OLAP

Z FITwiki

Obsah

OLAP vs. OLTP

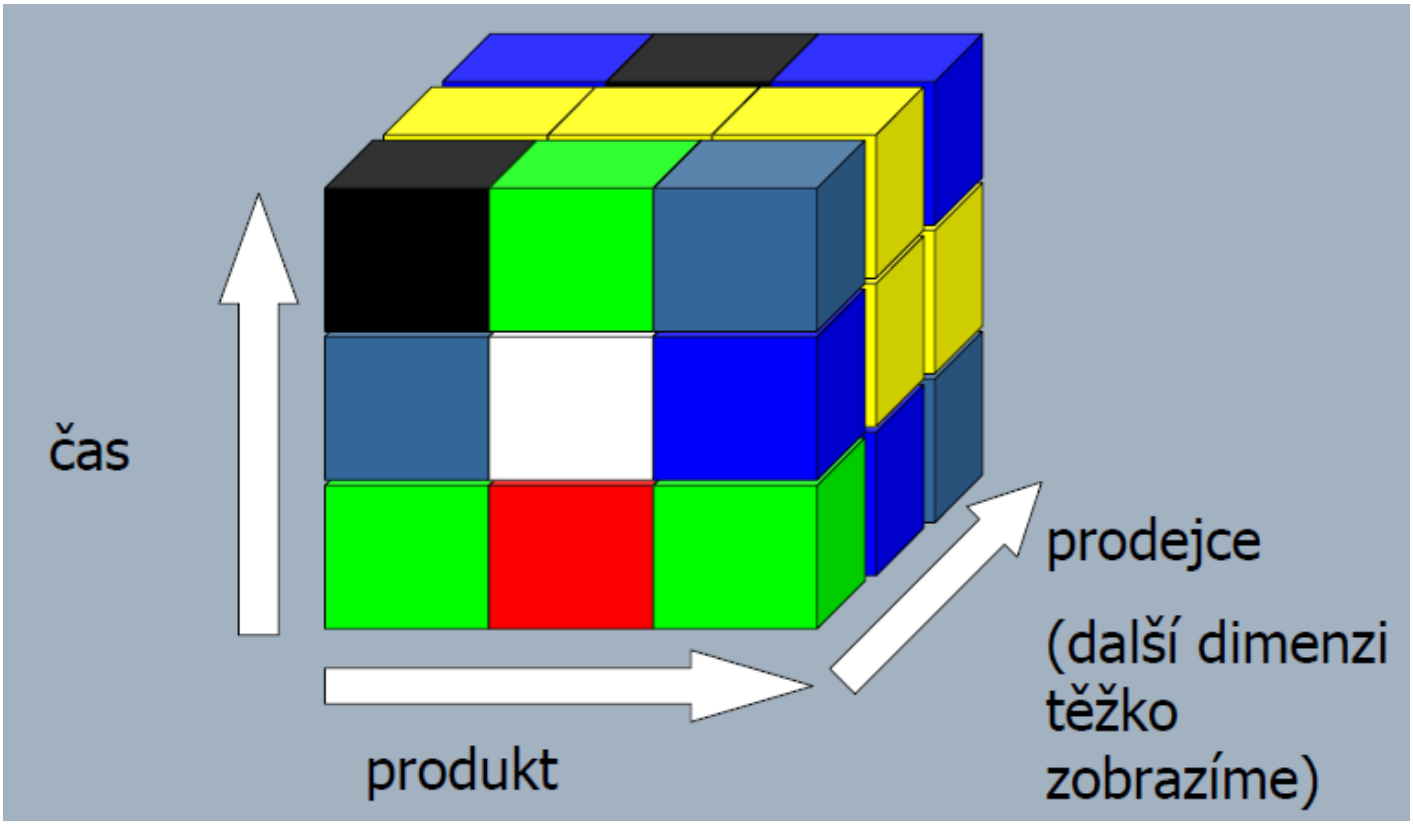
OLTP (On-Line Transaction Processing)

