NÁVRH A IMPLEMENTÁCIA DÁTOVÉHO SKLADU PRE POTREBY MTF STU TRNAVA

DESIGN AND IMPLEMENTATION OF DATA WAREHOUSE FOR FMT SUT TRNAVA

Michal KEBÍSEK, Pavol TANUŠKA, Michal ELIÁŠ

Autori: ***Ing. Michal Kebísek, Doc. Ing. Pavol Tanuška, PhD., Ing. Michal Eliáš***
Pracovisko: ***Ústav aplikovanej informatiky, automatizácie a matematiky, Materiálovotechnologická fakulta STU***
Adresa: ***Paulínska 16, 917 24 Trnava***
Email: ***michal.kebisek@stuba.sk, pavol.tanuska@stuba.sk, michal.elias@stuba.sk***

Abstract

V príspevku je popísaný proces návrhu a implementácie dátového skladu. Obsahuje návrh štruktúry dátového skladu, vytvorenie OLAP kocky a popis procesu ETL.

This contribution contains description of design and implementation of data warehouse. It contains structural design DW, creating of OLAP cube and description of ETL process.

Key words

dátový sklad, ETL proces, OLAP

data warehouse, ETL (Extraction Transformation Loading) process, OLAP (On-Line Analytical Processing)

Úvod

Mnoho organizácií a spoločností prešlo určitým vývojom svojich informačných systémov. Počas tohto vývoja bol zhromaždený veľký objem informácií o procesoch prebiehajúcich v organizácii, podnikaní, trhu, klientoch a podobne. Tieto dáta možno využiť nielen k prevádzkovým potrebám organizácie, ale takisto na iné účely – k rôznym analýzam v procese rozhodovania, ako súčasť manažérskych systémov a na vyťažovanie informácií, ktoré nie sú zreteľné „na prvý“ pohľad. Systémy, ktoré to umožňujú, dostali označenie dátové sklady alebo dátové tržnice.

„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 [1].

Účel vytvorenia dátového skladu

Navrhovaný dátový sklad bude vytvorený pre potreby študijného oddelenia Materiálovotechnologickej fakulty STU v Trnave v spolupráci s Ústavom aplikovanej informatiky, automatizácie a matematiky MTF STU Trnava. Má slúžiť na uchovávanie a vyhodnocovanie informácií o uchádzačoch hlásiacich sa na túto fakultu, jej aktuálnych študentoch, a taktiež o jej absolventoch. Na základe získaných informácií bude možné vyhodnotiť napr. regionálne zloženie uchádzačov a študentov, ich študijné výsledky na strednej škole, študijné výsledky na fakulte atď.

Analýza problémovej oblasti

Dáta ukladané do dátového skladu sa budú čerpať z dátových súborov študijného oddelenia fakulty, ktoré sú založené v databázovom systéme FoxPro. Každá entita v nich má svoj vlastný dátový súbor spolu s indexovým súborom.

Za účelom rozšírenia a zlepšenia získaných informácií sa medzi dáta vkladané do dátového skladu zahrnuli aj podrobné informácie o jednotlivých stredných školách. Tieto dáta nie sú uložené v dátových súborov FoxPro, ale nachádzajú sa v štandardných súboroch programu Microsoft Excel.

