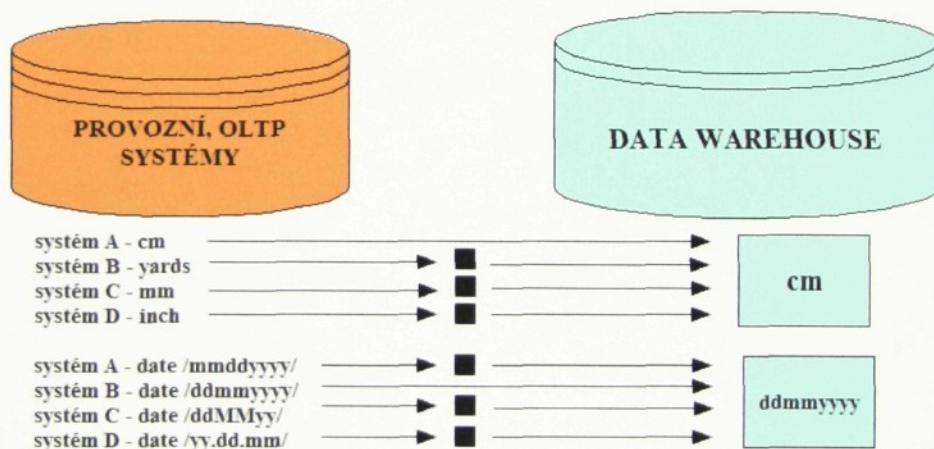
Data v datovém skladě jsou organizována kolem hlavních „předmětů“ podniku. Každý podnik, organizace pracuje s určitými pojmy, jako jsou např. ZÁKAZNÍK, PRODUKT, DODAVATEL, TRANSAKCE a pod. Tyto pojmy představují subjekty (předměty). Mezi těmito subjekty existují nejen určité vztahy, ale tyto je možné také nadále dělit jako např.: ZÁKAZNÍK = { Právnická

osoba, Fyzická osoba, Fyzická osoba - podnikatel} atd. Tato orientace způsobuje, že návrh datového skladu je datově řízen. Datový sklad by neměl obsahovat data, která nejsou používána v DSS procesech.

Integrovaný

Data, která jsou ukládána v datovém skladu, pochází z několika produkčních systémů podniku. Data jsou na základě určitých pravidel spojována tak, aby poskytla koncovému uživateli celopodnikový pohled na oblast jeho zájmu. Integrace může mít mnoho podob, jedná se o konzistentní názvosloví, konzistenci měřených veličin, konzistenci kódovací struktury, konzistenci formátů atributů apod. Jestliže data přicházejí do datového skladu z nekonzistentních provozních systémů, je zřejmé, že musí být transformována do formátu požadovaného konzistentním datovým skladem. Důležitost integrace lze ukázat na příkladech jmenných konvencí, konzistenci fyzických atributů dat apod., viz obrázek č. 11.

Obr. č. 11: Integrace dat z OLTP systémů do datového skladu



Pramen: ComputerWorld 25/99. str. 7

Stálá data

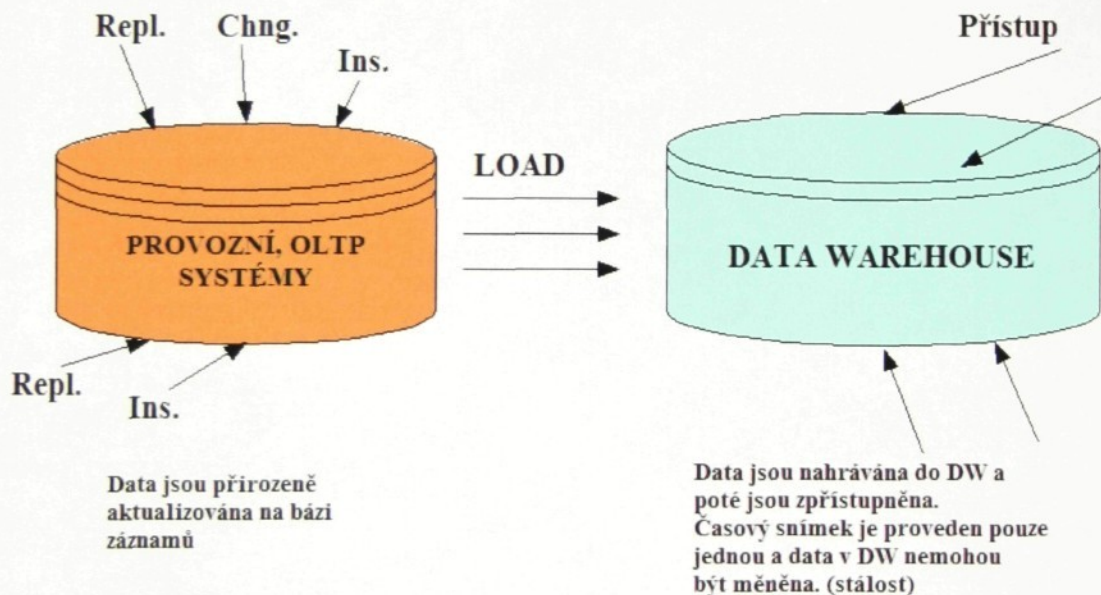
Tato vlastnost dat dělá z datového skladu statické prostředí. Oproti tomu jsou OLTP – produkční systémy dynamickým prostředím. Dynamičnost dat je zachycena na obrázku č. 12.

- Data v produkčních systémech podléhají změnám. Na operativní úrovni je dovoleno jednotlivé záznamy do databáze vkládat, mazat a případně je modifikovat.
- Oproti tomu existují pouze dva typy operací, které mohou být v datovém skladu realizovány – jedná se o naplnění datového skladu a přístup k datům.

Datový sklad neobsahuje všechny změny během dne, ale pouze ty konečné, tj. po definovaném okamžiku, kdy jsou data z produkčních systémů extrahována pro potřeby datového skladu. Data jsou obvykle extrahována po provedení tzv. uzávěrky dne. Jedná se tak o statická data vztahující se k jednomu, dobře definovanému okamžiku (lze se setkat s pojmem "snapshot").

Tento rozdíl mezi operativní databází a datovým skladem sebou nese důležitá fakta týkající se fyzického návrhu databáze. Na úrovni fyzického návrhu databáze datového skladu mohou být oproti operativní databázi použity volnější metody, které optimalizují především přístup k datům (normalizace / fyzická denormalizace). [i7]

Obr. č. 12: Dynamičnost prostředí OLTP a DW



Pramen: ComputerWorld 25/99. str. 7

Časově rozdílná data

Data v datovém skladu jsou ukládána po časových snímcích a tak vytváří časově rozdílnou řadu, historii. Tato data jsou přesná pouze k určitému momentu. Toto je rozdíl oproti provozním systémům, kde jsou data platná v momentě přístupu.

Z této vlastnosti plynou následující skutečnosti:

- Časový horizont je dlouhodobější (roky) oproti OLTP (dny, měsíce)
- Struktura klíčových atributů obsahuje také element času, který v OLTP nemusí být uvedený

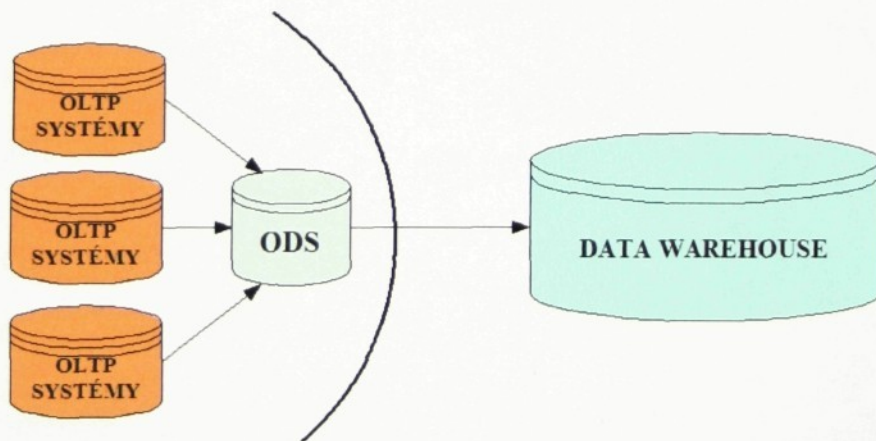
- V okamžiku, kdy je do datového skladu nahrán konkrétní snímek dat z provozních systémů, nemohou být už v datovém skladě modifikované

Operativní úložiště dat – ODS

ODS (Operational Data Store) je prostředkem, který představuje určité rozšíření datového skladu směrem k produkčním systémům. ODS tak představuje základ pro uchování dat OLTP systémů v krátkém časovém intervalu.

Definice ODS popisuje jako předmětově orientovanou, integrovanou a aktualizovatelnou množinu dat, která obsahuje aktuální data a která slouží pro podporu day-of-day procesů založených na detailních datech. Odlišnost mezi definicí datového skladu a ODS je tedy právě v možnosti aktualizace dat a jejich historické délce. Největším rozdílem mezi DW a ODS je fakt, že ODS obsahuje jen současná nebo blízka data, zatímco datový sklad obsahuje jak historická, tak i současná a historicky blízka data. ODS také uchovává výhradně detailní data, zatímco v datovém skladě jsou uložena jak detailní, tak i sumarizovaná (agregovaná) data. Objem dat v obou systémech je tak zcela rozdílný.

Obr. č. 13: Začlenění ODS do architektury datového skladu



Pramen: *ComputerWorld* 25/99. str. 7

Architektura celopodnikového, centralizovaného datového skladu s sebou přináší:

Výhody:

- zajištění celopodnikového pohledu na data
- dobře monitorovatelné prostředí
- vytvoření jednoho místa pro zajištění kvality dat

- snížení nákladů na HW, SW
- snížení redundantních dat, atd.

Nevýhody:

- Složitější realizace
- Pomalejší implementace

3.1.2 Sjedené data marty – Ralph Kimball

„Data warehouse není nic víc než sjedení všech konzistentních data martů“

Ralph Kimball

Tvrzení R. Kimballa neznamena nic jiného než postupné budování data martů. Logickým sjedením těchto data martů je pak datový sklad. Kimballova myšlenka budování data warehouse se stala velmi populární především z následujících důvodů

- důvodu rychlé implementace data martu,
- nižších počátečních investic.

Výsledná podoba datového skladu se nazývá tzv. BUS architektura.

BUS architektura

V každém podniku jsou definovány obchodní aktivity, které jsou prováděny jednotlivými zaměstnanci. Obchodní aktivitu lze definovat jako proces, který tvoří jednotlivé aktivity, které pak tvoří logické celky. Některé z obchodních aktivit přesahují definici individuálního procesu a mohou procházet napříč celou organizací. Proto je třeba zajistit určitou koherenci údajů používaných v rámci jednotlivých procesů. Od tohoto požadavku plynou následující dva pojmy:

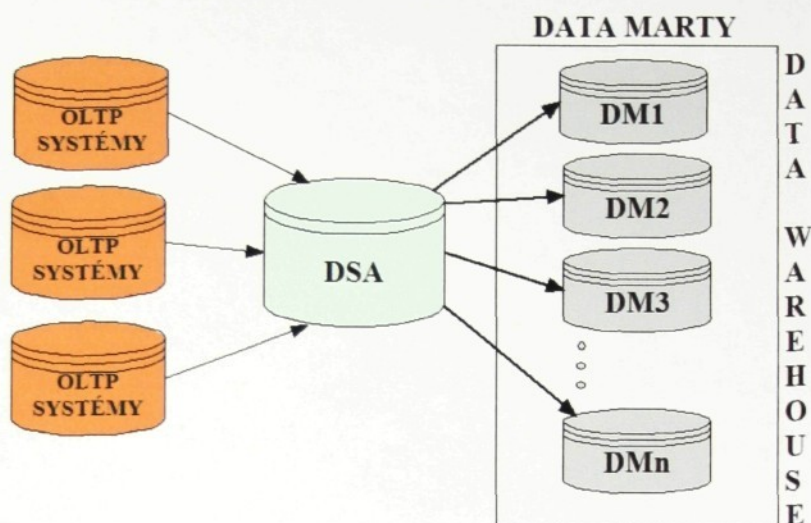
- CD – conformed dimensions = odpovídající dimenze
- CF – conformed facts = odpovídající fakta

Právě tyto dva pojmy tvoří základ tzv. BUS architektury. V praxi často podniky vytvářejí jednotlivé data marty, ale často právě opomíjejí zmíněné dva prvky BUS architektury, CD a CF. Jestliže tyto prvky nejsou integrovány do budování datového skladu, pak lze vybudovat data marty, které však nebude možné v budoucnu sjednotit do datového skladu. Znamená to totiž fakt, že data mezi jednotlivými data marty budou v takovýchto případech nekonzistentní. [11]

DSA – Data Staging Area – záchytná oblast

Data je potřebné před jejich umístěním do jednotlivých data martů extrahovat z produkčních systémů, vyčistit, formátovat, odstranit duplicity a kombinovat. K tomuto účelu slouží právě DSA. Ta je v ideálním případě umístěna na jediném serveru.

