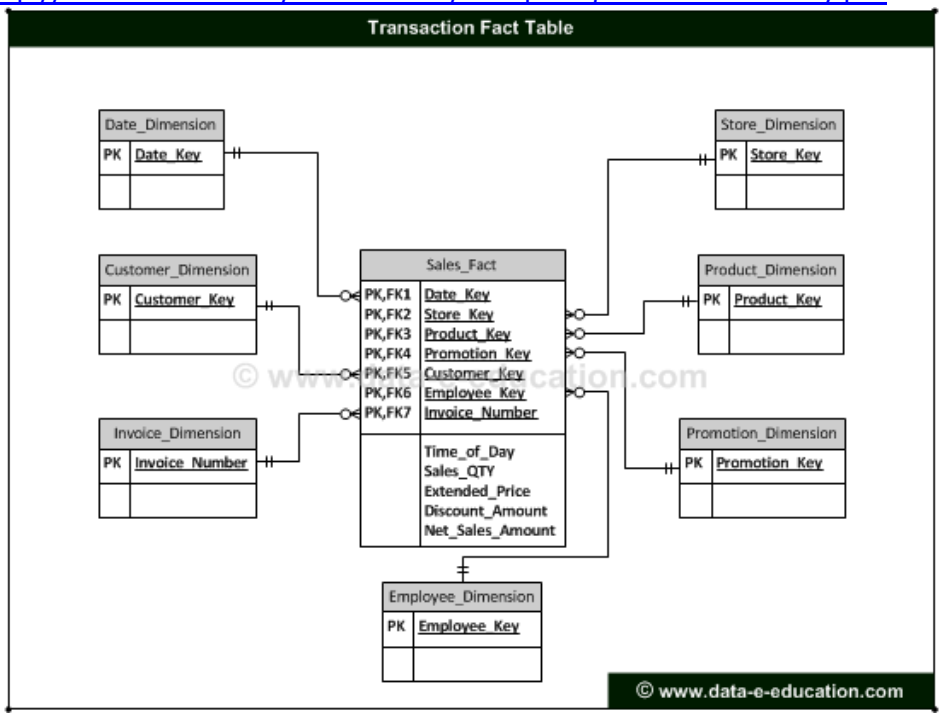
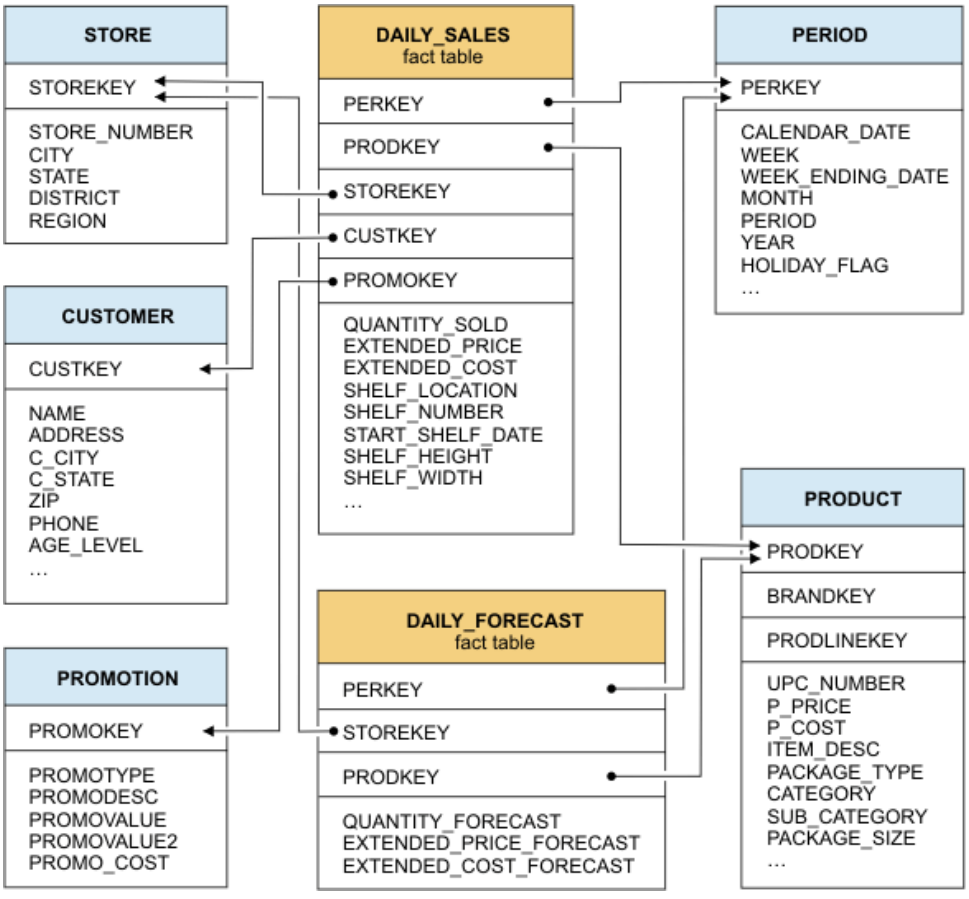
OLAP systémy, jejich význam a oblasti využití, základními principy, dimenze, agregace, extrakce a transformace dat, srovnání transakčních a analytických systémů (OLAP a OLTP technologií).