Návrh štruktúry dátového skladu

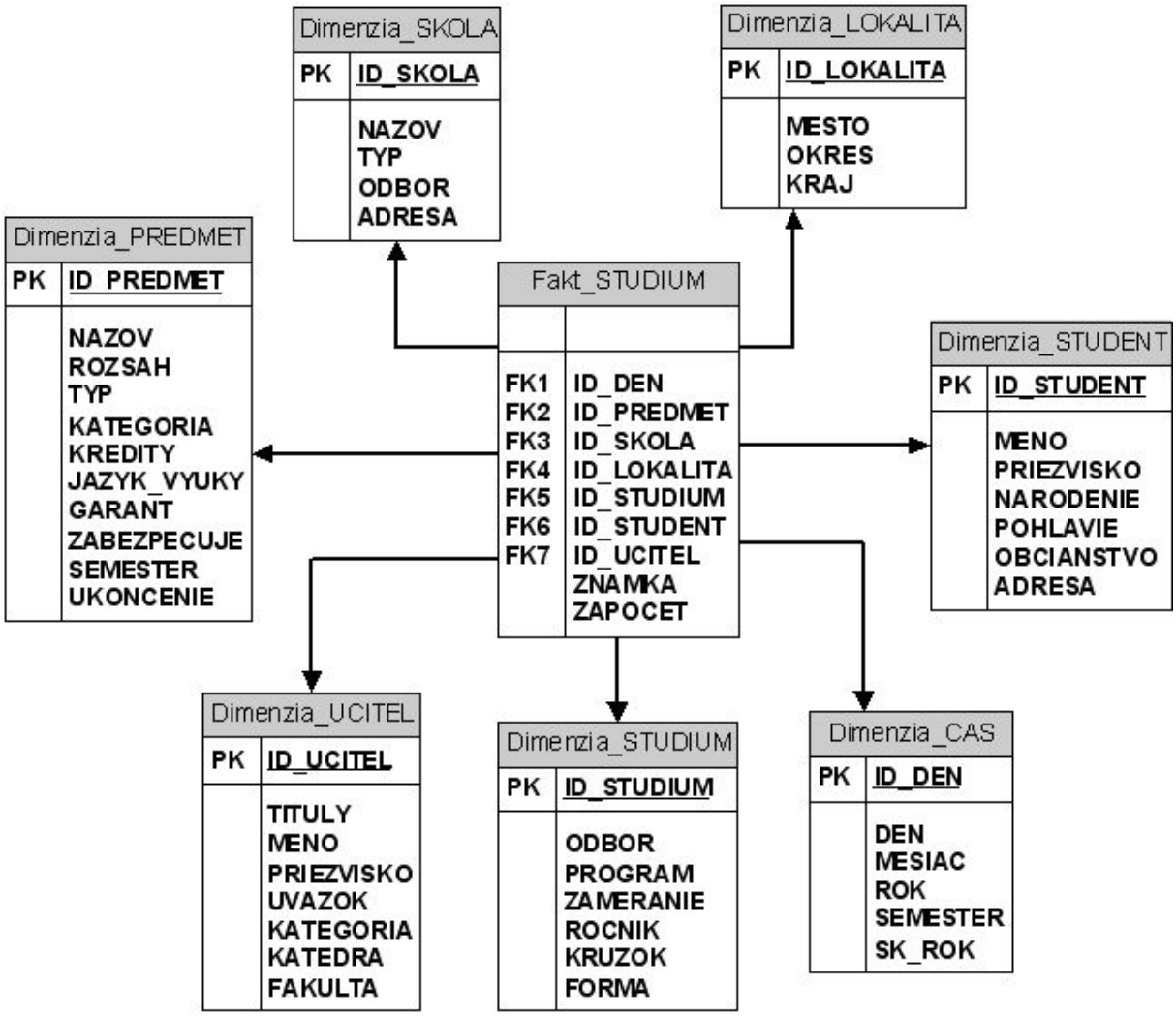
Medzi zdrojovými dátami neexistujú žiadne vzájomné väzby, tak je potrebné identifikovať vzájomné závislosti medzi jednotlivými objektmi.

Ako prvý krok je definícia dimenzií, ktoré bude obsahovať dátový sklad. Definovanie dimenzií ešte pred samotným získavaním zdrojových dát je definované z toho dôvodu, aby bolo možné zo zdrojových databáz extrahovať iba potrebné dáta, ktoré sa využijú v dátovom sklade.

Zdrojová databáza je typickým príkladom transakčnej bázy. Skladá sa z 25 entít a ku nim sú pridané vlastné entity – nová entita škola a upravená entita študent.

Pri návrhu štruktúry dátového skladu sme využili star schému, ktorá je na základe vstupných dát a požadovanej funkčnosti dátového skladu najvhodnejšia. Podstata star schémy spočíva v tom, že primárne kľúče z tabuliek dimenzií prechádzajú ako cudzie kľúče do tabuľky faktov. Centrum tvorí faktová tabuľka FAKT_STUDIUM, ktorá obsahuje referenciu do tabuliek dimenzií. Tabuľky dimenzií tvoria tabuľky obsahujúce dáta o študentoch, lebo práve ich dáta sú jeden z kľúčových faktorov dátového skladu. Ďalšie tabuľky dimenzií obsahujú dáta o predmetoch, stredných školách a regiónoch. Ako tabuľky dimenzií boli použité tabuľky Dimenzia_LOKALITA, Dimenzia_PREDMET, Dimenzia_SKOLA, Dimenzia_STUDENT, Dimenzia_STUDIUM, Dimenzia_CAS a Dimenzia_UCITEL.

Pri návrhu dimenzií sme zobrali do úvahy aj hierarchickú usporiadanosť dát charakteristickú pre každú dimenziu.



