VYSOKÁ ŠKOLA EKONOMIE A MANAGEMENTU

# DIPLOMOVÁ PRÁCE

2011

VÁCLAV HOBLÍK

**VYSOKÁ ŠKOLA EKONOMIE A MANAGEMENTU**

**Nárožní 2600/9a, 158 00 Praha 5**

# DIPLOMOVÁ PRÁCE

## MANAGEMENT FIREM

**Vysoká škola ekonomie a managementu**

+420 841 133 166 / info@vsem.cz / www.vsem.cz

# VYSOKÁ ŠKOLA EKONOMIE A MANAGEMENTU

**Nárožní 2600/9a, 158 00 Praha 5**

| **NÁZEV DIPLOMOVÉ PRÁCE** |
| --- |
| Datová kvalita, integrita a konsolidace dat v BI |

| **TERMÍN UKONČENÍ STUDIA A OBHAJOBA (MĚSÍC/ROK)** |
| --- |
| 10/2011 |

| **JMÉNO A PŘÍJMENÍ / STUDIJNÍ SKUPINA** |
| --- |
| Václav Hoblík / MF2 |

| **JMÉNO VEDOUCÍHO DIPLOMOVÉ PRÁCE** |
| --- |
| Doc. Ing. Jan Žižka, CSc. |

| **PROHLÁŠENÍ STUDENTA** |
| --- |
| Prohlašuji tímto, že jsem zadanou diplomovou práci na uvedené téma vypracoval samostatně a že jsem ke zpracování této diplomové práce použil pouze literární prameny v práci uvedené.<br><br>Datum a místo:<br><br>_____<br>podpis studenta |

| **PODĚKOVÁNÍ** |
| --- |
| Rád bych tímto poděkoval vedoucímu diplomové práce, za metodické vedení a odborné konzultace, které mi poskytl při zpracování mé diplomové práce. |

VYSOKÁ ŠKOLA EKONOMIE A MANAGEMENTU

# Datová kvalita, integrita a konsolidace dat v BI

Data quality, integrity and consolidation of data in BI

Autor: Václav Hoblík

**Souhrn**

V diplomové práci je zpracováno téma datové kvality, konsolidace a integrity dat v systémech Business Intelligence. Práce je rozdělena na dvě části. V první části se autor snaží vymezit danou problematiku a shrnout terminologii z jednotlivých oblastí. Pozornost je věnována problematice metadat, jejich uložení a použití jako integrujícího prvku firemní IT/ICT infrastruktury. Ve druhé části práce jsou demonstrovány získané poznatky na reálném projektu realizace BI pro farmaceutickou společnost. Zpracováno je zajištění datové kvality dat získaných z produkčních systémů a její udržení za provozu BI aplikace.

**Summary**

The theme of the work is data quality, consolidation and data integrity in Business Intelligence systems. The work is divided into two parts. In the first part the author attempts to define the issue and summarize the various areas of terminology. Attention is paid to the metadata, storage and use as an integrating element of corporate IT / ICT infrastructure. In the second part of the work is demonstrated acquired knowledge to real-BI implementation project for a pharmaceutical company. It is described the way how to ensure the data quality of data obtained from production systems and the maintenance of BI applications for operations.

**Klíčová slova:**
Business Intelligence, datový sklad, datová kvalita, datová integrita, datová konsolidace, metadata.

**Keywords:**
Business Intelligence, data warehouse, data quality, data integrity, data consolidation, metadata.

**JEL Classification:**

C8 - Data Collection and Data Estimation Methodology; Computer Programs

C88 - Other Computer Software

# Obsah

**Seznam zkratek**

| | |
|---|---|
| BI | Business Intelligence |
| DW, DWH | Datový sklad, Data Warehouse |
| ETL | Extraction Transform Load |
| OLAP | OnLine Analytical Processing |
| OLTP | OnLine Transaction Processing |

**Seznam tabulek**

Seznam grafů

**Seznam obrázků**

# 1 Úvod

Business Intelligence (BI) systémy se stávají v současné době součástí IT/ICT infrastruktury nejenom velkých společností, ale i menších a středních. S klesajícími bariérami na globálním trhu roste důraz na rychlé a správné rozhodování, podložené kvalitními informacemi. Zde je místo právě pro systémy Business Intelligence, jakožto analytické nástroje poskytující data v požadované kvalitě a čase konkrétnímu uživateli. Návrh a implementace takového systému ale není triviální záležitostí a řada projektů končí neúspěchem. Mezi jeden z hlavních faktorů neúspěchu patří nízká kvalita dat, která je „zděděna“ z produkčních systémů, nebo dokonce prohloubena špatnou konsolidací a integrací dat v rámci zpracování zdrojových dat do BI aplikace, popřípadě je již zpočátku zapříčiněna nevhodným návrhem. Důvodů neúspěchů BI aplikací může být jmenováno mnohem více, mimo jiné přílišná složitost pro uživatele, reporty, byť se správnými daty jsou dodány v nesprávný čas nebo nesprávným uživatelům. Vzhledem k rozsáhlosti tématu a zaměření práce je pozornost věnována datové kvalitě, integraci a konsolidaci.

Cílem práce je vymezit oblast datové kvality, způsob jejího měření a vyhodnocování, integrity a konsolidace dat. Dále na konkrétním příkladu ilustrovat poznatky získané v první části práce. Práce je strukturována do dvou hlavních částí: teoretické (kapitoly 2-7) a praktické (kapitola 8).

Druhá kapitola se věnuje vymezení termínu Business Intelligence. Popsány jsou jednotlivé prvky architektury BI systémů a jejich posloupnost při zpracování produkčních dat přes prostředky BI po výsledný výstup.

Třetí kapitola se zabývá daty a metadaty, tj. daty o datech. Mimo vymezení potřebné terminologie je v kapitole uvedena souvislost mezi daty, informacemi a znalostmi. U metadat je uveden návrh metadata repository, klíčového prvku architektury IT/ICT společnosti.

Ve čtvrté kapitole je zpracováno téma datové kvality. Po vymezení termínu kvalita dat a aspektů s problematikou spojených se práce zabývá řízením kvality dat. Metriky kvality dat udávají rámec kontroly a hodnocení kvality dat. Na konci kapitoly jsou uvedeny

možné zdroje nekvalitních dat a neméně důležitá podkapitola věnující se nákladům spojených s nekvalitními daty.

V páté kapitole navazuje na téma datové kvality integrita dat. V této kapitole jsou zpracovány jednotlivé stupně datové integrity: entitní, doménová a relační. Na konci kapitoly jsou uvedena možné porušení integrity.

Šestá kapitola rozšiřuje v rámci konsolidace dat v BI systémech téma ETL transformací popsané v druhé kapitole. Jednotlivé fáze transformačního procesu jsou podrobněji zpracovány.

Závěrečná, sedmá, kapitola teoretické části přináší přehled nejdůležitějších norem týkajících se datové kvality a to konkrétně ISO 8000 (datová kvalita) a ISO/IEC 25012:2008 definující model datové kvality.

V praktické části práce, shrnuté do kapitoly 8, je zpracován návrh aplikace BI pro farmaceutickou společnost. Aplikace je navržena jako konsolidující a integrující prvek IT/ICT společnosti. V úvodu kapitoly je uvedena vize řešeného problému. Následuje úvodní studie definující cíle projektu, očekávané přínosy a efekty projektu, analýzu požadavků a možnosti řešení. Navazuje vlastní návrh aplikace. Nejprve je navrhnuto metadata repository, jehož funkce je v první řadě určena pro potřeby BI aplikace. Do budoucnosti se počítá s rozšířením a začleněním po podnikové IT/ICT infrastruktury. V dalším kroku je uveden návrh vlastního datového skladu a ETL transformace. V rámci ETL transformace jsou popsány jednotlivé transformační kroky a aplikace pravidel pro kontroly kvality dat. Jako poslední krok návrhu jsou navržena řešení na kontrolu a opravy dat již uložených v datovém skladu.

# 2 Business Intelligence

První použití termínu Business Intelligence (BI) je přisuzováno **H. P. Luhnovi**[1] v článku "A Business Intelligence System" publikovaném v říjnovém vydání IBM Žurnálu roku 1958. Termín byl dále zpřesněn **Howardem J. Dresnerem**, analytikem společnosti Gartner Group v roce 1989 a postupně nahradil do té doby běžně používaný termín EIS[2] (Executive Information System). Co se skrývá pod Business Intelligence napoví definice České společnosti pro systémovou integraci:

*„Business intelligence (BI) je sada procesů, aplikací a technologií, jejichž cílem je účinně a účelně podporovat řídící aktivity ve firmě. Podporují analytické a plánovací činnosti organizací a jsou postaveny na principech multidimenzionálních pohledů na podniková data. Aplikace BI pokrývají analytické a plánovací funkce většiny oblastí podnikového řízení, tj. prodeje, nákupu, marketingu, finančního řízení, controllingu, majetku a řízení lidských zdrojů, výroby."* [3]

Úlohou BI je tedy nabídnout analytikům, popř. managementu, nástroje umožňující s minimálním technickou náročností a v co nejkratším čase získat podstatné a správné informace pro rozhodování v rámci strategických aktivit firmy. Důraz je dále kladen na možnost rychle a pružně měnit kritéria pro získání analýz.

Obrázek 1 zjednodušeně ilustruje tok dat v rámci firmy využívající nástroje BI, kdy BI vystupuje jako také jako integrující a konsolidující součást IS/ICT firmy pro analytické účely.

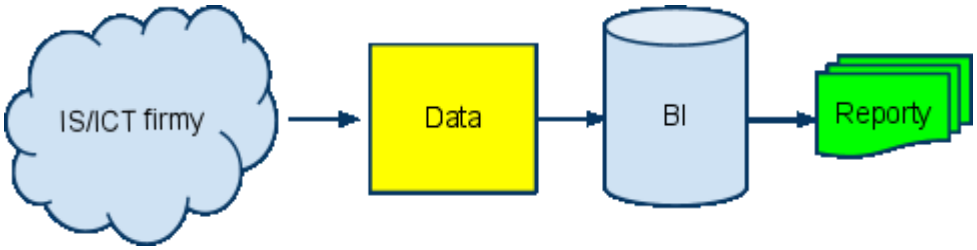
---

[1] H. P. Luhn (1896-1964) byl počítačovým expertem ve společnosti IBM. Luhnův algoritmus je dodnes používán např. pro ověření čísel platebních karet a KWIC (Key Words in Context).

[2] V současné době používáno pro aplikace navržené pro poskytování manažerských informací čerpajících data z BI.

[3] Gála, L., Pour, J., Toman, P. (2006) *Podniková informatika*. 1. vyd. Praha : Grada, str. 90

Obrázek 1 Schéma toku dat v rámci firmy využívající nástroje BI

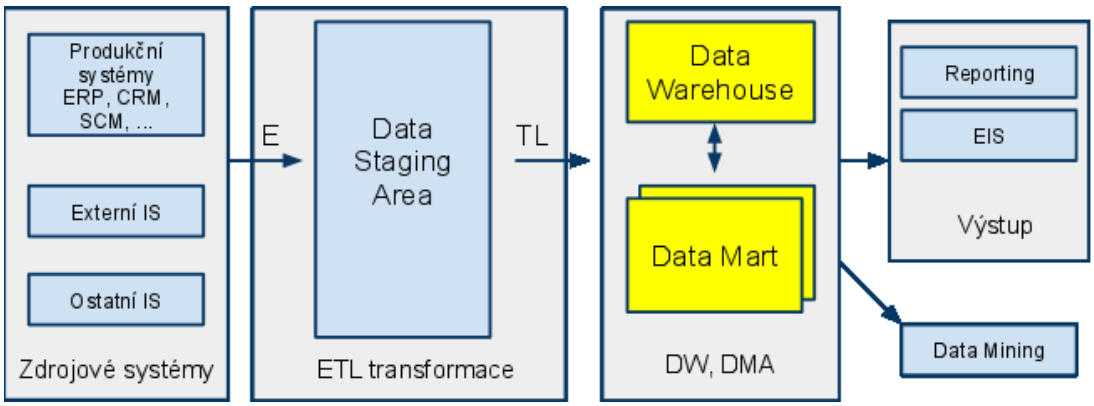



Zdroj: vlastní zpracování

## 2.1 Struktura BI

Na trhu existuje velké množství řešení BI s rozdílným zapojením jednotlivých technologií a postupů. Na následujícím obrázku je znázorněno obecné schéma zpracování dat v systému BI od získání zdrojových dat, přes transformaci a uložení dat v datovém skladu, až po výstup.

