IT Systems, specializovaná příloha Business Intelligence 2006

Jiří Bohuslav

Business Intelligence stojí na kvalitě dat
Metody a procesy čištění dat

Bouřlivý rozvoj informačních technologií způsobil, že systémy Business Intelligence již automaticky zaujímají své místo v podnikových systémech. Přesto se dá říci, že mnoho rozhodovacích procesů ještě není na těchto systémech založeno nebo informace, jež tyto systémy poskytují, nejsou zcela nebo správně využity.

Firmy téměř vždy uchovávají své informace v OLTP (On-Line Transaction Processing) systémech, které zachycují detailní jednotlivé operace. OLTP systém však není vůbec vhodný pro zodpovězení určitých analytických dotazů (Jaké byly trendy prodeje našich výrobků v minulých pěti letech? Jaký je procentuální nárůst prodeje výrobků mezi dvěma kvartály?), velmi složitě lze získat agregovaná data, obsahuje spoustu dat, které pro kvalitní rozhodování vůbec nepotřebujeme, navíc OLTP systém není optimalizován pro složité uživatelské dotazy.

Oblast business intelligence je specifický obor, který je zaměřen převážně na poskytnutí komplexního pohledu na data koncovému uživateli a získání informací potřebných pro správné rozhodnutí. Základním krokem v procesu získání dat z provozních systémů pro kvalitní rozhodování je analýza, konsolidace a čištění zdrojových dat. Z existujících dat (např. o chování zákazníků, objemech prodeje v různých regionech) je třeba vytěžit maximum informací, jež mohou při správném využití poskytnout výraznou konkurenční výhodu.

Obvyklým řešením pro získání kvalitních konsolidovaných dat je vybudování datového skladu (Data Warehouse), datových tržišť (Data Mart), případně nadstavbových OLAP („On-Line Analytical Processing“) aplikací, neméně důležité je zvolení vhodné technologie pro prezentaci dat uživateli (reporting), která navíc umožní intuitivním způsobem analyzovat data ve vícerozměrné struktuře, sestavovat ad-hoc dotazy. Datový sklad se tím pádem stane jedním velkým centralizovaným zdrojem informací pro celou firmu.

Pochopitelně je v zájmu firmy, aby data v datovém skladu byla všeobecně přijata jako „jedna verze pravdy“. Jde o to, že v datovém skladu se střetávají data z různých zdrojových systémů, ale i různá externí data, ručně udržované evidence a podobně. Přitom by měl být kladen velký důraz na zaručenou čistotu, tedy kvalitu dat. V datech přenášených do datového skladu se téměř vždy objevují duplicity a datové chyby, které není jednoduché odhalit. Tyto nepřesnosti způsobuje proměnlivé názvosloví („str. 56“/„stránka 56“), používání či nepoužívání diakritiky („František Novák“/„Frantisek Novak“), pravopisné a jiné chyby, které způsobují nekonzistenci a je nutné je rozpoznat. Většinou neexistuje korekce chyb ve zdrojovém systému, proto je třeba nekonzistence dohledat, opravit a záznamy logicky spojit do jednoho při plnění datového skladu (ETL proces). Zajímavým postřehem je, že tento proces čištění dat při plnění datového skladu může pak zpětně sloužit jako opravná zpětná vazba pro zdrojový provozní informační systém. Udává se, že až 15 % všech zdrojových dat je nekonzistentních nebo nesprávných.

Zkusme si teď nějaké metody a procesy čištění dat popsat podrobněji.

Name-Address Cleansing

Umožňuje čištění jmen a adres ve zdrojových datech. Nalézá a opravuje chyby a nekonzistence porovnáváním vstupních dat s knihovnami dat často dodávaných třetí stranou. Chyby a nekonzistence ve vstupních datech jsou eliminovány:

- analýzou a rozdělením jmen a adres vstupních dat na jednotlivé části,
- převedením jmen a adres na standardní formát podle konvencí dané země,
- doplněním jmen a adres o další údaje jako je pohlaví, kód státu, obchodní nebo geografické informace apod.

Při použití této metody se nejprve definují obecné parametry jako je národní prostředí, různá rozpoznávací kriteria, jako třeba typ a struktura adresy.

Dále je třeba definovat vstupní atributy a roli, jakou údaje ve vstupních datech hrají. Role indikuje, jaký druh dat daný atribut reprezentuje. Rolí může být několik, od těch obecnějších, jako je osoba, adresa, až po konkrétnější, jako je křestní jméno, oslovení, město, země apod.

