Fáza extrakcie a spracovanie dát zo systémov OLTP pre použitie v DWH

Phase of data extraction and processing from OLTP systems for use in DWH

Robert Halenár, Katedra aplikovanej informatiky, FPV UCM Trnava

Abstract: Deployment of data warehousing technologies always precedes the data collection phase. These collected data we obtain usually from the primary transaction systems OLTP (Online Transaction Processing). 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

Key words: Data warehouse, business intelligence, extraction, transformation, Loading, operational environment

Abstrakt: Nasadeniu technológií dátových skladov predchádza vždy etapa zhromažďovania dát. Tieto dáta získavame väčšinou z primárnych transakčných systémov OLTP (Online Transaction Processing). Údaje pre data warehouse pochádzajú väčšinou z rôznych nehomogénnych zdrojov, preto je príprava týchto údajov dôležitou fázou pri zavádzaní dátových skladov.

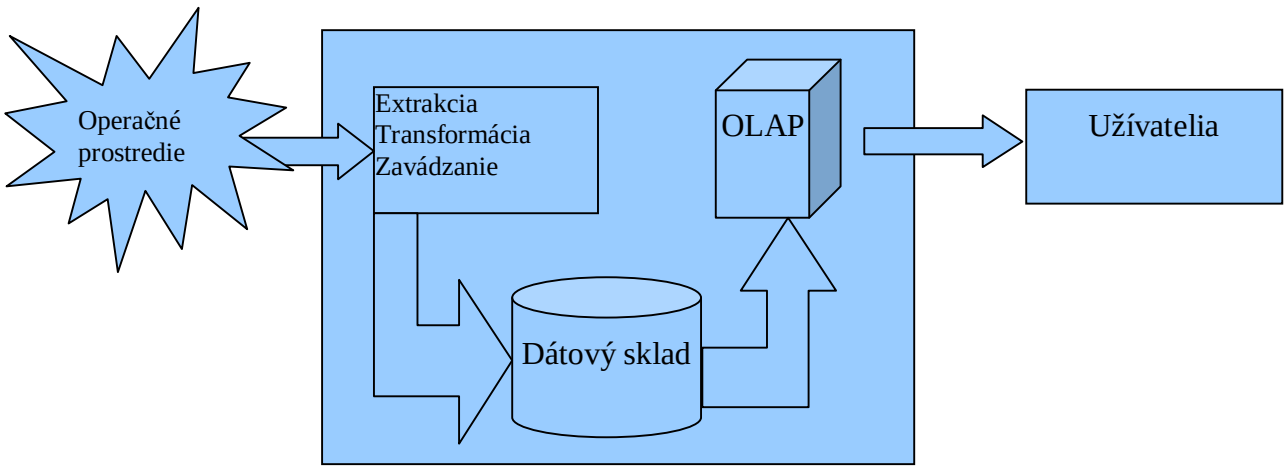
Kľúčové slová: Dátový sklad, business intelligence, extrakcia, transformácia, zavádzanie, operačné prostredie.

Fáza extrakcie údajov z OLTP a ich príprava pre OLAP

Informačné systémy využívajú transakčné systémy alebo analytické systémy. Podľa toho môžu pracovať s dvoma základnými typmi informácií - operatívnymi alebo analytickými. Prvý typ, operatívne informácie, slúži na realizáciu odchodných a ďalších transakcií v podniku. Tieto informácie sú uložené väčšinou v relačných databázach, zobrazujú aktuálny stav podniku a v priebehu jedného dňa sa môžu i niekoľkokrát meniť. Príkladom môže byť napr. účtovníctvo, dáta v dokumentoch obchodných prípadov a pod. Transakčné systémy realizujú ich spracovanie v reálnom čase a označujú sa ako systémy OLTP (On Line Transaction Processing). Vo vzťahu k analytickým aplikáciám sa dáta systémov OLTP chápu