Obr. 1. *Štruktúra navrhnutého dátového skladu*

Dáta, resp. dátové objekty dátového skladu, budú umiestnené v samostatnej schéme UMIS (University Management Information System). Táto schéma bude kľúčovou schémou dátového skladu a budú sa v nej nachádzať objekty dimenzií, tabuľky faktov a metadáta. Implementácia dátového skladu bude v systéme Oracle 9i R2.

Vytvorenie OLAP kocky

Prvým krokom ku vytvoreniu OLAP kocky je vytvorenie dimenzií. V navrhnutom dátovom sklade je potrebné vytvoriť jednotlivé tabuľky dimenzií. Samotné dimenzie sme pri vytváraní rozčlenili do hierarchicky usporiadaných stupňov.

Usporiadanie do levelov je vhodné z dôvodu uskutočnenia drill-down prehľadávania hlbšie cez dimenzie.

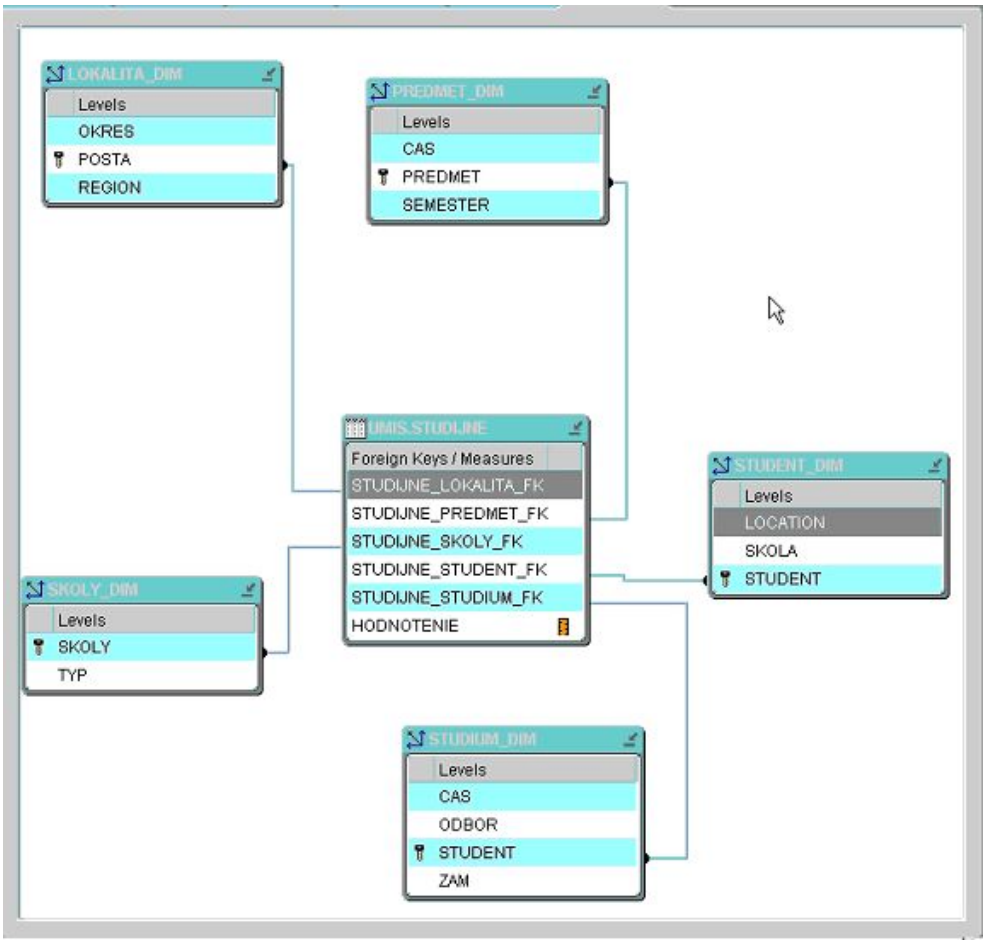
Príklad vytvorenia dimenzie STUDENT_DIM:

Dimenzia je vytvorená z dátového objektu UMIS_STUDENT. Dimenzia je hierarchicky usporiadaná do troch levelov. Najnižší level usporiadania je level STUDENT, ktorý obsahuje detailné informácie o študentoch. Druhým levelom je level SKOLA, odkiaľ študent prišiel na

vysokú školu. Najvyšším levelom je level LOKACIA, kde môžeme dáta získavať podľa lokácií. Zdrojový text dimenzie STUDENT_DIM je v nasledovnom výpise:

```
create dimension student_dim
    level student  is (umis_student.osc)
    level skola    is (umis_student.ss_kod)
    level location is (umis_student.nr_msto)
    hierarchy student_rollup
    (
      student child of
      skola   child of location
    )
    attribute student determines
    (
      umis_student.osc, umis_student.meno,
      umis_student.priezv, umis_student.rodc,
      umis_student.pohl, umis_student.jazyk,
      umis_student.dzapis, umis_student.dukon
    )
    attribute skola determines
    (
      umis_student.ss_kod,
      umis_student.ss_priem
    )
    attribute location determines
    (
      umis_student.nr_msto
    );
```

