Datový sklad

Ing. Roman Danel, Ph.D. **2010**

Co je to datový sklad a kdy se používá?

Pojmem datový sklad (anglicky Data Warehouse) označujeme zvláštní **typ databáze, určený primárně pro analýzy dat v rámci Business Intelligence** (oblast analýzy dat sloužící jako podklady pro manažerské rozhodování).

„Ke správnému rozhodování podnikového managementu a minimalizaci rizika špatných rozhodnutí je nutné rychle vyhodnocovat velké množství nesourodých informací z mnoha zdrojů. K tomuto úkolu byl stvořen datový sklad.“

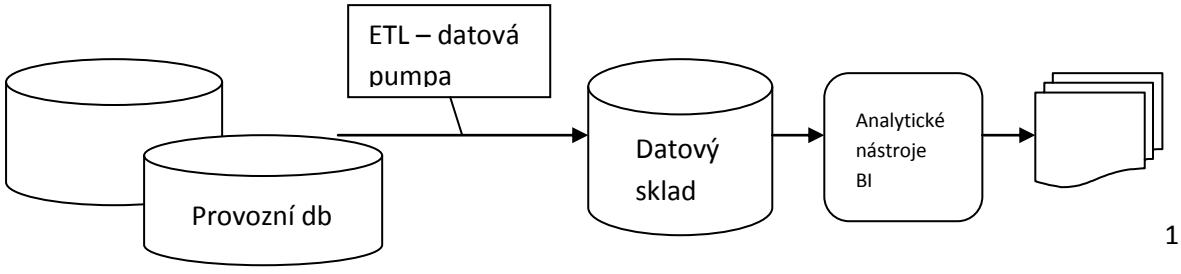
Definice

> ***Datový sklad je podnikový strukturovaný depozitář předmětově (subjektově) orientovaných, vzájemně provázaných, nepodléhající změnám, časově proměnných, historických dat používaný na získávání informací a pro podporu rozhodování. V datovém skladu jsou uložena detailní (atomická) i sumární data***
>
> **Bill Inmon, prezident Data Systems, „otec“ datových skladů**

Odkud získáme data?

Datové sklady nevznikají na zelené louce. Setkal jsem se u studentů s nesprávnou představou, že datový sklad je zdroj informací pro provozní databázi. Data, ukládaná a zpracovávaná v transakčních systémech neumožňují provádění kvalitních analýz, neboť provádění analýz na základě relacemi propojených tabulek je velmi obtížné. Z toho důvodu jsou data z provozních databází transformována do datového skladu, kde jsou uložena ve tvaru vhodnějšímu k analytickému zpracování.

Informace se do datových skladů ukládají **pomocí datových pump** z provozních databází (tím jsou myšleny relační databáze z podnikových informačních systémů – ERP, CRM atd.). Nástroje datové pumpy se také označují jako **ETL nástroje** (zkratka slov „Extraction“, „Transformation“ a „Loading“).



Ve fázi extrakce vybíráme data. Během transformace dochází k ověření a čištění dat (čištěním dat rozumíme doplnění chybějících hodnot, odstranění překlepů, převedení na shodné formáty, párování na jednotné číselníky/dimenze), datové konsolidaci a výpočtu agregací dle hlavních entit. Fáze „loading“ znamená přemístění dat do datového skladu.

Datové pumpy pracují dávkově a proces transformace dat z provozních databází do datového skladu může být časově náročný.

Příkladem nástroje ETL (datové pumpy) může být například DTS (Data Transformation Services) firmy Microsoft (je součástí instalace MS SQL Serveru 2000/2005) nebo Oracle Data Mart Builder.

Jaký je rozdíl mezi provozní relační databází a datovým skladem?

U **běžné relační databáze** je obvyklá snaha o co nejmenší redundanci uložení dat, které je dosahováno jejich normalizací a vnitřním provázáním jednotlivých logických funkčních celků.

V **datovém skladu** je naproti tomu řešení vždy vedeno snahou o jasnou vnitřní separaci jednotlivých funkčních celků - výsledkem je struktura, která je čitelnější pro uživatele (manažera, business analytika) za cenu zvýšených nároků na paměťový prostor. Při popisu struktury datového skladu mluvíme o multidimenzionální (vícerozměrné) struktuře uložených dat.

Do datového skladu se data dostávají ve větších dávkách a dále již obvykle nejsou modifikovány.

