

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

**VYSOKÁ ŠKOLA BÁŇSKÁ – TECHNICKÁ UNIVERZITA OSTRAVA
FAKULTA STROJNÍ**



INFORMAČNÍ SYSTÉMY

Business intelligence v informačních systémech

Ing. Roman Danel, Ph.D.

Ostrava 2013

© Ing. Roman Danel, Ph.D.

© Vysoká škola báňská – Technická univerzita Ostrava

ISBN 978-80-248-3051-3



Tento studijní materiál vznikl za finanční podpory Evropského sociálního fondu (ESF) a rozpočtu České republiky v rámci řešení projektu: CZ.1.07/2.2.00/15.0463, MODERNIZACE VÝUKOVÝCH MATERIÁLŮ A DIDAKTICKÝCH METOD

OBSAH

1	BUSINESS INTELIGENCE V INFORMAČNÍCH SYSTÉMECH	3
1.1	Datový sklad.....	4
1.2	OLAP.....	7
1.3	Dolování dat.....	10
2	KONTROLNÍ OTÁZKY.....	12
3	POUŽITÁ LITERATURA.....	13



1 BUSINESS INTELLIGENCE V INFORMAČNÍCH SYSTÉMECH



OBSAH KAPITOLY:

Datový sklad

OLAP

Dolování dat



MOTIVACE:

V této kapitole se seznámíte se software nazývaným business intelligence (BI). Jedná se o skupinu softwarových nástrojů určených především pro management, pro analýzu dat a podporu pro řízení a rozhodování. Postupně se seznámíme s datovými sklady a jejich vnitřní strukturou, s nástroji OLAP pro generování OLAP kostek a řekneme si, k čemu slouží dolování dat.



CÍL:

Po prostudování kapitol budete znát, co je to datový sklad a čím se liší od relační databáze, co je to olap analýza a dolování dat. Čím se liší dolování dat od OLAP a jaké dva základní modely dolování dat známe – deskriptivní a prediktivní. Budete znát, co znamená pojem business intelligence, jakou hraje roli v informačních systémech a kde a proč je využívána.



1.1 DATOVÝ SKLAD

1.1.1 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** (software sloužící jako podklady pro manažerské rozhodování).

Definice:

Datový sklad je podnikový strukturovaný depozitář předmětově (subjektově) orientovaných, vzájemně provázaných, nepodléhajících změnám, časově proměnných, historických dat používaný na získávání informací a pro podporu rozhodování. V datovém skladu jsou uložena detailní (atomická) i sumární data.

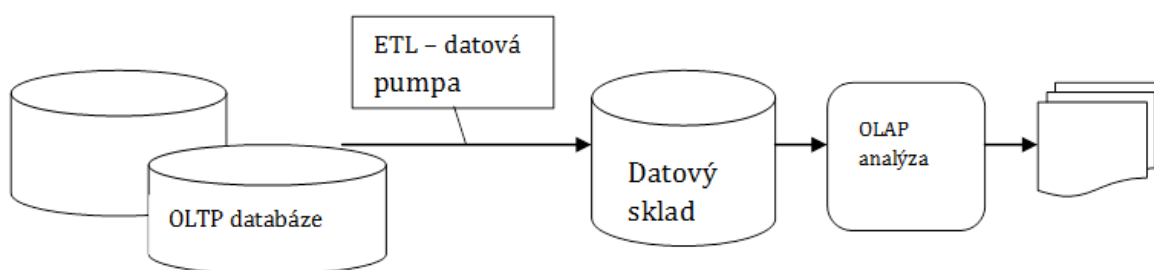
Bill Inmon, prezident Data Systems

1.1.2 Zdroje dat pro datový sklad

Data, ukládaná a zpracovávaná v transakčních systémech (OLTP) neumožňují provádění kvalitních analýz, neboť provádění analýz na základě dat uložených v relačních tabulkách je velmi obtížné. Z toho důvodu jsou data z provozních databází transformována do datového skladu, kde jsou uložena ve tvaru vhodnějším k analytickému zpracování.

