Možnosti tvorby datových skladů a metody dolování znalostí.

Thursday, May 30, 2013 8:22 AM

Základní problémy u běžných transakčních databázových systémů:

- nedosažitelnost dat skrytých v transakčních systémech
- dlouhá odezva při plnění komplikovaných dotazů
- složitá, uživatelsky nepříjemná rozhraní k databázovému softwaru
- cena v administrativě a složitost v podpoře vzdálených uživatelů
- soutěžení o počítačové zdroje mezi transakčními systémy a systémy podporujícími rozhodování

Cesta k řešení těchto problémů = datové sklady, tzv. Data Warehouse – DW

Datawarehouse

• 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.

• Klíčové: integrace vlastních + externích dat

Modely & Operátory warehousu

• Datové Modely
– relační
– hvězdice & vločky
– krychle
• Operátory
– slice & dice (řez & výřez)
– roll-up, drill-down (srolování, zavrtání)
– pivoting
– další

Komponenty DW

• **akvizice dat a jejich integrace** do DW (generátory kódu, replikace dat, middleware, 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 middleware, OLAP, multidimenzionální data, data řízená časem a událostmi)

Velká diskuze: E-R vs. multidimenzionální přístupy

2 přístupy k Datovému Modelování:

· **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.

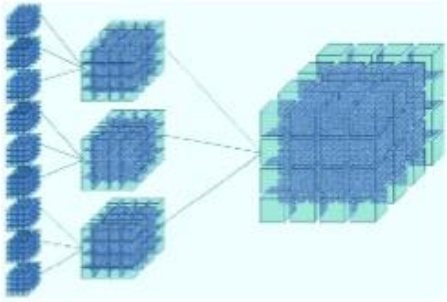
- samostatný informační systém postaven na již pořízených datech, určen především k jejich analýze
- architektura založená na ~~relačním~~ SŘBD, která se používá pro údržbu historických dat získaných z databází operativních dat, jenž byla sjednocena a zkontrolována před jejich použitím v databázi DW
- data z DW jsou aktualizována v delších časových intervalech, jsou vyjádřena v jednoduchých uživatelských pojmech a jsou sumarizována pro rychlou analýzu

- DW je obrovská databáze obsahující data za dlouhé časové období
- často slučuje data z více rozdílných zdrojů, které mohou obsahovat data různé kvality nebo používat nejednotné formáty a reprezentace
- objemově zabírá stovky GB až několik TB
- nemusí být databází v běžném smyslu, tj. pro přesné provádění transakcí
- je určen pro rychlé vyhledávání
- nejsou kladeny nijak důrazné požadavky na správnost a úplnost dat

Charakteristika

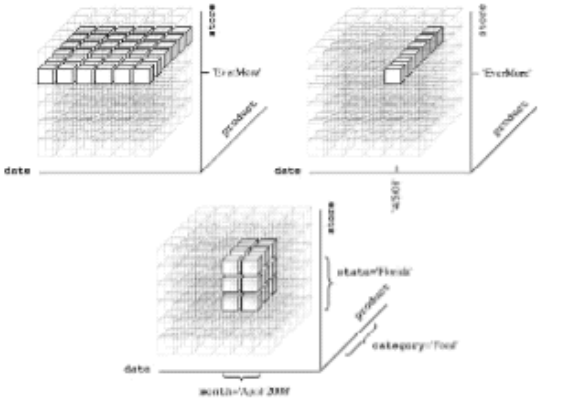
- data jsou uložena na různých místech ve formě relačních tabulek
  - uživatelé mohou tabulky jen číst
  - zapisovat může aktualizační program pravidelně udržující tabulky
- dotazy jsou většinou komplexní
  - podporují tzv. on-line analytické zpracování (OLAP)
    - výrazně se liší od on-line transakčního zpracování (OLTP)
    - operační databáze je přizpůsobena pro podporu OLTP
      - složité OLAP dotazy by vyústily do nepřijatelné odezvy
    - typické OLAP operace
      - **roll-up** (zvýšení stupně agregace)

Roll-up



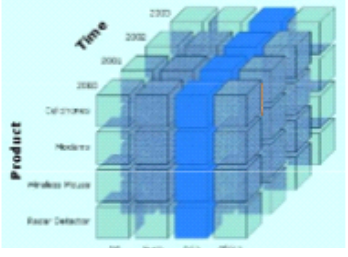
      - **drill-down** (snížení stupně agregace)

