Datové sklady

Obsah

1. Úvod - proč DW
2. DW v informačním prostředí organizace
3. Modelování DW
4. Problémy budování DW
5. Závěr

1. Úvod

Dvě kategorie přístupů k datům:

- Uživatelé přistupují k datům denně
  - OLTP (Online Transaction Processing)
  - Řízení změn dat v tabulkách během provádění obchodních procesů
    - Načítání určitých záznamů
    - Modifikace existujících záznamů
- Uživatelé hledají v množství dat agregované či odvozené informace
  - OLAP (Online Analytical Processing)
  - Sumární náhledy na data (reporty - sestavy)
  - Hledání vzorů v datech, rozhodování, udržení podniku na úrovni potřebné k úspěšnému přežití na trhu

Zdroj: historické data, multidimenzionální data

1. Úvod

Problémy s OLTP (online transaction processing):

- nedosažitelnost dat vytvořených či skrytých v transakčních systémech,
- dlouhé prodlevy, když se nedostatečně silné systémy snaží provést komplikované dotazy.

Teze: Datové sklady (Data Warehouses – DW) jsou dlouhodobou cestou k řešení těchto problémů

1. Úvod

Specifikace: co je DW?

- Skladování dat je kolekce technologií podporujících rozhodování, s cílem umožnit řídícímu pracovníkovi učinit lepší a rychlejší rozhodování

  Chaudhuri and Dayal, SIGMOD Record, March 1997

- DW označují db architekturu používanou pro údržbu historických dat, která jsou získána z jedné nebo více operativních db. Typicky, tato data jsou vyčištěna a restrukturována pro podporu dotazů, agregací a analýz

  Creative Data, Inc. 1997

1. Úvod

Klíčové: integrace vlastních + externích dat

Idea: návrh DW jako součást návrhu IS organizace

Cíle DW:

- S pokračující redukcí středních článků řízení, které vytvářely rozbory dat, by měl DW poskytnout podobné, nebo spíše kvalitnější služby.
- Poskytnout ne operativní data, ale tato data přeměněná ve strategické informace.

1. Úvod

Stádia dodávání dat v prostředí organizace

K. Sahin (1995): analogie s průmyslovou výrobou:

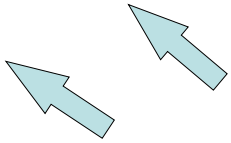
- výroba dat (OLTP ⇒ selektivní dotazy)
- skladování dat (DW, datová tržiště ⇒ dotazy intenzivní na data)
- prodej dat (OLAP)

Klíčové: multidimenzionalita, (multi)dimenzionální modelování (DM)

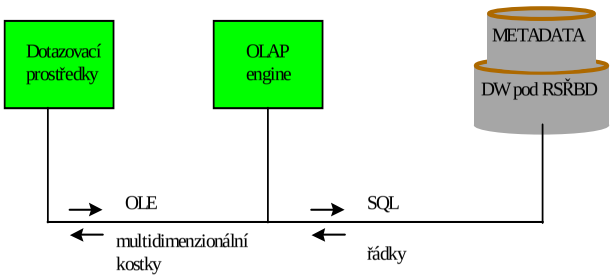
Standardizace: OLAP Council (1995) - Arbor Soft., Comshare, IRI Software, Pilot Software.

2. Problematika DW

- postavení DW v informační strategii organizace
- modelování DW
- konstrukce DW
- údržba DW
- použití DW



Komponenty DW



Komponenty DW

- akvizice dat a jejich integrace do DW (generátory kódu, replikace dat, middelware, kopírování)
- řízení dat (databázový server + služby: archivace, autorizace, zálohování a zotavení z chyb, provoz, monitorování a ladění, řízení zdrojů)
- slovník informací (metadata a přístup k nim)
- přístup k datům a komponenty dodání dat

  (db middelware, OLAP, multidimenzionální data, data řízená časem a událostmi)

Informační zdroje DW

- db pohled: integrace heterogenních dat

Přístupy:

  - přes globální schéma (neefektivní)
  - zvláštní databáze - DW (možnost využití replikací, automatické změny, průběžné výpočty agregací,...
- DW - architektury: FC, LC, LD, FD
- Dva hlavní problémy:
  - jak dostat data do DW
  - jsou-li tam, co s nimi dělat