Obrázek 2 Schéma BI s datovým skladem




Zdroj: vlastní zpracování

### 2.1.1 Zdrojové systémy

Systémy poskytující zdrojová (primární, produkční) data BI jsou založeny na transakčním zpracování a označují se **OLTP** (OnLine Transaction Processing). Data těchto systémů se ukládají převážně do **relačních databází**. Struktura dat je optimalizována pro běžné operace související s běžným životem firmy: **vkládání**, **úpravy** a **mazání** jednotlivých datových záznamů.

Příklady zdrojových systémů:

- **Enterprise Resource Planning** (ERP) – komplexní systémy zastřešující oblasti výroby, financí a účetnictví;
- **Customer Relationship Management** (CRM) – řízení vztahů se zákazníky;
- **externí informační systémy**.

### 2.1.2 ETL Transformace

V rámci ETL[4] (Extraction Transform Load) procesu se data transformují do podoby vhodné pro datový sklad s použitím dočasného úložiště (Data Staging Area). Dočasné datové úložiště se plní v první fázi - extrakci dat. Tato data se následně transformují a nahrají do datového skladu (viz. kapitola 6. Konsolidace dat).

### 2.1.3 Data Warehouse, Data Marts

Základním stavebním kamenem BI aplikací se zpravidla stává **datový sklad** (Data Warehouse, DWH, DW). Struktura datového skladu je optimalizována pro dotazování, nikoliv pro vkládání a aktualizaci dat, jako je tomu v OLTP aplikacích. Tyto aplikace pak bývají označovány OLAP (OnLine Analytical Processing).

[4] 80-90 procent úsilí při budování datového skladu je věnováno implementaci rozhraní mezi operativní databází a datovým skladem. (English 2004, str: 4)

Analytické systémy jsou charakterizovány následovně[5]:

- informace jsou poskytovány na základě informací z primárních systémů;
- data jsou uložena multidimenzionálně;
- obsahují různé úrovně agregace dat, podle hierarchické struktury dimenzí;
- zachycují faktor času a umožňují realizovat časové srovnání.

Data Marts (datová tržiště) jsou založena na podobném principu jako datové sklady, s tím rozdílem, že jsou určena pro omezený okruh uživatelů. Na složení datového skladu a datových tržišť existují dva různé pohledy. **R. Kimball** přistupuje k tvorbě datového skladu přes tvorbu jednotlivých datových tržišť, která jsou postupně budována podle potřeb společnosti. Jejich logickým sjednocením je datový sklad. Oproti tomu **B. Immon** upřednostňuje tvorbu centrálního datového skladu s ohledem na minimalizaci redundantních dat. Nad tímto skladem jsou budována jednotlivá datová tržiště. [6]

### 2.1.4 Výstup

Pro potřeby získání požadovaného výstupu jsou možné například následující alternativy:

- **Reporting** – přehledy získané dotazy do datových skladů (tržišť)
- **Executive Information System** (EIS) – aplikace navržené pro poskytování manažerských informací, které čerpají data z multidimenzionálních systémů

### 2.1.5 Dolování dat

Pod pojmem dolování dat (DM, Data Mining) si lze představit získávání strategických informací pomocí speciálních algoritmů. Slouží především k objevování nových skutečností a korelace mezi ekonomickými veličinami.

---

[5] Novotný, O, Pour, J., Slánský, D. (2004) *Business Intelligence. Jak využít bohatství ve vašich datech.* Praha: Grada, str: 21.

[6] Gála, L., Pour, J., Toman, P. (2006) *Podniková informatika*. 1. vyd.. Praha : Grada, str: 90.

Matematické a statistické metody používané pro dolování dat.[7]

- **rozhodovací stromy** – data zobrazena v podobě stromu, kde každý uzel vymezuje kritérium pro následné rozdělení dat;
- **neuronové sítě** – napodobení chování lidského mozku pomocí algoritmů pro nacházení podobností a vzorů pro tvorbu prediktivních modelů;
- **genetické algoritmy** – simulace biologické evoluce pro optimalizaci DW;
- **clustering a klasifikace** – rozdělení dat do skupin s podobnými vlastnostmi a jejich klasifikace.

V souvislosti s dolováním dat se uvádí termín **Knowledge Discovery in Databases**, kdy data mining bývá chápán jako jeho analytická část.

## 2.2 Důvody neúspěšnosti projektů BI

Úspěch realizace projektu BI je klíčový pro další rozvoj firmy. I úspěšně realizovaný systém poskytující chybná, zastaralá nebo neúplná data může mít fatální důsledky. Za úspěšné lze obecně považovat ty projekty, které splňují následující podmínky:

- projekt je dodán v požadovaném čase;
- nejsou překročeny vyčleněné zdroje – lidské a finanční;
- projekt je vnímán koncovými uživateli jako úspěšný.

S ohledem na provoz systému jsou podmínky úspěšného systému nastaveny následovně[8] :

- systém je provozován bez chyb a výpadků;
- řešení uspokojuje požadavky koncových uživatelů tak jak byly definovány na začátku projektu, popřípadě v průběhu zpracování změnových požadavků;

[7] Novotný, O, Pour, J., Slánský, D. (2004) *Business Intelligence: Jak využít bohatství ve vašich datech*. Praha: Grada, str: 21.

[8] Pour, J., Slánský D. (2010) *Efekty a rizika BI.*. CSSI.

– ekonomický efekt je stejný jako byl očekáván při definici projektu nebo vyšší.

Vybrané faktory které mají negativní dopad na kvalitu a úspěch BI:

– neustále se měnící, popř. protichůdné požadavky na funkčnost aplikace BI;
– nedostatečné řízení kvality dat;
– neexistence vize a strategie BI vedoucí k chaotickému řešení;
– nedostatečná motivace zainteresovaných osob;
– nedostatečná identifikace výsledných efektů.

Faktory s ohledem na data:

– nekonzistence dat;
– chybná architektura dat;
– neschopnost integrovat data z různých zdrojů;

Následující graf jednoznačně identifikuje nízkou kvalitu dat jako hlavní pojmenovaný faktor neúspěšnosti datových skladů.

Graf 1 Hlavních pět faktorů způsobujících neúspěch aplikací BI


Zdroj: McKendrick, J. (2008) Open Source Business Intelligence: A 2008 Progress Report., str. 2. (vlastní zpracování)

# 3 Data a metadata

## 3.1 Data

Obecně lze data definovat jako určitou posloupnost (řetězec) znaků. Obvykle reprezentují vlastnosti (atributy) objektů reálného světa a to ve formě vhodné k počítačovému zpracování. Jako příklad uvedeme řetězec 777456. Bez doplňujících informací nejsme schopni řetězec interpretovat, přiřadit k reálným objektům a vlastnostem, zároveň nelze určit, zda má správný formát, popř. neobsahuje chyby.

Data jsou „**surovinou**“[9] produkovanou a spotřebovávanou obchodními procesy s výhodou prakticky neomezené znovu použitelnosti. Všechny ostatní zdroje mají omezené použití (peníze, materiál, atd.).

## 3.2 Informace

Informace představují data zasazená do kontextu (použitelná data). Informace bývá definována jako funkce následujících komponent.

**Informace = fce (data, definice, prezentace)**[10]

Kvalita informace je měřena kvalitou těchto komponent, kdy každá rozhoduje o případném znehodnocení informace. Jasná a přesná **definice**[11] (data definition) je nutná k vymezení a jednotnému porozumění obsahu dat (lze použít příměr s výkladovým slovníkem - bez srozumitelného popisu slov nelze porozumět obsahu slov). Důležitou součástí je také dostupnost a prezentace[12] (presentation) dat příjemci informace.

---

[9] V literatuře lze s roztoucím významem dat nalézt dokonce snahy o zařazení dat mezi aktiva společnosti.

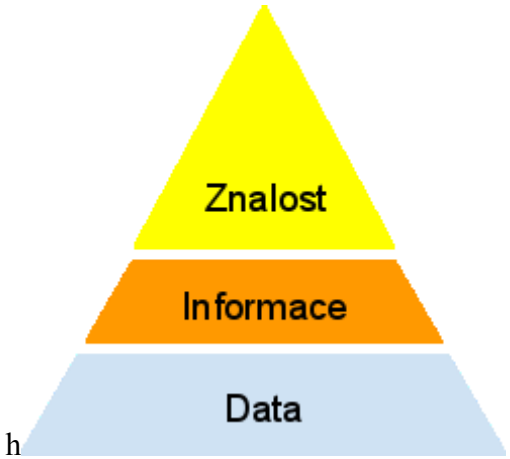
[10] English, L. (2004). *Improving Data Warehouse and Business Information Quality*. New York Wiley, str. 19

[11] Synonymem k pojmu **data definition** jsou **metadata**.

[12] V literatuře též označována jako interpretace.

Hodnotu získává informace až porozuměním, a tím vzniká **znalost**[13] (tj. porozumění obsahu informace). Na následujícím obrázku je ilustrována vztah mezi daty, informacemi a znalostmi. Data jsou základem pyramidy a jejich případná nekvalita ovlivňuje i horní stupně.

Obrázek 3 Vztah mezi daty, informacemi a znalostmi




h

Zdroj: vlastní zpracování

## 3.3 Metadata

Metadata je pojem, který lze definovat jako data o datech. Metadata pomáhají interpretovat význam popisovaných dat. Mimo informace o formátu dat, významu, původu obsahují také informace o relacích mezi daty. Metadata lze rozdělit následovně:

- struktura DW – DB model, reporting;
- operační metadata;
- metadata k ETL – metody použité při ETL, omezení, algoritmy;
- metadata pro koncové uživatele – informace o datech uvnitř DW.

[13] Definice vztahů mezi daty, informacemi a znalostmi je mimo jiné doménou **znalostního managementu** (Knowledge management). D. Tobin rozšířil model o další stupeň, a to **moudrost**.

Jedna z možných definic metadat zní následovně:

*„Metadata jsou data popisující jiná data. Jejich prostřednictvím se lze na data dotázat, jsme schopni data doplňovat, konsolidovat je a vzájemně je synchronizovat.“* [14]

Metadata velmi úzce souvisí s problémem datové kvality. Měření datové kvality probíhá formou porovnávání reality s očekávaným stavem, který je popsán právě metadaty. Ta v sobě obsahují popisy pravidel, kterými se řídí business procesy a v transformované formě i toky dat. Aplikací těchto pravidel je možné rozlišit vyhovující, popř. nevyhovující data. [15]

### 3.3.1 Metadata repository

Metadata repository, někdy též **data dictionary**, je centrální úložiště metadat, tj. data o datech. Metadata repository umožňuje následující operace s metadaty:

- vložení nových metadat;
- úprava stávajících metadat;
- dotazování nad metadaty.

Při vkládání a úpravě metadat je důležitá role **datového stevarda** (data steward), který nese odpovědnost za kvalitu metadat. Upravená, popř. nová metadata datový steward schvaluje, opravuje, popř. vrací odpovědnému pracovníkovi před vlastním nahráním do metadata repository.

Architekturu metadata repository je vhodné navrhovat s ohledem na možná budoucí rozšíření a úpravy. Flexibilní architektura minimalizuje možné budoucí náklady na úpravy spojené s možným rozšířením metadata repository.

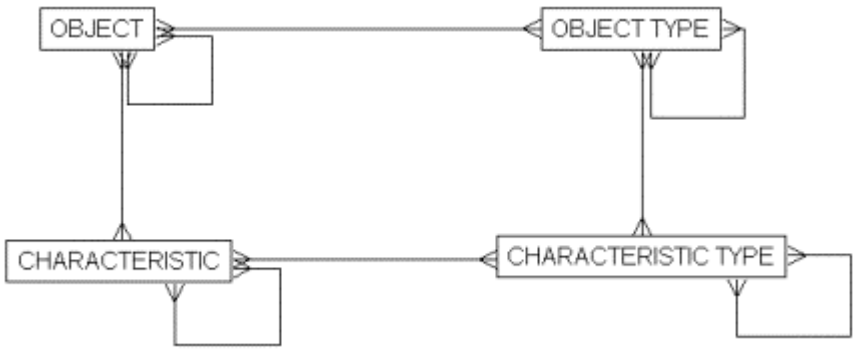
Vhodný základ architektury metadata repository je znázorněn na následujícím diagramu.

---

[14] Gála, L., Pour, J., Šedivá, Z. (2006). *Podniková informatika.* Praha: Grada. str. 62.

[15] Pavlis, D. (2010). *Pozice a úloha datového skladu v rámci informační strategie firmy*. CSSI

Obrázek 4 Abstraktní model metadata repository




Zdroj: Hoberman, S. (2002) Use Of Meta Data Entities : Part 3 Of A Series On Abstraction.