Thursday, May 30, 2013 8:24 AM

http://www.common.cz/attachments/118_petr_jasa_datove_sklady.pdf



From <http://www.bing.com/images/search?q=table+of+facts+table+of+dimensions&view=detail&id=CDCE4349212949F443E8EC07EA739665F19DCA98&FORM=IDFRIR>



From <http://www.bing.com/images/search?q=table+of+facts+table+of+dimensions&view=detail&id=9E9783F312C8C26698296EA1AFD9230AB0B4D1DB&FORM=IDFRIR>

Dimenze

A Dimension is a structural attribute of a cube that is a list of related names-known as Members-all of which belong to a similar category in the user's perception of a data. For example, all months, quarters and years make up a Time dimension; likewise all cities, regions and countries make up a Geography dimension. A Dimension acts an index for identifying values within a multi-dimensional array and offers a very concise, intuitive way of organizing and selecting data for retrieval, exploration and analysis. Dimensions are the business parameters normally seen in the rows and columns of a report.
Dimensions are lists of related terms used to organize your data. Thus, a natural Dimension name for the Members January, February and March might be Months. Dimensions, in turn, are used to construct Cubes, the multidimensional structures in which you store and model data.

From <http://olap.com/w/index.php/Dimension>

Význam a oblasti využití

BI - Business intelligence - pro podporu rozhodování, analýzy historických dat, procesní ukazatele

Extrakce, Agregace = asi ETL

- Extract Transform Load, způsob plnění datového tržiště daty. Data jsou vybrána z relačních databází kde se nachází (Extract), vyčištěna, upravena (Transform) a pak je s nimi naplněno datové tržiště (Load).

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

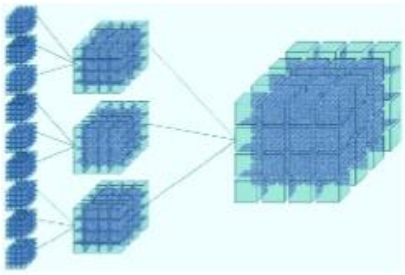
Datawarehouse

- 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

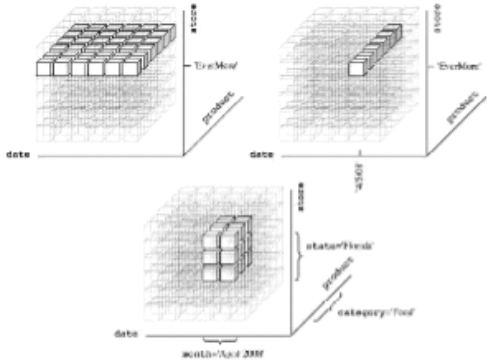
- data jsou uložena na různých místech ve formě relačních tabulek
  - uživatelé mohou tabulky jen číst
  - zapisovat může aktualizační program pravidelně udržující tabulky
- dotazy jsou většinou komplexní
  - podporují tzv. on-line analytické zpracování (OLAP)
    - výrazně se liší od on-line transakčního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

Roll-up



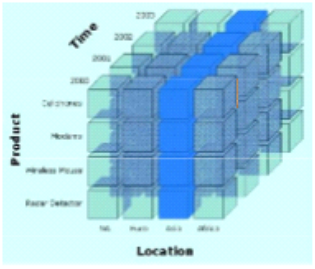
- **drill-down** (zanoření se do nižší úrovně dat pro získání více detailů)
- Drill-up (opak drill down, přechod o level výše pro skrytí detailů a získání lepšího celkového přehledu o datech)

Drill down

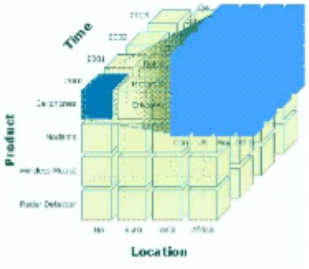


- **slice-and-dice** (selekce a projekce)
- **Slice** - v jednom rozměru krychle zvolíme pouze jednu hodnotu, což vytvoří novou krychli, která je "řezem" té původní
- **Dice** - vybereme konkrétní hodnoty v každé dimenzi krychle, vznikne menší krychle