Požadavky na funkčnost DW (1)

- **Dostupnost,** tj.DW musí být k dispozici uživatelům „neustále“. (Tento požadavek do jisté míry omezuje dobu, po kterou může být DW aktualizován)
- **Aktuálnost**, tj. DW by měl obsahovat co možná „nejčerstvější“ data z produkčních systémů. (Obyčejně je aktualizace DW aktualizován v noci, protože jde o proces, kdy HW je zatížen na maximum.)
- **Odezva,** tj. interaktivní zodpovídání rozmanitých analytických dotazů (OLAP) v „rozumné“ době (maximálně 3 minuty) i při značně rozsáhlých DW (až desítky GB)

Požadavky na funkčnost DW (2)

- **Čistota (kvalita) dat**, tj. jeden z fundamentálních požadavků (GIGO). Jde zejména o konsistenci „stejných“ dat vyskytujících se v různých produkčních systémů, jejich referenční a doménovou integritu. Právě zde vystupuje do popředí požadavek
  - jasně definovaného datového modelu
  - vybudování metainformačního systému (metadata – slovník a adresář)

  To je úkol zejména pro základní komponentu DW tj. program označovaný jako **ETL (extraction, transformation, loading**). Tato aplikace „sbírá“ data ze všech dostupných systémů, provádí jejich konsolidaci (čištění) a jejich následné nahrání do databáze DW.

ETL proces

- V datech přenášených do datového skladu se téměř vždy objevují duplicity a datové chyby, které není jednoduché odhalit. Tyto nepřesnosti způsobuje proměnlivé názvosloví („str. 56“/„stránka 56“), používání, či nepoužívání diakritiky („František Novák“/„Frantisek Novak“), pravopisné a jiné chyby, které způsobují nekonzistenci a je nutné je rozpoznat.
- Většinou neexistuje korekce chyb ve zdrojovém systému, proto je třeba nekonzistence dohledat, opravit a záznamy logicky spojit do jednoho při plnění datového skladu
- Tento proces čištění dat při plnění datového skladu může pak zpětně sloužit jako opravná zpětná vazba pro zdrojový provozní informační systém.
- Udává se, že až 15 % všech zdrojových dat je nekonzistentních nebo nesprávných.

Identifikace, kvantifikace, klasifikace a analýza chyb a problémů v datech

Cílem tohoto okruhu činností je poznat stav kvality dat a na jeho základě stanovit postupy, pravidla a standardy pro zvýšení a udržení kvality dat. Současně podle nastavených pravidel a mezí identifikuje záznamy a skupiny záznamů, které:

- systém automaticky opraví a upraví pro dosažení vyšší kvality,
- systém opraví a upraví na základě explicitního rozhodnutí (případně ověření),
- bude třeba ručně opravit z důvodu velmi nízké kvality, neúplnosti nebo zjevné nesprávnosti, již nelze automaticky opravit.

Zpracování dat pomocí čistících a unifikačních algoritmů

- **rozpoznání** (parsing) obsahu datových položek, opravy datových položek (odstranění překlepů, nesprávných zápisů, formátů apod.),
- **standardizace** – převod datových položek na jednotný formát, který je pak možno použít pro porovnání s rejstříky a číselníky a pro porovnání hodnot datových položek spravovaných různými systémy,
- **obohacení** – doplnění chybějících položek, pokud je to možné (např. chybějících částí adresy),
- **unifikace** – určení všech záznamů, které představují jeden konkrétní subjekt – např. nalezení a jednoznačné označení všech evidovaných záznamů o konkrétní osobě, adrese, vozidle atd.,
- **deduplikaci** – výběr nejlepšího záznamu, který bude nadále reprezentovat konkrétní subjekt,
- **identifikaci** – pro nové datové záznamy – určení konkrétního subjektu (například osoby), ke kterému záznam patří.

