Data management, kvalita dat

Využití asociačních pravidel pro zvyšování kvality dat

INS_2020_3. přednáška

Současná situace

- Nekvalitní data stojí americké firmy ročně 600 miliard dolarů (dle studie firem DataFlux a SAS)
- Na základě auditu jedna evropská firma objevila, že nevystavila fakturu na 4% objednávek – což představovalo 80 milionů dolarů (DM Review)
- V roce 1992 se vrátilo 96 000 daňových přeplatků zpět z důvodu nedoručitelné adresy
- Špatně uvedené ceny v databázi obchodních řetězců stojí ročně americké zákazníky na 2,5 miliard dolarů na přeplatcích
- Podle organizací jako Data Warehouse Institute, the Gartner Group a Meta Group – kvalita dat představuje jeden ze tří nejhlavnějších kriterií úspěchu datových skladů
- Středně velká firma může mít ve svých databázích, souborech, reportech 30 000 – 50 000 atributů (Platinum Technology)

Kvalita dat

- Kvalita dat je významným problémem a výzvou pro současné firmy
- Nekvalitní data mohou mít vliv na:
  - Nekvalitní řízení (např. rozhodování na základě nepravdivých dat)
  - Zpomalení rozhodovacích procesů (např. dlouhý čas k získání správných údajů)
  - Zhoršení image organizace (např. špatné informace na www)
  - Ztráta zákazníka (např. zaslání vyšší faktury)

Data Quality Assurance

- **Zajištění kvality dat** je proces profilování dat s cílem objevit nesrovnalosti a jiné anomálie dat a provádění aktivity čištění dat (např. odstranění odlehlých hodnot, chybějící údaje interpolace) ke zlepšení kvality údajů.
- Tyto aktivity mohou být realizovány v rámci datových skladů nebo jako součást databáze správy existující kus aplikačního software.

Kritika stávajících nástrojů a postupů

Hlavní uváděné důvody jsou:

1. Náklady projektu: náklady obvykle ve stovkách tisíc dolarů
2. Čas: nedostatek času zabývat se údaji ve velkém měřítku - čištění pomocí software
3. Bezpečnost: obavy o sdílení informací - přístup aplikací napříč systémy a dopad na data ve starších systémech

Definice pojmů

- Datová kvalita (Data Quality) – klasická definice
  - Data splňují následující atributy
    - Přesnost
    - Úplnost
    - Včasnost
    - Jedinečnost
    - Konzistentnost
- Datová kvalita je široký a těžko definovaný pojem
  - Zahrnuje nejen stav dat ale i procesy nakládání s daty
- O nekvalitních datech můžeme mluvit jestliže:
  - Data nesplňují dané specifikace
  - Nelze zaručit správnou interpretaci dat
  - Data nejsou vhodná pro řešení našich obchodních problémů

Datová kvalita

- Jsou tato data kvalitní (?):

Column 1
321453
212392
093255
214421
...

- Co z nich lze odvodit?

Datová kvalita

Datová kvalita = f(Definice + Data + Prezentace)

- Definice
  - Definice dat
  - Specifikace domény
  - Obchodní pravidla určující data
  - Procesy datové kvality
- Data (obsah)
  - Úplnost
  - Správnost
- Prezentace dat
  - Dostupnost
  - Včasnost
  - Jednoznačnost

Vybrané problémy v datech

- Obsah dat
  - Chybějící hodnoty
  - Chybná data
  - Překlepy
  - Data mimo danou doménu
  - Nelegální kombinace dat
- Strukturální
  - Entitní integrita
  - Referenční integrita
- Migrace/Integrace
  - Duplicitní záznamy
  - Chybějící záznamy
  - Konverze typů

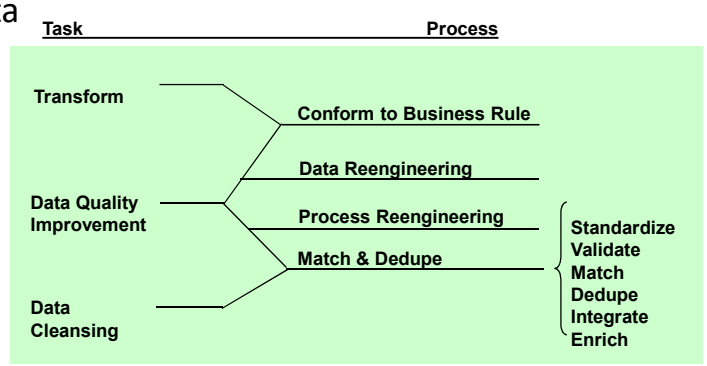
- Definice a standardy
  - Dvojznačné obchodní pravidla
  - Více formátů pro stejné atributy
  - Různý význam stejně pojmenovaných atributů
  - Více kódů se stejným významem
  - V jednom atributu více informací