Informace se do datových skladů ukládají z provozních databází **pomocí datových pump** neboli **ETL** (zkratka slov „Extraction“, „Transformation“ a „Loading“).

Ve fázi extrakce vybíráme data z OLTP databáze. 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. Ve fázi „loading“ ukládáme data do datového skladu.



Obr. 6.1 Schéma struktury SW Business Intelligence

Datové pumpy pracují dávkově a často automatizovaně. 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 MS SQL Serveru od verze 2000) nebo Oracle Data Mart Builder.

1.1.3 Rozdíl mezi provozní relační databází a datovým skladem

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



V **datovém skladu** je naproti tomu snaha o vytvoření struktury, která je čitelnější pro uživatele (manažera, business analytika) za cenu zvýšených nároků na paměťový prostor. Do datového skladu se data dostávají ve větších dávkách a dále již nejsou modifikovány.

Zatímco relační databáze popsána pomocí ERD (entitně-relační diagramy), datový sklad je navrhován pomocí dimenzionálního modelování.

Modelování relační databáze:

- Odstranění redundancí (opakování se dat)
- Normalizace (důsledkem je nižší srozumitelnost pro člověka)
- Optimalizace na vkládání, modifikaci a čtení dat

Modelování datového skladu:

- Optimalizován na čtení dat, vyhledávání a analýzy
- Důraz na srozumitelnost pro uživatele
- Denormalizace, redundance
- Rozšiřitelnost – přidání fakt a dimenzí bez dopadu na aplikace

1.1.4 Datové tržiště (Data Mart)?

Analytické nástroje BI mohou čerpat data pro analýzy buď z datového skladu 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.

Při tvorbě datového skladu z datamartů se jedná 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 síť" ("spider net" je architektura data warehouse založená na jednotlivých data martech).

Naproti tomu koncepcí, 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.

1.1.5 Operativní datové úložiště

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, poslané marketingové nabídky. To vyžaduje další komponentu datového skladu – **operativní datové úložiště** – ODS (Operating Data Store).



Pro účely maximální „operativnosti“ je operativní úložiště napojeno na datové zdroje prostřednictvím nástrojů EAI (Enterprise Application Integration). 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 platform, které zpracovávají události dávkově v předem stanoveném čase, EAI platformy reagují na jednotlivé události okamžitě.

1.1.6 Struktura datového skladu

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

Schéma obsahuje dva typy tabulek – faktové a dimenzionální.

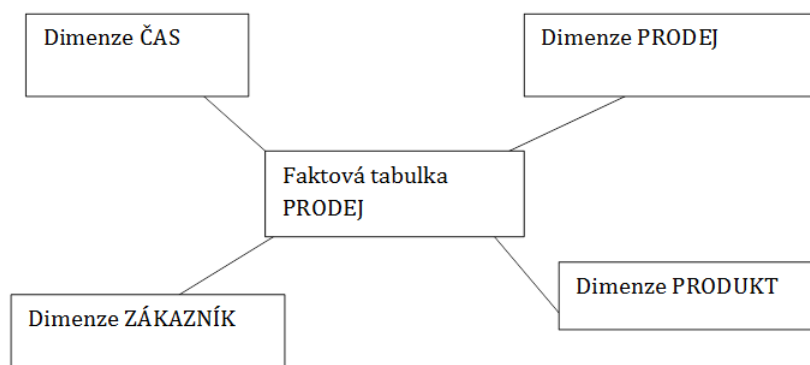
Jádro každého schématu tvoří jedna nebo několik **faktových tabulek**. V nich jsou uložena **vlastní analyzovaná data** – hodnoty, které jsou použity k analytickým výpočtům - agregacím, třídění apod.

Granularita faktové tabulky určuje úroveň podrobnosti v tabulce faktů. Čím nižší je úroveň granularity, tím detailnější jsou data určená k provádění matematických operací.

Faktové tabulky jsou 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.