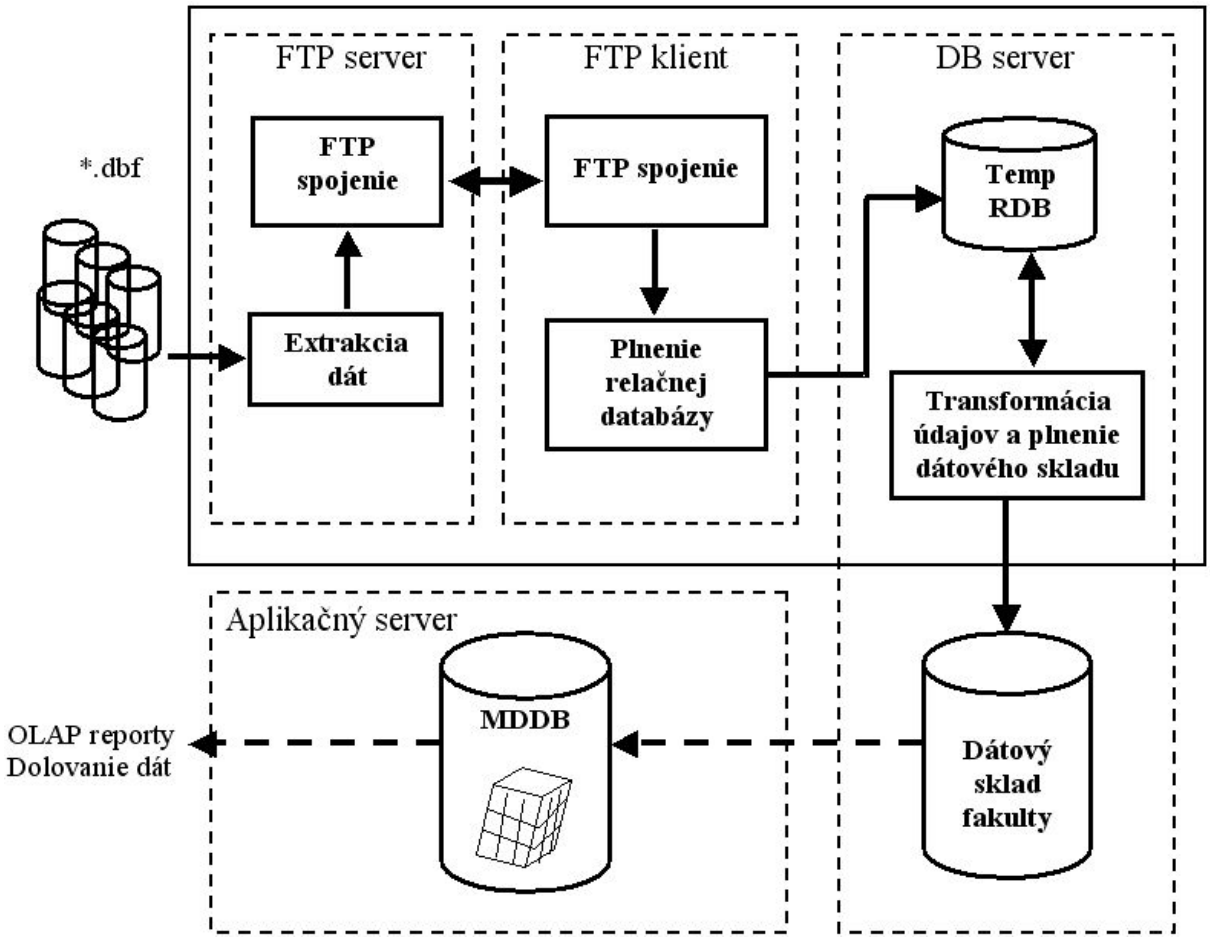
Po vytvorení dimenzií je možné vytvoriť OLAP kocku. Na obrázku je zobrazená topológia dátového skladu, kde v tabuľkách faktov sú zobrazené jednotlivé agregačné stupne.