Drill down

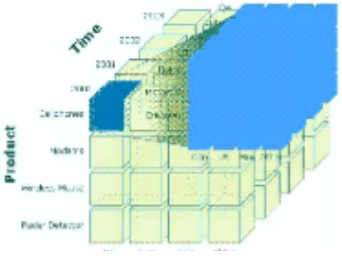


      - **slice-and-dice** (selekce a projekce)

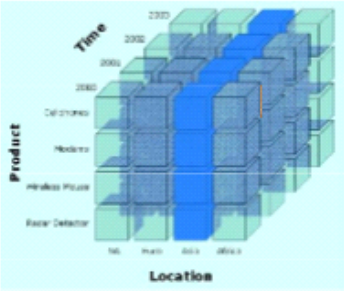
Slice



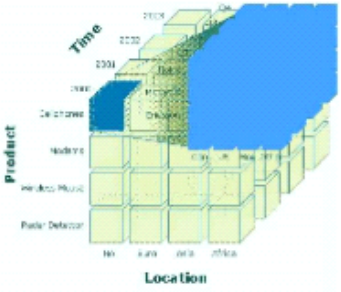
Dice



Slice

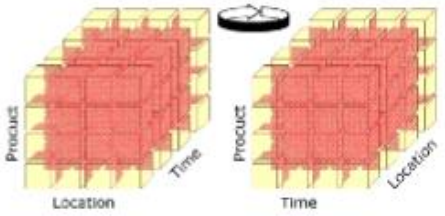


Dice



- **pivot** (přeorientování vícerozměrného pohledu na data)

Pivot



- na základě dotazu se pospojují potřebná data do vícerozměrné tabulky (nebo více tabulek), do kterých lze klást SQL dotazy
- pro častější dotazy si uchovávají předem připravené vícerozměrné tabulky
- zátěž je většinou způsobena složitými dotazy, jež přistupují k miliónům záznamů a provádějí množství operací
- data bývají modelována vícerozměrně
  - v obchodním data warehouse mohou těmito rozměry být např. čas prodeje, místo prodeje, prodavač, výrobek, …
  - rozměry mohou být i hierarchické např. čas prodeje jako den-měsíc-čtvrtletí-rok, zboží jako výrobek-kategorie-průmysl
  - spojení více tabulek pomocí odkazu na řádky jednotlivých tabulek
  - používají speciální organizaci dat, přístupové a implementační metody, jež obecně nejsou v komerčních databázových systémech určených pro OLTP podporovány

Databázový systém – OLTP (Online Transaction Processing Systems)

- zákaznicky orientovaný
- aktuální data -- lze považovat i za slabinu, při výpadku (chybě), vznikají ztráty pro byznys
- ER schéma
- sofistikované atomické transakce i přes několik systémů(bank, po síti,...)
- velikost DB až několik GB
- jednoduché a efektivní
- příkladem je bankomat

DataWarehouse – OLAP (Online analytical Processing)

- orientovaný na trh, rychlé (oproti OLTP) získání výsledků na analytické dotazy
- historická data, multidimenzionální datový model
- agregovaná data (nenormalizovaná=redundantní)
- schéma hvězdy či vločky
- převážně pouze čtení
- velikost až TB
- použití: byznys reporty o prodeji, marketing, management reporty, rozpočty, finanční předpovědi a reporty

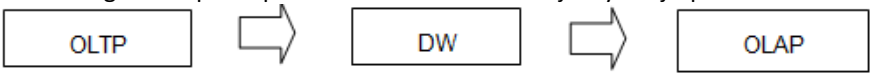
OLAP (**Online Analytical Processing**) je technologie uložení dat v databázi, která umožňuje uspořádat velké objemy dat tak, aby byla data přístupná a srozumitelná uživatelům zabývajícím se analýzou obchodních trendů a výsledků (Business Intelligence). Způsob uložení dat se svým zaměřením liší od běžněji užívaného **OLTP** (**Online Transaction Processing**), kde je důraz kladen především na snadné a bezpečné ukládání změn v datech v konkurenčním (víceuživatelském) prostředí.