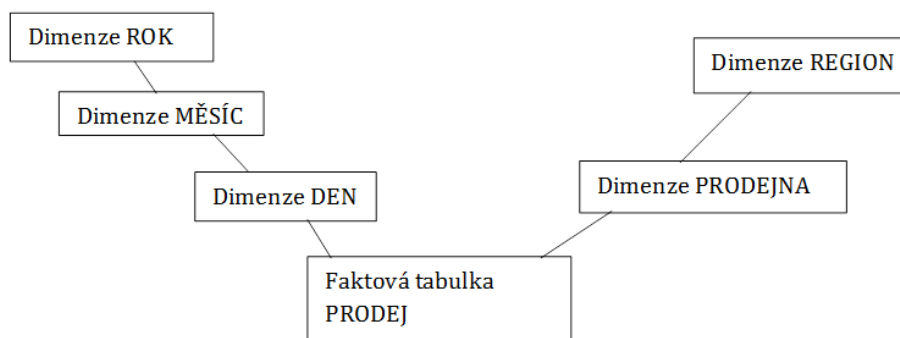
1.1.7 Schéma „hvězda“ a „sněhová vločka“

Schéma nám určuje strukturu datového skladu. Podle způsobu napojení dimenzí na faktovou tabulku rozlišujeme schéma typu **hvězda (star)** a schéma typu **sněhová vločka (snowflake)**.



Obr. 6.2 Schéma datového skladu typu „hvězda“

U schématu typu hvězda se jedná 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.



Obr. 6.3 Příklad – schéma datového skladu typu „sněhová vločka“

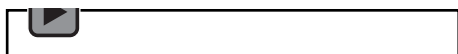
U schématu typu sněhová vločka jsou tabulky dimenzí normalizovány a mají relační strukturu. Faktová tabulka obsahuje cizí klíče do tabulek dimenzí.

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** (suhvězdy). Jedná se o soubor schémat typu „hvězda“, tj. více tabulek faktů se sdílenými dimenzemi.



Audio 6.1 Datový sklad



1.2 OLAP

1.2.1 Co je to OLAP a kdy se používá?

Pojmem OLAP (**On-line Analytical Processing**) se označují nástroje Business Intelligence pro on-line analýzu většího objemu dat. Termínem OLAP jsou také označovány databáze ve formě datového skladu. V širším pojetí je to pak označení informačního systému, který je založen na využití datového skladu a analytických nástrojů.