Běžná provozní aplikace nad relační databází řeší určitý specifický okruh úloh nad „svými“ specifickými daty. V datovém skladu je třeba naproti tomu shromáždit informace z mnoha různých zdrojů a seskupit je nikoliv podle původu, ale podle logického významu (úzce souvisí s **orientací na subjekt** - všechna data týkající se určité funkční oblasti potřebuji mít „na jedné hromadě“ bez ohledu na to, odkud pocházejí).

Zatímco relační databáze je navrhována pomocí ERD (entitně relační modelování), datový sklad je navrhován pomocí dimenzionálního modelování:

Modelování relační databáze:

- Odstranění redundantnosti (opakování se dat)
- Normalizace (důsledkem je nižší srozumitelnost pro člověka)
- Optimalizace na vkládání a úpravu dat

Modelování datového skladu:

- Optimalizován na čtení dat, vyhledávání a složité analýzy
- Důraz na srozumitelnost pro uživatele (dosaženo standardní strukturou – faktová tabulka, dimenze)
- Základní přístup – denormalizace, redundance
- Rozšiřitelnost – přidání fakt, dimenzí – bez dopadu na aplikace

Co znamená, že datový sklad je integrovaný?

„Integrovaný“ znamená, že 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.

Co znamená, že datový sklad je časově proměnný a nepodléhající změnám?

Data v produkčních systémech se mění (podléhají změnám). 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").

Pod pojmem "časově proměnný" rozumíme skutečnost, že data v datovém skladu jsou ukládána po časových snímcích a tak vytváří časově proměnnou řadu, historii.

Co je to datové tržiště (Data Mart)?

Analytické nástroje BI mohou čerpat data pro analýzy buď z datového skladu (obvykle celopodnikový) nebo z tzv. **datových tržišť** (anglicky Data Mart). Datové tržiště je tematicky orientovaný datový sklad určený ke zprostředkování informací pro určité oddělení podniku (např. marketing, finanční oddělení atd.).

Někdy může být datový sklad vytvořen sjednocením jednotlivých data martů. Data marty jsou ale budovány na základě požadavků jednotlivých útvarů společnosti. Z toho vyplývá:

· potřeba vlastních dat
· používání vlastních definice pojmů
· vlastní historie dat
· vlastní periodicita aktualizace dat.

Jedná se o tzv. **dvouvrstvou architekturu (koncept Ralfa Kimballa)** a volíme ji především tehdy, pokud je potřeba upřednostnit konkrétní oddělení či pobočku a dodat první výstupy datového skladu v relativně krátké době (v horizontu několika měsíců). Datový sklad se pak buduje postupně po jednotlivých datových tržištích a nejen výsledky, ale i finanční prostředky na vývoj jsou rozloženy v čase. Tímto způsobem vybudované prostředí pro podporu rozhodování však neposkytuje celopodnikový pohled na informace. Podíváme-li se na schematické znázornění architektury, odpovídající zmiňovanému přístupu, pak tato nám může připomínat "pavoučí sít" ("spider net" - obvykle používaný pojem pro architekturu data warehouse založenou na jednotlivých data martech).

Naproti tomu koncepce, s níž přišel Bill Inmon a Claudia Imhoff (obsahující dočasné úložiště, centrální úložiště a datová tržiště) se označuje jako třívrstvá **architektura datového skladu**. Jedná se o nejčistší

řešení, které ovšem vyžaduje vyšší počáteční náklady na analýzu a relativně dlouhou dobu na úplnou realizaci.

Co znamená v oblasti datových skladů termín ODS?

Často se také vyskytuje potřeba pracovat sice s konsolidovanými, ale hlavně aktuálními daty s minimální dobou odezvy – např. na call centru, kdy potřebujeme u každého zákazníka znát aktuální profil, jeho aktivované či objednané produkty, zařazení do segmentu, poslaných marketingových nabídek. To vyžaduje další komponentu datového skladu – **operativní datové úložiště** – ODS (Operating Data Store).

Pro účely maximální „operativnosti" je operativní úložiště často napojeno na datové zdroje prostřednictvím EAI (Enterprise Application Integration) platforem. Ty umožňují vzájemnou komunikaci mezi libovolnými dvěma aplikacemi v reálném čase a to bez nutnosti tyto dvě aplikace přímo vzájemně propojovat. Na rozdíl od ETL platforem, které zpracovávají události dávkově v předem stanoveném čase, EAI platformy reagují na jednotlivé události okamžitě.

Jaká je struktura datového skladu?

Data v datovém skladu jsou z logického (uživatelského) pohledu členěna do **schéma** (topologické uspořádání). Každé schéma odpovídá jedné analyzované funkční oblasti.