Definice pojmů

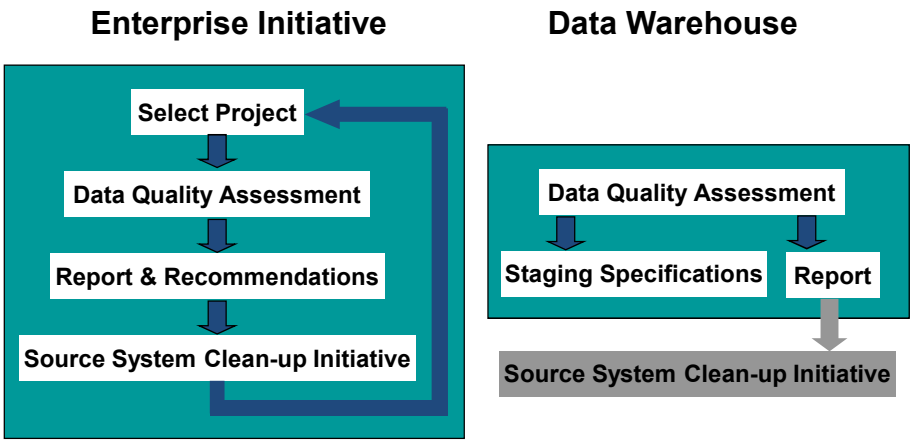
- Transformace dat (Data Transformation) – změna dat do konzistentní podoby podle integritních a obchodních pravidel
- Čištění dat (Data Cleansing) –proces transformace dat za účelem odstranění duplicitních a nekorektních záznamů v datech
- Zlepšování datové kvality (Data Quality Improvement) – proces zvyšování kvality dat na úroveň požadovanou pro podporu informačních potřeb organizace

Zvyšování kvality dat

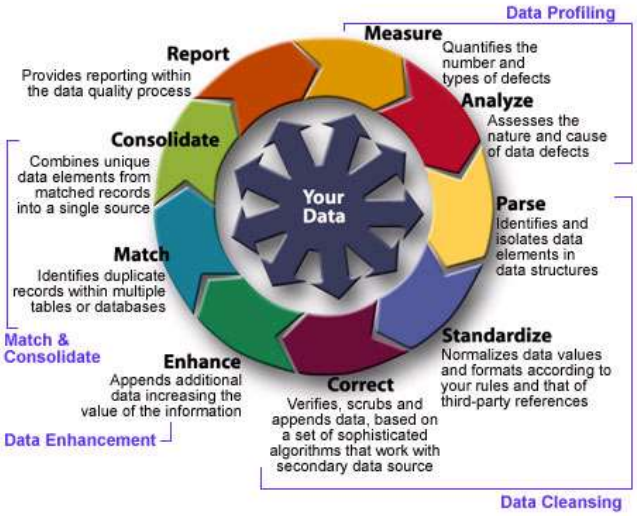
- Proces zvyšování datové kvality zasahuje:
  - Procesy
  - Data



Zvyšování kvality dat



Kroky zvyšování datové kvality



Změny

- Na základě posledního průzkumu (za rok 2018) byl zjištěn významný pokles využívání metrik v rámci masivních inspekcí.
- Současně došlo k nárůstu využívání metrik kvality dat v případě nutnosti nebo na základě požadavku (ad-hoc).

Výsledek výzkumu

Pořadí	2017	2018
	Dimenze	Dimenze
1	Accuracy (Přesnost)	Completeness (Úplnost)
2	Completeness (Úplnost)	Validity (Správnost)
3	Consistency (Konzistence)	Consistency (Konzistence)
4	Validity (Správnost)	Accuracy (Přesnost)
5	Timeliness (Včasnost)	Integrity (Integrita)
6	Integrity (Integrita)	Timeliness (Včasnost)
7	Accessibility (Dostupnost)	Accessibility (Dostupnost)
8	Currency (Aktuálnost dat – data representují realitu a odráží současný stav)	Lineage (Existence dokumentace informačního toku)
9	Precision (Přesnost)	Representation (Prezentace)
10	Lineage (Existence dokumentace informačního toku)	Precision (Přesnost)
11	Representation (Prezentace)	Currency (Aktuálnost dat – data representují realitu a odráží současný stav)