Obr. č. 14: BUS architektura datového skladu dle R. Kimballa



Pramen: ComputerWorld 26/99, str. 7

Výhody:

- Rychlejší přínos pro uživatele
- Nižší náklady

Nevýhody:

- Může dojít k nekonzistencím mezi jednotlivými data marty
- Dochází ke komplikovanějším načítacím procesům z důvodu existence většího počtu rozhraní mezi produkčními systémy a data marty.

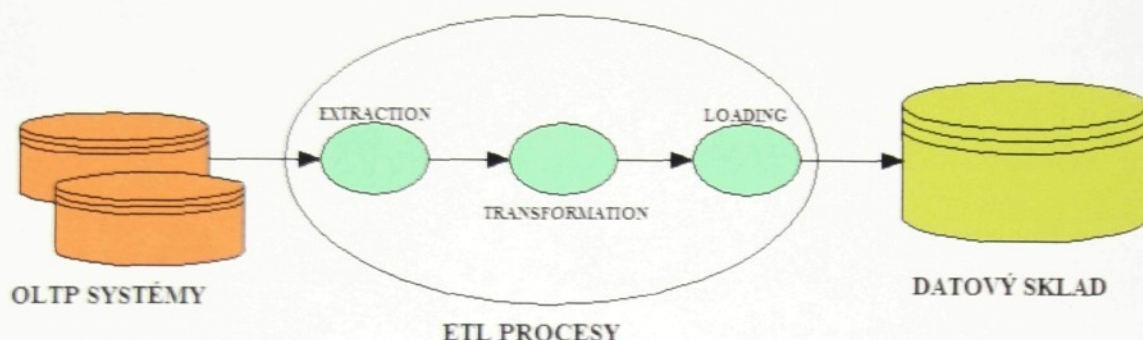
3.2 Datové pumpy – ETL procesy

Do datového skladu se data nezadávají, ale načítají se z různých provozních systémů či externích zdrojů. K tomuto účelu slouží nástroj zvaný datová pumpa. Někdy se také proces plnění datového skladu nazývá ETL proces (extraction $\Rightarrow$ tranformation $\Rightarrow$ loading). Datová pumpa má za úkol data čistit, protože tato data mohou být z globálního či lokálního hlediska nekonzistentní. Tato opravená a očištěná data mohou také sloužit jako určitá zpětná vazba pro provozní systémy.

3.2.1 Proces ETL

Proces načítání se většinou provádí v čase, kdy nejsou provozní systémy příliš zatíženy, aby se neprodlužovala doba odezvy pro uživatele těchto systémů. Základní schéma načítání do DS je znázorněno na následujícím obrázku č. 15.

Obr. č. 15: Proces ETL



Pramen: vlastní

Při plnění DW je nutné realizovat tyto hlavní kroky:

1. Extrakce vstupních dat - EXTRACTION
2. Transformaci vstupních dat - TRANSFORMATION
3. Načtení dat do DW – LOADING

3.2.2 Extrakce vstupních dat

Datová extrakce je prvním krokem v procesu ETL. Tento krok je však zároveň nejkritičtější, co se týče správného a informační hodnotu přinášejícího využívání datového skladu. Jedná se o schopnost převzít data z co nejširšího spektra datových zdrojů nejrozličnějšího charakteru, mezi které se řadí nesčetné databázové standardy, nedatabázové strukturované formáty, textové soubory, standardy elektronické pošty apod.

Souhrnně tedy můžeme extrakci chápat jako tu pracovní etapu, kdy usilujeme o přesné, rychlé, bezpečné, lehce kontrolovatelné a dobře říditelné načtení dat z co nejvíce externích datových zdrojů. Tato fáze se opakuje s periodicitou, která souvisí s využíváním daných informací. Po jejím skončení budou potřebná data načtena přímo do připravených zdrojových struktur pro extrahovaná data. Nástroje datového skladu musí umožňovat extrakci dat z provozních systémů, což ve většině případů znamená komunikovat určitým způsobem (ODBC, nativní drivery, textové soubory) s relační nebo síťovou databází či případně systémem souborů. V současnosti je obvyklé, že kvalitní

datové sklady mají předem definovanou konektivitu nejenom na databázové standardy, ale i přímo na konkrétní standardy z oblasti ERP - například SAP, BaaN, Oracle. Dalším z nepřehlédnutelných zdrojů informací se stávají webovské logovací soubory a protokoly, které zaznamenávají každé kliknutí na internetových stránkách na celém světě. Je zřejmé, že tady spočívá obrovský potenciál informací o chování uživatelů a tedy i potenciálních zákazníků.

3.2.3 Transformace vstupních dat

Data v datovém skladu jsou uložena jiným způsobem než ve zdrojových systémech. Provozní systémy používají v naprosté většině normalizovaný entito-relační datový model. U datového skladu se naproti tomu používá kombinace několika datových modelů (schéma hvězda, schéma sněhová vločka, normalizovaný entitně-relační model, denormalizovaný entitně-relační model, multidimenzionální datový model). Transformace dat se skládá z těchto dílčích operací: [8]

Validace – ověření správnosti dat, odhalení případných rozporů v datech

Čištění – odstranění či změna nesprávných dat

Integrace – dosažení konzistence dat pocházejících z různých systémů

Derivace – vytvoření derivovaných dat na základě vstupních dat

Denormalizace – snížení potřeby spojování tabulek při využívání datového skladu

Sumarizace – vytvoření požadovaných souhrnů z detailních dat

Prostředí DS musí poskytovat dostatek nástrojů a metod pro zvládnutí všech těchto kroků, vzhledem k povaze těchto procesů nepostačuje pouze SQL jazyk, ale je nutné používat specializované prostředky. Velmi důležitou fází je validace a čištění, protože základní podmínkou využitelnosti DW je to, že obsahuje důvěryhodná a správná data. Jakkoliv sofistikované metody na využití dat jsou v podstatě bezcenné, pokud nepracují se správnými daty. V této souvislosti se můžeme setkat s novým termínem ETL Quality. Transformaci dat lze tedy chápat jako proces získání co nejvyšší kvality dat, protože informace budou jen tak dobré, jak dobrá je kvalita dat.

3.2.4 Uložení dat

Uložení dat je poslední částí celého procesu, kdy jsou transformovaná data natažena do vlastního fyzického prostoru datového skladu a jsou přístupna pro vytěžování - pokládání dotazů. Data mohou být natahována ve stejném tvaru jaký mají cílové struktury, nebo mohou být natažena v předzpracovaném tvaru do takzvaných multidimenzionálních tabulek (kostek), které obsahují

předpřipravené podklady pro rychlé odezvy na dotazy zpracované podle jednotlivých dimenzí (hran kostky). Load je rovněž jako obě dvě předchozí fáze periodický. Data je možné ukládat na základě různých strategií:

- a) Pokaždé se uloží celý obsah DS znovu (použitelné pouze u velmi malých objemů dat či pro úvodní načtení)
- b) Ukládají se pouze přírůstky a změněná data (v tomto případě musí být k dispozici systém zajišťující rozpoznání změněných údajů).

3.3 Databáze datového skladu

Databáze datového skladu obsahuje konzistentní, pročištěná data OLTP systémů včetně jejich historie. Data jsou uložena ve formě, která jsou optimální pro reporting. Zde můžeme několik typů jako jsou tzv. hvězdicové schéma (Star schéma), schéma připomínající sněhovou vločku (Snowflake schéma), příp. tzv. galaxii (Constellation).

Star schéma

Hvězdicové schéma je v mnohém odlišné od standardní 3. normální formy relačního návrhu databáze. Skládá se ze dvou typů tabulek:

1. Tabulky faktů

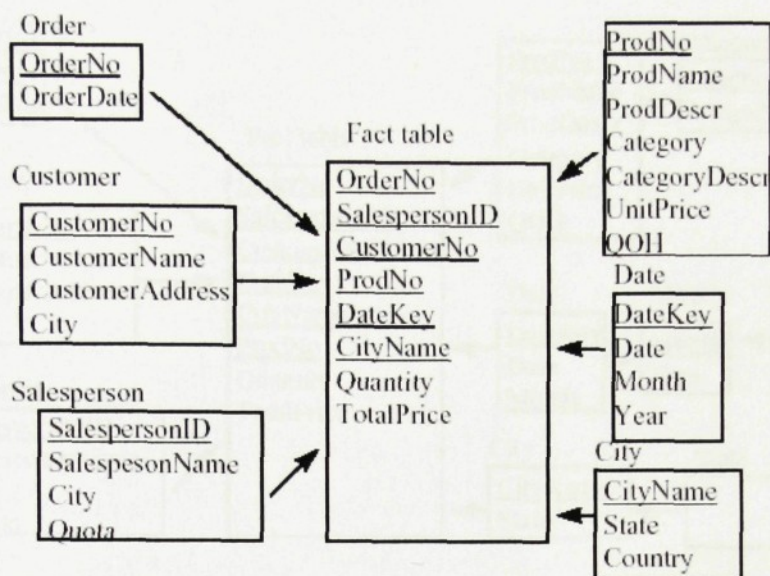
V těchto tabulkách jsou uloženy analyzované veličiny (počet prodaných kusů, prodej v Kč, počet telefonních hovorů, provolané sekundy atd.)

2. Tabulky dimenzí

Tyto tabulky obsahují hlediska, podle kterých fakta analyzujeme (čas, pobočky, druh produktu atd.). Dimenze mohou mít hierarchickou strukturu, např. rok, čtvrtletí, měsíc, den pro dimenzi času.

Tabulky faktů obsahují cizí klíče, které odkazují na nejnižší hierarchickou úroveň tabulek dimenzí. V tabulce faktu jsou tak uložena data s nejvyšší možnou granularitou a tabulky dimenzí určují, jak tyto data agregovat.

Obr. č. 16: Star schéma



Pramen: Data Warehouse Toolkit – Multidimensional Modelling – Kimball, Mergy, 2002, str. 125

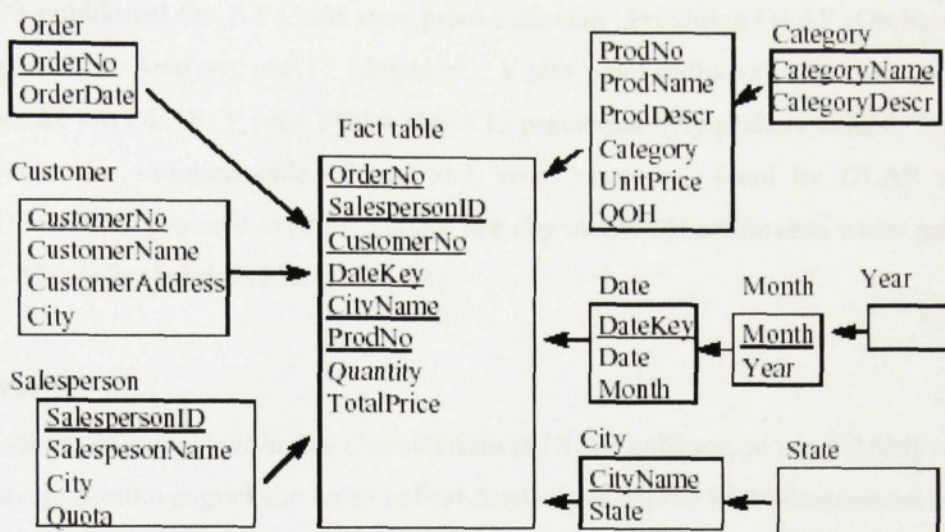
Star schéma na obrázku obsahuje jednu tabulku faktů a šest tabulek dimenzí. Primární klíč tabulky faktů je složený z cizích klíčů odkazujících do tabulek dimenzí. V ER diagramu je tak tabulka faktů centrem hvězdice, ze které vedou paprsky směrem k tabulkám dimenzí.