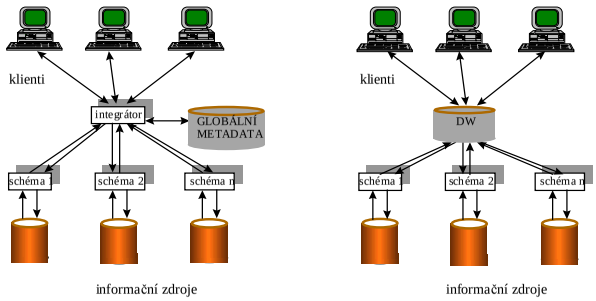
Vstupní data:

Row	FirstName	LastName	SSN	Address	Unit	Zip
1	Jane	Doe	NULL	123 MainStreet	NULL	22222
2	Jane	Doe	111111111	NULL	NULL	22222
3	J.	Doe	NULL	123 MainStreet	Apt 4	22222
4	NULL	Smith	111111111	123 MainStreet	Apt 4	22222
5	Jane	Smith-Doe	111111111	NULL	NULL	22222

Výstupní data (vyčištěná o duplicity a standardizována):

Row	FirstName	LastName	SSN	Address	Unit	Zip
1	Jane	Doe	111111111	123 MainStreet	Apt 4	22222

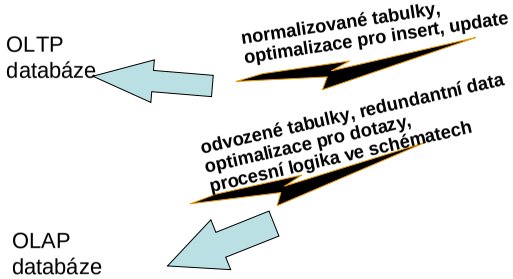
Dva přístupy k integraci dat



Dva přístupy k integraci dat

- virtuální pohled: je lepší, když se často mění informační zdroje
- materializovaný pohled: je lepší, když jsou informační zdroje stále a jsou třeba rychlé odpovědi

OLTP vs. OLAP



Vlastnosti DW

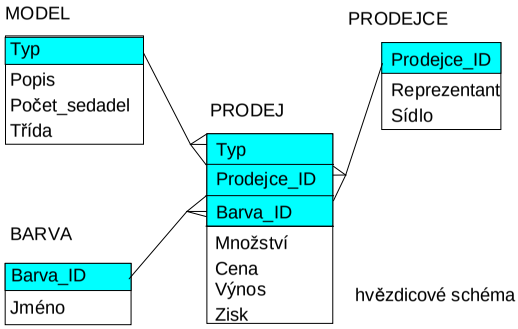
- poskytují metadata
- používají jednoduchá jména polí/sloupců
- používají multidimenzionální databáze (MDB)
- aktualizace jsou periodické (spíše v noci)
- prezentují data ve zjednodušeném, denormalizovaném tvaru
- Používají speciální logická schémata tabulek (hvězdice, sněhové vločky)
- podporují nástroje OLAP (Access/Excel, Safari, Cognos BI)
- odvozují data z (více) back-end OLTP systemů
- skladují historická data a mohou růst velmi rychle
- ukládají i agregovaná data

3. Modelování DW

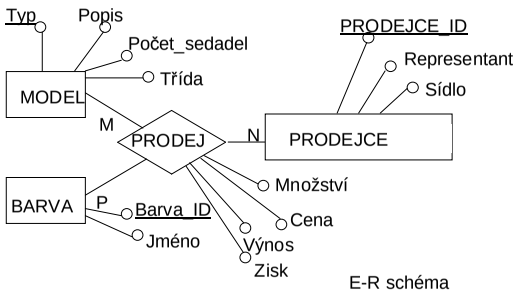
2 přístupy k DM:

- konceptuální struktury založené na *tabulkách* (dimenzionální a tabulky faktů) organizovaných do tzv. hvězdicových schémat,
- konceptuální struktury jsou založeny na hyperkostkách (kostkách, multidimenzionálních polích), které reprezentují data jako multidimenzionální strukturu.