Princip

- Hodnocení kvality jsou v podstatě náklady na nekvalitu, ne indexy

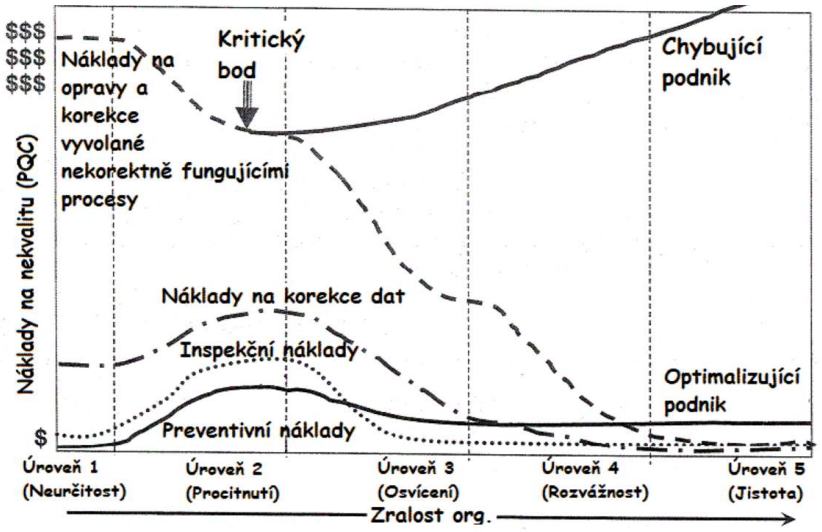
Vysvětlení:

- Rozhodneme-li se věnovat problematice kvality informací (dat), je nutno zvážit, co nekvalita informací znamená a jaké má dopady.
- Dopady jsou finanční i výkonnostní – negativní dopad do KPI (*Key Performance Indicator*).
- Firma by měla implementovat některý z modelů hodnocení nákladů nekvality (například PAF(***P****revention,* ***A****ppraisal,* ***F****ailure*) model, prostřednictvím kterého je možné objasnit, jak se z nákladového pohledu má firma dívat na nejčastěji implementovaný cyklus *inspekce* ↔ *opravy*.

Implementace modelu

- jedná se o manažerský nástroj pro řízení kvality,
- poskytuje celkový pohled na kvalitu (finanční),
- pomáhá k prioritizaci problémů,
- pomáhá hodnotit dopady implementovaných nápravných opatření.

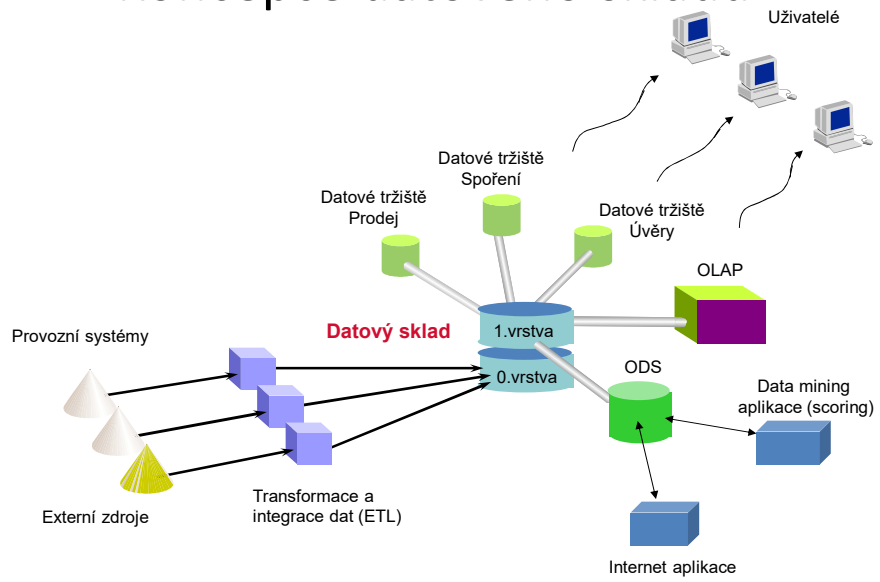
Trendy v jednotlivých kategoriích PAF



DM analýzy

- Kvalita dat je významným faktorem z hlediska analytického využití dat
  - 60 – 80 procent času DM projektů zabírá příprava dat
- Data pro pokročilé analýzy pocházejí většinou z datového skladu
- Zvyšování kvality dat
  - Během procesu načítání dat (ETL)
  - Během tvorby specializovaných datových tržišť