Relace dimenze : fakt je 1:N

Snowflake schéma

Snowflake (sněhová vločka) je určitým druhem hvězdicového schéma, ve kterém jsou tabulky dimenzí normalizovány, čímž se data rozdělují do dalších tabulek. Výsledné grafické schéma pak vytváří tvar podobný sněhové vločce. Hlavní rozdíl mezi těmito dvěma modely spočívá v tom, že tabulky dimenzí jsou normalizované, aby snížili redundance v uložených datech. Takováto tabulka je snadno udržovatelná a šetří diskový prostor. Ovšem tato úspora je zanedbatelná ve srovnání s typickou velikostí tabulky faktů. Navíc toto schéma může snižovat efektivnost analýz dat, neboť je zapotřebí provést více spojení tabulek, aby mohl být dotaz proveden. Proto může být výkon systému nepříznivě ovlivněn. Z tohoto důvodu není schéma sněhové vločky tak časté při návrhu datového skladu jako hvězdicové schéma.

Obr. č. 17: Snowflake schéma



Pramen: Data Warehouse Toolkit – Multidimensional Modelling – Kimball, Mergy, 2002, str. 145

Galaxie

Charakteristickým rysem je existence tabulek faktů, které sdílejí několik tabulek dimenzí. Toto schéma může být zobrazeno jako soubor hvězd a proto se nazývá „Fact Constellation“ (galaxie nebo souhvězdí)

3.4 OLAP

Proto aby bylo možné uložená data nějakým způsobem prohlížet a analyzovat, existují nástroje OLAP (On Line Analytical Processing). OLAP aplikace nejsou jediné, ale v oblasti datových skladů dominují. I mnoho odborníků považuje nesprávně termín OLAP za pokrytí celého procesu BI.

Nástroje OLAP (On Line Analytical Processing) jsou uživatelským rozhraním databáze datového skladu. Z názvu (on-line) je zřejmé, že se jedná o informační systémy, jejichž charakteristikou je rychlost a aktuálnost. Zpravidla obsahují standardní sadu reportingových nástrojů (grafy, tabulky, možnosti filtrování, zvýraznění extrémů..). Tyto nástroje jsou využívány především členy vrcholového vedení a podnikovými analytiky.

OLAP nástroje mohou pro své fungování využívat datový sklad, mohou však také fungovat bez něj. V minulosti byla samostatná řešení OLAP celkem rozšířená, dnes však OLAP využívá především datový sklad jako datovou základnu.

Definice dle E.F.Codda

V roce 1993 publikoval Dr. E.F.Codd svou práci s názvem „Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate“. V této práci definoval 12 vlastností, které jsou charakteristické pro OLAP. V roce 1995 k těmto 12 pravidlům byly přidány další 4. V současnosti existuje podle této definice sada 18 pravidel, resp. vlastností, které by OLAP systém měl splňovat.[i13] Těchto pravidel je příliš mnoho, než aby mohla být zmiňována v této práci – více o práci E.F.Codda je možné dozvědět se na [i14].

FASMI test

Jiným způsobem definice, kterým lze charakterizovat OLAP aplikace, je tzv. FASMI test. FASMI test představuje zkratku anglického termínu Fast Analysis of Shared Multidimensional Information. Test byl poprvé představen v roce 1995 a do dnešního dne nemusel být pozměněn. [i10]

FAST – systém je navržen pro poskytování většiny požadovaných odpovědí na dotazy jeho uživatelů během 5 sekund, s tím, že by ty nejjednodušší analýzy neměly trvat více než 1 sekundu. Pouze několik dotazů může trvat více než 20 sekund.

ANALYSIS – systém umí zvládat jakoukoli obchodní logiku či statistickou analýzu, kterou požaduje uživatel nebo aplikace

SHARED – systém musí zabezpečit veškeré bezpečnostní požadavky na důvěrnost dat. Jestliže umožňuje změny dat, musí současně vyřešit problémy aktualizace u vícenásobného přístupu k datům.

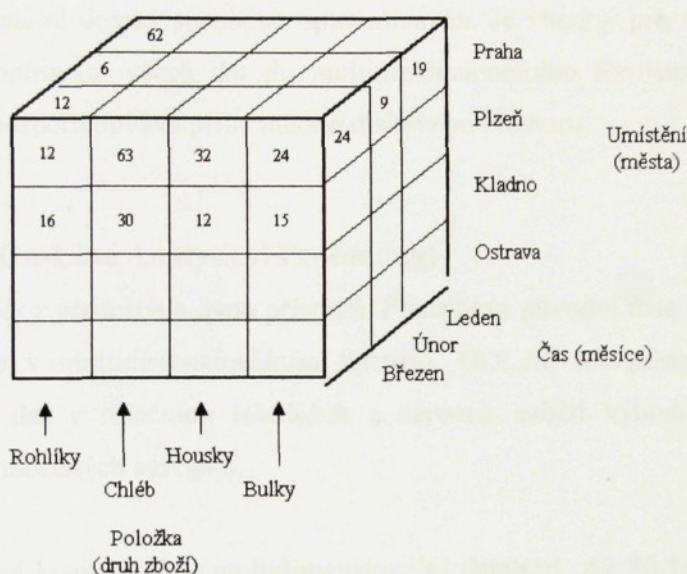
MULTIDIMENSIONAL – tento požadavek je označen jako klíčový. Systém musí poskytnout multidimenzionální konceptuální pohled na data, a to včetně plné podpory hierarchií resp. vícenásobných hierarchií tak, aby bylo možné analyzovat data přirozeným způsobem.

INFORMATION – pojem označuje veškerá data a odvozené informace, které jsou potřeba pro činnost aplikace. Měří se kapacita vstupních dat, která mohou být zpracována, ne kolik jich může být uloženo

3.4.1 Multidimenzionální architektura

Důležitou charakteristikou OLAPu je tzv. multidimenzionální architektura. Právě OLAP systémy poskytují multidimenzionální pohled na data, přitom nezáleží, jak jsou data v databázi fyzicky uložena. V praxi se multidimenzionální pohled vyjadřuje tzv. kostkou (cube), která je znázorněna jako obyčejná krychle. V ní pak hrany představují dimenze a buňky fakta.

Obr. č. 18: Multidimenzionální datová kostka



Pramen: www.adastra.cz

Tato datová kostka reprezentuje data ve třech dimenzích. A to dimenze Umístění, Času a Položky. Aktuálně zobrazenou úrovní dimenze Času je úroveň Měsíc. U Umístění je to Město a u Položky Druh zboží. Mírou tohoto zobrazení jsou Prodané kusy (v tisících). Potom např. hodnota „12“ udává, že v Praze v měsíci Březnu bylo prodáno 12 000 kusů Rohlíků. [i15]

Systémy OLAP se dají dělit do tří skupin podle toho, jak jsou data fyzicky v databázi uložena a zpracována. Setkáme se tak s termíny jako jsou ROLAP, MOLAP a HOLAP.

ROLAP (Relational On-Line Analytical Processing)

Při použití způsobu uložení dat ROLAP data zůstávají v původních relačních databázích, v tabulkách používajících star schéma. Nevytváří tak svou vlastní multidimenzionální kostku. Je nutné zdůraznit, že ROLAP by nesplnil podmínky zmíněného FASMI testu.

ROLAP určitým způsobem simuluje multidimenzionální přístup, jelikož relačně uložená data sám agreguje a předpočítává do požadovaného dotazu. Právě z tohoto důvodu trvají dotazy déle, jelikož agregace nejsou předpočítané. Přístup k datům se zajišťuje přes standardní prostředky SQL. ROLAP je více vhodný pro rozsáhlé databáze nebo na stará data, která nejsou často analyzována.

MOLAP (Multidimensional On-Line Analytical Processing)

MOLAP je multidimenzionální způsob uložení dat s vysokým výkonem. Proto přesně odpovídá definici OLAPU. MOLAP poskytuje nejlepší výkon ve fázi dotazování analýzy), neboť je právě pro mnohé dimenzionální dotazy speciálně optimalizován. Je vhodný pro malé až středně velké objemy dat, kdy kopírování všech dat do multidimenzionálního formátu nevyžaduje výrazně dlouhou dobu nebo nespotřebovává příliš mnoho diskového prostoru.

HOLAP (Hybrid On-Line Analytical Processing)

HOLAP slučuje prvky z předešlých dvou přístupů. Ponechává původní data v relačních tabulkách, ale ukládá agregace v multidimenzionálním formátu. HOLAP tak poskytuje propojení mezi rozsáhlými objemy dat v relačních tabulkách a zároveň nabízí výhodu rychlejšího výkonu multidimenzionálně uložených agregací.

Uživatel tak primárně komunikuje s multidimenzionální databází. Až 80 % dotazů vyžaduje jen agregované údaje a tyto dotazy tak rychlé obstará MOLAP. [8] Zbytek tak tvoří podrobné dotazy, které jsou vyřízeny pomocí ROLAP struktury. Dojde tak pouze k malému zpoždění.

Porovnání ROLAP a MOLAP systémů

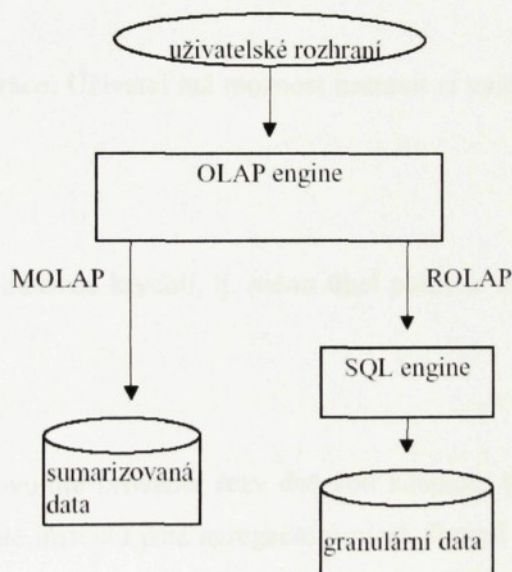
Z pohledu uživatele se dva zmiňované systémy chovají stejně. Stav, kdy jeden systém multidimenzionální přístup využívá a druhý ho jen simuluje, nepřináší žádné zásadní výhody či nedostatky. Jestliže však porovnáme tyto dva systémy z hlubšího pohledu, dojdeme k zásadním rozdílům. Rozdíly zachycuje následující tabulka.

Tab. č. 3 : Rozdíl ROLAP a MOLAP technologie

	ROLAP	MOLAP
Uložení, přístup	• Tabulky • Jazyk SQL • Agregace jsou předpočítány v agregčních tabulkách	• Pole, matice • Absence stand. jazyka • Uloženy jsou pouze buňky, které obsahují nějakou hodnotu
Využití	• Proměnlivý výkon • Relační prostředky	• Dobrý výkon • Multidimenzionální prostředky
Velikost databáze	• Gb až Tb • Velké místo pro indexaci • Snadný update	• Gb • % místa pro indexaci • Obtížný update

Pramen: <http://www.dc.uba.ar>

Obr. č. 19: Architektura MOLAP vs ROLAP



Pramen: *Data Warehouse Design Solutions – Christopher Adamson, Michael Venerable, str. 268*

Výkon

Lze konstatovat, že ROLAP technologie je vhodná pro větší kapacitu uložení dat. To však vede k větším nákladům na diskový prostor z důvodů indexace a tím i k větším nákladům na údržbu systému. Výhodou ROLAP je proces načítání dat, jelikož data pocházejí ze zdrojů, které mají podobnou strukturu. MOLAP systémy musí nejprve zdrojová data seřadit.