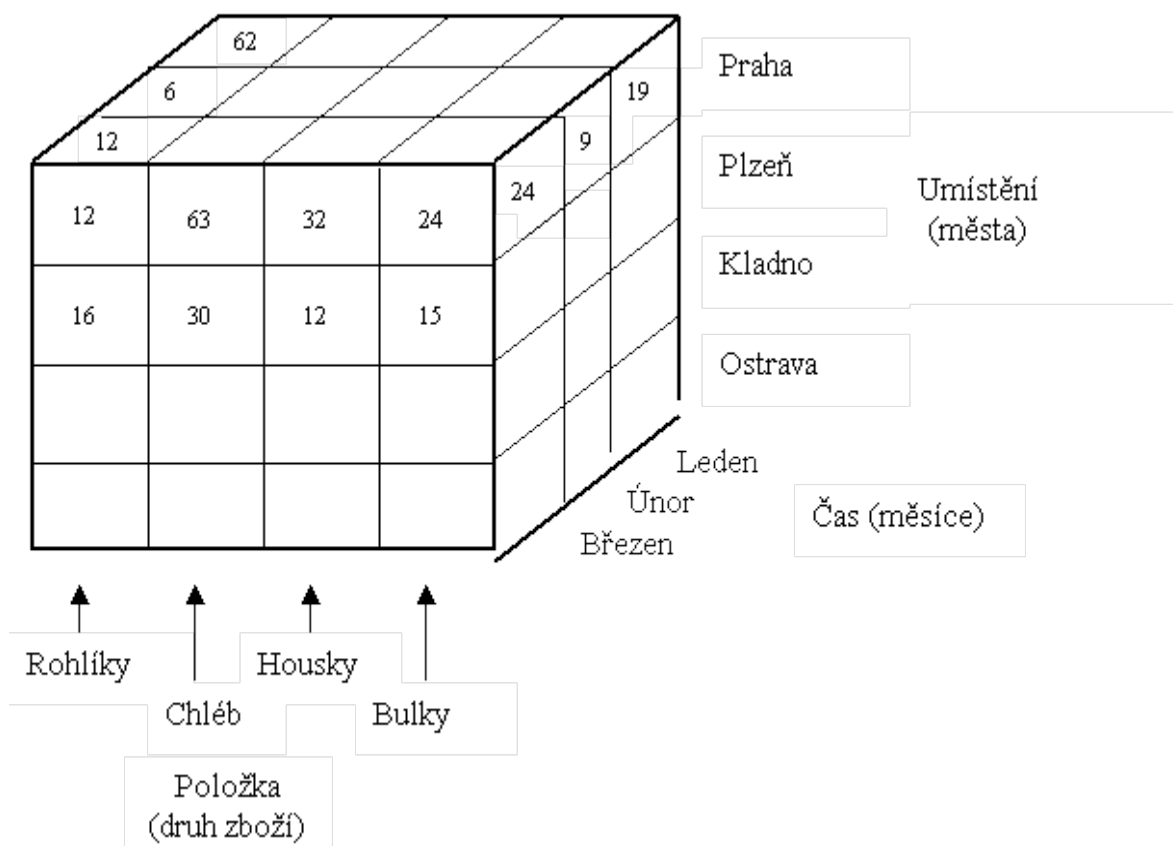
Definice:

Kategorie softwarových nástrojů umožňující analýzu dat uložených v databázi.

Edgar Codd

OLAP technologie pracuje s tzv. **multidimenzionálními daty**. Na rozdíl od dvourozměrného uložení dat v relačních databázích (sloupec, řádek), pracuje s tzv. **datovou kostkou** (anglicky cube, matematicky odpovídá datová kostka pojmu matice).





Obr. 6.4 Příklad datové kostky [zdroj: <http://datamining.xf.cz>]

Poznámka:

Datová kostka může mít více rozměrů než tři, např. Microsoft SQL Server umožňuje pracovat s kostkami až o 64 rozměrech.

1.2.2 ROLAP, MOLAP a HOLAP

Jedná se o rozlišení OLAP nástrojů podle toho, nad jakými daty je analýza prováděna.

Jestliže OLAP využívá k analýze přímo data uložená v relační databázi, hovoříme o **relačním OLAP** - **ROLAP**. OLAP v tomto případě v relační databázi vytvoří tabulky pro uložení agregací. Samotná analýza je náročná díky nutnosti používat standardní SQL příkazy k načtení dat z relacemi propojených tabulek. Zobrazení analyzovaných dat je již ovšem multidimenzionální.

Výhoda ROLAP: flexibilita, není potřeba budovat datový sklad.

Nevýhoda ROLAP: nároky na výkon databáze, pomalejší analýza.

Pokud je analýza prováděna nad datovým skladem, kde jsou data uložena v multidimenzionální struktuře, mluvíme o **multidimenzionální OLAP analýze** - **MOLAP**. Výstupy z analýzy MOLAP jsou oproti ROLAP rychlejší, protože odpadá nutnost používat složité SQL dotazy pro čtení dat z relacemi propojených tabulek. Data v multidimenzionální struktuře jsou uložena ve tvaru, který je vhodnější pro analytické zpracování; jsou seskupena dle logického významu. Oproti ROLAPu ovšem musíme data transformovat z relační databáze do datového skladu.

Výhoda MOLAP: rychlost analýzy.

Nevýhoda MOLAP: nároky na diskový prostor, data se musí transformovat do datového skladu.



