Zavádzanie údajov do dátového skladu a ich testovanie

Loading data into data warehouse and their testing

Robert Halenár, Katedra aplikovanej informatiky, FPV UCM Trnava

Abstract: Data for Data Warehouse mostly come from various non-homogenous sources, thus preparing the data is an important phase in the implementation of data warehouses. Data already stored in data warehouses is then required to undergo testing in order to avoid erroneous interpretations of data, eventually to detect errors caused by automatic export and import of the data. Tools and procedures for ETL (Extraction, Transformation, Loading) consists of certain phases about of this article discusses and also about the risks and mistakes in practice associated with this process.

Key words: Data warehouse, extraction, transformation, Loading

Abstrakt: Údaje pre data warehouse pochádzajú väčšinou z rôznych nehomogénnych zdrojov, preto je príprava týchto údajov dôležitou fázou pri zavádzaní dátových skladov. Dáta už uložené v dátových skladoch je následne nutné podrobiť testovaniu, aby sa predišlo chybným interpretáciám údajov, prípadne aby sme odhalili chyby vzniknuté pri automatizovanom exporte a importe údajov. Nástroje a postupy ETL (Extraction, Transformation, Loading) pozostávajú z určitých etáp o ktorých pojednáva článok a tiež o rizikách a chybách v praxi spojených s týmto procesom.

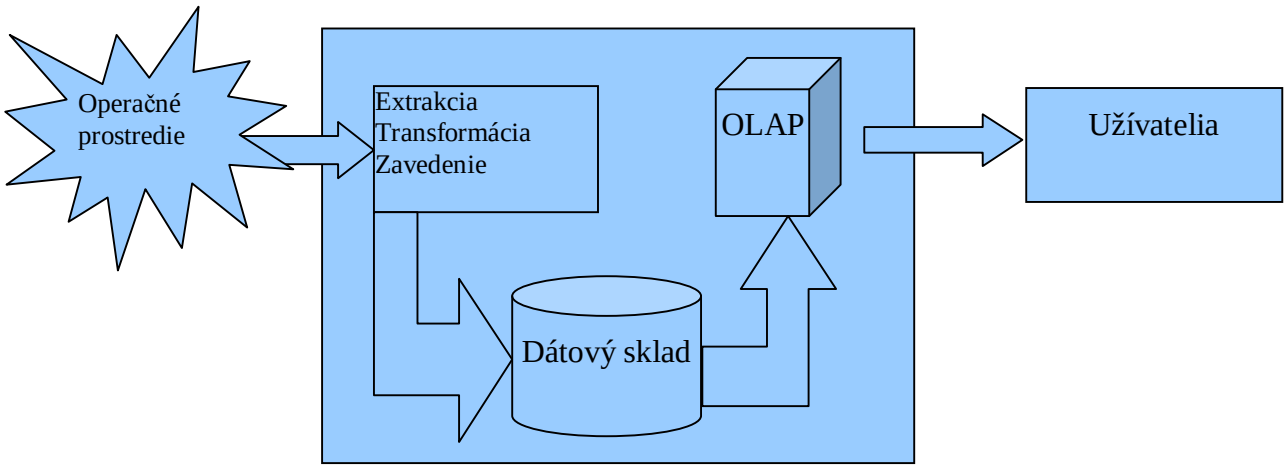
Kľúčové slová: Dátový sklad, extrakcia, transformácia, zavádzanie

Etapy ETL – Extraction, Transformation, Loading

Pred zavedením týchto technológií sa používajú údaje získavané z primárnych transakčných systémov OLTP – Online Transaction Processing. Je vhodné ak sú už spracované do tzv. Zostavy, ktoré sa potom spravidla ručne alebo pomocou softvéru typu Office spracovávajú do manažérskych podkladov pre účely rozhodovania. [1]

Takéto zostavy následne využívajú manažéri v praxi riadenia podniku, čo je tiež jeden zo spôsobov optimalizácie produkcie. [3]