Jestliže bude firma zvažovat, jakým způsobem bude řešit svůj přístup k OLAP technologiím, nelze jednoznačně doporučit konkrétně ani jeden z nich. Vždy bude záležet na předmětu podnikání firmy, na dynamice růstů dimenzí a dalších aspektech.

3.4.2 Základní operace v OLAP systémech

Jak již bylo zmíněno, OLAP aplikace poskytují sofistikovanou multidimenzionální analýzu uložených dat. K tomu, aby tato analýza mohla být prováděna, musí OLAP nástroje poskytovat určitou minimální množinu funkcí pro práci s daty. Tyto funkce jsou následující [15]:

Drill-down

Operace drill-down umožňuje uživateli získat nižší (jemnější) agregační úroveň, na data je možné koukat z více detailnějšího pohledu v dané dimenzi, lze ji uskutečnit pomocí sestupu v hierarchii dimenzí nebo přidáním nových dimenzí.

Roll-up

Zde jde o opak předešlé operace. Uživatel má možnost nastavit si vyšší (hrubší) agregační úroveň na datové kostce.

Pivoting

Pivoting umožňuje „otáčet“ datovou krychli, tj. měnit úhel pohledu na data na úrovni presentace obsahu datového skladu.

Slicing

Operace slicing (selekce) dovoluje provádět řezy datovou kostkou, tj. nalézt pohled, v němž je jedna dimenze fixována v jisté instanci jisté agregační úrovně. Jinými slovy tato dimenze aplikuje filtr na instance příslušné agregační úrovně dané dimenze. Např. čas = 1. čtvrtěk. Výsledkem je podmnožina datové kostky.

Dicing

Dicing (projekce) je obdobou „slicingu“, jenž umožňuje nastavit takový filtr pro více dimenzí. Např. čas = 1.čtvrtěk a země = USA. Výsledkem je opět jiná podmnožina datové kostky.

3.5 Data Mining

Dolování dat (data mining) je relativně novým pojmem, který se objevuje v oblasti podnikových informačních systémů. Pod tímto termínem si mnoho účastníků provozu běžných informačních systémů představuje různé termíny. Lze najít mnoho definic, pro tuto práci jsem zvolil dvě následující:

„Dolování dat je proces odhalování závislostí, vzoru a trendů proséváním velkých objemu dat s využitím technologií na rozpoznávání vzorů, statistických a matematických technik.“

Gartner Group, 1996

“Datamining je netriviální proces zjišťování platných, neznámých, potenciálně užitečných a snadno pochopitelných závislostí v datech”. Fayyad, 1996

Data mining není nějakým zázračným prostředkem, který je schopen po shromáždění veškerých dostupných dat z podniku najít řešení všech problémů. Je to však prostředek s mnoha úspěšnými využitími v nejrůznějších oborech, jehož postupy se po tisících projektů v celém světě ustálily a v němž existují spolehlivé metody, které mnoha zákazníkům přinesly konkurenční výhodu a úspěch. Data mining nevznikl jako jednotný vědní obor a metody, které používá, patří do nejrůznějších oblastí. Jedinou společnou charakteristikou je jejich cíl: vyhledávat závislosti v datech.

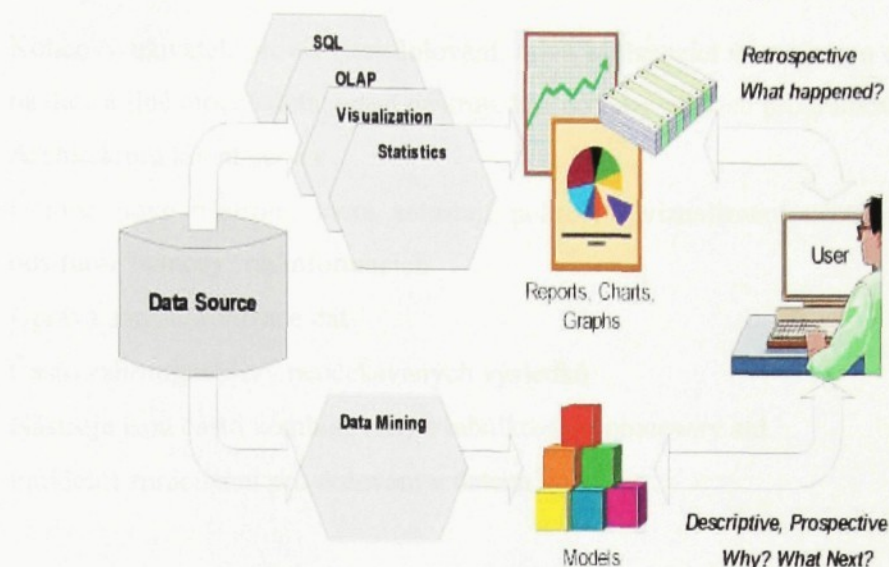
3.5.1 Zdroje dat

Jak již bylo zmíněno v první kapitole, dnešní svět zažívá explozi objemu generovaných dat a jejich následné ukládání do různých databází. Pro připomenutí lze zmínit některé oblasti, kde dochází ke zmíněnému pohybu různých typů dat:

- Bankovnínictví (bankovní transakce, žádosti o úvěr, historie splátek..)
- Telekomunikace (informace o telefonním provozu a platbách..)
- Státní správa (daňová přiznání, celní deklarace, žádosti o sociální podporu..)
- Pojišťovnictví (registrace pojistek a plnění..)
- Zdravotnictví (zdravotní záznamy, informace pro zdravotní pojišťovny..)

Tato data jsou ukládána do rozsáhlých podnikových databází, které nabývají formu datového skladu.

Obr. č. 20: Postavení a úlohy data miningu v procesu Business Intelligence



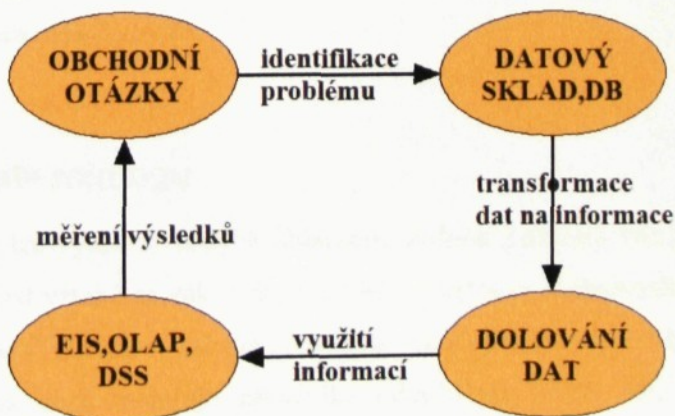
Pramen: www.tdan.com

3.5.2 Data mining a datové sklady

Kvalita vstupních dat přímo ovlivňuje správnost a přesnost výsledků dolování. To je také jeden z důvodů úzké návaznosti datových skladů a dolování dat. Sebedokonalejší modelovací technika či analýza nepřinesou očekávaný výsledek, pokud nejsou vstupní data očištěna od chyb, zkontrolována úplnost všech požadovaných údajů a sjednoceny formáty z různých systému.

Procesní charakter data miningu také vyžaduje průběžnou aktualizaci vstupních dat. Všem těmto požadavkům vyhovuje datový sklad. Nárůst zájmu o datové sklady je naopak jedním z hybných prvků rozvoje data miningových technologií.

Obr. č. 21: Virtuální cyklus dolování dat



Pramen: *Data Mining Techniques for Marketing, Sales and Customer Support*, Berry, Linhoff, str. 246

3.5.3 Hlavní charakteristiky a cíle dolování v datech

- Koncový uživatel, provádějící dolování, mívá k dispozici nástroje pro detailizaci pohledů na data a jiné mocné dotazovací nástroje bez potřeby znalostí programování
- Architektura klient/server
- Účinné nové nástroje, které zahrnují pokročilé vizualizační nástroje a pomáhají tak odstranit “nánosy” na informacích
- Úprava a synchronizace dat
- Často zahrnuje nálezy neočekávaných výsledků
- Nástroje jsou často kombinovány s tabulkovými procesory atd.
- Paralelní zpracování pro dolování v datech

3.5.4 Rozdíl mezi OLAP a dolováním dat

Pro mnoho lidí mohou termíny OLAP a data mining splývat do jedné definice. Zásadní rozdíly mezi nimi ilustruje následující obrázek.

Obr. č. 22: Rozdíl mezi OLAP a dolováním dat



Pramen: <http://e2.pef.czu.cz/pef/aktivita/>

3.5.5 Využití data miningu

Metody data mining lze využít v různých oblastech. Jednou z oblastí, kde je využití dolování dat zcela zásadní, je oblast vztahů se zákazníky. (CRM – Customer Relationship Management /řízení vztahu se zákazníky/). Údaje o zákaznících byly ceněny vždy, vzestup internetu a výhody členských karet však dnes motivuje zákazníky odevzdávat cenná data o svých nákupech a

demografických údajích více než kdy předtím, zbývá jen shromážděná data vhodným způsobem analyzovat. Data mining ve spojení s CRM je tím vhodným nástrojem.

CRM

Firmám zde data mining pomáhá pro analýzu stávajících i potenciálních zákazníků, pro analýzu jejich chování a kupního potenciálu. Z těchto dat se pak snaží získat nové skutečnosti a ty pak co nejlépe obchodně využít. K tomuto je však třeba alespoň základní orientace v základech statistických metod využívaných při dolování dat a také schopnost analyticky přemýšlet a správně výsledky interpretovat.

Kromě CRM může být dolování dat využito např. pro:

- Finanční sektor – předpovědi pomoci neuronových sítí
- Zdravotnictví - asociační analýza
- Bankovníctví – sekvenční a shluková analýza

3.5.6 Metody data miningu

Metody, se kterými data mining pracuje, jsou skutečně netriviální a v mnoha případech absorbují poslední poznatky z umělé inteligence a strojového učení. Zároveň však podstatným znakem data miningu je snaha zjištěné výsledky reprezentovat formou co možná nejpřístupnější uživateli. Výběr metody závisí na účelu zkoumání. Dle [i6] jsou uvedeny ty nejvíce využívané:

Prediktivní modelování

Přirozenou oblastí, ve které lze nalezené závislosti uplatnit, jsou například různé typy predikcí. Při nich předpokládáme, že pokud se data v minulosti chovala podle nějakého typu závislosti, budou se tak chovat i do budoucna. Pokud například pokles úrokové míry vedl v minulosti ze 70 procent k následnému vzestupu akciového kursu, můžeme tuto závislost předpokládat i do budoucna.

Rozhodovací stromy

Jejich výhodou je především snadná interpretace. Popis modelu pomocí rozhodovacího stromu je řadou jednoduchých rozhodovacích pravidel často prezentovaných ve formě grafu. Tyto grafy lze snadno interpretovat, bez širokých znalostí statistických metod.

Klasifikace

Klasifikační metoda je rozdělování dat do různých skupin podle jistých kritérií. Někdy jsou kritéria předem známa, ale mnohem častější je případ neřízené klasifikace, kdy výsledná kritéria předem

neznáme a úlohou klasifikace je jejich nalezení. Používanou technikou v takovýchto případech je shluková analýza.

Shluková analýza

Shluková analýza se používá k nalezení skupin podobných záznamů. Klasickým příkladem je segmentace trhu, která nalezne zákazníky s podobnými vlastnostmi a chováním. Pro každý segment lze spočítat průměrnou hodnotu zákazníka pro firmu. Obchodním výsledkem může být snaha udržet si výnosný segment zákazníků adresnou nabídkou lepších služeb.