- slouží pro každodenní operace
- zaměřené na přesné uchovávání aktuálních dat
- krátké jednoduché ACID transakce
- velké množství změn relativně jednoduchých dat
- strukturované a opakující se dotazy
- velikost - řádově jednotky až desítky GB
- metrika - průchodnost transakcí
- probíhá nad **operačními DB**
  - důraz na konzistenci a možnost zotavení
  - výkonné pro jednotlivé operace, pomalé pro komplexní agregované dotazy

OLAP (On-Line Analytical Processing)

- slouží k podpoře rozhodování
- pracuje s konsolidovanými agregovanými daty z heterogenních zdrojů (operační DB - i několik různých, pomocné soubory - text, CSV, excel)
- malé množství změn složitých dat
- velká intenzita jedinečných a složitých dotazů
- důraz na sumarizaci a konsolidaci dat, ne na přesnost jednotlivých záznamů
- průchodnost dotazů mnohem důležitější, nežli spolehlivé zpracování transakce
- velikost - řádově stovky GB až jednotky TB
- metrika - průchodnost dotazů, odezva
- pracuje nad **datovými skladišti**
  - databáze sloužící k podpoře rozhodování, která je uložena odděleně od operační databáze
  - data typicky ukládána multidimenzionálně (dimenze často hierarchické)
  - ukládají i všechna historická data
  - vyžadují speciální organizaci dat, přístupové a implementační metody

Konceptuální schéma OLAP



Multidimenzionální pohled

- oblíbený pohled na data
- množina *číselných měr* umístěných v multidimenzionálním prostoru
- dimenze např. čas prodeje, místo prodeje, prodejce, produkt
- dimenze hierarchické (čas prodeje může být organizován jako den-měsíc-čtvrtletí-rok, produkt jako produkt-kategorie-výrobce)
- každá dimenze může být hierarchicky popsána množinou atributů (například dimenze Výrobek může sestávat ze čtyř atributů: kategorie a resort, roku výroby a průměrného zisku)

Agregace

- agregace podle jedné nebo více dimenzí (např. výpočet a hodnocení celkového prodeje pro každou zemi (nebo na každý rok))

Porovnání

- porovnání několika hodnot agregovaných podle stejných dimenzí (např. prodeje a rozpočtu)

OLAP operace

Roll-up(vyrolování)
: vzrůst úrovně agregace (tj. další agregace nad daty)

Drill-down (zavrtání)
: snížení úrovně agregace a zvýšení detailu podle jedné nebo více dimenzí hierarchie

Slice & dice (seříznutí)
: výběr projekce (zobrazujeme jen část dat)

Pivoting (přetočení)
: přeorientování vícedimenzionálního pohledu na data

- počet pohledů je k! (k je počet dimenzí)

Reprezentace multidimenzionálních dat ve 2D

- například pro tisk, zobrazení na monitoru, ...

Dynamická tabulka (DT)

- od společnosti Vema
- rozdělena na sloupce dimenzí a sloupce faktů
- hierarchie dimenzí není – dimenze jsou ploché
- řádek tvoří uspořádanou n-tici dimenzních a faktových položek
- sloupce klíčů lze zaměňovat a třídit – pivoting
- zakrýváním klíčových sloupců lze zvyšovat a snižovat agregaci – roll-up, drill-down
- všechny sloupce lze zneviditelnit, případně nastavit filtry – slice & dice
- prostorově náročnější, avšak jsou vidět všechny agregační i dimenzní položky

Největší dlužníci

Výběr	1-12	Firma (>40000000)	Splatnost (<31.10)	Nákl. stř. nejvyšší (1)	Nákl. stř. vyšší	Nákl. stř. základní	Celkem	Zbývá (>0.00)
☐	1	45860951-Kristian s r.o.	24.06.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	61 244.00	61 244.00
☐	2	42035709-MEDIK s.r.o.	20.05.01	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	50 473.90	50 473.90
☐	3	46218487-BALLSTREET	04.03.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	59 187.78	9 467.78
☐	4	46471219-Johansson s.r.o.	18.04.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	7 870.45	7 870.45
☐	5	47952425-A B S tech.činnosti	22.04.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	5 390.00	5 390.00
☐	6	42587912-BSIO a.s.	30.03.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	3 856.70	3 856.70
☐	7	47526869-AKO s.r.o.	30.03.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	18 100.00	2 850.00
☐	8	99990001-Kutálek Petr, Ing.	15.01.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	1 428.00	1 428.00
☐	9	47935469-Computer office	06.02.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	1 439.60	707.60
☐	10	45872463-Renovace a.s.	07.04.04	1-Brno	3-výroba	31-výroba nábytku	121.40	121.40
☐	11	47952425-A B S tech.činnosti	20.03.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	16.50	16.50
☐	12	49385719-Ekolab s.r.o.	30.03.04	1-Brno	9-rozvahové pracoviště	91-rozvahové pracoviště	70 759.90	0.90
	12						**279 888.23**	**143 427.23**