Entita **objekt** (Object) obsahuje základní informace o metadata objektu. Mezi objekty je možná rekurzivní relace. Příklady:

- model objednávky;
- zákazník;
- příjmení zákazníka.

Entita **objektového typu** (Object Type) kategorizuje metadata. Mezi objekty je možná rekurzivní relace. Příklady:

- datový model;
- relace;
- entita;
- datový element.

Entita **charakteristika** (Charakteristic) obsahuje všechny popisné informace metadata. Příklady:

- definice;
- popis;

- možnost nabývat nulového hodnoty;
- datový typ;
- omezení datového typu.

Entita **typ charakteristiky** (Charakteristic) obsahuje popisné informace charakteristiky metadata. Příklady:

- definice;
- verze.

### 3.3.2 CWM

Jedním z používaných standardů pro modelování metadat v oblasti datových skladů je CWM (Common Warehouse Metamodel) od organizace OMG[16]. Hlavním cílem tohoto standardu je snadná výměna metadat mezi nástroji na modelování datových skladů a metadata repositáři v rozdílných heterogenních prostředích. V CWM lze nalézt následující specifikace pro:

- modelování metadat relačních i nerelačních systémů;
- modelování metadat ETL transformací;
- modelování metadat většiny ostatních součástí datového skladu.

Instance metamodelu jsou vyměňovány prostřednictvím dokumentů XMI (XML Metadata Interchange).

[16] OMG (Object management Group) je mezinárodní nezisková organizace zabývající se standardizací v oblasti počítačového průmyslu.

# 4 Datová kvalita (Data Quality)

Datová kvalita vypovídá o tom, v jaké míře data správně zobrazují objekty reálného světa. Nesprávná data po zpracování vedou k zavádějícím výsledkům. Mimo **objektivní parametry** (validní a neprázdné hodnoty) kvalita dat v rámci organizace obsahuje i **subjektivní rozměr**[17], který odráží potřeby a zkušenosti zainteresované osoby. V rámci organizace je proto nutné vytvořit přijatelný a obecně přijímaný model hodnocení datové kvality, který objektivně zhodnotí kvalitu dat.[18]

Jak bylo zmíněno výše, informace je mimo jiné funkcí dat. Z toho vyplývá že kvalita informace je úměrná kvalitě dat[19]. Nekvalitní informace může mít fatální důsledky. Např. management firmy dostane informaci, že na Plzeňsku je segment trhu pro její zboží v určitém rozsahu a pokud je tato informace založena na mylných datech, může dojít k naddimenzování / poddimenzování plánovaných prodejů.

Kvalita dat není izolovanou oblastí, ale integrální součástí správy dat v rámci organizace. Cena nízké kvality dat je vysoká. Málo kvalitní data narušují správný průběh firemních procesů. Náprava chyb v datech stojí dodatečné lidské zdroje: peníze, materiál, popř. další podnikové dodatečné zdroje.

**Kvalitní data jsou limitujícím faktorem BI aplikací**, kdy tyto systémy jsou vytvářeny za účelem poskytování informací pro strategické rozhodování s dopadem na celou organizaci[20].

**Informační kvalita** je definována pomocí následujících kritérií:

- **kvalita definice dat** (Data Definition Quality) – všichni uživatelé musí chápat a používat získaná data stejným způsobem;
- **kvalita hodnot dat** (Data Content Quality) – jak data odpovídají vlastnostem reálných objektů;

---

[17] Subjektivní vnímání kvality dat lze měřit podle dotazníků, popř. průzkumy mezi uživateli.

[18] Na kvalitu dat je možné také pohlížet způsobem, do jaké míry odpovídají obchodním požadavkům.

[19] Paralelu se surovinami v průmyslu. Nekvalitní vstup snižuje kvalitu výstupního produktu.

[20] Pavlis, D. (2010). *Pozice a úloha datového skladu v rámci informační strategie firmy.* CSSI

- **kvalita prezentace dat** (Data Presentation Quality) – jsou data prezentována uživatelům správným a intuitivním způsobem?

## 4.1 Řízení kvality dat (Data Quality Management)

Řízení kvality dat vyžaduje definici **rolí, odpovědností, politik a procedur** týkajících se získání, údržby a rozmístění dat. K úspěchu při realizaci řízení datové kvality je nutná kooperace mezi obchodním a technologickým (IT) oddělením. Obchodní oddělení je odpovědné za definici obchodních pravidel a ověření kvality dat. IT oddělení je pak odpovědné za realizaci a řízení realizace.

Přístupy:

- **proaktivní** – definice rolí a zodpovědností, dodržování obchodních pravidel;
- **reaktivní** – řešení existujících chyb v datech.

Role a odpovědnosti:

- **projektový manažer** (Project Manager) – řídí projekt, rozpočtování;
- **business analytik** (Business Analyst) – zpracovává požadavky na kvalitu a ty přenáší do návrhu aplikace, které jsou pak realizovány vývojovým týmem;
- **datový stevard** (Data Steward) – zodpovědný za správu dat.

### 4.1.1 Pilíře řízení kvality dat

Mezi pilíře řízení kvality dat lze zařadit následující činnosti:[21] [22]

- **profilování dat** (Data Profiling) – sběr statistických informací k určení informací o stávající kvalitě dat. Odrazový můstek pro další aktivity.[23] Zjištění zda data jsou úplná a přesná;

[21] Geiger J. G. (2010) *Data Quality Management.*

[22] Moss, L. T., Abai, M., Adelman, S. (2005) *How to improve data quality.* Informit.

[23] Při budování datového skladu většinou tato činnost nazývána **Source System Analysis**.

- **zkvalitnění dat** (Data Cleansing) – pokud je problém identifikován, je možné provést jednu z následujících operací: odstranění, přijmutí chybných dat pokud jsou v tolerovaném rozmezí, oprava dat nebo vložení defaultní hodnoty;
- **prevence vzniku chyb** (Data Defect Prevention);
- **integrace dat** (Data Integrity);
- **rozšíření dat** (Data Augmentation) – rozšíření dat o informace ze zdrojů třetích stran.

Následující tabulka zobrazuje jednotlivé varianty profilování dat.

Tabulka 1 Profilování dat

| Typ | Příklad |
|---|---|
| **Mimo akceptovaný rámec** | Věk zákazníka = 196 let |
| **Nestandardní data** | Americká tř., Americká třída, Americká T. |
| **Neplatné hodnoty** | Povolené hodnoty 1..10, Hodnota = 14 |
| **Rozdílná lokální pravidla** | 5/6/2010, 5.6.2010 |
| **Rozdílné formáty** | +420123456789, +420 123 456 789 |
| **Rozdílný vzhled** | petr nový, Petr Nový |
| **Verifikace** | Neodpovídající poštovní směrovací číslo |

Zdroj: Vlastní zpracování.

### 4.1.2 Přístupy k řízení kvality dat

V navazujících podkapitolách jsou uvedeny vybrané perspektivní přístupy k řízení datové kvality.

#### 4.1.2.1 Správa dat (Data Governance)

Pojem Data Governance je vnímán a vykládán různými způsoby. Pro použití v této publikaci byla vybrána následující definice:

*„Data Governance, je akční program pro řízení datové kvality. Jedná se o soubor organizačních a procesních opatření, který staví data a jejich kvalitu na významné místo v prioritách organizace. Aplikace programu Data Governance (byť postupná) je nezbytnou součástí každé vážně míněné snahy o konsolidaci dat, zakotvuje principy péče o data do struktur organizace a tím umožňuje trvalé zlepšování jejich stavu.“*[24]

Z výše uvedené definice vyplývá, že cílem Data Governance je zajistit důvěryhodná data a určit osoby zodpovědné za případné nedostatky. Důraz je kladen na postupnou evoluci procesů pracujících s daty v rámci organizace.

#### 4.1.2.2 Total Data Quality Management (TDQM)

Na MIT[25] vznikl program TDQM s cílem vytvořit teoretické základy v relativně nové oblasti kvality dat, které je možné prakticky využít v podnikání.

Proces se skládá ze tří základních částí:

- stanovení metrik a provedení subjektivního a objektivního hodnocení kvality dat;
- porovnání výsledků hodnocení, identifikace chyb a stanovení hlavních příčin vzniku odchylek;
- zjištění a realizace nezbytných opatření pro zlepšení.

#### 4.1.2.3 Optimální postup řízení datové kvality

V navazujícím přehledu jsou shrnuty a doplněny jednotlivé kroky, které vedou k rostoucí kvalitě dat ve společnosti. Je důležité podotknout, že se v žádném případě nejedná o časově omezenou činnost, ale o aktivity, které vyžadují kontinuální pozornost a zároveň respektování definovaných pravidel všemi zaměstnanci. Ideální je iterativní přístup, kdy v jednotlivých iteracích je pozornost věnována každému z bodů.

---

[24] Klaus M. (2007) Konsolidace databází a kvalita dat ve veřejné správě. Konference Internet ve státní správě a samospráv Hradec Králové: ISSS, str.116.

[25] Massachusetts Institute of Technology

1. Firemní strategie zajištění kvalitních dat.
2. Definice kritických míst vzniku potenciálních nekvalit dat.
3. Definice jednotlivých metrik a způsobu jejich měření.
4. Definice jednotlivých rolí a zodpovědností s ohledem na správu dat a metadat. Zajištění potřebné kvalifikace pro klíčové role, popř. najmutí externích zdrojů.
5. Analýza a hodnocení chyb a stanovení hlavních příčin.
6. Opatření k zamezení opakování nových chyb a korekce stávajících.

## 4.2 Metriky datové kvality

Základní moto řízení datové kvality lze označit následující větu:

*„Co nelze měřit, to nelze řídit.*“

Před vlastním měřením kvality dat je nutná specifikace metrik (charakteristik), které jsou očekávány uživateli. Ty se dělí na dvě skupiny:

### 4.2.1 Objektivní metriky

Objektivní metriky lze opětovně vypočítat z dat. Jedná se o charakteristiky kvality dat samotných bez závislosti na jejich interpretaci

- **úplnost** (Completeness) – do jaké míry jsou informace kompletní (počet neprázdných požadovaných atributů, použitelnost a vhodnost datových hodnot);
- **validita** (Validity) – stupeň v jakém data odpovídají doméně (množině hodnot, podrobněji zpracováno v kapitole věnované doménové integritě) a obchodním pravidlům (business rule). Příklad: hodnota pole PSČ odpovídá jednomu z údajů z množiny všech PSČ v rámci ČR;
- **přesnost** (Accuracy) – do jaké míry data korespondují s realitou (s autoritativním zdrojem informací, opět úzká souvislost s doménovou integritou).;
- **podrobnost** (Precision) – do jaké míry jsou data správně podrobná. Příklad: datový typ je celé číslo oproti požadovanému reálnému s dvěma místy za desetinou čárkou;

- **neduplicita** (Nonduplication) – vztah 1:1 mezi záznamy a objekty reálného světa. Souvisí s integritou dat.

### 4.2.2 Subjektivní metriky

Měří se pomocí dotazníků a průzkumů v pravidelných intervalech. [26] [27]

- **dostupnost** (Accessibility) – dostupnost dat na požádání, ve správný čas na správném místě.
- **použitelnost** (Usability) – data jsou použitelná a srozumitelná
- **správnost** (Rightness)
- **včasnost** (Timeliness)

## 4.3 Kontrola kvality dat

Vlastní kontrolu dat lze realizovat dvěmi způsoby:

- **v rámci návrhu a implementace řešení** – pozornost věnována i potenciálním problémovým místům;
- **analyticky** – analýza datových zdrojů a procesů (v optimálním případě se zapojením metadata repository).

Kontrola dat je realizována zpravidla na následujících místech:

- **produkční systémy** – uživatelská kontrola vstupních dat v rámci provozu aplikace;
- **návrh BI**;
- **vlastní aplikace BI** – statistické metody uvnitř DW;
- **ETL transformace**;

[26] Tobišek R. (2010) BI Kudy na řízení kvality dat. System online.

27 English, L. (2004). Improving Data Warehouse and Business Information Quality. New York Wiley, str. 19

## 4.4 Zdroje chyb

Data jsou do produkčních systémů vkládána uživateli (do BI se dostávají nepřímo přes ETL transformaci ). V následujícím přehledu jsou uvedeny možné příčiny vzniku chyb[28] „lidským faktorem“ seřazeny podle náročnosti jejich předcházení vzestupně.

- nedostupnost potřebných nástrojů;
- nedostatečná motivace;
- nedostatečná kvalifikace.

