DATABÁZOVÉ SYSTÉMY a METODY ZPRACOVÁNÍ DAT 2

Doc. Dr. Ing. Jana Klečková
kleckova@kiv.zcu.cz
UN 328

1

Předpokládané znalosti

- SŘBD, vlastnosti SŘBD. Konceptuální modelování. Relační model dat. Závislost atributů, normální formy. Základní principy transakčního zpracování.
- Dotazy v SQL, příkaz SELECT. Integritní omezení dle SQL. Oprávněnost přístupu k datům dle SQL.

2

Obsah předmětu

1.	Úvod, databázová technologie - analýza současného stavu, trendy vývoje, metody zpracování dat.
2.	Principy a nástroje Business Intelligence, vrstvy pro analýzu dat - reporting, datový sklad, datamining..
3.	Přístupy k vytváření datových skladů a datových tržišť, bussiness požadavky, projektový plán.
4.	Dimenzionální model - hierarchie dimenzí, rozdělení atributů na dimenze, fakta, atributy, definování typů vztahů, aditivity faktů, definice omezení dotazů.
5.	Problematika modelování dat v datových skladech, charakteristiky tabulek faktů a dimenzí, transformace mezi jednotlivými modely.
6.	Datová kostka – jiný pohled na dimenzionální modelování
7.	Architektura datového skladu.
8.	Metadata pro správu datového skladu, metadata zdrojových dat, metadata datového skladu, metadata datové pumpy, metadata pro data a funkce na pozadí datového skladu, metadata pro koncového uživatele, standardizace metadat.
9.	Datová pumpa, proces extrakce, transformace a vložení dat, metody čištění dat.
10.	Technologie implementace datového skladu, plnění datového skladu.
11.	Dolování dat (Data mining) - popis dolování dat, úlohy dolování dat, použité aplikace dolování dat.
12.	Techniky dolování dat- statistické metody, metody umělé inteligence.

3

Důležité termíny

- **Sledujte Courseware**

4

Podmínky absolvování předmětu

Zápočet

- Za zápočet je nutné obdržet alespoň 24 bodů, tj. 60% z maximálního počtu 40 bodů. K získání zápočtu je třeba úspěšně absolvovat zápočtový test a získat dostatečný počet bodů při osobním předvedení semestrální práce. Při nesplnění jedné z těchto nutných podmínek nemáte nárok na zápočet.
- **Zápočtový test** je úspěšně absolvován, pokud za něj obdržíte alespoň 14 bodů, tj. 60% z maxima 22 bodů. Pokud získáte méně bodů, máte možnost test maximálně jednou opakovat.
- Předvedení **semestrální práce** probíhá zpravidla v laboratoři **UC-333**, za předvedení programu včetně vytištěné dokumentace obdržíte maximálně 18 bodů.
- Zápočet je zapisován do katalogu (indexu) v den, kdy student splnil všechny podmínky, tj. úspěšně absolvoval zápočtový test a získal v součtu požadované minimum 24 počtu bodů za zápočtový test a semestrální práci.

5

Podmínky absolvování předmětu

Zkouška

- Na zkoušku se může napsat každý, kdo má v katalogu (indexu) zapsán zápočet.
- Celkem lze obdržet až 60 bodů, pro absolvování předmětu je nutné získat ze zkoušky 36 bodů.

- Důležité upozornění: Kombinovaná zkouška - se skládá z písemné práce a ústní části.

6

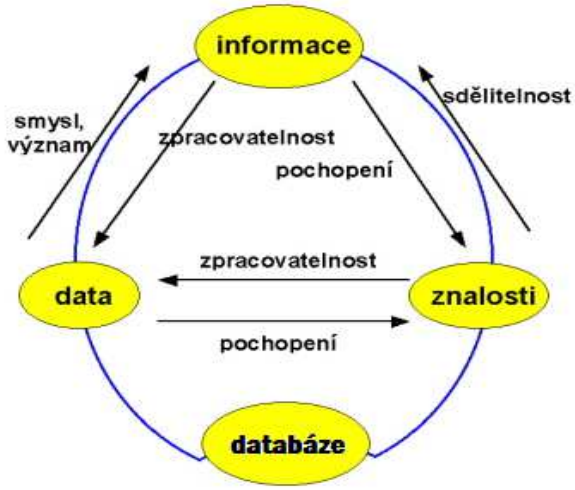
Časová náročnost

Aktivity	**Časová náročnost aktivity [h]**
Kontaktní výuka	65
Příprava na zkoušku [30-60]	40
Příprava na souhrnný test [10-40]	11
Vypracování seminární práce v mag. st. [40-50]	40
Celkem	**156**