Údaje pre proces business intelligence a data warehouse teda pochádzajú z rôznych nehomogénnych zdrojov. Môžu do byť údaje zo súborových databáz (Access, dBase,...), údaje z databáz spravovaných niektorým databázovým serverom (Oracle, Informix, Microsoft SQL Server, Sybase, Interbase, Ingres,...), mútu to byť údaje vyexportované nejakou databázovou platformou do tzv. flat súboru a podobne. Príprava a zavádzanie údajov je dôležitou súčasťou každého riešenia dátového skladu. Údaje z operačného prostredia je potrebné pred zavedením do dátového skladu vyextrahovať, vyčistiť, upraviť a až následne vo vhodnej forme do dátového skladu zaviesť. [1]



Obrázok 1: Dátový sklad [1]

Extrakcia

Údaje, ktoré chceme preniesť do dátového skladu, sú jednak umiestnené v rôznych nehomogénnych operačných prostrediach, hardvérových platformách (PC, mainframe, iMac...), operačných systémoch (Windows, Unix, Linux, Sun, Solaris...), databázových systémoch (MS SQL Server, Oracle, Informix, IBM DB2...), archívnych systémoch, podnikových systémoch (SAP, PeopleSoft, Baan...), a navyše sa môžu vyskytovať v rozličných formátoch. Úlohou extrakcie je získať údaje práve z takýchto zdrojov.

Kapitolu samu o sebe tvoria archívne dáta, ktoré obsahujú historické údaje. Hlavný rozdiel medzi dátovým skladom a archívom je v tom, že údaje v dátovom sklade sa pravidelne obnovujú. Údaje z archívov sú nezastupiteľným zdrojom historických dát pri prvom naplnení dátového skladu. Oproti tomu archívne dáta nepoužívame pre obnovu údajov v dátovom sklade.

Okrem interných údajov z nášho informačného podnikateľského prostredia je niekedy potrebné pracovať aj s externými údajmi. Tieto údaje môžeme získať analýzou

konkurenčného prostredia, zakúpením údajov o zákazníkoch, alebo aj stiahnutím údajov voľne dostupných na Internete. Z toho vyplýva, že tu nemôžeme periodicky odoberať vzorky, ako sme zvyknutý pri interných údajoch. Externé údaje preto vyžadujú nepretržité monitorovanie za účelom určenia, kedy sú dostupné.

Pre extrakciu sú k dispozícii rôzne postupy, nástroje a technológie. Môžeme vytvárať vlastné aplikácie vo vyšších procedurálnych jazykoch, C++, C# alebo v procedurálnych nadstavbách jazyka SQL (T-SQL, PL/SQL...). Pre menšie množstvo údajov je výhodné vytvoriť prístupovú bránu (gateway). Táto metóda ale pre väčšie objemy údajov zaťažuje sieť. Niekedy môžeme použiť výstupy z vlastných podnikových systémov, ktoré umožňujú konverziu a vyčistenie údajov. Pri správne navrhnutej etape ETL máme k dispozícii meta dáta pre všetky fázy tejto etapy. Tieto meta dáta obsahujú informácie o mieste, type, prístupových privilégiách a štruktúre zdroju údajov. [1]

Kvalita údajov pre analýzy

Ak sú dáta v dátovom sklade nekvalitné, znižuje to dôveru v takéto riešenia a dátový sklad sa oprávnene nevyužíva. Jadro problému je v tom, že nekvalitné dáta sa dosť často vyskytujú v zdrojových systémoch. Použitie nekvalitných údajov vedie k chybným alebo minimálne nepresným zostavám a následne k chybným obchodným rozhodnutiam. Dobre navrhnutý dátový sklad musí umožňovať riešenie jednoduchých ale aj zložitých požiadaviek, napríklad nájsť správnu cieľovú skupinu užívateľov pre marketingovú kampaň, zistiť, či určitý konkrétny zákazník alebo skupina zákazníkov kupuje príslušné produkty. To sú stále len typické úlohy. V živote firiem nepanuje ale len stereotyp, dochádza napríklad k akvizíciám iných spoločností a tým aj prevzatia jej zákazníkov a podobne. S postupom času je potom potrebné uskutočňovať potrebné zmeny v systémoch OLTP, aby sa zlepšila kvalita dát dátového skladu. [1]

Firmy využívajú pre svoju činnosť rôzne druhy ekonomického softvéru, napríklad účtovníctvo, skladové hospodárstvo, evidenciu pohybov tovaru a podobne, pričom samozrejme zhromažďujú dáta. Sčasti sú možno bezcenné, ale možno aj veľmi cenné, ale zostávajú nevyužité, pretože sú uložené vo forme, ktorá ich robí nedostupnými pre účely získavania informácií. Existencia dát totiž vôbec neznamená dostupnosť informácií.