Důležitým místem kontroly dat a prevence chyb vzniklých lidským faktorem jsou vstupní formuláře produkčních systémů. Vhodným nastavením validačních kritérií je možné zabránit vstupu nekvalitních dat. Tato oblast bývá také často označována jako **Data Quality Firewall**.

Jako další zdroj chyb lze uvést vlastní strukturu IT řešení. IT řešení ve velkých společnostech bývá zpravidla sestavena z několika heterogenních systémů od různých dodavatelů. I když v rámci jednotlivých subsystémů mohou být kvalitní data, přenos mezi jednotlivými systémy, popř. sjednocení dat do jednotlivých entit může být problematické. Jako příklad lze uvést rozdílnost velikosti datových typů, ať už programovacích jazyků, tak i vlastních běhových prostředí.

## 4.5 Náklady spojené s datovou kvalitou

S otázkou datové kvality velmi úzce souvisí i náklady. Tyto náklady lze rozdělit na dvě obecné kategorie:

[28] Mezi tyto chyby lze zařadit překlepy nebo záměnu vstupních polí.

1. Náklady spojené se zvyšováním datové kvality
   - **proaktivní** – prevence vzniku chyb (např. definice validačních kritérií vstupních formulářů).
   - **reaktivní** – náklady na čištění systémů BI. Tyto náklady lze dále rozdělit na náklady týkající se detekce chyb a dále náklady na jejich opravu.
2. Náklady vyvolané nekvalitními daty[29]
   - **ztráta příjmů** (např. zákazníkům nejsou účtovány využívané služby vzhledem k chybám k aplikace);
   - **plýtvání penězi** (např. rozesílání nabídek nebo katalogů na neplatné popř. neúplné adresy);
   - **promarněná příležitost** – na základě špatných dat je učiněno rozhodnutí, které znamená nižší příjmy (než v případě, že by data byla dostatečně kvalitní);
   - **ztráta dobrého jména firmy** – nespokojení a frustrovaní zákazníci odejdou ke konkurenci.
   - **zásah do soukromí zákazníků** – nekvalitní data mohou způsobit škody zákazníkům (např. zákazníkovi nemusí vyhovět požadavkům pro poskytnutí úvěru kvůli nekvalitním datům);

Dále jsou uvedeny některé příklady týkající se nákladů spojených s kvalitou dat. Phillip Crosby odhadl náklady na zpracování nekvalitních dat mezi **15 a 20 procenty z příjmů**. Joseph M. Juranův odhad se pohybuje mezi **20 a 40 procenty z prodejů** zahrnující náklady spojené se stížnostmi zákazníků, opravou vad a vyřazením vadných výrobků.[30]

[29] Won, K., Byoungju, Ch. Towards (2003). *Quantifying Data Quality Costs.Journal of Object Technology.*

[30] English, L. (2004). *Improving Data Warehouse and Business Information Quality.* New York Wiley, str. 19

The Datawarehouse Institute odvodil, že nekvalitní data stojí ročně kolem **600 milionů dolarů.** Dále poukázal na další znepokojivý fakt, a to, že mezi **15-20 procenty dat v rámci organizace je nepoužitelných**. [31]

[31] Geiger J. G. (2010) . *Data Quality Management.*

# 5 Integrita dat

Integritou dat rozumíme celistvost, úplnost a spolehlivost dat, kdy musí existovat úzká shoda mezi uloženými daty a objekty reálného světa. Zároveň je nutné zabránit neautorizovaným změnám dat po celou dobu jejich života (vznik, přenos, uložení a čtení). Integrita dat je v podstatě o důvěře v naše data. [32]

## 5.1 Entitní integrita

Entitní integrita je součástí metriky datové kvality - **neduplicita**. Základním pravidlem je existence **unikátního** a **nenulového** (lépe, nenabývá hodnoty NULL) **primárního klíče**. Každá entita je jednoznačně identifikována[33].

Omezení definované entitní integritou může být definováno jedním nebo složeným primárním klíčem.

## 5.2 Referenční integrita

Referenční integrita pomáhá udržovat vazby mezi tabulkami v relačních databázích. Zpravidla je definována cizím klíčem mezi dvěmi tabulkami, popř. pouze nad jednou u rekurzivních struktur.

*Příklad: Pokud má tabulka A definovaný cizí klíč (constraint) na atribut tabulky B (stejného datového typu), není možné přidat do tabulky A záznam, ve kterém je neexistující hodnota cizího klíče atributu z tabulky B.*

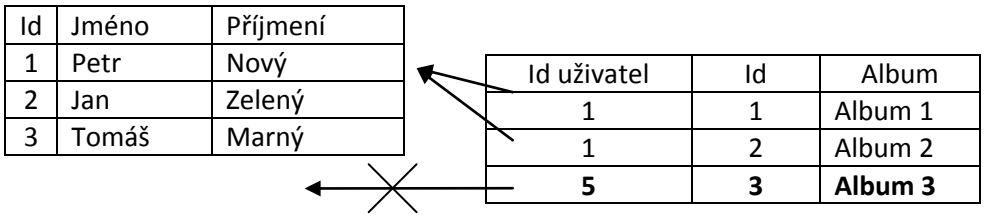
Při mazání záznamů z tabulky, která se je referencována cizím klíčem z jiné tabulky, není možné mazat záznamy na které existuje cizí klíč z jiné tabulky. Odstranění záznamů je nutné provádět tzv. kaskádovým mazáním, kdy jsou nejprve odstraněny

[32] Na integritu dat lze pohlížet nejenom z logického hlediska, ale i z fyzického, kdy je potřeba data zabezpečit např. před přírodními katastrofami.

[33] V relačních databázích bývá pravidlem při operacích INSERT nebo UPDATE kontrola unikátnosti a formátu primárního klíče.

záznamy s cizím klíčem z jedné tabulky, teprve potom je možné bez porušení referenční integrity (a vyvolání chyby) smazat záznam v referencované tabulce.

Obrázek 5 Příklad referenční integrity (vazba mezi autorem a alby)




Zdroj: Vlastní zpracování

## 5.3 Doménová integrita

Doménová integrita znamená, že hodnota atributu leží v daném oboru hodnot povolených pro daný atribut. Po každý atribut (sloupec) lze specifikovat následující doménové informace:[34]

- **datový typ** – mezi základní datové typy se řadí: integer (celá čísla), decimal (čísla s desetinou čárkou) a character (znak nebo řetězec znaků);
- **délka** – množství číslic nebo znaků povolených pro datový typ;
- **datový formát** – formát datového typu (např. formát datumu dd/mm/yyyy nebo dd.mm.yyyy [35]);
- **omezení** (constraints) – omezení na určité hodnoty;
- **prázdná** (NULL) **hodnota atributu** – zda může atribut nabývat nulových hodnot;

[34] DatabaseDev. (2010) *Domain Level Integrity.*

[35] **dd** - den v roce včetně 0 na začátku, pokud se jedná o jedociferné číslo , **mm** - měsíc v roce včetně 0 na začátku, pokud se jedná o jedociferné číslo, **yyyy** - rok

– **defaultní hodnota** – hodnota, která se použije při vynechání nastavení hodnoty při vytváření (popř. úpravě) entity.

## 5.4 Možné porušení integrity

Porucha integrity může nastat z následujících příčin:

– neautorizovaná modifikace dat;
– konkurenční přístup více vláken. Řeší se použitím synchronizačních primitiv; [36]
– aktivní útok na porušení integrity dat;
– špatné ošetření vstupních dat;
– chybný návrh databáze;
– kolaps databázového (aplikačního) serveru.

[36] Synchronizační primitiva jsou prostředky na ošetření konkurečního přístupu ke shodnému zdroji; většinou jsou zabezpečena na úrovni operačního systému.

# 6 Konsolidace dat

Konsolidace dat je termín používaný pro definic procesu spojení velkého množství dat z různých heterogenních zdrojů na jedno místo do požadované struktury. Tím lze zajistit tvorbu výstupu, které je v produkčních systémech obtížné zajistit. V systémech BI je datová konsolidace realizována pomocí ETL transformace, kdy jsou data:

- vyčištěna od duplicit a nekonzistencí;
- doplněna o informace ze zdrojů třetích stran;
- agregována.

## 6.1 ETL

ETL proces se skládá ze tří částí: extrakce (extraction), transformace (transormation) a nahrání dat (load). Celý ETL process je spravován s využitím metadata repository.

### 6.1.1 Extraction

V rámci extrakce jsou získávána data z rozdílných produkčních zdrojů, ať už se jedná o relační nebo objektové databáze, datové soubory (xml, csv, xls) nebo načítání dat z externích systémů. Často se jedná o nejnáročnější část ETL transformace. Zdrojových systémů může být velká řada a navíc nemusí být vůbec nebo špatně zdokumentovány. U zdrojových systémů zpravidla nejsme schopni ovlivnit dostupnost a výkon pro potřeby BI aplikace. V první řadě je nutné nalézt odpovědi na následující otázky: [37]

- Jakou metodu vybrat pro extrakci?
- Jakým způsobem poskytnout data k dalšímu zpracování?

Extrakci lze provádět:

[37] Oracle (2011) Oracle9i Data Warehousing Guide. *(kapitola 11. Extraction in Data Warehouses)*

- **inkrementálně** – ve specifikovaný okamžik jsou získána data, která se změnila od poslední extrakce;
- **celkovou extrakcí** – v daný okamžik jsou získána všechna data ze zdroje.

Extrakci lze dále provádět:

- **online** – data jsou získávána přímo ze zdrojové aplikace;
- **offline** – data jsou získávána z exportních souborů (databázové dumpy, logy, datové soubory).

### 6.1.2 Transformace

Během ETL transformace se nahraná zdrojová data transformují do podoby odpovídající struktuře datového skladu, tj. jsou připraveny dimenze a tabulky faktů. Pro jednotlivé záznamy jsou vyhodnoceny metriky datové kvality. Nevyhovující data jsou potom dle zvolené strategie buď automaticky odmítnuta nebo postoupena pověřenému uživateli k opravě. V rámci transformace je nutné vhodným způsobem definovat:

- transformační pravidla – jaký atribut z daného zdroje odpovídá atributu požadovaného formátu;
- průběh transformace – posloupnost jednotlivých kroků transformace.

### 6.1.3 Nahrání dat

Finální nahrání transformovaných dat do aplikace BI a to většinou přímo do tabulky faktů a dimenzí datového skladu. Nahrávaná data mohou být:

- **vložena** – neexistující záznamy v BI;
- **aktualizována** – stávající data jsou přehrána novější verzí.

# 7 Standardy a normy

## 7.1 ISO 8000, Datová kvalita

Jednotlivé části specifikace jsou vyvíjeny a postupně publikovány technickým výborem TC 184 (Automation systems and integration), podvýborem SC 4 (Industrial data). Specifikace popisuje standardizovanou výměnu dat.

Tabulka 2 Části specifikace ISO[38] 8000, Datová kvalita

| Specifikace | Stav | Popis |
|---|---|---|
| **ISO/DTS 8000-1** | Ve vývoji | Datová kvalita – Část 1: Přehled<br>(Data quality - Part 1: Overview) |
| **ISO/DTS 8000-2** | Ve vývoji | Datová kvalita – Část 2: Klíčová data: Výměna charakteristických dat: Slovník<br>(Data quality -- Part 2: Master data: Exchange of characteristic data: Vocabulary) |
| **ISO/TS 8000-100:2009** | Publikováno | Datová kvalita – Část 100: Klíčová data: Přehled<br>Data quality -- Part 100: Master data: Overview |
| **ISO 8000-102:2009** | Publikováno | Datová kvalita – Část 102: Klíčová data: Výměna charakteristických dat: Slovník<br>Data quality -- Part 102: Master data: Exchange of characteristic data: Vocabulary |
| **ISO 8000-110:2009** | Publikováno | Datová kvalita – Část 2: Klíčová data: Výměna charakteristických dat: Syntaxe, semantika a schoda se specifikací<br>Data quality -- Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification |
| **SO/TS 8000-120:2009** | Publikováno | Datová kvalita – Část 2: Klíčová data: Výměna charakteristických dat: Původ<br>Data quality -- Part 120: Master data: Exchange of characteristic data: Provenance |
| **ISO/TS 8000-130:2009** | Publikováno | Datová kvalita – Část 2: Klíčová data: Výměna charakteristických dat: Přesnost<br>Data quality -- Part 130: Master data: Exchange of characteristic data: Accuracy |
| **ISO/TS 8000-140:2009** | Publikováno | Datová kvalita – Část 2: Klíčová data: Výměna charakteristických dat: Kompletnost<br>Data quality -- Part 140: Master data: Exchange |

[38] International Organisation of Standardisation

| | | |
|---|---|---|
| | | of characteristic data: Completeness |
| **ISO/CD 8000-150** | Ve vývoji | Datová kvalita – Část 2: Klíčová data: Rámec řízení kvality<br>Data quality -- Part 150: Master data: Quality management framework |

Zdroj:
http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=54158

## 7.2 ISO/IEC 25012:2008

Norma byla publikována spojeným technickým výborem JTC 1. (Joint Technical Comittee) podvýborem SC 7 (Software and systems engineering).