Vloženo z <http://cs.wikipedia.org/wiki/OLAP>

Použití DW

- prezentace dat
- testování hypotéz
- objevování nových informací

Architektura DataWarehouse

- tři úrovně:
- klient
- OLAP server (MOLAP/ROLAP server)
- databázový server DW
- data lze organizovat v tzv. multidimenzionálním datovém modelu
  - odlišný od modelu relačního
  - odpovídá mu specializovaný software, multidimenzionální SŘBD (MDD)
  - model připomíná techniku spreadsheet ve více než dvou rozměrech
  - data jsou implementována pomocí vícerozměrných polí, jejichž dimenze odpovídají dimenzím podnikání organizace
- navržení a vytvoření DW je proces skládající se z následujících bodů:
  - definovat architekturu, umístění a rozčlenění dat a fyzickou organizaci
  - naplánovat kapacitu, vybrat OLAP servery a nástroje
  - spojit servery, klientské nástroje, zdroje přes gatewaye, drivery ODBC, ...
  - navrhnout schéma a pohledy, přístupové metody, některé složité dotazy
  - mít skripty pro získávání, čištění, transformaci, ukládání a aktualizaci dat
  - vytvořit koncové uživatelské aplikace
  - spustit data warehouse i aplikace
- vytvoření je složitý proces trvající mnohdy i několik let
- mnoho organizací proto používá Data Mart umožňující rychlejší práci
- 

Datová tržiště (Data Mart)

- DW slouží jako základna pro extrakci množin dat, resp. jejich agregaci do dílčích (replikovaných) MDD (Multidimenzionální DB)
  - MDD může pro DW sloužit ve dvou rolích
    - ”front-end” pro DW a poskytovat uživateli služby pro realizaci analytického zpracování (DW/OLAP)
    - “front-end” jednomu (několika) systémům OLTP - alternativa za DW, tj. poskytnout uživateli s OLTP data analytickým způsobem (OLTP/OLAP) – jde vlastně o datové tržiště

Systém OLAP (OnLine Analytical Processing)

- na databázové stroje jsou kladeny specifické požadavky
- objem zpracovávaných dat
- transakční systém o velikosti gigabajtů dosáhne použitím jen jedné dimense velikosti desítek či stovek gigabajtů
- rychlost odezvy analytického systému je důležitá
- počet uživatelů současně pracujících s databází není zajímavý
  - počet pracovníků vyššího managementu je omezen
  - pro pracovníky nižších stupňů bývají údaje z datových skladů převedeny do menších specializovaných databází – datových tržišť
- s těmito omezeními se vyrovnává dvojím způsobem
- uzpůsobení stávajících systémů pro práci s vícerozměrovými daty
- přidáním modulu, který to zajišťuje a prostředků pro jeho ovládání
- v lepším případě mění způsob uložení dat, v horším“překládá” operace s

vícedimenzionálními daty na operace s daty relačními
- vytvoření speciálního systému správy dat, určeného pouze pro OLAP
- umožňuje provést maximum optimalizací vzhledem k nárokům, jež jsou kladené analytickým způsobem práce - převažující způsob

Programy pro vytváření a plnění databáze (ETL - viz SI)

- převodní programy
  - načtení data z několika databází, či souborů a udělat z nich novou databázi, agregace se musí naprogramovat
- systémy znázorňující převodu dat graficky a administrátor dat namapuje zdrojová data do struktur vytvářeného datového skladu
  - výsledkem jsou buď programy (scripty) nebo přímo vykonání funkce
- moduly pro plánování jednotlivých akcí

Nástroje pro práci s daty - poslední trendy v architektuře klient/server

- nabízejí variantu tenkého klienta v podobě HTML prohlížeče

Reporting, monitorování, ad-hoc dotazy

- programy umožňující kladení dotazů a formátování odpovědí
  - nejčastěji jde o vizuální dotazovací nástroje
  - makra v tabulkovém procesoru
  - uživatelské rozhraní různě propracované:
    - zadání seskupení výsledku podle různých kritérií
    - formální kontrola dotazů
    - vytváření slovníků a metadat

MOLAP - Multidimenzionální OLAP

- datová krychle (obsahuje fakta)
- hierarchické dimenze (částečné či totální uspořádání)
  - **vločkové schéma** -- hlavní tabulka faktů je v relaci s dimezionálními tabulkami, přes cizí klíče, dimenzionální tabulky mohou být také v relaci s dalšími subdimenzionálními tabulkami podobně jako hlavní tabulka faktů; vytváří hierarchie dimenzí
  - **hvězdové schéma** -- je speciální případ vločkového, dimenzionální tabulky již nejsou v relaci s dalšími subdimenzionálními tabulkami; žádné hierarchie, jednodušší

ROLAP – Relační OLAP

- na relační architektuře založený model DW strukturou propojených DB tabulek - Relační OLAP (ROLAP) – pomalejší zpracování než MOLAP
- užívá relační nebo rozšířený relační DBMS, např server METACUBE Informix, pracuje s relačními tabulkami uspořádanými do hvězdy/vločky, adresuje pomocí klíče, data jsou neagregovaná)