Specifickým případem pak je kombinace obou přístupů, označována jako **hybridní OLAP - HOLAP**. Analýza se zde provádí nad relační databází, ale tabulky s agregacemi jsou ukládány do multidimenzionálních struktur v datovém skladu.

Výhody HOLAP: přístup k rozsáhlým datům při současně rychlé agregaci.

Nevýhody HOLAP: nutnost udržovat data na dvou místech.

Speciálním typem OLAPu je dynamický OLAP (DOLAP). Při tomto řešení OLAPu je multidimenzionální matice (kostka) budována jako virtuální v RAM paměti. Z toho plyne základní výhoda – neomezená flexibilita. Nevýhodou jsou vysoké nároky na RAM paměť a nutnost budovat kostku pokaždé znovu.

1.2.3 Základní operace v OLAP analýze

- **Drill-down** – umožňuje uživateli ve zvolené instanci jisté agregační úroveň 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 úroveň nastavuje vyšší (hrubší) agregační úroveň (menší detail v hierarchii dimenzí).
- **Pivoting** – umožňuje „otáčet“ datovou krychli, 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 instanci jisté agregační úrovně. Jinými slovy tato dimenze aplikuje filtr na instance příslušné agregační úrovně dané dimenze.

1.2.4 OLAP definice podle Codd

OLAP je definován pomocí 12 pravidel stanovených dr. Coddem (1993). OLAP lze také definovat pěti slovy - *Fast Analysis of Shared Multidimensional Information* (FASMI, 1995).

12 pravidel OLAP databáze:

- **Multidimenzionalita.** Pohled analytika na realitu je typicky multidimenzionální, tomuto pohledu musí odpovídat podstata datového modelu.
- **Transparentnost.** Skutečnost, zda multidimenzionální nástroje jsou součástí uživatelského běžného prostředí (tabulkového kalkulátoru apod.) nebo nejsou, musí být uživateli transparentní (skryté). Takto lze pohlížet také na data. Jejich skutečné uložení je pro uživatele nepodstatné.
- **Dostupnost** k datovým zdrojům (Accessibility). Odkud data pocházejí je problém systému, nikoliv uživatele. Systém by měl poskytovat pouze požadovaná data.
- **Konzistentní výkon** (Consistent Reporting Performance). Výkon systému nesmí záviset na počtu dimenzí.
- Architektura **client-server**.
- **Rovnoprávnost dimenzí** (Generic Dimensionality).
- **Dynamické řízení „řidkých matic“** (Dynamic Sparse Matrix Handling).
- **Víceuživatelská podpora** (souběžné vyhledávání a aktualizace, zabezpečená integrita a bezpečnost).
- **Neomezené operace napříč dimenzemi** (mezidimenzionální vazby).
- **Intuitivní ovládání.**
- **Flexibilní reporting.**
- **Neomezený počet dimenzí** a úrovní v hierarchiích.

Jako jednoduchý OLAP nástroj lze použít MS Excel - „**kontingenční tabulka**“ (pivot table).



Práce s OLAP kostkou je v podstatě práce s kontingenční tabulkou - na základě volených dotazů se kostka upravuje pro kvalitnější a rychlejší práci s daty.

1.3 DOLOVÁNÍ DAT

1.3.1 Co je to dolování dat (Data Mining)?

Data Mining (neboli dolování dat) je **proces hledání neznámých informací a znalostí ve velkém objemu dat**. Pro Data Mining se využívají technologie rozpoznávání vzorů, statistické a matematické metody.

Definice Data Mining

Data Mining je proces výběru, prohledávání a modelování ve velkých objemech dat, sloužící k odhalení dříve neznámých vztahů mezi daty za účelem získání obchodní výhody.

Fayyad

V odborné literatuře je určitá nejednotnost v používání pojmu dolování dat (Data Mining) a **dobývání znalostí z databází** (Knowledge Discovery in Databases, **KDD**). Někteří autoři tyto pojmy používají jako synonyma. Podle první mezinárodní konference KDD (Montreal, 1995) je dolování dat součástí KDD.