Kontingenční tabulka (KT)

- od společnosti MS
- řádky i sloupce jsou dimenze (hierarchické)
- fakta jsou průsečíky dimenzí
- fakt je proto v zásadě pouze jediný, více faktů se v průsečíku hůře zobrazuje, řeší se to záložkami nebo více sloupci, což je nepřehledné
- dimenze lze přesunovat a zaměňovat – pivoting
- zakrýváním hierarchie dimenzí lze zvyšovat a snižovat agregaci – roll-up,drill-down
- všechny sloupce lze zneviditelnit nebo nevyužívat – slice & dice
- prostorově méně náročná, avšak nejsou vidět všechny faktové položky a hierarchie dimenzí může být nepřehledná

Součet z celkem		nákl. nejvyšší	nákl. vyšší	nákl. základní							
		1					Celkem z 1	2		Celkem z 2	Celkový součet
		11		Celkem z 11	12	Celkem z 12		21	Celkem z 21		
splatnost	firma	111	112		121			211			
1.1.2002	A	44567		44567			44567				44567
Celkem z 1.1.2002		44567		44567			44567				44567
1.1.2003	B		11000	11000			11000				11000
	C				12000	12000	12000				12000
Celkem z 1.1.2003			11000	11000	12000	12000	23000				23000
1.2.2003	A	84732		84732			84732				84732
Celkem z 1.2.2003		84732		84732			84732				84732
12.4.2003	A	76355		76355			76355				76355
Celkem z 12.4.2003		76355		76355			76355				76355
1.1.2004	D							13000	13000	13000	13000
Celkem z 1.1.2004								13000	13000	13000	13000
Celkový součet		205654	11000	216654	12000	12000	228654	13000	13000	13000	241654

Získávání dat

- **data** - surová data (8)

- **informace** - interpretovaná data (8 - den v měsíci)
- **znalost** - informace zařazená do souvislostí (8 je den v měsíci kdy prodal konkrétní produkt)

- data se získávají z mnoha různých zdrojů (různá kvalita, různě reprezentace, různé formáty)
- spojují se data z mnoha operačních databází a externích zdrojů
- data musí být sjednocena (získání, čistění, integrování dat)
- data se obnovují periodicky (pravidelný přesun dat k archivaci v datovém skladišti)

Získávání (Extraction)

- datové extrakce z cizích zdrojů obvykle implementovány pomocí tzv. gateways a standardních rozhraní

Čištění (Transformation)

- ošetřuje nekonzistentnost dat
  - různé typy polí
  - různé názvy polí
  - různá sémantika polí (stejná pole různé sémantiky)
  - chybějící položky (povinné/volitelné položky)
  - různá integritní omezení

Data migration

- prostředky data migration dovolují aplikovat jednoduchá pravidla transformace (např. přepiš řetězec gender na sex)

Data scrubing (drhnutí dat)

- užívají specifické znalosti o dané doméně
- aplikují syntaktický pohled na problém
- často provádějí syntaktickou analýzu dat a vyhledávací techniky.