DM pomocí tabulek a vztahů



DM pomocí tabulek a vztahů

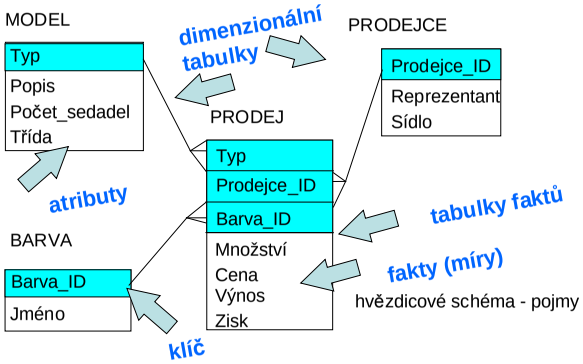


E-R schéma

DM pomocí tabulek a vztahů

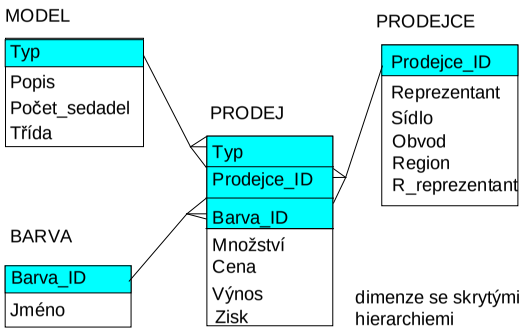
- DM je technika logického návrhu používající relační model dat s důležitými omezeními.
- Základní komponenty DM:
  - fakty, tabulky faktů
  - dimenze, dimenzionální tabulky
  - atributy,
  - řídkost.
- Výhody: jednoduché k chápání, jednoduché hierarchie, redukce počtu spojení, jednoduchá údržba, velmi jednoduchá metadata

DM pomocí tabulek a vztahů



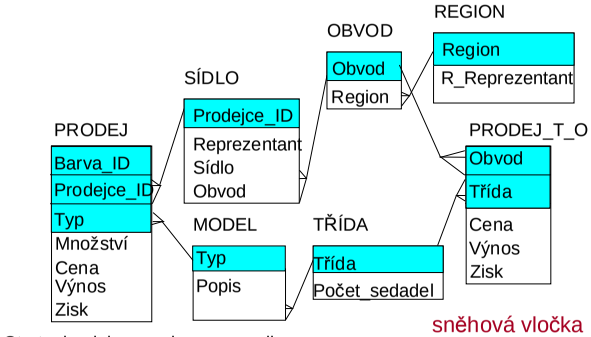
hvězdicové schéma - pojmy

DM pomocí tabulek a vztahů



dimenze se skrytými hierarchiemi

DM pomocí tabulek a vztahů

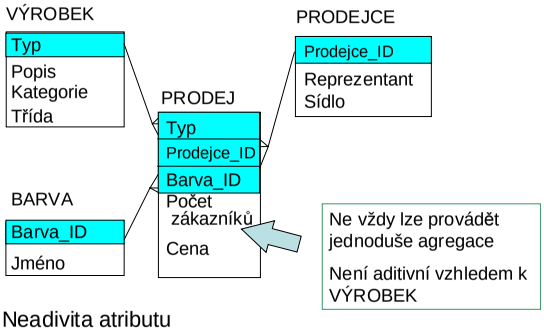


Strategie: dekompozice a normalizace dimenzionálních tabulek, více faktových tabulek

sněhová vločka

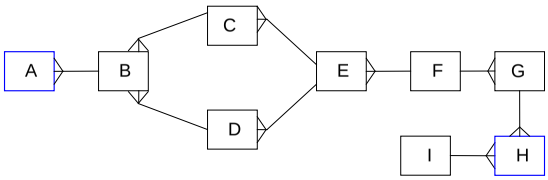
souhvězdí

DM pomocí tabulek a vztahů



Neadivita atributu

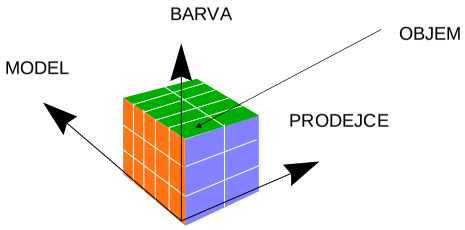
DM pomocí tabulek a vztahů



- členy hierarchií úrovně 0

hierarchie dimenzí

DM pomocí kostek



OBJEM - *množina dat*
další pojmy: *pozice*, *buňky*

(hyper)kostka

DM pomocí kostek

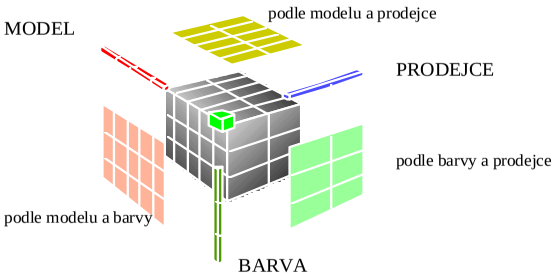
schéma MDB je množina vícerozměrných polí