ako primárne, zdrojové alebo produkčné. Na druhej strane systémy pracujúce s analytickými informáciami využívajú primárne dáta vytvorené v systémoch OLTP. Pre tieto systémy sa vzhľadom na spôsob uloženia dát a operácie s dátami vžil v osemdesiatych rokoch minulého storočia názov OLAP (On Line Analytical Processing). Avšak so zavedením pojmu Business Intelligence, ktorý vo svojej podstate kopíruje uvedený význam výrazu OLAP, a súčasne s rozvojom nástrojov a technológií na podporu analytických činností v organizácii sa výraz OLAP trochu zúžil. Väčšina odborníkov v súčasnosti chápe pojem OLAP v užšom význame, ktorý definuje OLAP čisto technologicky, teda ako „informačnú technológiu založenú predovšetkým na koncepcii multidimenzionálnych databáz, ktorej hlavným princípom je niekoľkodimenzionálna tabuľka umožňujúca rýchlo a pružne meniť jednotlivé dimenzie a tak meniť pohľady používateľov na modelovanú ekonomickú realitu“. V ďalšom texte budeme pracovať s užším - technologickým významom výrazu OLAP. [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,...), môžu to byť údaje vyexportované nejakou databázovou platformou do tzv. flat súboru a podobne. Príprava a zavádzanie údajov je dôležitou súčasťou každého riešenia dátového skladu. [1]



Obrázok 1: Dátový sklad [1]

Mali by sme mať istotu, že údaje boli predtým zabezpečené, že nedošlo k ich poškodeniu, či úmyselne alebo nie. A to sa týka nie len bežných transakčných databáz, ktorých bezpečnosť je už dostatočne vyriešená, ale aj údajov získaných z rôznych automatizovaných systémov riadenia.[6]

Údaje zo zabezpečeného operačného prostredia je potrebné pred zavedením do dátového skladu vyextrahovať, vyčistiť, upraviť a až následne vo vhodnej forme do dátového skladu zaviesť. [1]

„Dátový sklad je podnikovo štruktúrovaná úschovňa subjektovo orientovaných, integrovaných, časovo premenlivých, historických dát použitých na získavanie informácií a podporu rozhodovania.“ V dátovom sklade sú uložené atomické a sumárne dáta. [4]

Extrakcia, transformácia, zavedenie

Nástroje a postupy ETL (Extraction, Transformation, Loading) sú veľmi dôležitou súčasťou každého projektu dátového skladu. Celý proces ETL je kompletný a vo väčšine prípadov časovo pomerne náročný. U niektorých implementácií môže zabrať aj viac než polovicu celkového času úsilia a teda aj veľká časť nákladov potrebných na vytvorenie dátového skladu.

Jednotlivé etapy procesu ETL sú:

- Extrakcia – výber dát prostredníctvom rôznych metód
- Transformácia – overenie, čistenie, integrovanie a časové označenie dát
- Loading – premiestnenie dát do dátového skladu