Norma ISO/IEC 25012:2008 definuje obecný model kvality dat uchovávaných ve strukturované podobě v počítačových systémech. Normu lze použít ke stanovení požadavků na kvalitu dat, plánování a vyhodnocování.

Příklady použití:

- defince a vyhodnocování kvality dat produkčních systémů;
- identifikace kritérií datové kvality použitelné pro hodnocení a zlepšování kvality dat;
- vyhodnocení souladu dat s právními předpisy

ISO/IEC 25012:2008 rozděluje atributy kvality dat do patnácti charakteristik rozdělných na dva typy:

- vnitřní
- závislé na systému

Tabulka 3 Atributy datové kvality podle ISO/IEC 25012:2008

| Atribut | Popis |
|---|---|
| **Vnitřní** | |
| **Přesnost** (Accuracy) | Do jaké míry data korespondují s realitou. |
| **Úplnost** (Completeness) | Do jaké míry jsou informace kompletní. |
| **Konzistence** (Consistency) | Do jaké míry jsou data konzistentní s ostatními daty v kontextu. |
| **Důvěryhodnost** (Credibility) | Do jaké míry jsou data důvěryhodná pro uživatele. |
| **Aktuálnost** (Currentness) | Do jaké míry jsou data aktuální. |
| **Vnitřní a závislé na systému** | |
| **Dostupnost** (Accessibility) | Dostupnost dat na požádání, ve správný čas na správném místě. |
| **Shoda** (Compliance) | Do jaké míry mají data vlastnosti, které se řídí standardy a předpisy v daném kontextu |
| **Důvěrnost** (Confidentiality) | Do jaké míry mohou přistupovat k datům jenom autorizovaní uživatelé. |
| **Efektivnost** (Efficiancy) | Do jaké míry mají data vhodné atributy pro zpracování a poskytnutí požadované úrovně výkonu. |
| **Přesnost** (Precistion) | Do jaké míry jsou data přesná. |
| **Sledovatelnost** (Traceability) | Do jaké míry jsou data dostupná k případným auditům a změnám. |
| **Srozumitelnost** (Understandability) | Do jaké míry je možné data číst a interpretovat uživateli. |
| **Závislé na systému** | |
| **Dostupnost** (Availability) | Do jaké míry jsou data dostupná autorizovaným uživatelům nebo aplikacím. |
| **Přenosnost** (Portability) | Do jaké míry je možné data přesunout nebo nahradit při zachování stávající datové kvality. |
| **Obnovitelnost** (Recoverability) | Do jaké míry je možné zachovat úroveň datové kvality při možných poruchách. |

Zdroj: Guerra-Garcí, C., Caballero, I., Piattini, M. Capturing Data Quality Requirements for Web Applications by means of DQ_WebRE.

# 8 Návrh BI aplikace a analýzy kvality dat

Tématem datových skladů a aplikací poznatků na vybranou farmaceutickou společnost se autor zabýval ve spolupráci s Bc. Jindřiškou Novotnou a Bc. Petrou Škorpovou v práci **Návrh multidimenzionálních modelů řízení pro vybranou firmu** (VŠEM, 2010). Byl navrhnut model vlastního datového skladu, jednoduchý ETL proces a reporty. Diplomová práce návrh datového skladu významně rozšiřuje o další prvky: zajištění datové kvality, integrity a konsolidace, dále pak o metadata repository a jeho provázání s BI aplikací. Autor uvádí poznatky získané z předchozí práce pouze v nezbytné míře, většinou v případech, kdy došlo k úpravě, opravě nebo rozšíření.

V následujícím textu je provedena analýza[39] aplikace BI pro vybranou farmaceutickou společnosti. Aplikace je navržena s ohledem na datovou kvalitu a integritu. Analýza je realizována nad daty mezinárodní farmaceutické společnosti, která má ve svém portfoliu přípravky na předpis. Prodeje jednotlivých léků jsou realizovány přes distributory se speciálním povolením.

V textu návrhu BI aplikace je nejprve uvedena vize projektu a dále následuje úvodní studie[40]. Další podkapitoly jsou věnovány analýze a modelování metadata repository, BI aplikace včetně ETL procesu.

Projekt vlastní implementace BI aplikace bude realizován s podporou metodiky Open UP, kdy metatada repository a datový sklad budou budovány iterativně, v jednotlivých fázích.

[39] Analáza je základní část vývoje jakéhokiliv softwaru. Výstupem je pak **plán vývoje,** někdy též nazývaný **technivký projekt**. Pro zdůraznění důležitosti bývá pouřívána paralela se stavební dokumentací.

[40] Novotný, O., Pour, J., Slánský, D. (2004) Business Intelligence : Jak využít bohatství ve vašich datech. str. 66-77

## 8.1 Vize

Řešeným problémem je rostoucí počet zpracovávaných souborů ve formátu .dbf, .xls a csv., které obsahují záznamy jednotlivých prodejů. V rámci projektu je nutné nalézt řešení umožňující konsolidaci dat z heterogenních zdrojů a zajistit jejich dostatečnou kvalitu a integritu.

Každý měsíc jsou zpracovávány údaje od jednotlivých prodejců, kteří dodávají zboží do 2600 lékáren. Jednotliví distributoři poskytují výstupní soubory v různých formátech: .xls, .dbf a nebo .csv s rozdílnými formáty datových záznamů. Hlavním problém je časová náročnost manuálního zpracování, kdy je nutná pečlivá kontrola všech dodaných souborů, dále data konsolidovat do jednotné formy k dalšímu zpracování. S rostoucí časovou náročností roste i pracnost, chybovost a nepřesnost dat. Náročnější se stává i kontrola. V případě chyby je nutné projít všechny záznamy a zrevidovat je.

Cílem projektu je vytvořit systém, který zjednoduší a sjednotí zpracování importních souborů. Jako vhodný systém byla vybrána BI aplikace založená na centrálním datovém skladu. Nad rámec BI aplikace je navrhnuto a realizováno metadata repository, jakožto centrální úložiště metadat a definována role datového stewarda, který má za správu metadat zodpovědnost.

## 8.2 Úvodní studie

Úvodní studie mapuje situaci v rámci společnosti a má zásadní vliv na výslednou kvalitu projektu, v tomto případě metadata repository a BI aplikace. Hlavním účelem je zmapovat prostředí uvnitř společnosti a návaznost navrhovaného řešení na ostatní součásti IT/ICT infrastruktury v rámci společnosti.

### 8.2.1 Cíle projektu

Cíle projektu jsou následující:

- vytvoření systému, , který zjednoduší a sjednotí zpracování importních souborů;
- vytvoření analytické BI aplikace;
- vytvoření metadata repository;
- definice rolí a zodpovědností za správu metadat;
- definice rolí a zodpovědností za datovou kvalitu;
- konsolidace analytických dat do jednoho systému;
- zrychlení zpracování dat;
- zajištění bezpečnosti dat;
- odstranění závislosti na zpracovateli dat;
- zajištění integrity dat a odstranění duplicit z výsledných reportů.

### 8.2.2 Očekávané přínosy a efekty projektu

Od projektu se očekávají následující přínosy:

- zvýšení produktivity práce, kdy lidské zdroje nebudou vázány manuálním zpracováním velkého počtu informací;
- optimalizace a redukce nákladů;
- v BI aplikaci uložena kvalitní data;
- vyšší kvalita dodávaných reportů;

- možnost operativně generovat dotazy pro konkrétní požadavky (ad-hoc dotazy)
- vyšší samostatnost managementu;
- reporty dodávané v předem očekávaných termínech a ne podle toho, kdy se povede zpracovat všechna data.

### 8.2.3 Katalog uživatelů

Vymezení uživatelů pro vstup do navazující analýzy a specifikace požadavků:

- **ředitel** (profesní orientace: management, počet: 1);
- **vedoucí marketingového oddělení** (profesní orientace: management, počet: 1);
- **vedoucí analytického oddělení** (profesní orientace: management, počet: 1);
- **pracovník analytického oddělení** (profesní orientace: ekonomika, počet: 1);

### 8.2.4 Analýza požadavků

Požadavky jsou rozmístěny u návrhů jednotlivých částí (metadata repository a BI aplikace, ve které jsou obsaženy i požadavky na ETL proces) navrhovaného systému. Jsou organizovány pomocí FURPS analýzy, kdy jsou požadavky členěny následovně:

- **funkcionalita** (functionality);
- **použitelnost** (usability) – návod, dokumentace a celková použitelnost;
- **spolehlivost** (reliability) – možnost obnovení systému, dostupnost;
- **výkon** (performance) – odezva systému;
- **rozšiřitelnost** (supportability) – možnost rozšíření.

### 8.2.5 Oblasti řešení a priority

Jak již bylo uvedeno, oblasti řešení jsou v zásadě tři. V následujícím seznamu jsou seřazeny podle priorit:

1. metadata repository;

2. datový sklad;
3. ETL transformace.

Další oblastí řešení je **datová kvalita,** která jde napříč zmíněnými oblastmi a je neméně důležitá.

### 8.2.6 Analýza možností řešení

Analýza možností řešení se zabývá alternativami, které jsou k dispozici při řešení navrhovaného systému. Možné alternativy jsou shrnuty v následujících podkapitolách.

#### 8.2.6.1 Existující řešení

Kompletní řešení realizováno jedním externím dodavatelem. V rámci systému jsou řešeny metadata repository, vlastní datový sklad a ETL transformace.

V rámci tohoto řešení je možné volit mezi komerčním systémem (Microsoft, Adastra) a open source (Pentaho). U open source je možné si zaplatit podporu a nasazení aplikace jako službu.

Výhody komerčního řešení:

- zkušenost s realizací podobných projektů;
- podpora při řešení problémů;
- profesionální přístup.

Nevýhody komerčního řešení:

- cena;
- nemusí být možné integrovat s ostatními systémy v rámci architektury;
- rozšíření systému s velkou pravděpodobností budou stát dodatečné náklady.

Výhody open source:

- cena;
- zdrojové kódy k dispozici;
- podpora v komunitě.

Nevýhody open source:

- podpora pouze v rámci komunity pokud není zaplacena od poskytovatele;
- většinou není snadné je zaintegrovat do firemní IT/ICT infrastruktury.

#### 8.2.6.2 Vlastní implementace

Vlastní implementace lze opět realizovat několika způsoby, a to s různým stupněm outsourcingu. Je možné buď outsourcovat celý projekt, nebo například pouze vývoj.

Výhody:

- řešení šité na míru;
- možnost ovlivnit funkcionalitu pečlivou formulací požadavků.

Nevýhody:

- vyšší časové nároky na zaměstnance (specifikace požadavků, ověřování a testování kvality implementace);
- vyšší odborné znalosti IT zaměstnanců v případě participace při vývoji.

#### 8.2.6.3 Kombinace řešení

Řešení je možné kombinovat různými způsoby. Jako příklad lze uvést metadata repository dodané jako existující systém a jeho napojení na vlastními silami implementovaný datový sklad.

### 8.2.7 Návrh architektury

Architektura navrhovaného systému se skládá ze dvou aplikací. První aplikace je metadata repository. Tato aplikace je sice řešena v rámci projektu BI, nicméně má přesah i do ostatních částí IT/ICT infrastruktury společnosti. Navazuje implementace BI aplikace. Implementované součásti jsou datový sklad a ETL transformace.

Implementace datového skladu v této fázi je vyvolána potřebami na analýzy prodejů a proto by možná bylo korektnější použit přesnější terminologii datové tržiště. Analýza je podrobněji rozpracována v navazujících kapitolách.

### 8.2.8 Rizika

V následující tabulce jsou zpracována rizika identifikovaná v rámci studie. Dopad může nabývat hodnot A-E, kdy A je fatální a E nejméně kritický.