Posledním krokem je definice výstupních komponent. Každá komponenta reprezentuje konkrétní část jména nebo adresy, jako je oslovení, titul, standardizované křestní jméno, prostřední jména, jméno ulice, číslo domu apod., nebo obecnější entitu, jako adresu, jméno atd. Dále mohou být přidány výstupní komponenty reprezentující příznaky a chybové hlášky zpracování jednotlivých záznamů. Příznaky indikují, zda byl daný záznam nebo jeho část úspěšně analyzován a nalezen v databázi srovnávacího software. Tyto komponenty mohou sloužit k oddělení úspěšně zpracovaných záznamů od záznamů obsahujících chyby.

Match-Merge

Pomocí této metody se na základě předem definovaných pravidel dají určit záznamy, které reprezentují stejná data a podle dalších pravidel se tyto záznamy dají sloučit do jediného. Tato metoda spolu s metodou Name-Address poskytuje nástroj pro jednoznačnou identifikaci kontaktních informací o zákaznících. Prvním krokem při použití metody Match-Merge je definice porovnávacích pravidel. Pravidel může být více, záznamy jsou označeny jako identické, pokud splní jedno z následujících pravidel.

1) Porovnání založené na podmínce

Podmínkami, podle kterých lze provést porovnání, mohou být například:

- přesná shoda – záznamy musí být identické,
- standardizovaná přesná shoda – záznamy musí být identické, ignorují se malá/velká písmena, mezery a nealfanumerické znaky,
- podobnost - lze specifikovat procentuální hodnotu podobnosti, kterou musí záznamy splňovat, aby byly označeny za identické,
- standardizovaná podobnost,
- částečný název – jeden záznam se vyskytuje jako podřetězec ve druhém záznamu počínaje prvním slovem,
- detekce zkratek a akronym.

2) Porovnání jmen

Lze použít u záznamů reprezentujících jména. Jednotlivým atributům lze přiřadit role jako je oslovení, křestní jméno apod. Na základě přiřazených rolí lze definovat pravidla typu:

- srovnání iniciál se jmény, kdy se za identické považují záznamy jako „P.“ a „Petr“,
- srovnání podřetězců se jmény, záznamy např. „Rob“ a „Robert“ se považují za identické,
- podobnost – lze specifikovat procentuální hodnotu podobnosti, kterou musí záznamy splňovat, aby byly označeny za identické,
- detekce složených jmen,
- srovnání oslovení vůči křestnímu jménu a příjmení – např. „p. Novák“ a „Martin Novák“,
- detekce chybějících spojovníků,
- detekce přehozeného křestního jména a příjmení.

3) Porovnání adres

Používá se u záznamů obsahujících adresy. Jednotlivým atributů lze přiřadit role jako je ulice, město, stát. Podle typu role lze pak definovat různá srovnávací pravidla.

4) Porovnávání váhy atributů

Záznamy jsou potom označeny za identické pokud váhový součet podobnosti jejich atributů překročí určitou, zadanou hodnotu.

Příklad porovnání a sloučení záznamů reprezentujících stejná data pomocí metody Match-Merge

Vstupní data:

Row	FirstName	LastName	SSN	Address	Unit	Zip
1	Jane	Doe	NULL	123 Main Street	NULL	22222
2	Jane	Doe	111111111	NULL	NULL	22222
3	J.	Doe	NULL	123 Main Street	Apt 4	22222
4	NULL	Smith	111111111	123 Main Street	Apt 4	22222
5	Jane	Smith-Doe	111111111	NULL	NULL	22222

Výstupní data (vyčištěná o duplicity a standardizována):

FirstName	LastName	SSN	Address	Unit	Zip
Jane	Doe	111111111	123 Main Street	Apt 4	22222