Schéma obsahuje dva typy tabulek – faktové a dimenzionální.

Jádro každého schématu tvoří jedna nebo několik **faktových tabulek**. V nich jsou uložena **vlastní analyzovaná data** - veličiny, které sledujeme; hodnoty, které jsou použity k analytickým výpočtům - agregacím, třídění apod. Většina paměťového místa v datovém skladu zabírají faktové tabulky, které obsahují detailní údaje ze všech zdrojů - tedy řádově více údajů než ostatní tabulky.

S faktovou tabulkou je spojena **granularita**. Granularita určuje úroveň podrobnosti v tabulce faktů. Čím nižší je úroveň granularity, tím detailnější jsou data určená k provádění matematických operací.

Faktové tabulky jsou pomocí cizích klíčů spojeny s **dimenzemi**. Dimenze jsou tabulky, které obsahují seznamy hodnot sloužících ke kategorizaci a třídění dat ve faktových tabulkách (atributy, prostřednictvím kterých se „díváme" na data). Je to vlastně **číselník**, podle kterého chceme data analyzovat.

Vlastnosti dimenzí:

a) Dimenze určují úhel pohledu – čas, produkt, zákazník…
b) Dimenze udržují hierarchie (vztah 1:N)
c) Vztah mezi faktovou tabulkou a dimenzemi je 1:N

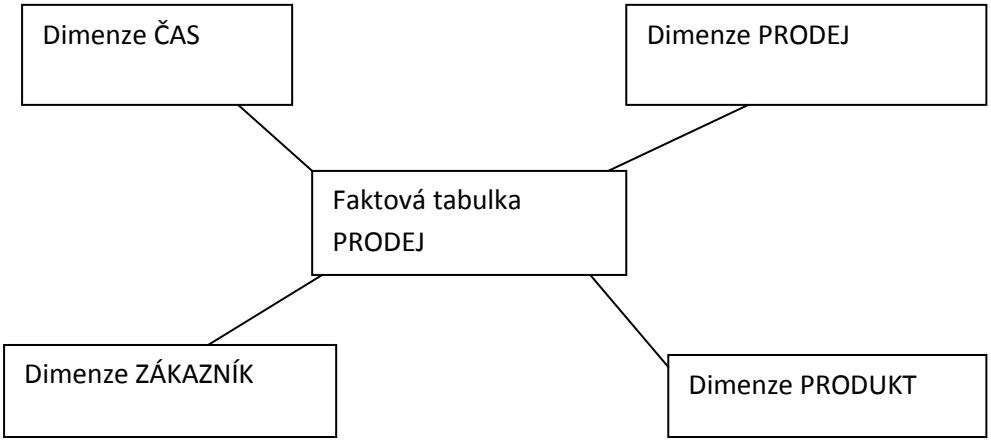
Co znamená, že datový sklad je multidimenzionální?

Datový sklad je založen na multidimenzionálním datovém modelu. S takto uloženými daty lze následně pracovat jako s tzv. datovou kostkou (cube). Datová kostka může mít větší množství rozměrů (dimenzí).

Co je to schéma „star“ a „sněhová vločka“?

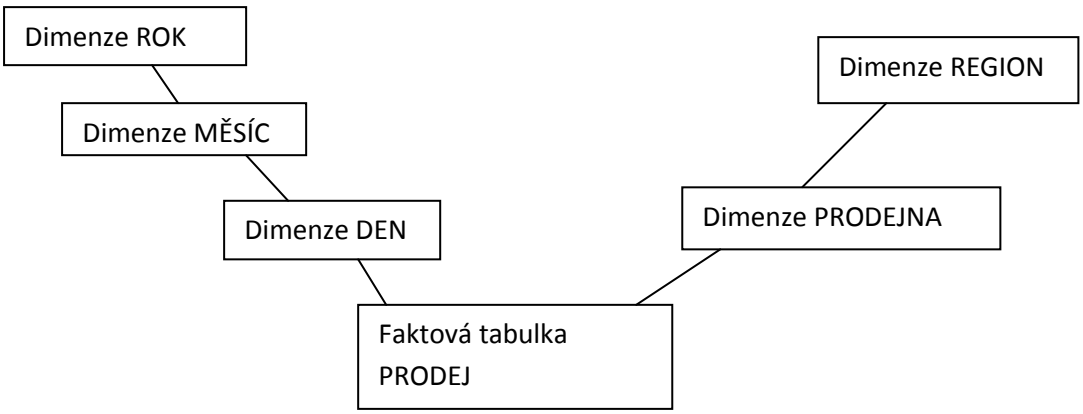
Schéma nám určuje strukturu datového skladu. Podle napojení dimenzí na faktovou tabulku rozlišujeme schéma typu **hvězda (star)** a schéma typu **sněhová vločka (snowflake)**. U schématu typu sněhová vločka jsou tabulky dimenzí normalizovány a mají relační strukturu. Faktová tabulka obsahuje cizí klíče do tabulek dimenzí.