Tabulka 4 Rizika

| Riziko | Dopad | Opatření na snížení rizika |
|---|---|---|
| Nedostupnost služeb | A | Monitoring stavu systému a mechanismy jeho uvedení do funkčního stavu. |
| Nekvalitní data | A | Kontrola dat v rámci ETL transformace, data profiling |
| Nekvalitní návrh systému | A | Ověření návrhu prototypem aplikace. |
| Neautorizovaný přístup a změna dat | A | Definice striktních přístupových práv. Každý uživatel pouze práva nezbytná pro bezproblémovou práci se systémem v rámci své role. |
| Nedostatečná práva uživatel | C | Korektní definice přístupových práv pro danou roli. |
| Neplánovaný růst nákladů spojených s vývojem systému | B | Kvalitní návrh, smluvní ošetření. |
| Přílišná komplikovanost použití BI | B | Tvorba manuálů a dokumentace. |

Zdroj: Vlastní zpracování

## 8.3 Návrh metadata repository

Kapitola obsahuje kompletní návrh metadata repository. Po vymezení požadavků jsou v kapitole uvedeny i ostatní náležitosti návrhu.

### 8.3.1 Požadavky [41]

Požadavky na metadata repository jsou shrnuty do následující tabulky. Požadavky slouží jako podklad pro vlastní návrh aplikace a zároveň pro analýzu řešení, tj. výběr vhodné technologie.

Tabulka 5 Požadavky na metadata repository

| Identifikátor | Priorita | Popis |
|---|---|---|
| F001 | 1 | Metadata repository umožní správu (vložení, aktualizace, smazání) metadat. |
| F002 | 1 | Jednotlivé záznamy metadat jsou verzovány. |
| F003 | 1 | Struktura repository bude navržena s ohledem na maximální znuvupoužitelnost a na postupnou integraci procesů v rámci společnosti. |
| U001 | 2 | Bude vyhotovena podrobná dokumentace k metadata repository a srozumitelný uživatelský manuál. |
| R001 | 1 | Je nutné zajistit maximální spolehlivost aplikace a definovat zodpovědnou osobu za správu dat. |
| P001 | 2 | Vzhledem k použití v rámci výpočetně náročných procesů je nutné zajistit maximální možnou odezvu. |
| S001 | 2 | Aplikace je rozšiřitelná dle potřeby a přenositelná mezi různými aplikačními prostředími. |

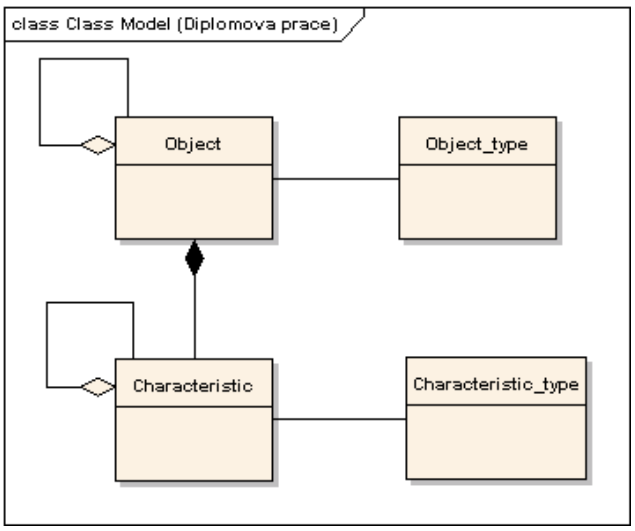
Zdroj: Vlastní zpracování

[41] Prefix identifikátoru obsahuje typ požadavku z FURPS analýzy. Tj. například F001 označuje funkční požadavek (Functionality) s číslem 1 doplněným nulami zleva na trojmístný formát, tj. 001.

### 8.3.2 Diagram tříd

Návrh metadata repository vychází z kapitoly **3.3.1 Metadata repository**. Na následujícím obrázku je navržen upravený diagram tříd. Hlavním objektem je **Object**, který má navázán jednotlivé charakteristiky (Charecteristic). Objekty mohou mít agregovány další objekty a to samé platí pro charakteristiky.

Obrázek 6 Diagram tříd metadata repository



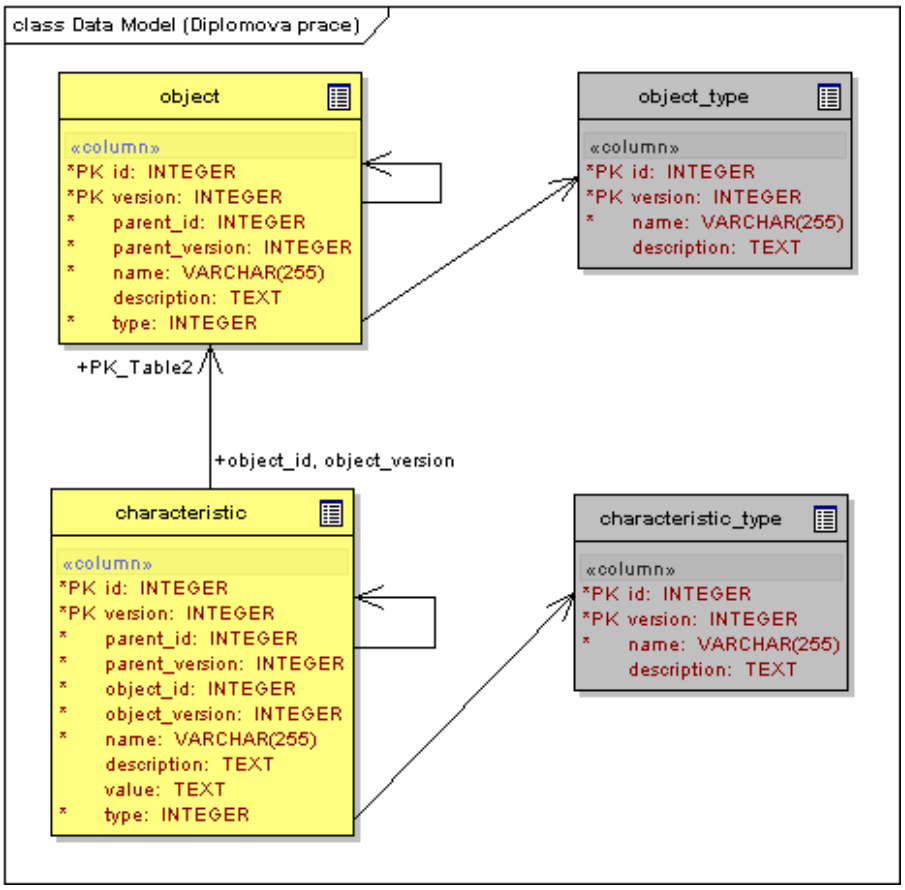

Zdroj: Vlastní zpracování

K zajištění správné funkčnosti a udržení integrity dat je nutné, aby při aktualizaci nadřazeného objektu došlo k aktualizaci všech podřízených, tj. pokud změním nadřazený objekt, je nutné vytvořit nové verze podobjektů a k nim i rekurzivně nové verze charakteristik. Pokud upravím nadřazenou charakteristiku, musím upravit pouze podcharakteristiky. Aktualizace objektu není nutná.

### 8.3.3 Logický datový model

Logický datový model zachycuje jednotlivé tabulky s atributy a relace mezi tabulkami. Požadavek na verzování je promítnut do složeného primárního klíče sestávajícího se z identifikátoru  a čísla verze, kdy jako aktuální je brána nejvyšší verze. Reference mezi tabulkami je **Object** a **Characteristic** je potom realizována pomocí kombinace atributů **id** a **version**.

Obrázek 7 Logický datový model metadata repository



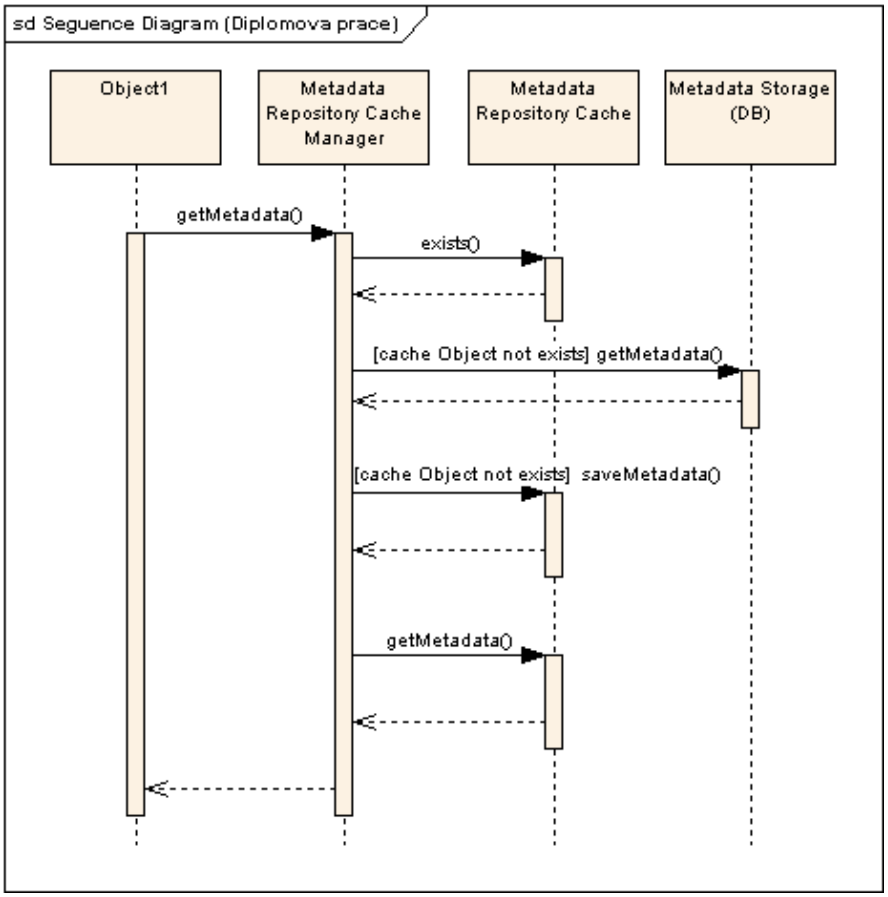

Zdroj: Vlastní zpracování

### 8.3.4 Návrh rozhraní (API)

Při návrhu metadata repository je prioritou, mimo samotnou správnost dat, i rychlost odezvy. Proto je nutné zapojit cachovací mechanismus, na který zajistí odezvu v dostačujícím čase. Metadata jsou předávána v XML. Příklad metadatového XML lze nalézt v příloze 1.

Na následujícím diagramu je znázorněno dotazování metadata repository. Klientská aplikace si vyžádá z Metadata Repository Cache Managera metadata. Ty jsou buď k dispozici v Metadata Repository Cache nebo jsou získaná z Metadata Repository Storage, nahrána do Cache a vrácena klientské aplikaci.

Obrázek 8 Sekvenční diagram získání metadat klientskou aplikací




Zdroj: Vlastní zpracování

## 8.4 Návrh datového skladu

### 8.4.1 Požadavky

Souhrn nejdůležitějších požadavků na BI aplikaci je uveden v následující tabulce.