7

Databázové technologie

Analýza současného stavu, trendy

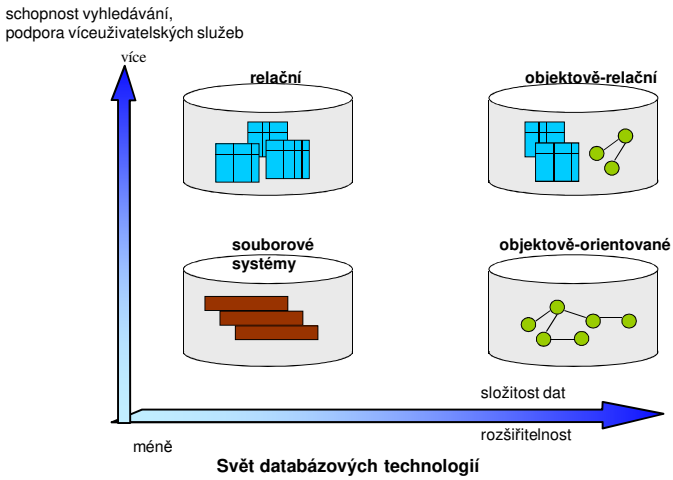


9

E-R model

- E-R model dává možnost konceptuálního modelování se systematickým přístupem k výslednému návrhu schémat relací.
- E-R model má mnoho odpůrců - přesto je tato koncepce ve svých četných variantách de facto standardem ve světě strukturovaných metodologií návrhu nejen databází, ale i obecnějších systémů.
- Na této metodologii návrhu jsou vybudovány i prostředky objektové.

10



Svět databázových technologií

11

Provoz a údržba

- Většina organizací se minimálně stará o údržbu podnikových aplikací stávajících systémů. To způsobuje, že se používají "špatná" databázová schémata a obecně dochází k "úpadku databází" [10].
- Autoři tvrzení vychází z diskusí s téměř dvaceti správci databází (DBA) u tří velkých podniků. Databáze se mění v závislosti na podmínkách byznysu, běžně jednou za čtvrtletí i více.
- Heterogenní a dynamické datové prostředí vede k tomu, že často mizí role centrálního správce a objevuje se spíše decentralizovaný přístup s více skupinami DBA zabývajícími se databázemi v podniku.

12

Provoz a údržba

- Databáze jsou většinou relační. Od zavedení relačního modelu dat bylo sice zavedeno několik databázových modelů, jako je objektově-orientovaný (OO), objektově-relační (OR), XML či RDF.
- OO a OR SŘBD reagovaly na objektové přístupy v softwarovém inženýrství z 90. let. Tyto prostředky, však nikdy na trhu nebyly skutečně konkurenceschopné. Důvody by mohly být v nedostatku jejich teoretických základů a omezené výkonnosti.

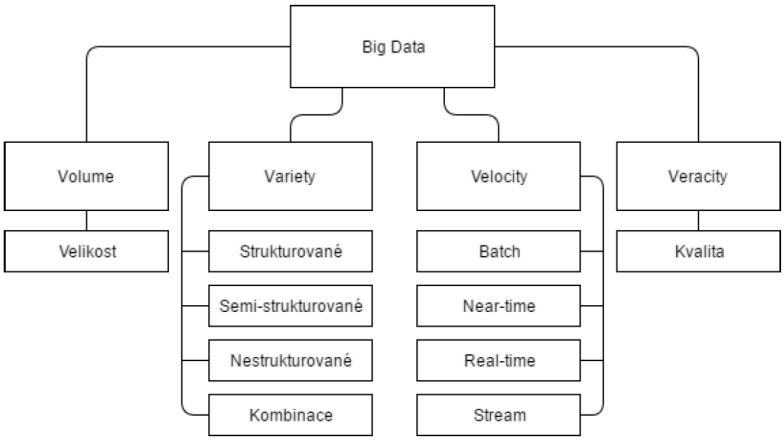
13

Aktuální problém 1– nárůst dat

- Současnou situaci v databázovém světě ovlivňují tzv. Big Data.
- Základní charakteristiky Big Data, tzv. *V*-charakteristiky jsou 3V:
  - objem (Volume),
  - rychlost (Velocity,
  - různorodost (Variety).
- V poslední době rozšířeno o další charakteristiku, označováno jako 4V:
  - věrohodnost (Veracity), v podstatě kvalita dat.