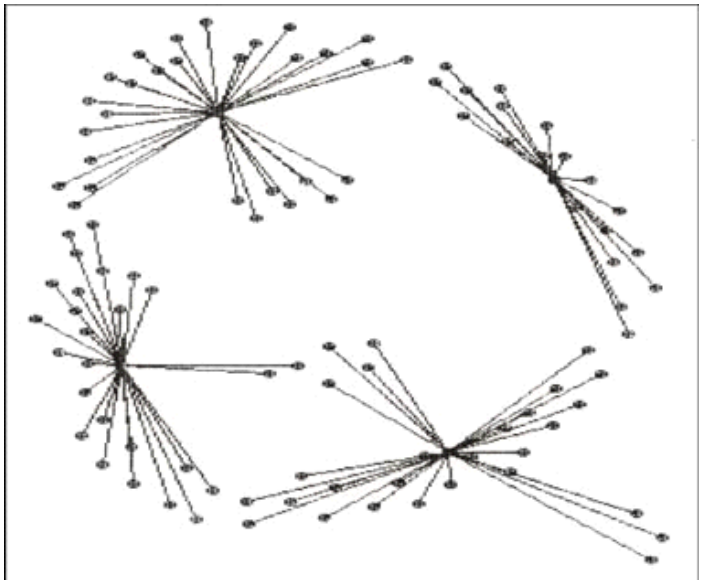
Metody dolování znalostí

Asociace

- Klasické – mezi dvěma podmnožinami atributů
- Transakční – v rámci množiny atributů
- Agregované – mezi podmnožinou atributů a jejich charakteristikami

Algoritmy generující asociace:
- Triviální
- Uspořádané generování pravidel
- Vzorkování
- K-množiny
- …

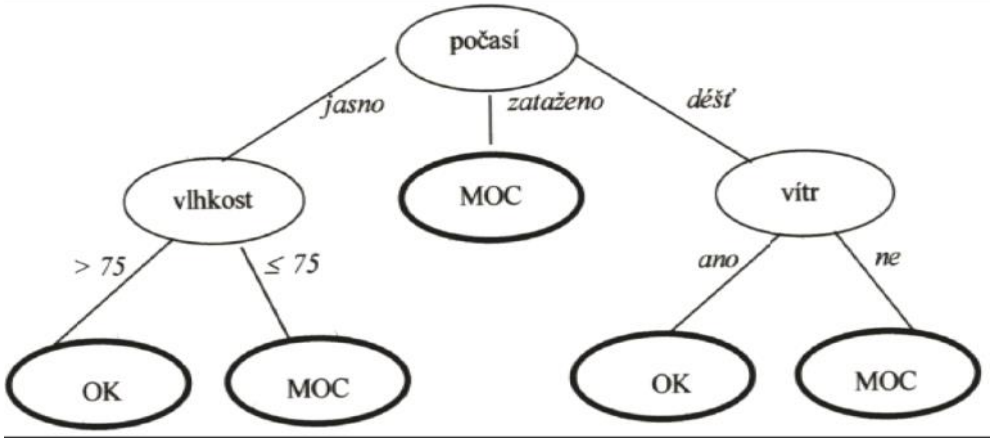
Shlukování



-analýzuje, zda se množina objektů přirozeně rozpadá na výrazné podmnožiny (shluky) objektů vzájemně si podobných a přitom nepodobných objektům podmnožin ostatních
- případně dále analyzuje, zda existuje celá hierarchie takových rozkladů
- pokud shluky existují, čím jsou charakteristické
- jak se případné další objekty zařadí do již definovaných shluků
Shluková analýza netvoří ucelenou teorii, ale je to řada metod založených na různých principech (různorodost řešených problémů, požadovaných typů výsledků, velká data, neurčitost definice shluku)

Metody dle cíle shlukování
- **hierarchické –** produkující hierarchii rozkladů, kde každý rozklad je zjemněním předcházejícího
- **nehierarchické –** produkující prostý rozklad objektů na podmnožiny
Metody dle typu výsledných shluků
- shluky kulové, body soustředěny kolem svého těžiště
- shluky obecné tvoří souvislé husté oblasti nejrůznějších tvarů
Metody dle typu rozkladu
- shluky disjunktní
- shluky překrývající se
Algoritmy
- **nehierarchické** (optimalizační k-středové, analýzy módů, fuzzy k-středové, neuronové sítě)
- **hierarchické** (aglomerativní, divizivní)
- **vzorkování**

Rozhodovací stromy



Poznámky :
metoda k-středové = k-means

From <https://d.docs.live.net/e3534876709763a3/Dokumenty/ZCU/Statnice/Statnice.docx>

Metody dolování (jiný zdroj, ne z přednášek)