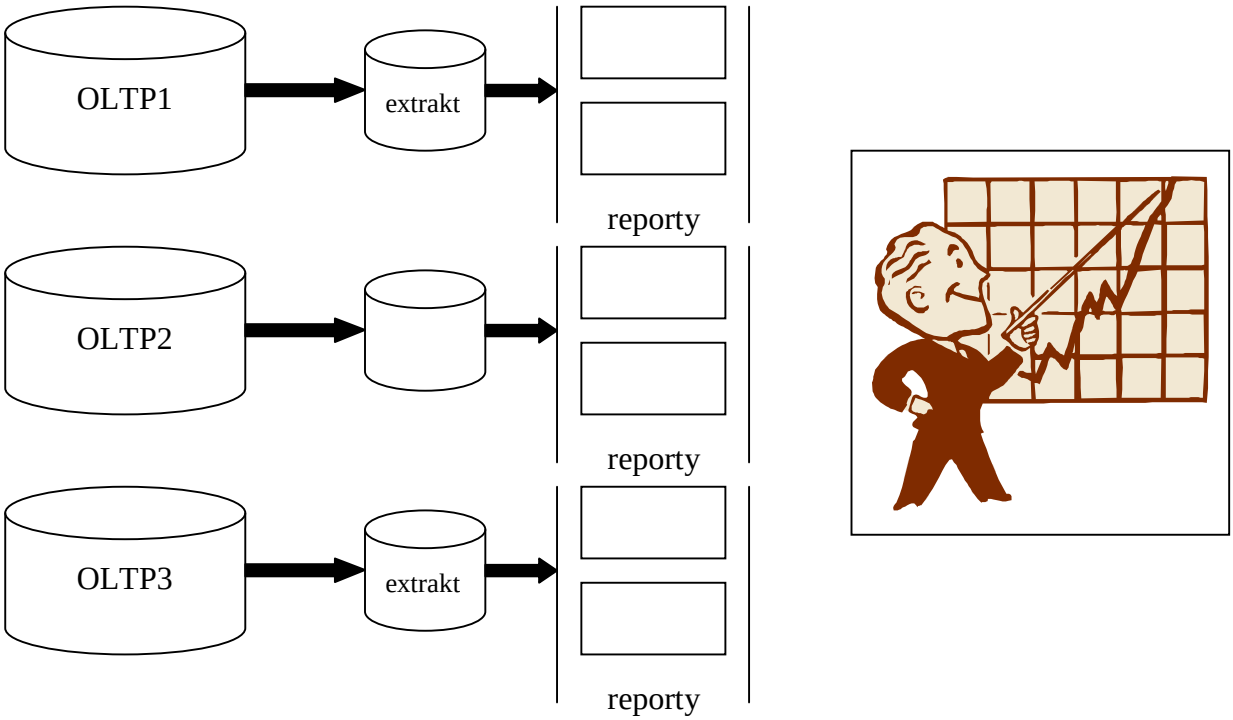
Hlavným cieľom etapy ETL je centralizácia údajov, tzn. ich zhromaždenie z mnohých spravidla nehomogénnych zdrojov z databáz OLTP a naplnenie dátového skladu určenými údajmi v požadovanom čase. Pojem ETT (Extraction, Transformation, Transport) je v inej terminológii ekvivalentný pojmu ETL. Oba pojmy popisujú rad procesov, ktorých úlohou je extrakcia údajov zo zdrojových systémov, ich transformácia a vyčistenie a prenos údajov do dátového skladu. Údaje sa teda v tejto etape nielen prenášajú, ale aj spracovávajú, napríklad indexujú, sumarizujú, zisťujú sa prípadné zmeny štruktúr zdrojových údajov potrebných pre dátový sklad, podľa potreby sa mení štruktúra kľúčov a udržujú sa meta dáta, teda dáta o dátach, v tomto prípade predpisov a definícií pre prenos a spracovanie údajov.

Počiatočným naplnením dátového skladu údajmi z operačných databáz úloha ETL samozrejme nekončí, pretože dátový sklad sa v pravidelných intervaloch plní aktualizovanými dátami. Je potrebné si uvedomiť, že na úrovni ETL sa pracuje s údajmi, z ktorých sa neskôr stanú informácie. Údaje zavedené do dátového skladu by mali byť kvalitné, presné, k veci a aktuálne, otestované vhodným spôsobom [5], aby mohli jednotlivý užívatelia

dátového skladu, napríklad manažéri, obchodníci a analytici, používať sklad účinne a efektívne. [1]

Spracovanie dát z operačného prostredia

Dáta je samozrejme možné prerobiť na informácie a následne analyzovať aj v operačnom prostredí, kde tieto dáta vznikajú. Taký postup je možný len pri málo vyťažených transakčných systémoch. Inak dochádza k neúmernému znižovaniu výkonu týchto systémov. Problém znižovania výkonu je možné čiastočne vyriešiť výberom (extrakciou) a prenesením dát z jedného prostredia do prostredia iného. Problém sa čiastočne vyriešil použitím techník spracovania extraktu, ktoré vyberajú dáta z jedného prostredia a prenášajú ich do iného prostredia, kde sa spracujú na inom hardvéry. [2]



Obrázok 2: Extrakcia dát [2]

Dáta pre extrakciu sa vyberajú podľa určitých kritérií a následne sa umiestnia spravidla do súborov alebo databáz v inom operačnom prostredí. Tieto vyextrahované dáta a výsledky analýz získaných nad týmito dátami sú potom k dispozícii analytikom a pracovníkom, ktorý riadia a rozhodujú.