MDB je dána vícerozměrnými množinami dat uloženými v těchto polích

výhody:

- pole nabízí přímo jisté informace např. počet pozic v každé dimenzi
- jednodušší vyhledávání (místo v řádcích tabulky se hledá v dimenzích a pak se přímo určí pozice buňky pole
- v poli se přirozeně seskupují data (⇒ řezy kostkou)

DM pomocí kostek



DM pomocí kostek

DM databáze pomocí tabulky

MODEL	BARVA	PRODEJCE	OBJEM
Škoda	zelená	Autosalon W	3
Škoda	zelená	Šebek	10
Škoda	červená	Autosalon W	1
Škoda	červená	Šebek	3
Škoda	bílá	Autosalon W	2
Škoda	bílá	Šebek	8
Mustang	zelená	Autosalon W	1
Mustang	zelená	Šebek	5
Mustang	červená	Autosalon W	2
Mustang	červená	Šebek	4

OBJEM	MODEL	BARVA	
Mustang	bílá	Autosalon W	8
Mustang	bílá	Šebek	1
Mirafiori	zelená	Autosalon W	2
Mirafiori	zelená	Šebek	3
Mirafiori	červená	Autosalon W	7
Mirafiori	červená	Šebek	1
11	Mirafiori	bílá	Autosalon
3	Mirafiori	bílá	Šebek

DM pomocí kostek

tabulka faktů:

prodej	prodejce	model	mn
	p1	c1	12
	p2	c1	11
	p1	c3	50
	p2	c2	8

kostka:

	c1	c2	c3
p1	12		50
p2	11	8	

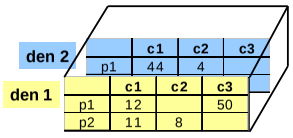
dimenze = 2

DM pomocí kostek

tabulka faktů :

prodej	prodejce	model	datum	mn
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

kostka:



dimenze = 3

DM pomocí kostek

Datová analýza kostek: inspirace v RMD

Kdy je vhodné DM?

Př.: OBJEM_PRODEJE(MODEL, PRODEJCE, BARVA, OBJEM)

Pozorování: mezi klíčovými atributy neexistuje žádná funkční závislost.

Př.: PRODEJCE(ROD_Č, JMÉNO, ADRESA, VĚK)

Dimenze: 100 prodejců, 80 jmen, 100 adres ⇒ 800 000 buněk. Uloženo bude ale jen 100 hodnot věku.

Nejvýhodnější situace: mezi dimenzemi existují multizávislosti.

Př.: Ve schématu OBJEM_PRODEJE platí:

DM pomocí kostek

DM vyšel z pohledu na data podle populárních programů typu spreadsheet.

Operace podporované na MDB:

- Agregace
- Pivoting (rotating) - změna vizualizace dat kostky
- Roll-up: od prodeje podle města k prodeji podle obvodu
- Drill-down: od prodeje podle obvodu k prodeji podle města
- Operátor Cube
- Slice_and _dice (redukce dimenzionality dat)

ROLAP, MOLAP

- ROLAP = OLAP s přímými relačními dotazy
  - Např. na materializovaných pohledech
  - Nebo hvězdicových schématech v DW
- MOLAP = OLAP s multidimenzionální databází (MDB)
  - MDB je speciální druh db
  - vidí data jako velký, rychlý spreadsheet

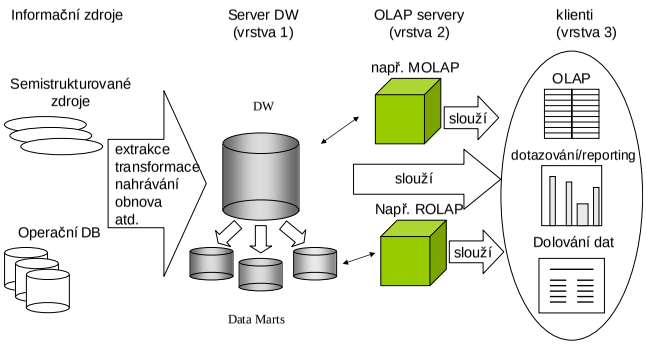
4. Problémy budování DW

- heterogenita
- migrace
- metadata
- prezentace výsledků

Koncepce: dvě vrstvy, tři vrstvy,

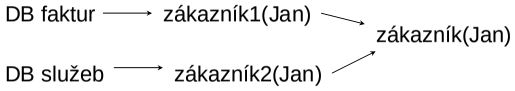
- Možnosti:
  - více menších specializovaných datových tržišť (např. pro jeden byznys proces)
  - operační data store