Data auditing

- umožňují odhalovat pravidla a vztahy (nebo signalizovat porušení stávajících pravidel) při prohlížení dat
- aplikují sémantický pohled na problém
- jsou vlastně variantou prostředků dolování dat
- například může takový prostředek odhalit podezřelý vzorek (založený na statistické analýze)

Natažení (Loading)

- typicky používány dávkové utility
- Problém: sekvenční natažení vyžaduje dlouhý čas (dny, týdny), ale datové skladiště může být off-line pouze po krátkou dobu (přes noc)

Preprocessing při natažení

- kontrola integritních omezení,
- třídění,
- sumarizace, agregace
- budování indexů a nebo např. materializovaných pohledů

Plné natažení dat

- nová data se natahují kompletně do nové DB
- během budování nové DB je stále k dispozici stará DB
- po natažení je stará DB nahrazena novou

Checkpointy
: při havárii během natažení je možné proces restartovat od posledního checkpoitu

Obnova (aktualizace) dat

- šíření změn zdrojových dat na data ve skladišti (otázkou je jak často)

Replikace
pokud zdrojová DB podporuje replikaci je možné toho využít pro šíření změn do datového skladiště

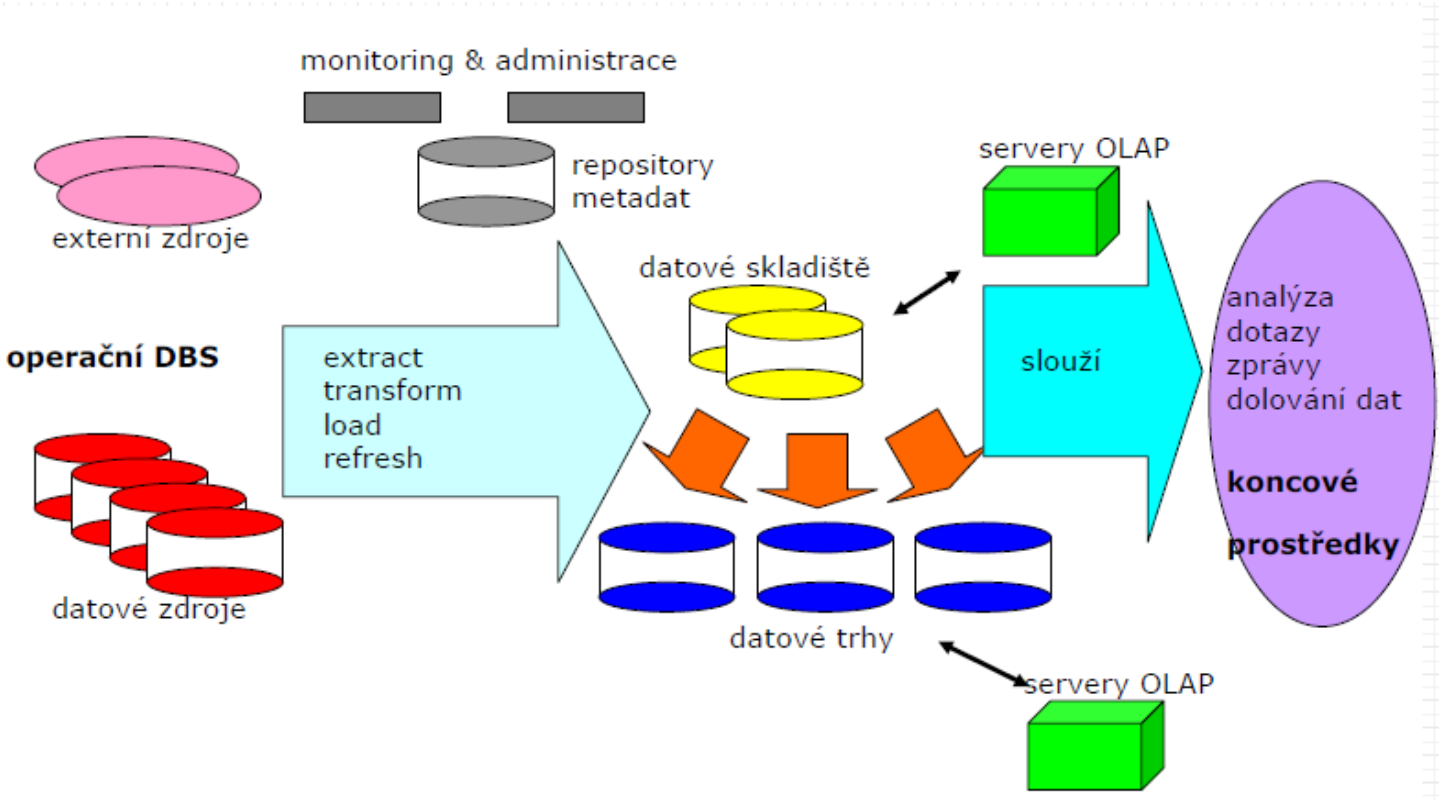
- **Data shiping** - replikace při níž je tabulka datového skladiště obsluhována jako vzdálený snímek zdrojové tabulky. Existuje trigger reagující na změnu zdrojové tabulky, který provede přesun dat do tabulky v datovém skladišti.
- **Transaction shipping** - Na straně zdroje je sledován logovací soubor transakcí, aby se zjistily změny v replikované tabulce a takové záznamy jsou přeneseny do replikačního serveru, který zabalí odpovídající transakci tak, aby byla schopna změnit replikovanou stranu. Nepotřebuje triggery (nezatěžuje zdrojovou DB).