Tabulka 6 Požadavky na BI aplikaci

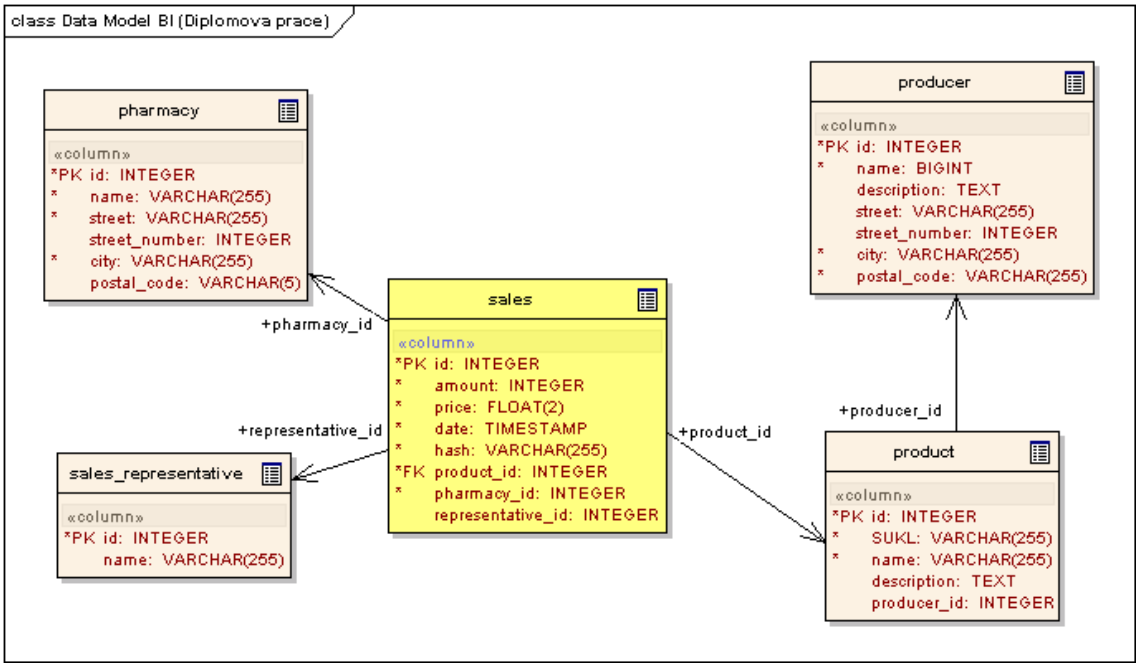
| Identifikátor | Priorita | Popis |
|---|---|---|
| F101 | 1 | Aplikace umožní automatizované zpracování vstupních datových souborů. |
| F102 | 1 | Data jsou konsolidována a je zajištěna integrita dat k zamezení duplicit, odstranění nulových záznamů, dále pak referenční a doménová integrita. |
| F103 | 1 | Na data jsou aplikovány jednotlivé metriky datové kvality. |
| F104 | 1 | Nekvalitní data, která se nepodaří automatizovaně doplnit (z číselníků nebo ze zdrojů třetích stran) budou zařazena do reportu, který bude k dispozici ke kontrole pověřené osobě. |
| F105 | 1 | Sjednocení a konsolidace číselníků lékáren a prodejců. |
| F106 | 1 | Autentizace a autorizace uživatelů. |
| U001 | 2 | Podrobná dokumentace a srozumitelný uživatelský manuál. |
| R001 | 1 | Maximální spolehlivost aplikace. |
| P001 | 2 | Optimalizace výkonu BI aplikace pro velký počet záznamů. |
| S001 | 2 | Aplikace je rozšiřitelná dle potřeby a přenositelná mezi různými aplikačními prostředími. |

Zdroj: Vlastní zpracování

### 8.4.2 Logický datový model

Logický datový model zachycuje strukturu vlastního datového skladu. Žlutou barvou je podbarvena tabulka faktů. Jednotlivé atributy obsahují i návrh datových typů, popř. omezení.

Obrázek 9 Logický datový model BI aplikace




Zdroj: Vlastní zpracování

Do tabulky faktů se ukládají záznamy vlastních prodejů. Mimo reference na dimenze jsou v tabulce obsaženy datové atributy:

- **počet kusů** (amount) – počet kusů prodaných lékárnou za den;
- **částka** (price) – částka za kterou byl realizován prodej produktu (celková suma za produkt / den);
- **datum** (date) – datum prodeje;

- **hash** – řetězec znaků vypočítaný jednocestnou hashovací funkcí (md5), k jednoznačné identifikaci jedinečnosti záznamu (entitní integrity). Použitím tohoto atributu se předejte duplicitním importům stejných dat.

Dimenze jsou potom rozděleny podle schématu **sněhové vločky** (Snow Flake). První dimenzí je produkt (product), který obsahuje informace o prodávaném produktu ke kterému se vztahují prodeje. Zde je klíčovým atributem SUKL, který identifikuje produkt kódem zaregistrovaným u Státního ústavu pro kontrolu léčiv. Jednotlivé produkty mají reference na své výrobce.

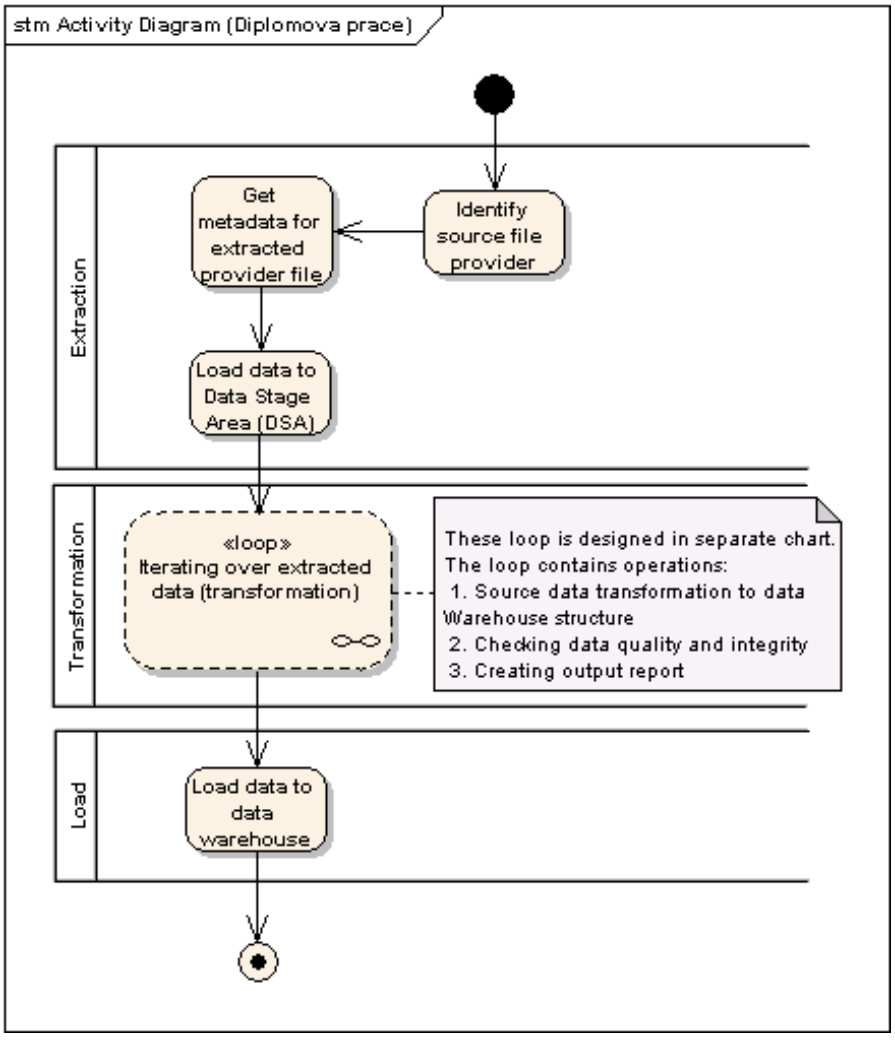
Další dimenzí je lékárna (pharmacy), kde byl realizován prodej. Atributy lékárny jsou název a její poloha jak je naznačeno v datovém modelu. Poslední dimenzí z modelu je obchodní zástupce (sales representative), který dodal produkt do vybrané lékárny, resp. pod kterého lékárna spadá.

Takto strukturované dimenze umožní tvorbu reportů dle zadaných požadavků, kdy klíčové pro marketingové oddělení jsou agregované statistiky za vybrané časové úseky pro vybrané entity (lékárna, obchodní zástupce)

## 8.5 Návrh ETL transformace

V navazujícím textu je zpracována analýza ETL transformace. Jak bylo uvedeno v předchozím textu, skládá se ze tří částí: extrakce zdrojových dat, transformace a nahrání upravených dat. Při ETL transformaci hraje důležitou roli metadata repository, které poskytuje informace o zdrojových datech, transformaci a struktuře dat v datovém skladu. Následující activity diagram prezentuje základní průchod ETL transformace.

Obrázek 10 Activity Diagram ETL transformace




Zdroj: Vlastní zpracování

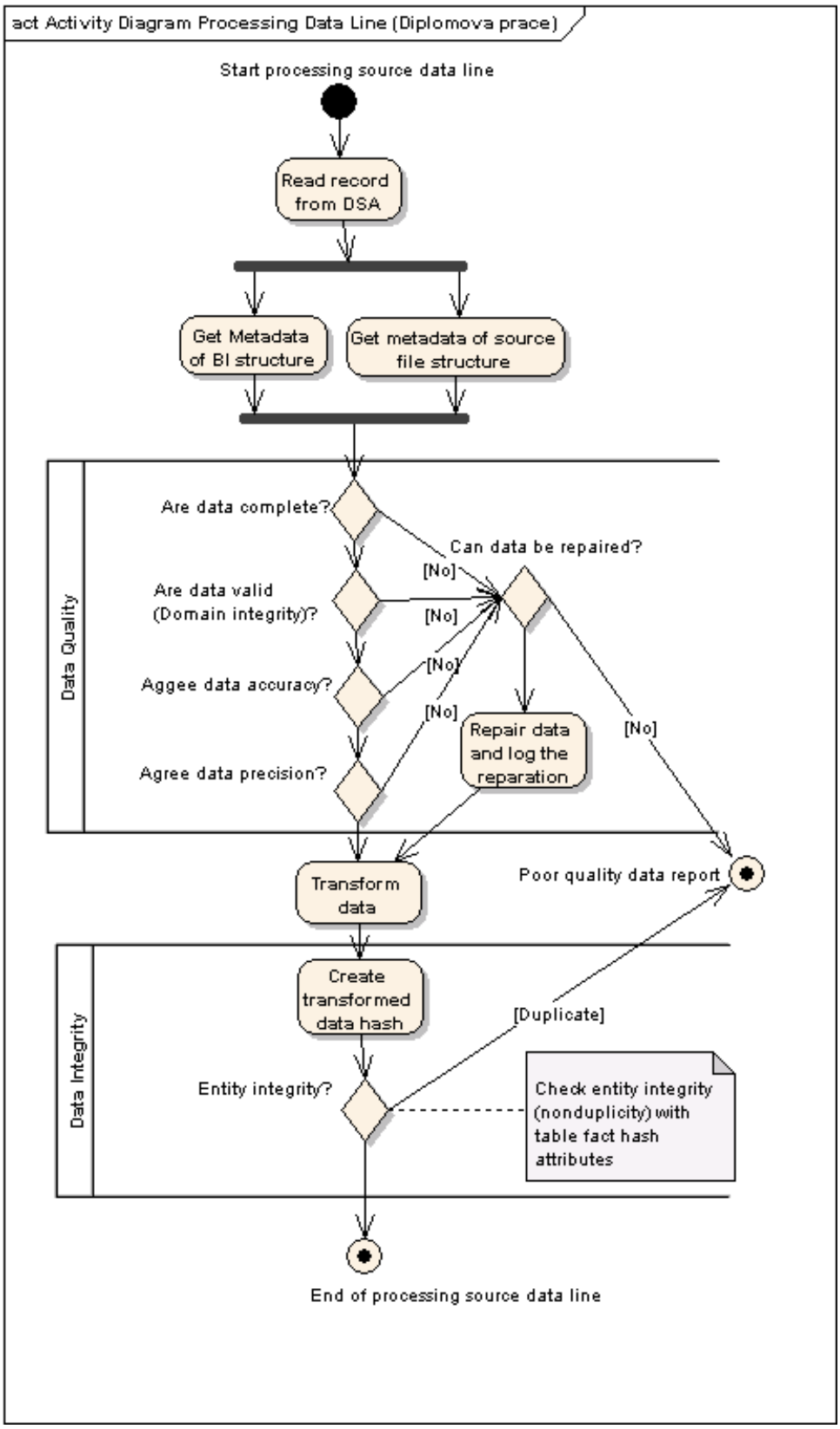
Ve fázi extrakce jsou zdrojová data čtena ze souborů uložených v předem definovaném úložišti (síťový disk), kam je může nahrát pověřená osoba. Vzhledem k tomu, že zdrojové soubory pocházejí od různých obchodních zástupců a mají rozdílný formát, je prvním krokem identifikace poskytovatele souboru. To je realizováno jednoduchým mechanismem prefixování souborů podle obchodního zástupce a časového údaje ve formátu DDMMYYYY (př. xyz_02022011.dbf).

Pokud je identifikován zdroj souboru, jsou donačtena transformační metadata z metadata repository. Tato obsahují informace o zdrojovém souboru, mapování zdrojových dat na strukturu datového skladu a samozřejmě strukturu datového skladu. Metadata dále obsahují informace o vlastní transformaci, jednotlivé kroky a jejich návaznost.

Vlastní transformace jednotlivých datových záznamů je zobrazena na obrázku 10 . Po získání metadat následuje sekce ověření datové kvality vstupních souborů. Postupně jsou ověřovány jednotlivé metriky. Pokud data nevyhovují, je realizován automatizovaný pokus o opravu, kdy v některých případech je možné data doplnit o informace uložené v metadatech. Informace o automatické opravě je pak zalogována pro kontrolu pověřeným uživatelem. V případě nemožnosti opravy poškozených dat, je celý řádek umístěn do reportu obsahujícího nekvalitní data. Tento report musí být následně manuálně zkontrolován a data v něm opraveny pověřenou osobou.