Ak dáta usporiadame do multi – dimenzionálnej štruktúry, budeme pracovať s omnoho menšími rozmermi dimenzií a dáta potom budú mať oveľa väčšiu výpovednú schopnosť. Cez dimenzie sa potom dokážeme dostať k príslušnému faktu, ktorý leží na priesečníku dimenzií.

Snažíme sa, aby multi – dimenzionálna informácia bola uložená tak, aby bola orientovaná na predmet podnikania, a nie viazaná na konkrétny systém pre zber údajov, odkiaľ predmetný údaj pochádza. [2]

Čistenie údajov

Údaje z externých zdrojov majú určitú kvalitu, ktorá je buď postačujúca, alebo nepostačujúca pre ich zavedenie do dátového skladu. Často dokonca býva kvalita údajov značne premenlivá. Pre čistenie údajov sa v anglickej terminológii používa niekoľko rovnocenných výrazov – cleaning, scrubbing alebo cleansing.

Čistenie údajov môže byť niekedy veľmi náročné, a teda aj nákladné. V niektorých prípadoch dokonca ani nemá zmysel čistiť údaje s vysokými nákladmi, ak je prínos pre podnikanie zanedbateľný. Dokonca keď aj systémy OLTP obsahujú kvalitné údaje, tieto údaje nemusia byť zárukou kvalitného dátového skladu. Systémy OLTP totiž neobsahujú historické údaje. [1]

Transformácia

Transformácia samotná je súbor úloh a úkonov, ktoré vedú ku zvýšeniu kvality údajov, hlavne k odstráneniu anomálií. Anomálie nie sú v systémoch OLTP spravidla na závadu. Vývoj niektorých systémov trvá pomerne dlho, obmieňajú sa verzie softvéru, mení sa vývojové prostredie a technologické platformy, na ktorých sa tento softvér vyvíja. Menia sa aj operačné systémy (napr. MS DOS na Windows).

Okrem technických záležitostí pôsobí aj ľudský faktor, napríklad pravopisné chyby, poradie zadávania a pod. Počas čistenia dát sa teda zjednocuje formátovanie údajov, priraďovanie dátových typov, jednotiek miery a peňažných mien. Údaje v databázach OLTP často obsahujú rôzne kódované údaje, alebo aj primárne kľúče, ktoré sa skladajú z viacerých častí. Pri zavedení týchto údajov do dátového skladu je potrebné rozložiť tieto údaje na atomické hodnoty. [1]

Najčastejšie sa vyskytujúce problémy pri transformácii údajov:

- Nejednoznačnosť údajov – napríklad údaje o pohlaví zákazníka môžu byť uložené rôznym spôsobom (Female, male, F, man, woman, F, female...), po transformácii by mal byť takýto stĺpec v jednotnom tvare (F alebo M)