Příklad – schéma typu „hvězda“:



Jedná se o nejjednodušší způsob jak převést data z relační databáze do multidimenzionální struktury. Grafické vyjádření připomíná tvarem hvězdu, odtud vzniklo pojmenování tohoto schématu. Každá dimenze je reprezentována právě jednou dimenzionální tabulkou.

Příklad – schéma typu „sněhová vločka“:



U schématu sněhové vločky jsou dimenzionální tabulky normalizovány. Tím se šetří diskový prostor, protože se snižuje redundance dat.

Vytvoření hvězdicového modelu (kde jsou data v nenormalizované podobě) může být zdlouhavé, ale poskytuje vysoký dotazovací výkon. Ve schématu sněhové vločky jsou dimenzionální tabulky relačně propojené – zavedení údajů je rychlejší, ale dotazovací výkon je nižší, protože při dotazech je nutné opět propojovat více tabulek.

Další možnou formou schématu je **constellation** (souhvězdí). Jedná se o soubor schémat typu „hvězda“, tj. více tabulek faktů se sdílenými dimenzemi.

Operace s datovým skladem v OLAP analýze

Drill-down – umožňuje uživateli ve zvolené(-ých) instanci(-ích) jisté agregační úrovně nastavit nižší(jemnější) agregační úroveň. Jedná se o navigaci v hierarchii dimenzí směrem k většímu detailu.

Roll-up – jde o opak předešlé operace. Ve zvolených instancích jisté agregační úrovně nastavuje vyšší (hrubší) agregační úroveň (menší detail v hierarchii dimenzí).

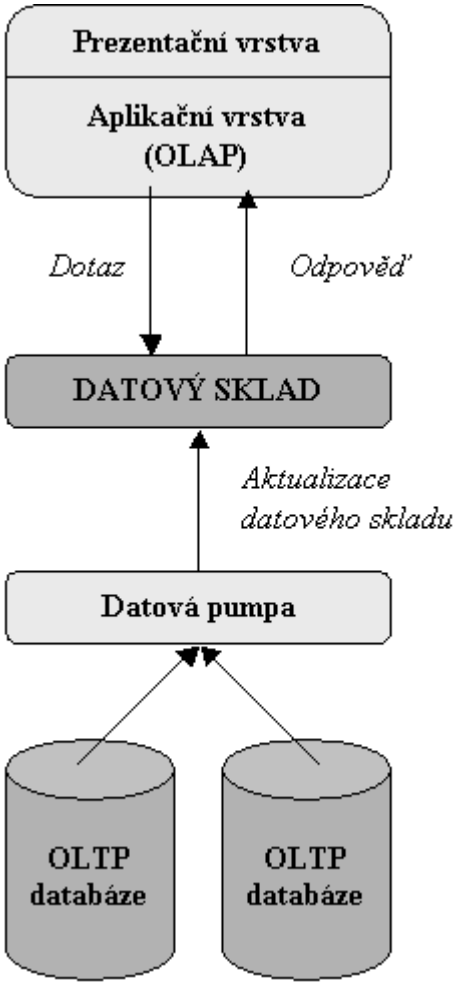
Pivoting – umožňuje „otáčet“ datovou krychlí, tj. měnit úhel pohledu na data na úrovni prezentace obsahu datového skladu.

Slicing – dovoluje provádět řezy datovou kostkou, tj. nalézt pohled, v němž je jedna dimenze fixována v jisté(-ých) instanci(-ích) jisté agregační úrovně. Jinými slovy tato dimenze aplikuje filtr na instance příslušné agregační úrovně dané dimenze.

Dicing – je obdobou „slicingu“, jenž umožňuje nastavit takový filtr pro více dimenzí.

Využití datových skladů

Datový sklad je hlavní zdroj dat pro systémy Business Intelligence. Máme-li vytvořen datový sklad, informace v datovém skladu můžeme dále analyzovat. Pro následnou analýzu se používají například technologie OLAP (základní analýza) nebo techniky „data miningu“ (pokročilejší analýza).



Zdroj: http://datamining.xf.cz