Příklad architektury



Čištění dat

- migrace (např. Euro → dolar)
- fúze (např. seznam emailových adres, seznam zákazníků)



- odhalování pravidel, vztahů (jako dolování dat)

Nahrávání dat

- Inkrementální vs. refresh
- Off-line vs. on-line
- Frekvence nahrávání
  - v noci, 1x za týden/měsíc, spojitě
- Paralelně/po částech

Implementace pomocí nových typů indexů

- GK index v Informixu

Příklad: časté dotazy na dodávky výrobků jisté třídy

- CREATE INDEX PRODEJ_TŘÍDA
  ON PRODEJ
  (SELECT AS KEY V.Třída
  FROM PRODEJ P, VÝROBEK V
  WHERE P.Č_výrobku = V.Č_výrobku)

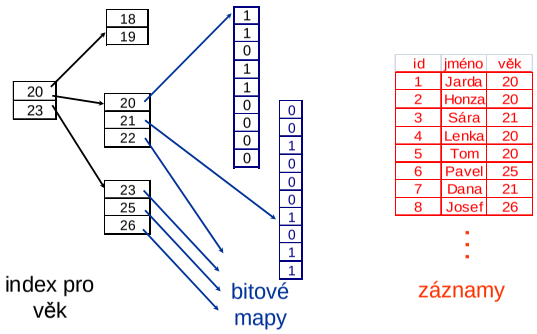
- Teorie: jde o speciální případ vícetabulkového spojovacího indexu

  Implementace: bitové mapy v listech B-stromu

Klasifikace DW

Vlastnosti	operační zdrojové systémy	Operační Data Store	DW	Datové tržiště
obsah	detailní data	detailní data + vhodné agregace	sumarizované informace + vhodné detaily	specializovaná funkčnost - sumáře
časová osa	současnost	skoro současnost	body v čase	body v čase
aktualizace	kontinuální	častá	periodická	periodická
požadavky na výkon	vyladěno na aktualizace	vyladěno na produkční prostředí	vyladěno na dotazování	neurčité
nestálost obsahu	velmi nestálý	nestálý	stálý	stálý
množství přistupovaných dat	nízké	optimalizováno na výkon	může být vysoké	mírné

Bitové mapy



id	jméno	věk
1	Jarda	20
2	Honza	20
3	Sára	21
4	Lenka	20
5	Tom	20
6	Pavel	25
7	Dana	21
8	Josef	26

záznamy

Bitové mapy

Použití bitových map:

- Query:
  - Get people with age = 20 and name = “fred”
- List for age = 20: 1101100000
- List for name = “fred”: 0100000001
- Answer is intersection: 0100000000000
- Good if domain cardinality small
- Bit vectors can be compressed

Proces zavádění DW

- založit organizační model,
- založit model řízení dat/informací,
- založit model vývoje DW,
- založit tým pro návrh a budování DW,
- založit technické procesy pro DW na základě operačních zvyklostí v podniku.

Společné omyly týkající se DW

- *DW jsou repozitáře pro všechny data podniku*

  spíše: více specializovaných DW (datová tržiště)
- *DW požadují relační databázi*

  existují i nerelační MDD, speciální servery
- *DW jsou vždy veliké*

  tendence: strategická data ne velkých objemů, vysoký stupeň agregace ⇒ rychlejší odezva na dotazy, vyšší nároky

5. Závěr

- Trendy: Směrem ke komplexním řešením business intelligence (BI)
- Software pro BI – sada nástrojů umožňujících uživateli přistupovat k datům podniku pomocí reportů, OLAP, kostek, grafů, ad-hoc dotazů a interaktivních analytických panelů (dashboards)