Asociační analýza

Hledání asociačních pravidel, známé spíše jako analýza nákupního košíku, je vyhledávání hodnot, které se v jedné transakci vyskytují často zároveň. [8] Problém se vyskytuje především v oblasti prodeje, kde údaj o produktech, které se často kupují zároveň, může pomoci při umisťování produktů v prodejně (pokud se produkty umístí vedle sebe, zákazník si je spíše oba koupí). Cílem je odhalit pravidla typu: při nákupu zboží A, spotřebitelé výrazně častěji nakupují zboží D a B. Odhalení takových kombinací také pomůže marketingovým odborníkům v organizování nabídky či společných balíčků produktů nebo pro vytváření různých zvýhodněných balení.

Tab. č. 4: Vybrané techniky ve vztahu k úlohám data miningu a jejich použití v praxi

Model	Techniky	Úlohy	Popis	Použití v praxi
PREDIKCE	Lineární a nelineární regrese	Odhady hodnot	Pracná a časově náročná technika, výhodou je srozumitelná interpretace	Hodnocení rizika úvěru v bankovníctví
	Rozhodovací stromy	Klasifikace	Výhodou je jednoduchá laická interpretace	Hodnocení marketingové kampaně
	Neuronové sítě	Odhady hodnot, klasifikace, predikce	Modelování, velká variabilita možných modelů a snadná modifikace	Předpověď vývoje finančních a akciových trhů
DESKRIPCE (objevování znalostí)	Asociační analýza	Analýza vztahů	Typickým příkladem je analýza nákupního košíku	Definice komplementárních produktů pro zkoumané segmenty zákazníků
	Shluková analýza	Segmentace	Schopnost zpracování extrémně velkých modelů	Analýza chování klientů bankovních ústavů na základě provedených transakcí
	Neuronové sítě	Segmentace	Neuronové shlukování	Analýza nákupního chování zákazníků

Pramen: www.e-biz.cz

3.5.7 Nasazení projektu dolování dat

Jestliže se firma rozhodne implementovat data mining do své podnikové architektury, musí akceptovat určitá specifika. V poslední době se sice objevila řešení, která nejsou finančně tak náročná jako v minulosti, ale implementace není pouze v zakoupení daného softwaru a napojení jej do podnikové databáze. Cesta od okamžiku zakoupení produktu po finální výstupy data miningových nástrojů je mnohem složitější.

Nejdříve je třeba přesně specifikovat oblast, která reprezentuje pro management palčivý a pro data mining řešitelný problém a následně definovat datové zdroje. Datové zdroje pro data mining by měly být konzistentní a měly by obsahovat čistá data, proto faktorem, který nasazení data miningu podstatně urychlí, je přítomnost datového skladu. Pokud ve firmě datový sklad není dosud zaveden, bývá právě data miningový projekt dobrým důvodem jej vybudovat. Ani zde však práce s daty nekončí. Obvykle vzniká i u čistých dat potřeba další transformace a navazujících činností, které hledání závislostí napomohou.

Data mining umí zjistit závislosti v datech, není však schopen nahradit odborníka při posuzování vydolovaných pravidel. Ta mohou být triviální a zřejmá i laikovi, jiná pravidla může za triviální označit znalec poměrů, které v tom daném podniku panují.

4 Nástroje pro řešení a využití DW

4.1 Přehled řešení datových skladů v ČR

Náklady na vybudování rozsáhlé datové základny mohou být relativně vysoké. Proto si kompletní a unikátní řešení od velkých dodavatelů mohou dovolit pouze firmy, které jsou si jisty návratností velkých investic. Na tyto společnosti se zaměřuje pozornost velkých dodavatelů jako jsou Microsoft, Oracle, Microstrategy atd..

Technologie datových skladů však není určena pouze pro bohaté a velké společnosti. I malé a střední podniky musí znát a hodnotit své podnikání a postavení na trhu. Z tohoto důvodu existuje mnoho přednastavených řešení a standardních nástrojů, což značným způsobem snižuje náklady a také urychluje implementaci řešení. Řešení jednotlivých dodavatelů lze charakterizovat jako univerzální, pokrývající všechny oblasti průmyslu či služeb. Existují zde i nabídky datových skladů, které jsou speciálně zaměřeny na určitou oblast např. zdravotnictví, státní správa atd.

V rámci implementace celého procesu BI existují řešení, které svým rozsahem pokrývají veškeré procesy. Velké firmy mají ve svém portfoliu produktů nástroje, které řeší ETL procesy, uložení dat, OLAP nástroj, nástroje pro vizualizaci či data mining, a to vše v rámci jednoho produktu. Oproti tomu existuje celá řada řešení, resp. firem, které se specializují pouze na určitou fázi procesu BI. Tzn. existují firmy, jejichž produkty jsou ryze zaměřeny na ETL procesy či na data mining.

Na českém trhu je etablováno mnoho firem, které ve svém portfoliu produktů mají řešení datových skladů. Do širšího popisu jsem zařadil čtyři velká řešení (určené pro velké společnost) a jedno řešení menší (vhodná pro nasazení do firem střední až menší velikosti).

Je nutné poznamenat, že dle mého názoru podceňují firmy prezentaci svých produktů na internetových stránkách, které jsou pro celou řadu uživatelů, příp. i potencionálních zákazníků tím nejdostupnějším informačním zdrojem. Konkrétní informace o produktech jsou tak často velmi kusé a neúplné, lze se také setkat ze situací, kdy firma o svém řešení neprezentuje na této úrovni prakticky nic.

Tab. č. 5: Vybraná řešení v oblasti datových skladů

Název řešení	Dodavatel	Podrobnější informace
SAMAC MIT/400	Aimtec a.s.	www.aimtec.cz
MIS Charlie	AŘ MANAŽÉR a.s.	www.armanazer.cz
DSB Analyzer	Data software Brno s.r.o.	www.dsb.cz
Microsoft SQL Server	Microsoft s.r.o.	www.microsoft.cz
Oracle BI & DW Platform	Oracle Czech s.r.o.	www.oracle.cz
Datos Tools	Pražská softwarová s.r.o.	www.praguesoft.cz
DWH	PVT a.s.	www.pvt.cz
K2 OLAP	Q.gir s.r.o.	www.qgir.cz
SAP Business Information Warehouse	SAP BW s.r.o.	www.sap.cz
SAS DW Technologie	SAS Institute ČR s.r.o.	www.sas.cz
Datový sklad COMPEKON	Compekon s.r.o.	www.compekon.cz
Adaptive Server IQ	Sybase ČR, s.r.o.	www.sybase.cz

Pramen: Vlastní

4.1.1 ORACLE

Firmě ORACLE patří vedoucí místo v oblasti data warehousingu. Konkrétní výzkum udává až 37,5 % podíl na trhu. (IDC 2001). Firma ORACLE má dalekosáhlé reference z implementace své databázové technologie. Z tohoto důvodu může těžit ze své minulosti, kdy její robustní řešení byla nasazována masově ve velkých firmách v ČR.

Tab. č. 6: Vybrané vlastnosti produktu DW

Název:	ORACLE BI & DW PLATFORM
Poslední verze:	Oracle 10g
Web	www.oracle.cz
Odvětvová orientace:	universální
Počet instalací v ČR:	stovky
Určeno pro firmy:	Malé, střední, velké
Podpora databází:	Oracle 10g
Podpora oper.systémů:	Windows (NT,2000,XP), Unix, Linux

Pramen: vlastní

Popis:

Firma ORACLE poskytuje pro oblast DW robustní řešení, skládající se z několika modul, které řeší jednotlivé oblasti BI. Jedná se Oracle BI Applications, které je modulem pro konkrétní odvětví. Spolu s tím je dodáván Oracle Daily BI, který řeší otázku analýz konkrétních dat.

- Předdefinované a design vlastních analytických pohledů (tabulky a grafy - Oracle Discoverer, Oracle Reports), intuitivní ovládání
- Ad-hoc reporty nad Oracle databází realizované přímo z webového prohlížeče
- Přístup do aplikací odkudkoliv pomocí internetového prohlížeče, z PDA a mobilních telefonů včetně hlasového ovládání

4.1.2 MICROSOFT

Microsoft SQL Server 2000 je komplexní databázový a analytický software pro rychlou tvorbu rozšiřovatelných řešení nové generace pro elektronický obchod, skladování dat i operativní řízení. Analytické služby SQL Serveru 2000 poskytují podnikům funkcionalitu potřebnou k lepšímu rozhodování.

Podle výroční zprávy OLAP [i6] za rok 2001 vzrostl v roce 2001 podíl společnosti Microsoft na více než třímiliardovém trhu z 11,7 na 21,3 %, a Microsoft tak získal náskok oproti svým největším konkurentům v tomto oboru - společnostem Oracle a IBM.

Tab. č. 7: Vybrané vlastnosti produktu DW

Název:	MICROSOFT SQL SERVER
Poslední verze:	SQL Server 2000 Enterprise Edition
Web	www.microsoft.cz
Odvětvová orientace:	universální
Počet instalací v ČR:	stovky
Určeno pro firmy:	Malé, střední, velké
Podpora databází:	MS SQL server, Oracle 9i, Sybase IQ, IBM DB2, Caché
Podpora oper.systémů:	Windows (NT,2000, XP, Server)

Pramen: vlastní

Popis:

V podání firmy Microsoft je datový sklad volitelnou komponentou relačního databázového nástroje SQL Server. Součástí jsou ETL nástroje, integrované nástroje a algoritmy pro dolování dat, podpora přístupu přes MS Office.

Produkt MS SQL Server 2000 je možné rozdělit do několika částí:

- Microsoft SQL Server – relační databázový transakční (OLTP) systém
- Microsoft Data transformation Services – nástroj pro ETL procesy
- Microsoft OLAP Server – systém pro On-line Analytical Processing
- Microsoft Data Mining services – nástroj pro dolování dat

4.1.3 AIMTEC

Působení společnosti je úzce spjato s dynamickým rozvojem automobilového průmyslu v České a Slovenské republice posledních let. Společnost na svých internetových stránkách charakterizuje své zaměření způsobem: „Aimtec a.s. realizuje své projekty u zákazníků, kterými jsou jak globální dodavatelské firmy s výrobními aktivitami na území obou států, tak i lokální podniky, jejichž výrobní program směřuje do oblasti dodávek dílů a komponent evropským automobilkám“¹. I z toho důvodu je řešení této společnosti dostupné pouze pro platformu IBM eServer iSeries .

Tab. č. 8: Vybrané vlastnosti produktu DW

Název:	SAMAC MIT/400
Poslední verze:	--
Web	www.aimtec.cz
Odvětvová orientace:	universální
Počet instalací v ČR:	8
Určeno pro firmy:	Střední, velké
Podpora databází:	IBM DB2
Podpora oper.systémů:	OS/400

Pramen: vlastní

Popis:

Páteří řešení datového skladu společnosti SAMAC je multidimenzionální OLAP server MIT/400, SAMAC MIT/400 podporuje přístup k podnikovým datům v podobě datových tržišť. Server MIT/400 OLAP nabízí programovací nástroje pro vytváření rozhraní pro generování a výběr dat na různých úrovních. Pro vytváření jednotlivých dimenzí datových krychlí mohou být použity příkazy

¹ www.aimtec.cz

SQL. Pro generování komplikovanějších dimenzí mohou být vytvořeny procedury použitím programovacích jazyků AS/400 jako je RPG/400, COBOL nebo ILE/C.

Serverové aplikace

MIT/400 OLAP Server - administrace multidimenzionálních OLAP datových tržišť

MIT DynaServer/400 - relační databázový server

Udávané výhody[i7]:

- procesně uspořádaný přístup k datům v data martech
- otevřenost a nezávislost na aplikačním software
- snadná orientace v datech skladu pomocí hierarchické navigační struktury
- názorná prezentace v klientském, grafickém prostředí EXCEL, LOTUS
- krátká doba implementace - typicky hodiny pro jednu kostku, desítky minut na novou sestavu

4.1.4 SYBASE

Společnost Sybase ČR, s.r.o. je jedním z předních dodavatelů softwarové infrastruktury, aplikačního softwaru a IT služeb. Orientuje se na řešení zejména v těchto oblastech jako jsou Business Process Integration, Enterprise portály a další webová řešení, Mobile Computing a Business Intelligence. [i4]

Do produktového portfolia společnosti patří software pro operativní a analytické CRM, databázové servery pro OLTP a datové sklady, vývojové a case nástroje, portálové a aplikační servery, nástroje pro ETL atd.