14

Charakteristika Big Data



15

Zpracování Big Data

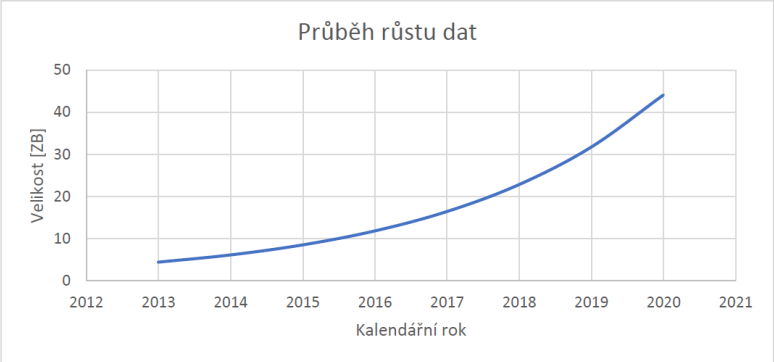
- V literatuře [Po1]se uvádí dokonce 14 takových *V*, které zásadně ovlivňují infrastrukturu ukládání a zpracování Big Data.
- Efektivní využívání systémů zahrnujících zpracování velkých objemů dat vyžaduje v mnoha aplikačních scénářích odpovídající nástroje pro jejich ukládání na nízké úrovni a analytické nástroje ve vyšších úrovních.
- Z pohledu uživatele je nejdůležitějším aspektem zpracování velkých objemů dat na počítači jejich analýza - Big Analytics.

Zdroj:

[Po1]Pokorný, J.: Big Data Storage and Management: Challenges and Opportunities. In: Proc. of 12th IFIP WG 5.11 Int. Symp. on Environmental Software Systems, IFIP AICT 507, Springer (2017)

16

Predikce růstu dat



Graf 1 Předpovídané množství dat od roku 2013 do 2020 (zdroj: IDC)

17

Big Data – Big Analytics

- Velké kolekce dat obsahují data v různých formátech, např. relační tabulky, XML data, textová data, multimediální data nebo RDF trojice.
- Vznikají potíže při jejich zpracování algoritmy pro dolování dat (DM).
- Zvyšující se objem dat v úložišti a počet jeho uživatelů vyžaduje spolehlivé řešení škálování v těchto dynamických prostředích a pokročilejší prostředky pro zajištění vysokého výkonu, než nabízejí tradiční databázové architektury.
- Big Analytics se provádí i nad velkým množstvím transakčních dat rozšířením metod používaných v datových skladech (DW) – tato technologie DW je spíše zaměřena na strukturovaná data ve srovnání s bohatší variabilitou typů dat, což je aktuální pro Big Data.
- Analytické zpracování velkých objemů dat proto vyžaduje nejen nové databázové architektury, ale také nové metody pro analýzu dat.

18

Možná řešení problému Big Data

- Pro ukládání a zpracování Big Data lze dnes volit tradiční SŘBD, paralelní DBS, distribuované souborové systémy (např. HDFS), datová úložiště typu klíč-hodnota (NoSQL databáze) a nové databázové architektury (NewSQL databáze).
- **Pro volbu technologie jsou rozhodující aplikace**, které mohou být jak transakční, tak analytické.
- Požadují obvykle různé architektury software i hardware, často v jedné infrastruktuře.

19

Vývoj databází v 90. letech

- Snaha integrovat heterogenní data v podniku a rozšiřovat možnosti SŘBD o další datové typy (ukládat do databáze všechno, tj. možné i nemožné)
- Jedním ze směrů technického řešení těchto problémů bylo rozšířit relační tabulky SQL o objekty (netriviální rozsáhlé objekty typu text, audio, video atd.
- Bylo nutno vyvíjet nové dotazovací jazyky, které umožnily nejen nové typy dotazů (např. najdi k objektu v prostoru jeho nejbližšího souseda)
- Došlo i k přehodnocení dotazování