Data Profiling

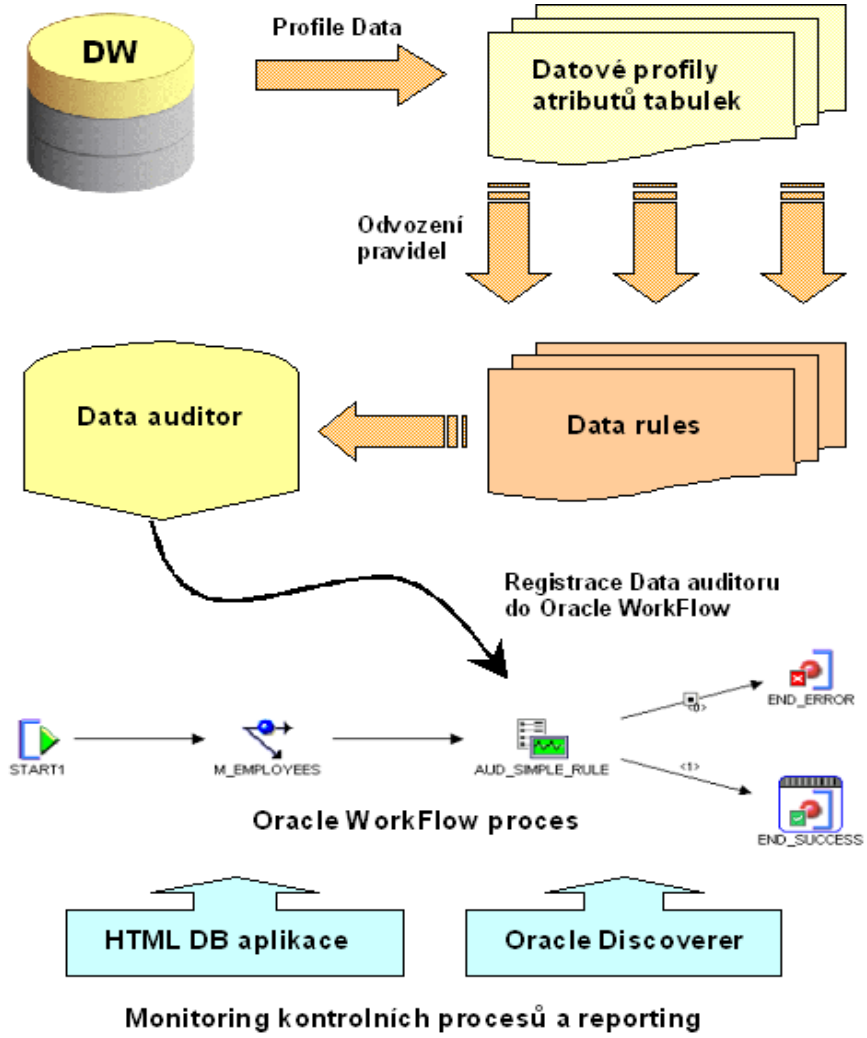
Metoda, která umožňuje zkoumat data z hlediska struktury, identifikovat anomálie, provádět analýzu atributů, referenční analýzu, případně tzv. custom analýzu, kde si pravidla definuje sám uživatel.

Řešení profilace a čištění dat je realizováno v následující posloupnosti kroků:

- definice zdrojových dat – tabulek,
- spuštění procesu profilace dat,
- na základě analýzy zdrojových dat je třeba navrhnout pravidla pro jednotlivé atributy (data rule),
- tato pravidla lze převést na tzv. korekce,
- korekce je třeba aplikovat na vstupní data/atributy,
- pro ETL proces lze nastavit strategii pro čištění dat jako například remove (data, která nesplňují pravidla se nepřenesou dále, jsou odstraněna), match (očištění dat na základě nejbližší shody hodnoty atributů - odstranění překlepů, velká/malá písmena, apod…) nebo custom (libovolná vlastní logika čištění dat).

Jako příklad produktu pro tvorbu ETL procesů, který nabízí širokou škálu nástrojů pro potřeby kontroly a čištění dat, lze zmínit Oracle Warehouse Builder. V poslední verzi OWB se skrývá také mnoho komponent pro správu kvality dat a pro data profiling.

Na obrázku je znázorněn typický postup při řešení Data Profilingu pomocí nástrojů Oracle.



Transformace dat zajišťující čištění atributů se v některých případech nedají dostatečně jednoduše zobecnit, proto se v případě, že žádnou z uvedených metod a nástrojů nelze použít, neobejdeme bez psaní vlastních transformací. Pro řešení složitějších situací, kdy je potřeba při čištění dat využívat robustnějších metod, existuje řada dalších, cenově samozřejmě mnohem náročnějších specializovaných produktů (např. Trillium Software od společnosti Harte-Hanks, SAS Data Quality Solution, Firstlogic – součást Business Objects, Melissa Data).