Slice

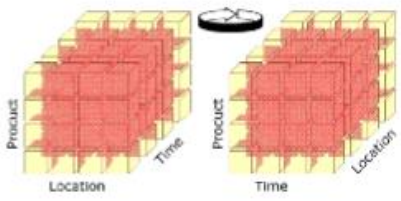


Dice



- **Pivot** (přeorientování vícerozměrného pohledu na data, prostě prohození os)

Pivot

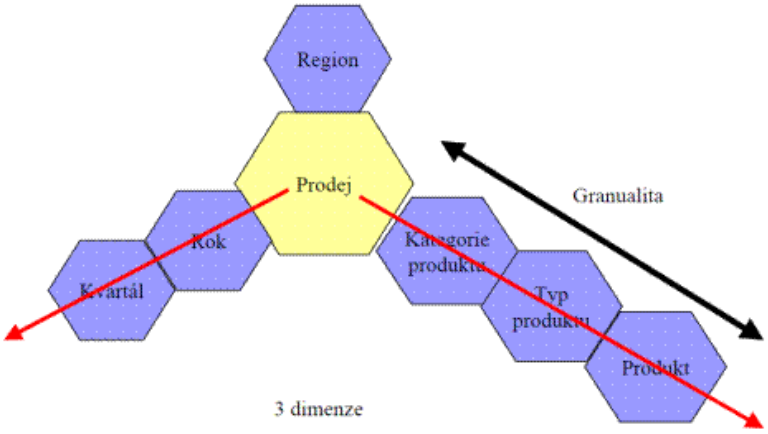


- na základě dotazu se pospojují potřebná data do vícerozměrné tabulky (nebo více tabulek), do kterých lze klást SQL dotazy
- pro častější dotazy si uchovávají předem připravené vícerozměrné tabulky
- zátěž je většinou způsobena složitými dotazy, jež přistupují k miliónům záznamů a provádějí

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

Základní myšlenka multidimenzionálního modelování



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

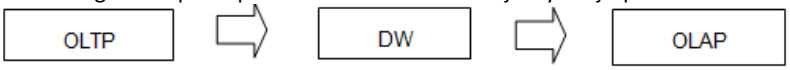
Použití DW

- 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

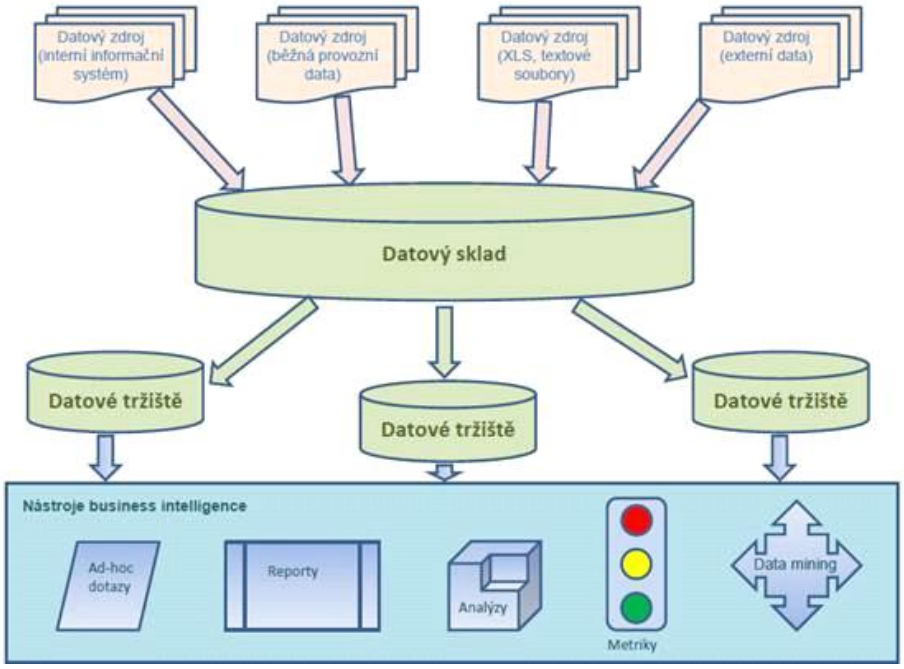
- vytvoření je složitý proces trvající mnohdy i několik let
- mnoho organizací proto používá Data Mart umožňující rychlejší práci
- 

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 kladeným analytickým způsobem práce - převažující způsob



Programy pro vytváření a plnění databáze

- 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 kriterií
    - 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

Pozn: dle mého názoru do MOLAP patří jen multidimenzionální krychle (proto MOLAP). Vločka a hvězda jsou ROLAP.

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)

Srovnání OLAP a OLTP

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

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