- Chýbajúce hodnoty a duplicitné záznamy – duplicita údajov je menší problém, ak je niečo naviac, tak sa to dá odstrániť. V niektorých prípadoch to môže byť časovo náročné. Väčší problém predstavujú chýbajúce údaje. V takom prípade máme viac možností. Pri malom objeme chýbajúcich údajov ich môžeme ignorovať. Niekedy môžeme chýbajúce hodnoty doplniť z iných zdrojov.
- Konvencia názvov pojmov a objektov – ak zlučujeme údaje z rôznych zdrojov, ktoré popisujú v podstate rovnaký jav, ale majú jednotlivé entity vedené pod rôznym názvom, tak musíme zlúčiť terminológiu, a vytvoriť jednotnú konvenciu názvov.
- Rôzne peňažné meny – suma 125,50 znamená niečo úplne iné v eurách než v českých korunách. V prechodnom období prechodu na euro sa uvádzajú ceny v miestnej mene aj v eure.
- Formáty čísiel a textových reťazcov – údaje sú v relačných databázach a súboroch uložené v rôznych druhoch formátov. Najväčší problém je s ukladaním číselných údajov. Najčastejšie sa pre tieto údaje používajú numerické a reťazcové dátové typy. Do numerického dátového formátu sa ukladá číslo ako numerická hodnota, do reťazcového dátového typu sa číslo ukladá ako postupnosť číslic a iných znakov, napríklad desatinných čiarok a medzier. Problém ale môže nastať napríklad v rodnom čísle v tvare 7707027416, pretože ho v tom tvare môžeme uložiť aj ako číslo aj ako textový reťazec. V častejšie uvádzanej podobe 770702/7416 môžeme toto rodné číslo uložiť len ako textový reťazec. Podobne je to aj s poštovými smerovacími číslami. V podobe 91701 ho môžeme uložiť aj ako číslo aj ako textový reťazec. Numerické formáty pre uloženie PSČ sa nepoužívajú aj z iného dôvodu. Veľa z nich totiž začína nulou a tak sa päťmiestne PSČ 08701 zmení na štvormiestne 8701, pretože nuly pred prvou platnou číslicou sa v numerických formátoch vynechávajú.
- Referenčná integrita – okrem hodnôt sú v údajoch skryté aj rôzne vzťahy, napríklad master – detail, organizačná štruktúra firmy, hierarchická štruktúra zamestnancov a podobne. Ale údaje sú dynamické, organizačná štruktúra sa mení, často bez dokumentácie a adekvátnych zmien v databázach OLTP. Ak sa zruší nejaké oddelenie a zostanú po ňom nejaké záznamy, tieto môžu potom skresliť údaje a teda nepriaznivo ovplyvniť kvalitu údajov.
- Chýbajúci dátum – čas plní v dátovom sklade významnú úlohu. Od neho sa všetko odvíja a skoro každá analytická databáza má časovú dimenziu. V mnohých

transakčných systémoch sa údaje neoznačujú časovými údajmi, v iných je naopak čas dôležitou veličinou. Napríklad dátum objednávky transakcie a podobne. Časový údaj musí byť prítomný v dátach pred ich zavedením do dátového skladu, alebo sa musí určiť a pridať pri zavádzaní dát. Je potrebné dobre uvážiť, kedy a kde sa transformácia uskutoční. Môžeme ju vykonávať sériovo alebo paralelne so zavádzaným údajov. Pri sériovom spôsobe sa transformácia vykoná pred zavedením dát do dátového skladu. Pri paralelnej metóde sa tento proces vykonáva súbežne so zavádzaním. [1]

Prenos

Zavŕšením etapy ETL je prenos údajov z pamäte zdrojových dát alebo prechodnej vynášanej oblasti do dátových skladov. Prenos spočíva v presune údajov a ich uloženia do databázových tabuliek. Prenos by mal byť plánovaný a automatizovaný. Pri prvotnom naplnení dátového skladu môže ísť o obrovské množstvo údajov. Potom sa údaje zavádzajú v pravidelných časových obdobiach, napríklad každý deň, a to v takých objemoch, koľko údajov za dané obdobie v databázach OLTP vznikne.

Zavádzanie dát by malo byť plánované a hierarchizované. Tiež by malo byť automatizované na najvyššiu možnú mieru. Po zavedení údajov spravidla prebieha ich indexovanie, aby bol prístup k nim optimalizovaný. Pre jednoznačnú identifikáciu údajov sa používajú aj umelo vytvárané kľúče, s ktorých pomocou zaistíme jednoznačnosť každého riadku v tabuľke. Dáta dátového skladu sú totiž často kombináciou mnohých transformovaných záznamov, ktoré nemajú žiadne prirodzené kľúče, ktoré by sa dali použiť pre jednoznačnú identifikáciu. [1]

Chyby a problémy etapy ETL

Proces ETL vždy prebehne úspešne. Problémy môžu byť so spoľahlivosťou úložiska údajov (disky sú mechanické zariadenia, ktoré sa opotrebovávajú), môže dôjsť k výpadkom spojenia, zdroje údajov sa môžu meniť, napríklad pri upgrade systémov OLTP, ktoré sa nedokumentujú v metadátach. Dôležité je overenie údajov, pretože ak údaje nie sú overené, tak môže dôjsť k problémom pri extrakcii a transformácii. Podľa závažnosti zlyhania je potrebné začať nanovo alebo môžeme pokračovať od miesta zlyhania. Nepresné alebo neúplné údaje môžu byť príčinou nepresnosti výsledkov analýzy, čo následne môže viesť k nesprávnym obchodným, alebo ešte horšie k strategickým rozhodnutiam. [1]