Koncepce datového skladu



Metadata

- Pro kontrolu a zvyšování kvality dat je třeba disponovat kvalitními metadaty (integritní a obchodní pravidla)
- Zvyšovat kvalitu dat lze:
  - Zlepšováním procesů pracujících s daty
  - Využít a aplikovat definovaná integritní a obchodní pravidla
  - Automatickou detekce nekvalitních dat + automatická tvorba metadat
    - Např. využití DM algoritmů (regrese, Decision Tree, NN) pro doplnění chybějících hodnot

Využití asociačních pravidel

- Myšlenka výzkumu: využít asociační pravidla pro automatické objevení chyb v datech a jejich nápravu
- Využít rozšíření asociačních pravidel a všech možností 4FT kvantifikátorů
- Definovat nové typy asociačních pravidel vhodné pro oblast kvality dat

Současné kvantifikátory

- Využití kvantifikatorů
  - Implikační
  - Dvojitě implikační
  - Ekvivalenční
  - Další (Average)

Co lze řešit

- Pravidla lze aplikovat:
  - Na tabulku
  - Na databázi (více tabulek)
- Nalezená pravidla mohou pomoci řešit následující problémy v datech:
  - Chybějící hodnoty
  - Chybná data
  - Nelegální kombinace dat
  - Stejný význam různě pojmenovaných atributů
  - Různý význam stejně pojmenovaných atributů
  - Více kódů se stejným významem
  - Validace stávajících obchodních pravidel

Nové typy pravidel

- Nové typy pravidel např.:
  - 1. Matematické pravidla
    A * B = C, kde * může nahrazovat řadu aritmetických operací
  - 2. Pravopisná a konverzní pravidla
    V atributu JMENO se vyskytuje hodnota DAVID v 25 záznamech, 3 záznamy mají podobnost < než daný práh
- Nová pravidla tak mohou řešit:
  - Překlepy
  - Duplicitní záznamy
  - Různé měrné jednotky

Shrnutí

- Datová kvalita je obsáhlým problémem
- Zvyšování datové kvality zahrnuje kontrolu a změny:
  - Vlastních dat
  - Procesů pracujících s daty
- Základem zajištění datové kvality jsou správná a kompletní metadata (integritní a obchodní pravidla)
- Rozšíření asociačních pravidel může přinést významnou pomoc pro indikaci a odstranění chyb v datech
- Implementace technologií, technik, metod kvality informací trvá přibližně 2 roky, implementace prostředí plně podporujícího řízení kvality informací trvá cca 4-5 let

Zdroje

- Dasu, Tamraparni, Johnson Theodore: Exploratory data mining and data cleaning, Hoboken : Wiley-Interscience, 2003
- http://web.mit.edu/tdqm
- http://www.dataquality-research.com
- D. Myers, 2018 Anual Report on the Dimensions of Data Quality, DQ Matters Data Quality eLearning, 2018.
- D. B. Laney, Infonomics: How To Monetize, Manage, and Measure Information as an Asset for Competitive Advantage, Bibliomotion, Inc., 2018.
- http://dimensionsofdataquality.com/alldimensions

- Děkuji za pozornost

Literatura

- Dasu, Tamraparni, Johnson Theodore: Exploratory data mining and data cleaning, Hoboken : Wiley-Interscience, 2003
- http://web.mit.edu/tdqm
- http://www.dataquality-research.com

- Kimball Raplh: The Data Warehouse Toolkit, John Wiley & Sons, 2002
- Kimball Raplh: The Data Warehouse Lifecycle Toolkit, John Wiley & Sons, 1998
- Lacko Luboslav: Databáze: datové sklady, OLAP a dolování dat s příklady v MS SQL Serveru a Oracle, Computer Press, 2003
- Humphries M., Hawkins M. W. : Data warehousing : návrh a implementace, Computer Press, 2002
- Berry M. J., Linoff G.: Data Mining Techniques for marketing, sales and customer support, John Wiley & Sons, 1997
- Rud Olivia Parr: Data mining, Computer Press, 2001
- Berka Petr: Dobývání znalostí z databází, Academia, 2003