1. popis dat (asi asociace?)
   - najít charakteristiky dat
   - často ve spojení s dalšími metodami, např. segmentací (hledám charakteristiky segmentů)
2. shlukování (clustering)
   - hledání a vytváření kategorií, do kterých jsou data uspořádána
   - nemusí být disjunktní, jeden objekt klidně ve více skupinách
   - tím se shlukování liší od klasifikace!!
3. klasifikace
   - rozdělování objektu do jednotlivých tříd podle cílového atributu (nespojitá veličina)
   - předpokládá se, že jednotlivé třídy, do kterých objekty rozděluji, jsou předem známé
4. regresní analýza
   - obdoba klasifikace; cílový atribut je spojitá veličina
5. analýza závislosti
   - hledám takový model, který popisuje vztahy mezi daty
   - např. analýza spotřebního koše (co se s čím často prodává)
   - hrozí, že naleznu neexistující závislost

Zavádění datového skladu

strategie velkého třesku (hub architecture)

- buduji celý sklad naráz
- vytvoření celopodnikového datového modelu
- datová tržiště modelována dimenzionálně a napojena na samotný DW

strategie postupného budování datových tržišť (bus architecture)

- z nich pak dohromady vznikne datový sklad
- výhodou je jednoduchost a srozumitelnost datového modelu, iterativnost procesu

Dva způsoby budování datového skladu

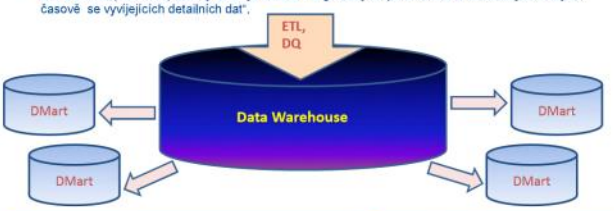
1. **Data warehouse jako množina data martů (bus architecture)**
   Ralph Kimball: "Data warehouse není nic jiného než sjednocení data martů„.

Data marty je možné sjednotit pouze za předpokladu tzv. "všeobecně přijatých" dimenzí a faktů (conformed dimension). V opačném případě není možné DM spojovat do 1 celku resp. pokud by se spojovaly, výsledkem budou špatná data!

Plus	Minus
Rychlá implementace data martů.	Redundance dat.
Nízké počáteční náklady.	Každý DM má vlastní historii, ETL, dimenze, řešení datové kvality.
	Hůř monitorovatelné procesy, vyšší HW i SW nároky na údržbu.

Dva způsoby budování datového skladu

2. **Centrální Data Warehouse (hub architecture)**
   Bill Inmon: „(Centrální) datový sklad je soubor integrovaných, předmětově orientovaných, stálých, časově se vyvíjejících detailních dat".



Plus	Minus
Centrální datový sklad plněný jednotným ETL postupem, použito 1 řešení datové kvality a vytvořeny společné dimenze.	Vyšší náklady na návrh a implementaci centrálního skladu.
Minimalizována redundantnost dat.	Delší „přípravný" čas.
Možnost centrálního monitorování datového skladu	

ODS je soubor integrovaných, předmětově orientovaných, nestálých aktuálních detailních dat vytvořená pro aktuální potřeby uživatelů.

Porovnání transakčních systémů (OLTP) a analytických systémů (OLAP)

Znak	OLTP	OLAP
Charakteristika	Provozní zpracování	Informační zpracování
Orientace	Transakční	Analytická
Uživatel	Běžný uživatel, databázový administrátor	Znalostní pracovník (manažer, analytik)
Funkce	Každodenní operace	Dlouhodobé informační požadavky, podpora rozhodování
Návrh databáze	Entitně-relační základ, aplikačně orientovaný	Hvězda/sněžná vločka, věcná orientace
Data	Současná, zaručeně aktuální	Historická
Sumarizace dat	Základní, vysoká podrobnost dat	Shrnutá, kompaktní
Náhled	Detailní	Shrnutý, multidimenzionální
Jednotky práce	Krátké, jednoduché transakce	Komplexní dotazy
Přístup	Číst, pořizovat a aktualizovat	Pouze číst
Zaměření	Vkládání dat	Získávání informací
Počet dostupných záznamů	Desítky	Miliony
Počet uživatelů	Stovky – tisíce.	Desítky – stovky.
Velikost databáze	100 MB až GB	100 GB až TB
Přednosti	Vysoký výkon, vysoká přístupnost	Vysoká flexibilita, nezávislost koncového uživatele
Míry hodnocení	Propustnost transakcí	Propustnost dotazů a doba odezvy