Proces extrakcie bol logickým krokom od systémov OLTP k systémom na podporu rozhodovania. Dáta sa presúvajú z transakčného systému do klientskych systémov určených pre analýzu. Zdalo by sa, že extrakcia dát a spracovanie takto získaných extraktov je ideálne

riešenie, ale dochádza k mnohým problémom. Jednak môže dochádzať k mnohonásobnému vetveniu tým spôsobom, že extrahované dáta sa stanú zdrojom ďalšej extrakcie. Extrakcia dát môže úplne zamestnať kapacitu IT oddelenia organizácie, čo je tiež nežiaduce. Dochádza tiež k duplicitám, kedy sa pri extrakcii a spracovávaní extraktu zakaždým pristupuje k rovnakým údajom. Tiež flexibilita extrakcie je veľmi obmedzená. Pretože extrakty obsahujú len dáta, a nie metadáta, teda dáta o dátach, je ťažké prispôsobiť extrakciu zmenám v predmete a spôsobe podnikania. Taktiež chýba jednotná časová základňa, jednotlivé algoritmy pre transformáciu dát a výpočet požadovaných hodnôt. Prístup k externým údajom býva nekonzistentný a nie je správne definovaná ani granularita externých údajov. Zostavy vygenerované na základe extrahovaných dát tak vo väčšine prípadov obsahujú skôr dáta než informácie. [2]

Oblasť vynášania údajov

Jedná sa o akúsi „základňu“ dátového skladu, ktorou je oblasť pre vynášanie údajov. Z hľadiska implementácie sa môže jednať napríklad o pamäť operačných dát, adresár textových, alebo flat súborov, tabuľky v relačnej databázy alebo vlastné štruktúry údajov, ktoré používajú nástroje určené pre vynášanie dát. Z hľadiska princípu môžeme použiť dva modely vynášania:

- Model lokálneho vynášania
- Model vzdialeného vynášania

Výber modelu závisí na užívateľských požiadavkách a samozrejme na objeme údajov a kvalite rýchlosti pripojenia. [1]

Model lokálneho vynášania

Proces úpravy a transformácie údajov sa v tomto prípade vykonáva najskôr, a to lokálne v operačnom prostredí a až potom sa prenáša do vynášanej oblasti. Tento spôsob môže značne zaťažovať operačnú pamäť počítača, takže je ho možné využívať u menej zaťažených systémoch. [1]

Model vzdialeného vynášania

V tomto prípade sa „surové“ dáta najskôr prenesú z operačného prostredia do vynášanej oblasti, alebo dokonca priamo do prostredia dátového skladu, a až potom sa spracujú. [1]

Záver

Problematika prenosu dát z operačných prostredí je neustále aktuálna, pretože s vývojom jednotlivých prostredí a vývojom nástrojov na extrakciu je nutné riešiť aj problematiku extrakcie údajov v nových podmienkach a vyšších nárokoch na kvalitu získaných údajov. Získané dáta sú vhodné na ďalšie spracovanie pre OLAP systémy a ako ukazujú niektoré nové riešenia [7], je možné ich použitie následne aj v riadiacich systémoch a systémoch na podporu rozhodovania (DSS).

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) Čarnický Š.: Základné princípy a hlavné komponenty riešení Business Intelligence, Medzinárodná vedecká konferencia Semafor 2007, str. 53-65, (21.9.2007, Spôsob prístupu: http://semafor.euke.sk/zbornik2007/pdf/carnicky.pdf)

(4) Kebísek M., Tanuška P., Eliáš M.: Návrh a implementácia dátového skladu preotreby MTF STU Trnava - Vega 1/4078/07. Publikované v: Materials Science and Technology [online]. - ISSN 1335-9053. - Roč. 8, č. 7 (2008).

(5) Tanuska, Pavol - Vazan, Pavol - Schreiber, Peter: The Partial Proposal of Data Warehouse Testing Task. In: ISCCC 2009 : Proceedings of the 2009 International Symposium on Computing, Communication and Control, October 9-11, 2009, Singapore.IACSIT Press, 2009. - ISBN 978-9-8108-3815-7.

(6) HALENÁR, I. - KEBÍSEK, M. - KUNÍK, S.: Safety risks of information processes in automation. In: 7th International Scientific - Technical Conference - PROCESS CONTROL 2006 Kouty nad Desnou. Pardubice : University of Pardubice, 2006. ISBN 80-7194-860-8

(7) TRNKA Andrej - Proposal of application datawarehouses into control process. (Abstrakt príspevku je uverejnený v zborníku International Doctoral Seminar 2009 : Abstracts. s. 20). In: International Doctoral Seminar [elektronický zdroj] :

Proceedings : Smolenice / May 17-19, 2009 / editorial: Alena Sucáková, Dagmar Cagánová. - Trnava: AlumniPress, 2009. - ISBN 978-80-8096-088-9. - S. 343-348.

Adresa autora:

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