Datová skladiště

Datovým skladem nazýváme technologii

- **natažení**
- **uložení** a
- **poskytování**

dat pro podporu rozhodování prováděnou analýzou informací



Architektura

- **Datové trhy (Data mart)** - podmnožiny datového skladiště poskytující data jednotlivým oddělením

- extrahování dat z mnoha operačních databází a externích zdrojů
- čištění, transformování a integrování těchto dat
- periodická obnova datového skladiště tak, aby odrážel změny ve zdrojích a mazání dat z datového skladiště, nejčastěji do pomalejší archivní paměti

- Prezentace multidimenzionálního pohledu na data různým koncovým prostředkům:
  - dotazovací prostředky
  - generátory zpráv
  - analytické prostředky

    - prostředky dolování dat (data mining)
- Repository pro ukládání a správu metadat (katalogu) a pro monitorování a administraci systému datového skladiště

Servery datových skladišť

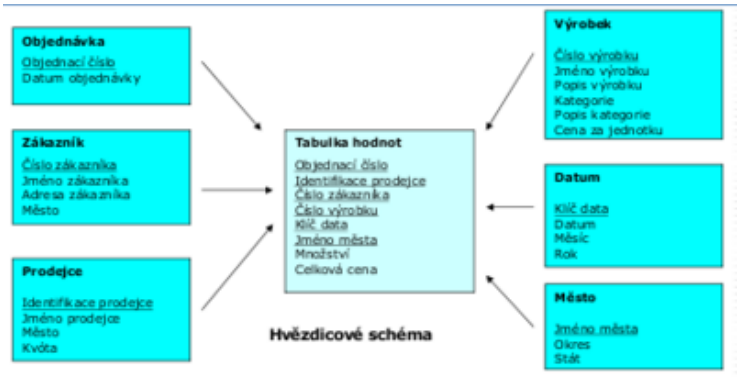
- kvůli zrychlení užívají redundantní struktury, jako jsou:
    - indexy
    - materializované (datově uložené) pohledy

Relational OLAP (ROLAP)

- datová skladiště nad klasickými nebo rozšířenými relačními DB
- data uložena v relačních DB
- speciální SQL podpora pro multidimenzionální dotazování

Hvězdicové schéma

- užívá většina datových skladišť k reprezentaci multidimensionálního modelu
- jedna tabulka hodnot
- jedna tabulka pro každou dimenzi (sloupce odpovídají atributům dimenze)
- Každá n-tice v tabulce hodnot sestává z ukazatele do každé dimenze, který poskytuje jeho multidimensionální souřadnice.



Sněhová vločka (snowflake)

- hierarchie dimenzí je explicitně reprezentována normalizováním tabulek dimenzí

Sumarizační tabulky

- obsahují předagregovaná data
- v nejjednodušším případě odpovídají předagregovaná data agregování tabulky hodnot podle jedné nebo více vybraných dimenzí

Multidimensional OLAP (MOLAP)

- ukládají data přímo multidimenzionálně
- používají speciální datové struktury (řídké matice)
- velké paměťové nároky

Kategorie: Státnice 2011 | Pokročilé informační systémy

- Stránka byla naposledy editována 19. 6. 2014 v 15:12.