Tab. č. 9: Vybrané vlastnosti produktu DW

Název:	SYBASE IQ
Poslední verze:	Sybase IQ
Web	www.sybase.cz
Odvětvová orientace:	universální
Počet instalací v ČR:	stovky
Určeno pro firmy:	Malé, střední, velké
Podpora databází:	Sybase
Podpora oper.systémů:	Windows (NT,2000,XP), Unix, Linux

Pramen: vlastní

Popis:

Řešení firmy Sybase se skládá z několika navzájem navazujících částí, které řeší proces BI. První ze součástí je Adaptive Server IQ, což je databázový server navržený pro ukládání a zpřístupňování dat pro aplikace datových skladů a business intelligence. Oplývá speciální patentovanou technologií vertikálního ukládání dat a zkracuje tak dobu odezvy i pro velmi složité dotazy nad velmi rozsáhlými daty až o několik řádů.[8] Další částí je komponenta Informatica, což je systém pro extrakci, transformaci a čištění dat vstupujících do datového skladu (ETL procesy). Nedílnou součástí je i PowerDesigner Warehouse Architekt. Tato část je určena pro návrh a modelování datových skladů. Pokrývá oblast extrakce zdrojových dat, modelování a generování databází datových skladů, návrh struktury analytických modelů, multidimensionálních a hierarchických struktur a generování metadat a aplikací pro celou řadu analytických a presentačních nástrojů.

4.1.5 SAS Institute

Společnost SAS je dalším z velkých a úspěšných společností působících v oblasti business intelligence. Produkty firmy jsou určeny pro průmyslové společnosti a blízce spolupracují se systémy pro řízení výroby a nabízí tak komplexní pohled na výrobní procesy

Tab. č. 9: Vybrané vlastnosti produktu DW

Název:	SAS Data Warehousing
Poslední verze:	----
Web	www.sas.com
Odvětvová orientace:	universální
Počet instalací v ČR:	40
Určeno pro firmy:	Malé, střední, velké
Podpora databází:	MS SQL server, Oracle 9i, Sybase ASE, IBM DB2,
Podpora oper.systémů:	Windows (NT,2000,XP), Unix, Linux, OS/390

Pramen: vlastní

SAS Process Intelligence se skládá z pěti částí. Information Store zajišťuje jednotný pohled na data týkající se výrobního procesu. Navigator je rozhraní, které slouží jako prostředník mezi koncovými uživateli a Information Store pro zadávání dotazů a generování odpovědí. Discoverer je webové rozhraní, které umožňuje regionálním nebo globálním organizacím analyzovat historická data. Agent funguje jako systém včasného varování, neustále sleduje data a identifikuje potenciální problémy dříve, než ovlivní kvalitu. Collaborative Space je prostředí, v němž si jednotlivé složky organizace vyměňují získané zkušenosti.

5 Specifika řešení DW a DM

V této kapitole se zaměřuji na využití datových skladů v prostředí internetu. Internet se postupem času stává čím dál více dominantním médiem pro oblast zábavy, vzdělávání, obchodu, komunikace a dalších mnoha činností. Je velmi pravděpodobné, že tuto svou pozici bude v budoucnosti stále více posilovat.

Už dnes poskytuje celosvětová síť dobré podmínky pro elektronické obchodování a dle různých názorů bude do několika let primárním kanálem pro zjišťování nákupních požadavků, vyřizování objednávek, reklamací a dalších souvisejících hospodářských činností. S touto tezí nemusí někdo souhlasit, ale zcela určitě musí být brána v potaz. Proto i podnikové datové sklady musí s internetem jako s dalším zdrojem dat počítat a měly by být takto navrženy a implementovány.

5.1 Využití datového skladu v prostředí internetu

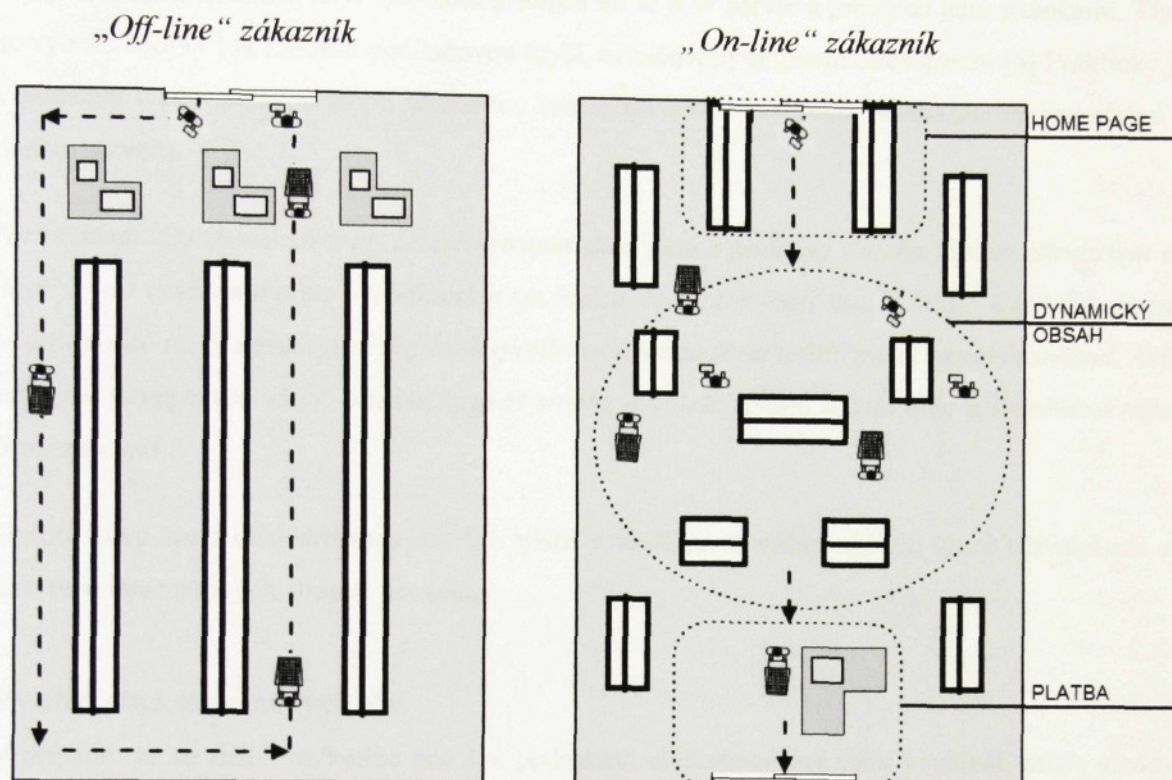
Absence osobního kontaktu zákazníka s prodávajícím, tak jak je známý z klasického podnikání, se dá považovat za jednu z největších nevýhod elektronického podnikání. Na druhou stranu však elektronický způsob podnikání může přinést celou řadu výhod, ke kterým se v klasickém případě těžko lze dopracovat.

Internet je médiem založeným na komunikaci mezi člověkem a technikou. Z toho plyne, že provozovatelé komerčních serverů mají možnost poznat „svého“ zákazníka o mnoho více, než jejich kolegové v kamenných obchodech. Internet se tak stává zdrojem velice cenných dat. Jestliže tato data firma vhodným způsobem zpracuje a použije je např. pro zkvalitnění své internetové nabídky zboží, dosáhne velké konkurenční výhody oproti firmám, které tento zdroj dat pro své podnikání nevyužívají.

5.1.1 Zákazník

Zákazníka lze definovat jako nositele poptávky po určitém produktu. Z tohoto hlediska se neliší klasický zákazník od svého „internetového“ kolegy. Lze je však odlišit podle prostředků, kterými své potřeby uspokojují. Klasický zákazník před sebou v obchodě tlačí nákupní košík, jeho interakce je závislá na prodávacích. Oproti tomu zákazník využívající elektronické nakupování, uspokojuje své potřeby z pohodlí domova, resp. z místa, kde má přístup k počítači a internetu. (viz obrázek č. 23)

Obr. č. 23: Rozdíl mezi klasickým a „internetovým“ zákazníkem



Pramen: *The Data Webhouse Toolkit – Kimball, Merz, 2000, str. 42-43*

Jedno internetové pravidlo říká, že vzdálenost on-line konkurence je jediné kliknutí myši. Virtuální hustota navzájem si konkurujících internetových serverů vytváří obrovské konkurenční tlaky. Z tohoto důvodu je nutné nasazení výkonných marketingových nástrojů, které zajistí, že potenciální zákazník nakoupí právě u toho „správného“ internetového prodejce. Základním kamenem v konkurenčním boji je sice kvalita výrobku či poskytované služby, avšak v dnešní době už toto není dostatečné. Důležitým nástrojem je budování vztahů k zákazníkům resp. vybudování dlouhotrvajícího partnerství založeného na oboustranné důvěře.

K budování takového partnerství potřebuje provozovatel internetového obchodu dostatek individuálních a detailních dat. Důležitou veličinou jsou nejen informace o zákaznické spokojenosti, ale také informace o jeho případné nespokojenosti. Pokud zákazník provedl nákup příp. jinou transakci, je stanovení úrovně jeho spokojenosti poměrně snadné. V případě, že k této situaci nedojde, bude získání konkrétních informací velmi obtížné. Z těchto důvodů získává stále více na důležitosti zpracování tzv. clickstreamových dat. Nástrojem, které toto zpracování provádí je tzv. clickstreamová analýza. Lze se také setkat s pojmy jako je web housing a web mining.

5.1.2 Clickstreamová data

K jejich získání dochází, když zákazník přichází na WWW server a prochází jeho stránkami. Tímto se vytváří spojitý tok kliknutí počítačovou myší, označovaný pojmem clickstream.[6] Prakticky jde v podstatě o uživatelskou stopu interakce, kterou po sobě návštěvník nesmazatelně zanechává na daném serveru.

Na příkladu lze demonstrovat, že clickstream data jsou z podstaty daleko lepším záznamem než např. často využívaná data o telefonních spojeních, tzv. CDR (call detail data). CDR data mohou ukázat pouze to, že strana A se úspěšně spojila se stranou B na určitý počet minut či sekund. Avšak už to nic nevypovídá např. o faktech, proč strana A volala stranu B, zda byla uskutečněna nějaká transakce atd..

Oproti tomu lze z clickstreamových dat zjistit rozsáhlou množinu aktivit, které uživatel při své návštěvě internetových stránek provedl.

Využití clickstreamových dat

V případě, že se firma rozhodne pro své podnikání clickstreamová data využívat, může sledovat celou řadu ukazatelů a zlepšovat tak svou pozici na poli elektronického obchodování. Mezi nejdůležitější udávané [6] veličiny, které firma pro své potřeby analyzuje, patří:

- Chování zákazníků
- Zákaznická loajalita
- Efektivita marketingu
- Efektivita prodeje
- Efektivita obsahu www stránek

Zdroje clickstreamových dat

Clickstreamová data lze získávat pomocí různých prostředků, metod a způsobů. Tyto prostředky se navzájem liší finanční náročností, nutností využití dalších výpočetních prostředků a samozřejmě i vypovídající schopností. Uvedené tři způsoby jsou v podstatě nejvyužívanější a nejdostupnější.