Opravovanie chybných dát priamo v dátovom zdroji

V ideálnom prípade sú všetky chybné dáta, nájdené a opravené priamo v dátovom zdroji, prípadne priamo v operačnom prostredí nahradené správnymi hodnotami. Táto prax zaisťuje, že získané údaje na oboch úrovniach – prevádzkových aj rozhodovacích budú ťažiť z čistých dát. Skúsenosti však ukázali, že oprava údajov pri zdroji, môže byť obtiažna z nasledovných dôvodov:

- Prevádzková zodpovednosť - zodpovednosť za aktualizáciu dátového zdroja sa samozrejme dostáva do rúk zamestnancov, ktorí nemusia byť ochotní prijať zodpovednosť za správnosť dát v ďalšom spracovávaní a znova opravovať dáta.
- Správne dáta nie sú známe – aj keď ľudia v prevádzke vedia, že dáta sú zlé, nemusí byť jednoduché určiť správne hodnoty údajov. To platí najmä pri zákazníckych údajoch (napr. číslo sociálneho poistenia zákazníka). Ľudia v prevádzke nemajú inú možnosť, ako sa zákazníkov po jednom pýtať a získavať správne informácie. Je to nudné, časovo náročné a spôsobuje to nepríjemnosti v komunikácii s ich zákazníkmi. [5]

Testovanie etapy ETL

Etapu ETL je potrebné dôkladne otestovať najskôr na simulovaných a neskôr na ostrých údajoch. V etape testovania sa už prejaví aj to, či sme etapu ETL dobre a podrobne zdokumentovali.

Testovanie je nutné prevádzať podľa určitých zásad, ktoré umožnia rýchle a efektívne odhalenie prípadných nedostatkov. [4]

Ani po otestovaní a plnom nasadení ETL nemáme istotu, že všetko bude stále fungovať. Objem dátového skladu rastie rýchlo a metrika zavádzania a granularity dát vyžaduje pravidelnú revíziu. Na vykonanie procesov ETL je možné použiť špecializované nástroje, brány a dátové pumpy medzi databázovými systémami, a interne vyvinuté alebo dodávateľské nástroje. Z konkrétnych sú to napríklad Microsoft DTS (Data transformation Services), DPS (Data Pipeline Services) a ďalšie. [1]

Zoznam bibliografických odkazov

(1) Lacko, L.: Datové sklady analýza OLAP a dolování dat, Computer Press, Brno, 2003., 486 s., ISBN 80-7226-969-0

(2) Lacko, L.: Business Intelligence v SQL Serveru, Computer Press, Brno, 2005., 391 s., ISBN 80-251-1110-5

(3) Schreiber, P., Važan, P., Tanuška, P.: Production optimalization by using genetic alorythms and simulation model of production system, ANNALS OF DAAAM FOR 2008 & PROCEEDINGS OF THE 19TH INTERNATIONAL DAAAM SYMPOSIUM Pages: 1229-1230 Published: 2008, Spôsob prístupu: http://apps.isiknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=Y2gN41bIB@lkdBN82F@&page=1&doc=3&colname=WOS

(4) Tanuška P., Trnka A.: Základné zásady testovania dátových skladov – The fundamental principles of data warehouse testing. In: Materials science and technology - ISSN 1335-9053. - č. 2 (2008, 1. 4. 2008, Spôsob prístupu: http://www.mtf.stuba.sk/docs/internetovy_casopis/2008/2/obsah.htm

(5) Humphries M., Hawkins M., Day M.: Data Warehousing – Architecture and implementation, Journal of database management, 2000, 360 s., ISBN 0130809020

Adresa autora:

Robert Halenár, Ing., PhD.
Katedra aplikovanej informatiky, FPV UCM
Nám. J. Herdu 2
917 01 Trnava
robert.halenar@ucm.sk