Poslední fáze je nahrání transformovaných do datového skladu. To je v tomto případě realizováno **přírůstkově**. Pokud data existují již v datovém skladu, jsou odmítnuta a zaznamenána jako duplicitní. Ověření entitní integrity se provádí přes jednocestnou hashovací funkci. Po transformaci dat je vypočítán hash a ten je porovnán se záznamy uloženými v datovém skladu.

Obrázek 11 Activity diagram zpracování jednoho datového záznamu




Zdroj: Vlastní zpracování

## 8.6 Návrh řešení kvality dat

Návrh řešení kvality dat obsahuje **analytickou** a **návrhovou (implementační)** rovinu.[42] Analytická rovina definuje zdroje a příčiny chyb, požadavky na organizační zajištění. Implementační rovina obsahuje návrh ETL procesů, kontrolních a opravných procedur.

### 8.6.1 Role a zodpovědnosti

V rámci řešení projektu byly identifikovány role definované v následující tabulce.

Tabulka 7 Role a zodpovědnosti

| Role | Zodpovědnost | Oprávnění |
|---|---|---|
| Datový stevard | – analýza a kontrola reportů;<br>– návrh oprav. | – oprava chyb v BI aplikaci;<br>– správa a úprava metadata repository; |
| Analytik | – definice reportů;<br>– komunikace s koncovými uživateli. | – spouštění ad-hoc dotazů. |
| Vlastník dat | – definice pravidel a významu dat. | |
| Sponzor | – zajištění prostředků pro provoz řešení. | – kontrola realizace projektu. |

Zdroj: Vlastní zpracování

[42] Novotný, O, Pour, J., Slánský, D. (2004) *Business Intelligence: Jak využít bohatství ve vašich datech*. Praha: Grada, str: 137.

### 8.6.2 Mapa datových zdrojů

Zdrojová data jsou všechna ve formě souborů poskytových jednotlivými prodejci. Následující tabulka obsahuje vybraný datový zdroj z celkové mapy datových zdrojů. V tabulce 9 je uvedena analýza kvality dat daného datového zdroje.

Tabulka 8 Vybraný datový zdroj

| | |
|---|---|
| **Název** | PHARCB<ROK>_<MESIC>.DBF |
| **Typ** | dBase |
| **Kódování** | DOS/OS2-852 |
| **Velikost** | 100 – 200 KB |
| **Počet záznamů** | 500 – 1000 |
| **Původ** | Externí prodejce, soubor nahrán na vymezený prostor na FTP serveru. |
| **Periodicita** | Jednou měsíčně, zpravidla první týden nového měsíce. (Možná odchylka + týden) |
| **Zodpovědnost** | Vlastník dat |
| **Formát datové věty** | C_LEK – identifátor léku u prodejce<br>LEKARNA – název lékarny<br>UL_CIS – ulice lékárny s číslem popisným<br>MĚSTO – město lékárny<br>PSČ – poštovní směrovací číslo lékarny<br>VYROBCE – název výrobce<br>NAZ_LEKU – název léku<br>SUKL – kód SUKL<br>KUSY – počet prodaných kusů<br>CASTKA – celková částka za prodané kusy<br>DATUM – datum prodeje<br>ICO – IČO lékárny |

Zdroj: Vlastní zpracování

Tabulka 9 Analýza kvality dat

| **Metrika** | Priorita | Počet | Popis | Příklad |
|---|---|---|---|---|
| **Úplnost** | 1 | 3 | Chybějící SUKL | SUKL chybí u: **Lipobase Repair krém na ruce 30g tuba** |
| **Úplnost** | 2 | 1 | Chybějící PSČ | PSČ chybí u lékárny: **Lékárna U svaté Anny** |
| **Validita** | 2 | 1 | Nevalidní adresa | UL_CIS: **neuvedeno** |
| **Validita** | 1 | 5 | Místo lékárny uvedeno jméno provozovatele, název lékárny součást ulice | LEKARNA: **Mgr.Marie Novotná** UL_CIS: **Palackého 351,Lék.U b.jednorož** |
| **Přesnost** | 2 | 2 | Chybějící háčky a čárky | UL_CIS: Dobrovodsk**a** 2093 |
| **Přesnost** | 3 | 2 | Velká a malá písmena | **N**ám. Svobody 1 |

Zdroj: Vlastní zpracování

Tabulka 9 Obsahuje informace o analýze kvality dat souboru dat PHARCB. Analýza byla provedena v programu Open Office 3.2.1 nad vzorovým souborem dat obsahujícím 260 záznamů. Nejprve byla zkontrolována úplnost záznamů a následně ostatní metriky datové kvality.

### 8.6.3 Vlastní zajištění kvality dat v BI aplikaci

V předchozích kapitolách byla popsána místa kontroly kvality dat před vlastním uložením v datovém skladu (návrh BI aplikace, návrh ETL transformace). Přes všechna opatření je možné přenesení chybných dat do BI aplikace, a proto je nutné navrhnout mechanismus řešení kontroly kvality a případných oprav.

Po analýze produktů na trhu byl zvolen nástroj **Talend Open Profiler**[43] ke kontrole kvality dat v rámci BI aplikace. Výhody vybraného nástroje jsou:

- cena – open source nástroj pod licencí GPL[44];
- podpora komunity;
- možnost přímé úpravy chybných dat;
- analýza jednotlivých atributů a tabulek;
- podpora regulárních výrazů a SQL vzorů pro pro analýzy a monitoring dat;
- analýza redundancí;
- zobrazení výsledků ve formě grafů;
- intuitivní GUI.

[43] Produkt společnosti Talend (http://www.talend.com)

[44] Případně lze upgradovat na placenou verzi Talend Data Quality, obsahující mimo jiné i přímou podporu.

# 9 Závěr

V rámci práce je zpracováno téma datové kvality, integrity a konsolidace. Problematika je ilustrována na návrhu datového skladu a metadata repository. Nejprve je uvedena vize a úvodní studie pro přesné vymezení řešeného problému. V rámci práce je navržena architektura metadata repository včetně logického datového modelu a cachování dat pro zajištění co nejrychlejší odezvy (s ohledem na budoucí začlenění do IT/ICT infrastruktury společnosti). Návrh datového skladu obsahuje logický datový model a formou diagramů analyzovaný ETL proces. Součástí nárhu je definice rolí a zodpovědností, očekáváných efektů a rizik projektu, mapa datavých zdrojů a návrh zajištění kvality dat ve vlastní BI aplikaci.

V práci, vzhledem k zaměření, není zpracována analýza reportování, dolování dat a ad-hoc dotazů.

# Literatura

### Monografie

Gála, L., Pour, J., Toman, P. *Podniková informatika*. 1. vyd.. Praha : Grada, 2006. 484 s. ISBN 80-247-1278-4.

Gála, L., Pour, J., Šedivá, Z. *Podniková informatika*, 2. přepr. a aktual. vyd. Praha : Grada, 2006. 484 s. ISBN 80-247-1278-4.

Novotný, O., Pour, J., Slánský, D. *Business Intelligence : Jak využít bohatství ve vašich datech*. 1. vyd. Praha : Grada, 2004. 256 s. ISBN 80-247-1094-3.

Basl, J., Blažíček, R. *Podnikové informační systémy*. 2. vyd. Praha : Grada, 2008. 283 s. ISBN 978-80-247-2279-5.

English, L. *Improving Data Warehouse and Business Information Quality*. New York : Wiley, 1999. 518 s. ISBN 0-471-25383-9.

Bureš, V. *Znalostní management a proces jeho zavádění* : *průvodce pro praxi*. Praha : Grada, 2007. 212 s. ISBN 80-247-1978-9.

### Internetové zdroje

OMG: *DataWarehousing MOF/CWM.* [online]. [s.l.] : OMG, 2010 [cit. 2010-09-18]. Dostupné z WWW: <http://www.omg.org/technology/documents/modeling_spec_catalog.htm>.

Polášek, M. *Metadata a datové sklady*. [online]. [s.l.] : System online, 2010 [cit. 2010-09-18]. Dostupné z WWW: <http://www.systemonline.cz/clanky/metadata-a-datove-sklady.htm>

Tobišek R. *BI Kudy na řízení kvality dat*. [online]. [s.l.] : System online, 2010 [cit. 2010-10-09]. Dostupné z WWW: <http://www.systemonline.cz/business-intelligence/kudy-na-rizeni-kvality-dat.htm>

Pour, J., Slánský D. *Efekty a rizika BI*. [online]. [s.l.] : CSSI, [2010] [cit. 2010-09-18]. Dostupné z WWW: <http://www.cssi.cz/cssi/system/files/all/SI_04_2_pour.pdf>

Pavlis, D. *Pozice a úloha datového skladu v rámci informační strategie firm*y. CSSI, [2010] [cit. 2011-02-11]. Dostupné z WWW: <http://www.cssi.cz/cssi/system/files/all/SI_04_4_pavlis.pdf>

Gerard, N. *Data Modeling : Entity Integrity*. [online]. [s.l.] : [s.n.], 2010 [cit. 2010-09-28]. Dostupné z WWW: <http://gerardnico.com/wiki/data_modeling/entity_integrity>

DatabaseDev. *Domain Level Integrity*. [online]. [s.l.] : DatabaseDev, 2010 [cit. 2010-09-28]. Dostupné z WWW: <http://www.databasedev.co.uk/domain_integrity.html>

Pipino, L L., Lee Y. W., Wang R. *Data Quality Assessment*. [online]. [s.l.] : TDQM MIT, 2010 [cit. 2010-10-09]. Dostupné z WWW: <http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf>

Geiger J. G. *Data Quality Management*. [online]. [s.l.] : [s.n.] 2010 [cit. 2010-10-09]. Dostupné z WWW: <http://www2.sas.com/proceedings/sugi29/098-29.pdf>

Moss, L. T., Abai, M., Adelman, S. *How to improve data quality.* [online]. [s.l.] : INFORMIT. 2005 [cit. 2010-02-27]. Dostupné z WWW: <http://www.informit.com/articles/article.aspx?p=399325&seqNum=4>

Hoberman, S. *Use Of Meta Data Entities* : *Part 3 Of A Series On Abstraction*. [online]. [s.l.] : 2002 [cit. 2010-02-27]. Dostupné z WWW: <http://www.ewsolutions.com/resource-center/rwds_folder/rwds-archives/rwds-2002-10/meta-data-entities-3>

McKendrick, J. *Open Source Business Intelligence: A 2008 Progress Report.* [online]. [s.l.] : 2002 [cit. 2010-06-12]. Dostupné z WWW: <http://www.stratebi.es/todobi/nov08/osbi-2008-progress-report%5B1%5D.pdf>

Klaus M. *Konsolidace databází a kvalita dat ve veřejné správě. Konference Internet ve státní správě a samosprávě.* [online]. Hradec Králové : ISSS, 2007 [cit. 2010-07-03]. Dostupné z WWW: <http://http://www.isss.cz/archiv/2007/download/isss2007.pdf>

Won, K., Byoungju, Ch. *Towards Quantifying Data Quality Costs.Journal of Object Technology.* [online]. Zurich : ETH, 2003 [cit. 2010-07-15]. Dostupné z WWW: <http://www.jot.fm/issues/issue_2003_07/column6.pdf>

Guerra-Garcí, C., Caballero, I., Piattini, M. *Capturing Data Quality Requirements for Web Applications by means of DQ_WebRE*. [online]. [s.l] : 2011 [cit. 2010-07-22]. Dostupné z WWW: <http://www.edbt.org/Proceedings/2011-Uppsala/papers/workshops/beweb_workshop/a7-garcia.pdf>

Oracle *Oracle9i Data Warehousing Guide.* [online]. [s.l] : [2011] [cit. 2010-07-22]. Dostupné z WWW: <http://download.oracle.com/docs/cd/A97630_01/server.920/a96520/toc.htm>

# Přílohy

Příloha 1 Příklad XML při dotazování metadata repository

```
<?xml version="1.0" encoding="windows-1250"?>
<object>
  <id><id>
  <name><name>
  <version></version>
  <parent_object_id></parent_object_id>
  <charakteristics>
    <characteristic>
          <id><id>
          <name><name>
          <version></version>
          <value></value>
          <type>FOREING_KEY</type>
      </characteristic>
  </charakteristics>
  <child_objects>
      <child_object>
          <id><id>
          <name><name>
          <version></version>
          <order>1</order>
      </child_object>
      <child_object>
          <id><id>
          <name><name>
          <version></version>
          <order>2</order>
      </child_object>
  </child_objects>
</object>
```