1. Logové soubory

Logové soubory představují nejrozšířenější způsob získávání dat. Logové protokoly produkuje prakticky každý webový server, jejich snadná dostupnost je tedy značnou výhodou. Výhodou je také standardní a snadno čitelný formát záznamu. Naopak nevýhodou je jejich omezený informační

obsah, protože tyto záznamy nejsou primárně určené pro detailní záznam interakce uživatele se serverem. [i10]

2. Odposlech komunikace

Druhou alternativou pro získání clickstreamových dat je odposlech síťové komunikace mezi uživatelem a serverem. Proces odposlechu se provádí na straně serveru. Tento záznam je obdobou logového souboru, obsahuje však více komplexnějších údajů včetně informací, které mohou pomoci zvýšit efektivitu a přesnost zpracování clickstreamových dat. Je nutné použít dodatečné zařízení a vhodný software, které odposlech provádí. Toto zařízení snižuje zatížení komunikačních linek a navíc odlehčí serverům, které tak nemusejí vytvářet komplikované logové záznamy.

3. Aktivní obsah

Poslední uvedenou alternativou pro sběr dat je vkládání aktivního obsahu do internetových stránek. V praxi to funguje tak, že součástí sledovaných internetových stránek je jednoduchý programový kód, který shromažďuje požadované informace přímo v prohlížeči uživatele a pak je následně odesílá na server. Spornou otázkou tak zůstává určité narušení soukromí uživatelů.

Uživatelské akce

Uživatelské akce lze definovat jako množinu určitých aktivit, které může uživatel na dané internetové stránce vykonat. Akce internetového uživatele mohou být motivovány různými potřebami. Tyto potřeby se mohou v jediném momentě změnit a vyvolat jinou akci. Zde je dle [7] výčet nejobvyklejších akcí, které může návštěvník www serveru učinit.

Vyhledávání – vyhledání určitého produktu, služby nebo informačních zdrojů

Sběr informací – srovnání výrobků, cen, pročítání FAQ

Studium – využívání manuálů, on-line příruček

Nakupování a objednávání

Komunikace – účast v diskuzích, využívání e-mailu

Sledování průběhu – zjištění stavu objednávky, zjištění skladových zásob

Download – stažení obrázků, audia, videa, softwaru

Náhodné akce – kliknutí na špatný objekt, chyby URL

5.1.3 Clickstream analýza

Clickstream analýza představuje rozsáhlou škálu analytických prostředků, aplikovaných na informace získané provozem obchodních, či jinak zákaznický orientovaných řešení v prostředí internetu. [i15]. Těchto dat je následně použito při analýze, která má za cíl pochopit chování zákazníka a těchto poznatků využít k prospěchu online komerčních aktivit.

Clickstream analýza produkuje dvě množiny charakteristik:

Provozní charakteristiky

Z těchto charakteristik lze vyčíst informace o počtu unikátních IP adres přistupujících na server, informace o celkovém počtu shlédnutých stran apod. Všechny běžné internetové servery umožňují sběr takového typu informací. V minulosti právě tato data byla předmětem zkoumání. Primárním účelem byla analýza provozu webových stránek, zjištění chyb a jejich následná náprava. Proto existuje mnoho nástrojů, které umožňují zmíněna data zpracovat a analyzovat.

Tyto nástroje však nelze považovat za nástroje clickstream analýzy. Jsou jakýmsi vedlejším produktem. Za ideální využití provozní charakteristiky je možno považovat proces optimalizace prodejní aplikace a odstranění jejích chyb. Typickými chybami, které mohou provozní charakteristiky odstranit, jsou nefunkční odkazy nebo stránky časově náročné na zobrazení v prohlížečích.

Charakteristiky komerční povahy

Tyto charakteristiky slouží k získání vzorců chování zákazníků online komerčních aplikací. Výsledky analýz obchodního charakteru mají napomáhat k tomu, aby se z návštěvníka serveru stal zákazník, případně aby stávající zákazník přinesl obchodníkovi vyšší obrát. Patří sem zejména predikce chování uživatelů na serveru, analýzy opuštěných nákupních košíků, cross-selling analýzy apod. Cross-selling je proces využívání znalosti, kdy víme, že nějaká část zákazníků kupuje produkt A v důsledku koupě produktu B. [i10] Právě využití takového poznatku vede k možnostem individuálních nabídek zákazníkům ve stejné skupině.

Typickým příkladem je např. při on-line nákupu oznámení typu „Zákazníci, kteří nakupující zboží A (o které jste jako zákazník projevil zájem např. přidáním do košíku) současně nakupují také zboží B a C“.

5.1.4 Návrh schématu datového skladu pro clickstream analýzu

V této kapitole je nastíněno, jakým způsobem se liší návrh klasického datového skladu od skladu určeného pro clickstream analýzu. Nejlepším způsobem, jak pochopit odlišnosti mezi nimi, je postupný návrh schématu vycházejícího z klasického datového skladu.

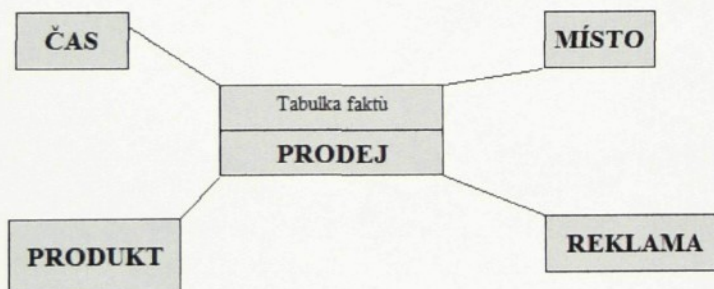
I. fáze

Na následujícím obrázku je staré schéma datového skladu pro klasický „kamenný obchod“. Pro zjednodušení je orientován pouze na sledování prodeje a marketingu.

Toto schéma obsahuje centrální tabulku faktů PRODEJ, která je obklopena tabulkami dimenzí ČAS, PRODUKT a REKLAMA. Tabulka dimenzí REKLAMA je určena pro sledování efektivnosti jakékoli podpůrné marketingové akce. To pak znamená, že podnik může dostat odpovědi např. na tyto otázky:

- *Jaké jsou nejlepší prodejní výsledky za rok 2002 oproti roku 2001 ?*
- *Jakým způsobem zvýšila naše marketingová kampaň prodej v Evropě za I.čtvrtletí ?*

Obr. č. 24: Zjednodušené schéma DW orientovaného na marketing a prodej.



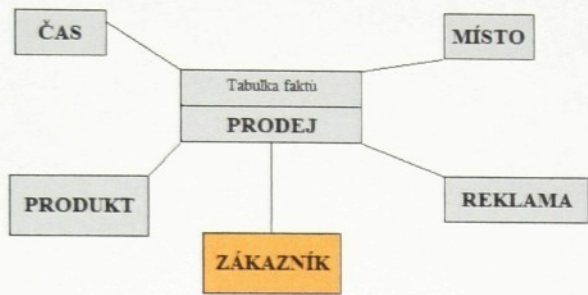
Pramen: vlastní

II. fáze

S příchodem trendu orientace na zákazníka (CRM) však podnik bude sledovat i své zákazníky. Proto bude do schématu přidána tabulka ZÁKAZNÍK. Poté bude moci zodpovědět následující možné otázky:

- *Kteří zákazníci z Evropy přibyli po marketingové kampani za IV. čtvrtletí ?*
- *Jaký je rozdíl mezi výdaji zákazníků, kteří přišli po marketingové kampani za III. resp. IV. čtvrtletí ?*

Obr. č. 24: Zjednodušené schéma DW orientovaného na CRM



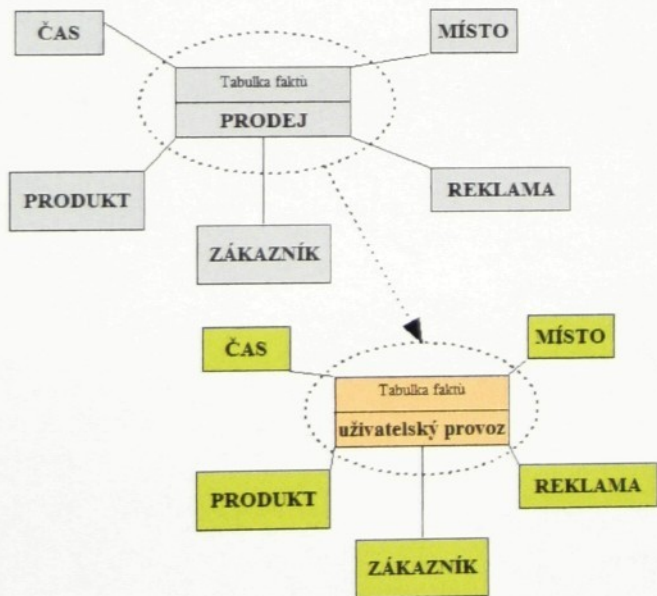
Pramen: vlastní

III. fáze

Nyní je možné postupně návrh schématu klasického datového skladu přetransformovat na datový sklad vyhovujícím potřebám elektronického obchodování, resp. určený pro pozdější clickstream analýzu.

Tabulka faktů PRODEJ se pro prostředí internetu změní na tabulku faktů UŽIVATELSKÝ PROVOZ. Veškerý uživatelský provoz, nejen pouze prodej zákazníkovi, může být v prostředí internetu zaznamenán. Pro připomenutí: může se jednat o náhodné procházení stránek, vyhledávání, veškerý strávený čas na stránkách apod.

Obr. č. 24: Tabulka faktů PRODEJ se stává tabulkou faktů UŽIVATELSKÝ PROVOZ

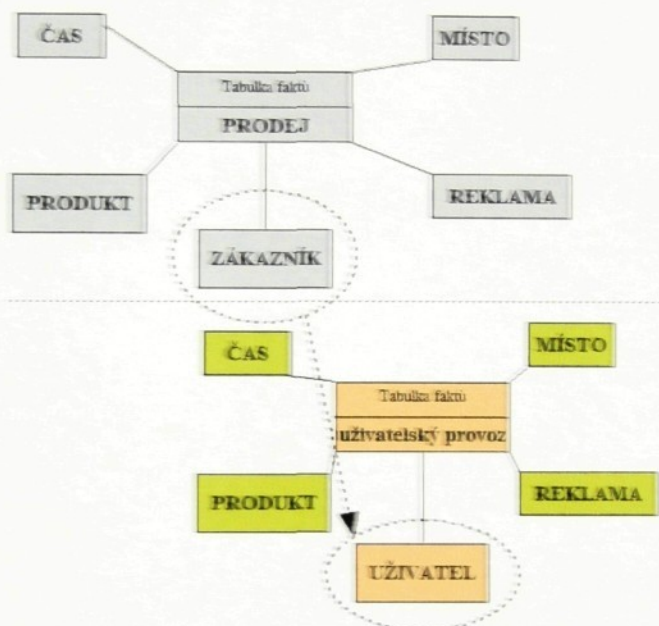


Pramen: vlastní

IV. fáze

V další fázi se tabulka dimenzí ZÁKAZNÍK transformuje do tabulky UŽIVATEL, zde je myšlen internetový uživatel. Uživatelé mohou být rozdělení do mnoha kategorií, jako jsou např. pravidelní zákazníci, občasní návštěvníci, náhodní návštěvníci, návštěvníci, kteří byli přesměrováni na stránky pomocí banneru, pomocí vyhledávače atd.

Obr. č. 25: Tabulka ZÁKAZNÍK se stává tabulkou dimenzí UŽIVATEL



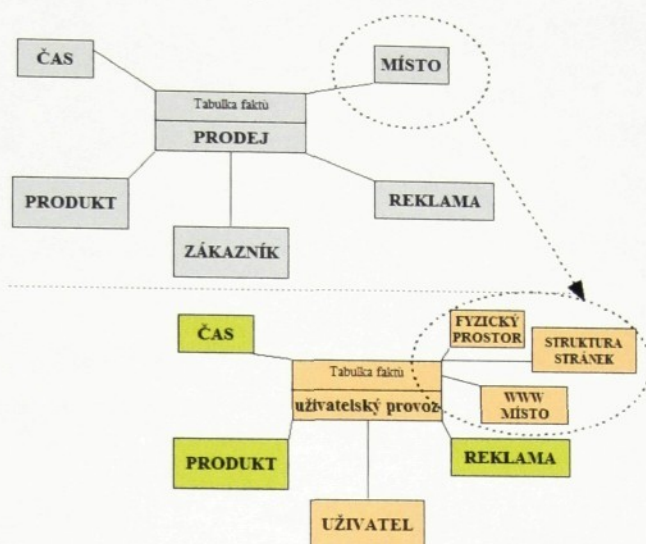
Pramen: vlastní

V. fáze

Tabulka MÍSTO se pro prostředí internetu rozpadá do dalších, nových dimenzí. Tabulka FYZICKÉ MÍSTO je analogií původní tabulky MÍSTO a udává tak fyzické lokální umístění, odkud návštěvník na dané stránky přichází.