Definice KDD

Knowledge Discovery In Databases (KDD) je proces netriviálního objevování implicitních, dopředu neznámých a potenciálně použitelných znalostí v datech.

Fayyad

1.3.2 Metody dolování dat

Dnes užívanými metodami dolování dat jsou například:

- regresní metody (lineární regresní analýza, nelineární regresní analýza, neuronové sítě)
- klasifikace (diskriminační analýza, logistická regresní analýza, rozhodovací stromy, neuronové sítě),
- segmentace – shlukování (shluková analýza, genetické algoritmy, neuronové shlukování – Kohonenovy mapy)
- analýza vztahů (asociační algoritmus pro odvozování pravidel typu „if X then Y“)
- predikce v časových řadách (Boxova-Jenkinsonova metoda, neuronové sítě, autoregresní modely, ARIMA)
- detekce odchylek

1.3.3 Modely dolování dat

Deskriptivní model – popisuje nalezené vzory a vztahy v datech, které mohou ovlivnit rozhodování (Př. Analýza prodeje zboží v supermarketu, na jejímž základě je pak umístěno zboží v regálech).

Prediktivní model – umožňuje předvídat budoucí hodnoty atributů na základě nalezených vzorů v datech (Př. Analýza zákazníků, u kterých je vysoká pravděpodobnost, že budou reagovat na písemnou reklamní nabídku...)

1.3.4 Metodologie dolování dat

Nejpoužívanější metodologie jsou: **SEMMA** (SAS), **5A** (SPSS) a **CRISP-DM**.



1.3.4.1 CRISP-DM

CRISP (CRoss Industry Standard Proces for Data Mining) je souhrnná metodologie dobývání znalostí z databází, umožňuje provádět rozsáhlé projekty dobývání rychleji, efektivněji a méně nákladně prostřednictvím osvědčených postupů.

Základní etapy procesu dobývání jsou:

- Co řešit (Business understanding) – porozumění problematice, formulování úlohy
- Kde vzít data (Data understanding)
- Jak data připravit (Data preparation)
- Jak data analyzovat (Data modelling)
- Co jsme zjistili (Evaluation) – porozumění výsledkům
- Jak výsledky využít (Deployment)

1.3.4.2 SEMMA

Podle metodologie SEMMA spočívá proces dobývání v těchto krocích:

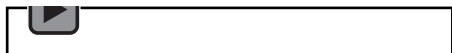
- **Sample** - vybírání vhodných objektů
- **Explore** - vizuální explorace a redukce dat
- **Modify** - seskupování objektů a hodnot atributů, datové transformace
- **Model** - analýza dat
- **Assess** - porovnání modelů a interpretace

1.3.4.3 5A

- **Assess** - posouzení potřeb projektu
- **Access** - shromáždění potřebných dat
- **Analyze** - provedení analýz
- **Act** - přeměna znalostí na akční znalosti
- **Automate** - převedení výsledků analýzy do praxe



Audio 6.2 Dolování dat



2 KONTROLNÍ OTÁZKY

1. Co to je Business Intelligence?
2. Co to je datový sklad a čím se liší od relační databáze?
3. Jak se dostávají data do datového skladu?
4. Jaká je struktura datového skladu a co je to schéma datového skladu?
5. Co obsahuje v datovém skladu faktová tabulka a co obsahuje tabulka dimenzí?
6. Co je cílem analýzy OLAP?
7. Co je to dolování dat?
8. Jaký je rozdíl mezi deskriptivním a prediktivním modelem u dolování dat?
9. Jaký je rozdíl mezi OLAPem a dolováním dat?



3 POUŽITÁ LITERATURA

- [1] Pour, J. – Maryška, M. – Novotný, O.: Business intelligence v podnikové praxi. Professional Publishing, 2012. ISBN: 80-7431-065-2.
- [2] Lacko, L.: Business Intelligence v SQL Serveru 2008. Computer Press Brno, 2009. ISBN: 80-251-2887-9.