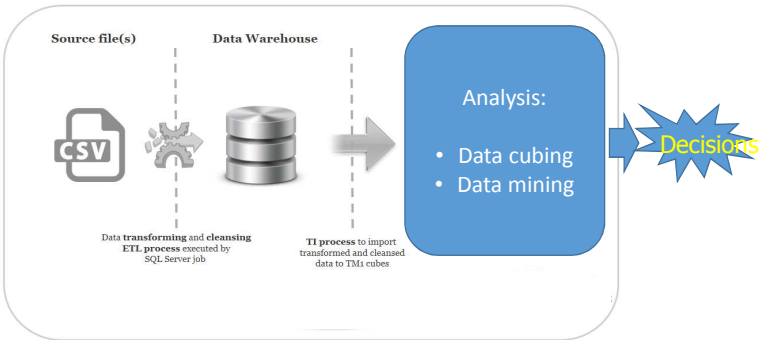
20

Vývoj databází v 90. letech

- Vznikaly tzv. univerzální severy s ad hoc přidávanými novými datovými typy.
- Integrace podnikových dat „ve velkém“ vedla k řadě architektur vycházejících z původních idejí distribuovaných databází řešených v 80. letech.
- Podstatou byl přístup zdola-nahoru k řešení distribuované databáze, založený na (ruční) integraci dílčích databázových schémat.
- Značné úsilí bylo věnováno řešení sémantických konfliktů mezi daty několika databází a transakcím nad více databázemi.
- Neúspěšnost těchto architektur v rámci IS podniku nakonec vedla k vývoji datových skladů (DW - Datawarehouse)
  - **Integrace dat se v DW provádí tak, že se potřebná data „vypumpují“ z operačních databází,vyčistí a uloží do databáze speciální.**
  - **Datová pumpa – proces ETL – nezbytný prostředek používaný i mimo DW.**

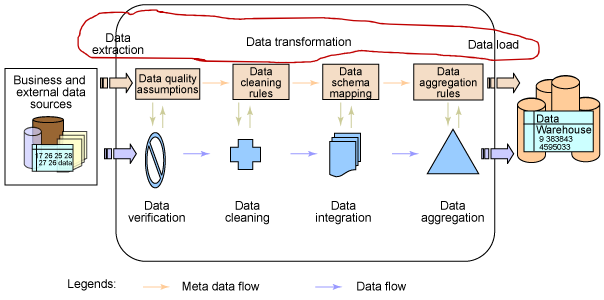
21

Analytics based on Data Cubing – System Architecture



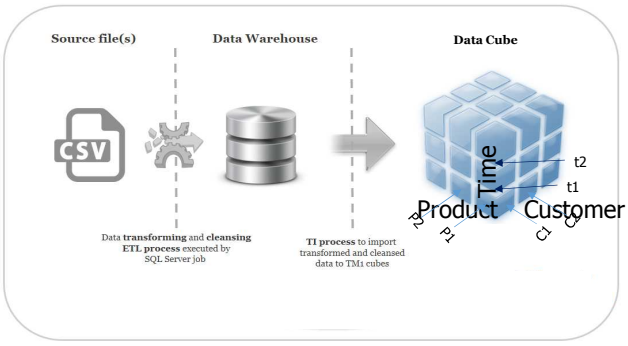
22

Analytics based on Data Cubing – ETL Process



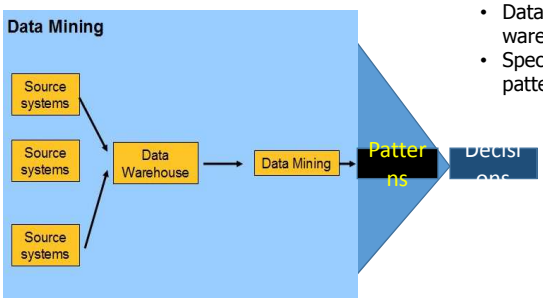
23

Data Warehouse - Analysis via Cubing



24

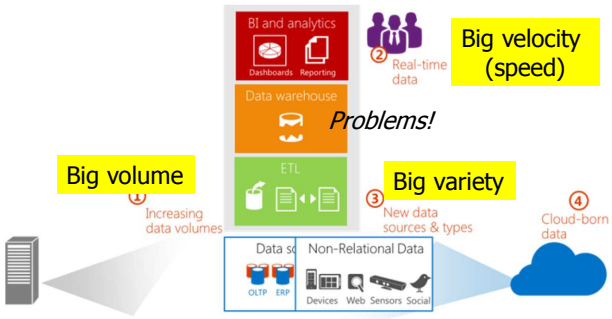
Data Mining



- Data mining (DM) uses data warehouse with differing data types
- Specialized software searches for patterns in data:
  - relationships in time
  - clusters based on similarities
  - dependent-independent variables (causation, prediction)
  - classification (breakdowns)
  - evolution analysis (prediction)
  - summaries

25

Big Data Challenge



26