Nová tabulka dimenzí WWW MÍSTO sleduje vstupní i výstupní webové stránky, včetně informací o externích odkazujících stránkách. Tento fakt umožňuje sledovat provoz, odkud uživatelé na stránky přicházejí a kde je opouštějí. Další novou tabulkou dimenzí je STRUKTURA STRÁNEK, která sleduje veškeré vztahy typu potomek / rodič mezi veškerými stránkami, vytváří tak dynamickou mapu stránek.

Obr. č. 26: Rozpad tabulky MÍSTO do nových tabulek dimenzí

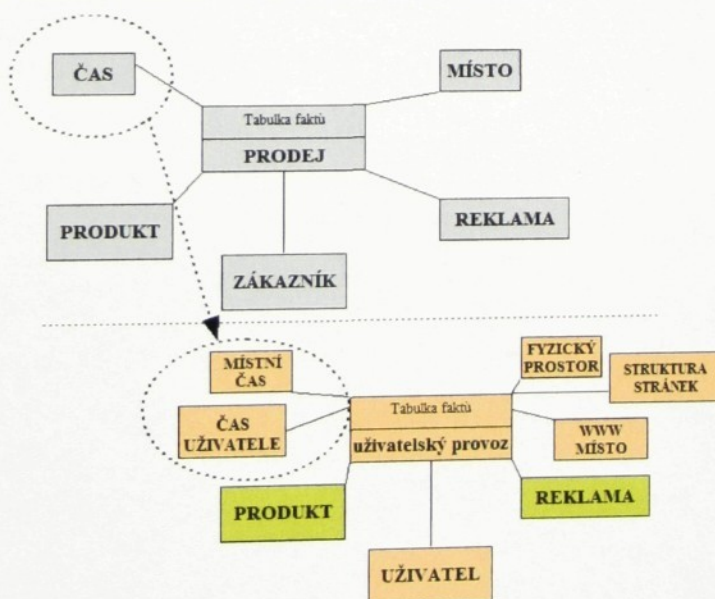


Pramen: vlastní

VI. fáze

Další tabulkou, kterou pro potřeby internetu je třeba upravit, je tabulka - dimenze ČAS. Rozdělí se teda na dvě nové tabulky dimenzí, MÍSTNÍ ČAS a ČAS UŽIVATELE. Dimenze MÍSTNÍ ČAS je opět analogií původní tabulky ČAS. Nová tabulka ČAS UŽIVATELE sleduje místní čas v zemi, odkud uživatel na stránky přichází. Tabulka tak kromě jiných věcí umožňuje analýzu uživatelů podle různých časových pásem, různých typů svátků apod.

Obr. č. 27: Rozpad tabulky ČAS do dalších dvou tabulek dimenzí



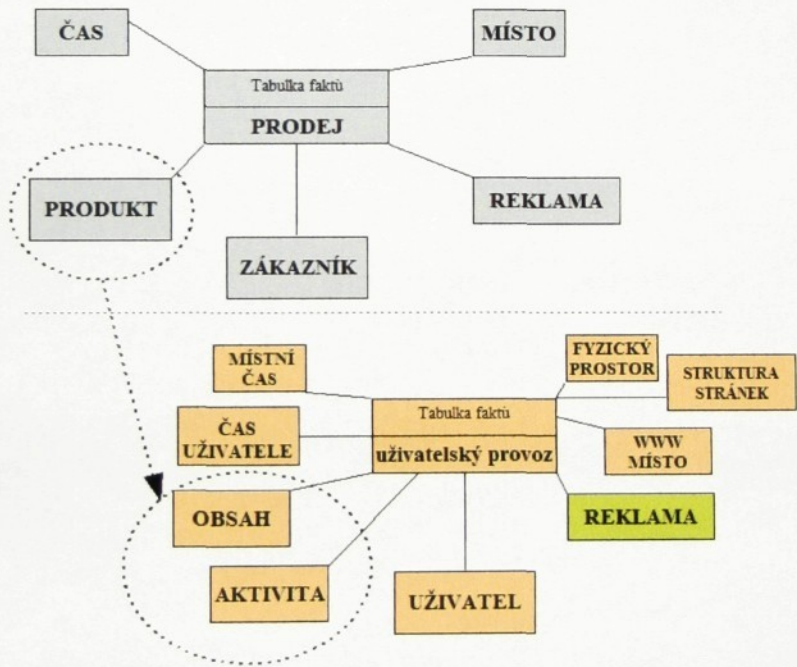
Pramen: vlastní

VII. fáze

Zbývá dimenze PRODUKT, kterou lze pro potřeby internetu rozdělit do dvou dimenzí např. s názvy OBSAH a AKTIVITA. Webové stránky slouží primárně jako zdroj informací, proto OBSAH vyjadřuje to, co je na stránkách uvedeno a AKTIVITA představuje určitou množinu akcí, které lze na daný obsah stránek aplikovat.

Např. jestliže provozuje výrobně orientovaná firma své stránky, pak jednou podmnožinou obsahu bude elektronický katalog jejich produktů. Webové stránky ale také obsahují další typy obsahu, jako jsou různé registrační formuláře, reklamační podmínky, soubory ke stažení, elektronické příručky apod. Pro každý typ obsahu pak lze definovat specifickou množinu aktivit uživatele.

Obr. č. 28: Rozpad tabulky PRODUKT

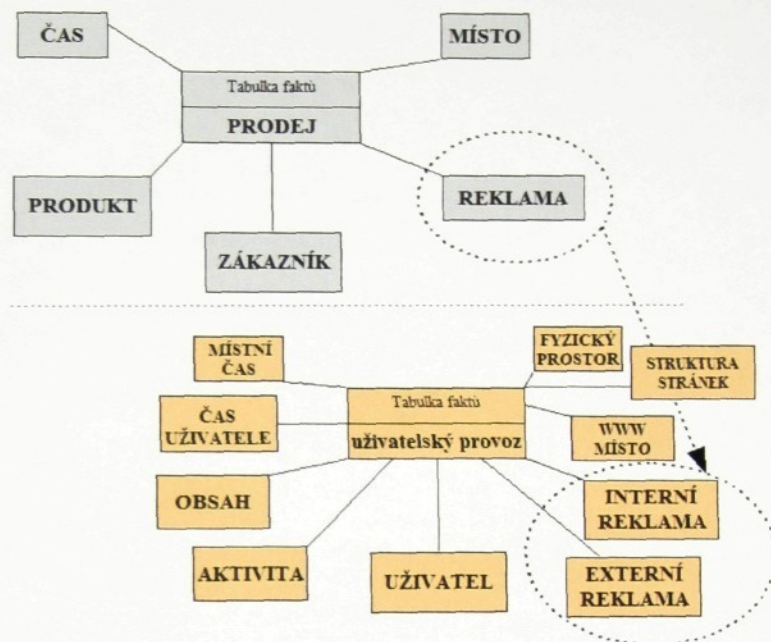


Pramen: vlastní

VII. fáze

V poslední fázi lze rozdělit tabulku REKLAMA do dvou dimenzí INTERNÍ, resp. EXTERNÍ REKLAMA. Právě specifické prostředí internetu umožňuje provést toto dělení. Na jedné straně existují reklamní akce v rámci vlastních internetových stránek, které mají za cíl přilákat např. nové nakupující. Na straně druhé existují jejich externí protějšky, jako jsou bannery, placené odkazy ve vyhledávacích apod. Toto dělení pak umožní sledování a vyhodnocování efektivity, resp. porovnání přínosů a nákladů, které tyto akce přinášejí.

Obr. č. 29: Rozpad tabulky REKLAMA



Pramen: vlastní

Toto navržené schéma lze považovat za pouze orientační. Klade si za cíl demonstrovat způsob, jakým způsobem by firma mohla postupovat při návrhu svého datového skladu určeného pro clickstream analýzu. V konkrétním návrhu podnik samozřejmě může použít jiné názvy dimenzí, lze použít jiné rozdělení původních dimenzí na vyšší či nižší počet. Prioritou musí být jasné vyjádření faktů, které chce firma v rámci provozu svých internetových stránek sledovat.

6 Závěr

Pro rozhodování není důležité velké množství dat a informací, ale schopnost abstrakce a odvození znalosti, která již těsně souvisí se schopností předvídat. V oblasti myšlení a abstrakce jistě žádný počítač a žádné automatické zpracování nemůže člověka nahradit. Nicméně tam, kde se jedná o velké množství vstupních dat nebo o analýzy opírající se o složité výpočty, může počítač člověku v zorientování se velmi pomoci.

Cílem této práce bylo popsat a vysvětlit úlohu datových skladů, které v rámci informačního systému hrají. Jejich role bude stále vzrůstat spolu s explozí objemu generovaných dat. Informace a znalosti budou či spíše dnes již jsou tím klíčovým místem, kde se rozhoduje o konkurenceschopnosti podniku. První část byla věnována teorii, konkrétně popisu jednotlivých součástí datového skladu a srovnání dvou různých přístupů a pojetí k této problematice. Druhá část pak byla praktická, věnovaná firmám, které mají řešení datových skladů ve svém produktovém portfoliu. Další praktickou částí je kapitola zaměřená na odlišnosti řešení datových skladů v prostředí internetu.

Hlavní přínosy práce jsou v nastínění úlohy, kterou dnes datové sklady hrají v rámci celopodnikového informačního systému. Nejsou pouhým, byť sofistikovaným úložištěm dat, ale stávají se důležitou součástí celého procesu business intelligence a zásadním způsobem tak ovlivňují celý chod podniku. Dalším přínosem je zmapování současného trhu existujících řešení problematiky data warehousingu, a to se zvláštním důrazem na Českou republiku. V neposlední řadě je také přínosem popis specifických datových skladů pro prostředí internetu spolu s návrhem konkrétního datového skladu.

7 Seznam literatury

- [1] Olivia Parr Rud, Data Mining, Computer Press 2001
- [2] Dohnal Jan, Pour Jan, Architektury informačních systémů, Ekopress 1997
- [3] Adamson Ch., Michael Venerable, Data Warehouse Design Solutions –, 2001
- [4] Kimball R, Margy R., The Data Warehouse Toolkit – MD Modelling, 2002
- [5] Kimball R., Merz R., The Data Webhouse Toolkit, 2000, Wiley
- [6] SKLENÁK, Vilém a kol. Data, informace, znalosti a Internet. 1. vyd. Praha, 2000
- [7] HUMPRIES, Mark a kol. Data warehousing, návrh a implementace, 2003
- [8] Hrnčířek Aleš, Datové sklady, diplomová práce, VŠE, 2002
- [9] Chaudhuri S. ,An Overview of Data Warehousing and OLAP Technology, 1997
- [10] E.F. Codd: The Relational Model for Database Management, 1990
- [11] Inmon W. H., Building the Data Warehouse, 2001

7.1 Internetové zdroje

- [i1] <http://www.ipesoft.sk/index.html>
- [i2] <http://www.dc.uba.ar>
- [i3] www.adastra.cz
- [i4] <http://mstr.insightstrategy.cz>
- [i5] www.systemonline.cz
- [i6] www.olapreport.com
- [i7] www.aimtec.cz
- [i8] www.sybase.cz
- [i9] www.microsoft.cz
- [i10] www.oracle.cz
- [i11] www.sybase.cz
- [i12] http://www.databaseanswers.com/codds_rules.htm
- [i13] <http://www.olapreport.com/fasmi.htm>
- [i14] www.tdan.com
- [i15] <http://lisp.vse.cz/~berka/docs/>
- [i16] www.e-biz.cz
- [i17] www.lupa.cz