Obr. 2. *Topológia dátového skladu*

Proces ETL

Proces napĺňania dát do dátového skladu je jedným z najdôležitejších procesov pri budovaní dátového skladu. Všetky dáta, ktoré sa ukladajú do dátových skladov, sú získavané z primárnych prevádzkových systémov, kde nie sú dáta tematicky usporiadané a poskytujú popis len pre jednotlivé transakcie. Údaje z transakčného prostredia je potrebné pred zavedením do dátového skladu vyextrahovať, vyčistiť, upraviť a následne vo vhodnej forme zaviesť do dátového skladu.



Obr. 3. *Proces ETL*

Na začiatku je potrebné vytvoriť pohľad do zdrojových tabuliek jednotlivých databáz s určitým stupňom agregácie zdrojových dát a potrebné objekty, do ktorých sa budú transformovať zdrojové dáta. Na tvorbu objektov je vytvorený balík zdrojových kódov, ktorý zabezpečuje vytváranie cieľových objektov, trigrov a indexov, a taktiež i zapísanie potrebných údajov do meta tabuliek. Do meta tabuliek sa zapíšu hodnoty, ktoré budú slúžiť na získavanie a transformácie dát zo zdrojových databáz do cieľových schém. Proces transformácie zdrojových dát prebieha cez množstvo procedúr, funkcií a trigrov. Po naplnení interfaceových štruktúr zo zdrojových databáz budú dáta pripravené na presun do dátového skladu. Proces presunu dát medzi interfaceovou schémou a schémou dátového skladu je založený na príznaku transakčného statusu. Každý nový záznam, ktorý sa naplní do dátového skladu, dostane príznak nový. Pri úspešnom spracovaní záznamu cieľovou schémou sa príznak záznamu zmení na spracovaný. V prípade, že sa pri spracovaní záznamu vyskytne chyba, v zázname sa nastaví príznak chyby. Tento spôsob spracovania je jednoduchý a rýchly na spracovávanie záznamov a detekciu chýb, ktoré sa vyskytnú počas spracovania.

Po pripravení dát zdrojovou schémou bude možné uskutočniť proces plnenia dát do dátového skladu. Proces, ktoré dáta sa do dátového skladu plnia a ktoré nie, zabezpečí meta tabuľka, v ktorej sa budú nachádzať dáta o všetkých zdrojových dátach i s dátumom posledného zberu dát. Princíp prepočítavania, ktoré dáta sa budú, a ktoré sa nebudú zbierať, spočíva v prepočítavaní posledného zberu dáta s frekvenciou zberu a porovnaním s aktuálnym dátumom. V momente, keď systém pripraví dáta v zdrojovej schéme, budú dáta pripravené na zber do cieľových štruktúr dátového skladu. Zber zdrojových dát zabezpečí ďalší balík

zdrojových kódov. Celý proces zberu a kontroly, ktoré dáta je potrebné zbierať, bude zabezpečený jednou plánovanou databázovou úlohou. V prípade, že zdrojový balík zistí, že nie sú žiadne dáta, ktoré by sa mali v danom cykle zbierať, tak zdrojový balík už ďalej nepokračuje v zbere a celý proces sa ukončí. Táto funkcionalita zabezpečí, aby sa zbytočne nealokovali systémové prostriedky.

Celý systém zberu dát bude veľmi variabilný. Možnosť variability spočíva v tom, že nebude potrebné dopĺňať zdrojový kód, ak je nutné pridať nové dáta pre zber, ale nové dátové štruktúry sa zapisujú iba do metadát a odtiaľ sa vytvoria potrebné dátové objekty, ktoré zabezpečujú samotné plnenie dátového skladu.

Vzhľadom na povahu spracovávaných dát nebude potrebné vykonávať zber dát pre dátový sklad v kratších intervaloch, ako je jeden mesiac.

Po zavedení dát do cieľových štruktúr dátového skladu bude už možné uskutočňovať nad dátami potrebné analýzy a vytvárať reporty.

Zhodnotenie navrhovaného riešenia

Problematika dátových skladov sa v súčasnosti stáva stále rozšírenejšou a dátové sklady nie sú len doménou veľkých bánk a finančných inštitúcií, ale dostávajú sa do poľa pozornosti aj stredne veľkých firiem a vzdelávacích inštitúcií.

Navrhovaný dátový sklad umožní fakulte získať nové poznatky o prihlásených uchádzačoch, aktuálnych študentoch ale aj o jej absolventoch.

Nad dátovým skladom bude možné vytvárať prehľady napríklad o regionálnom rozložení a typoch absolvovaných stredných škôl s ohľadom na dosiahnuté študijné výsledky. Zistiť, z ktorých regiónov majú študenti lepšie výsledky v jednotlivých vedných odboroch (technické predmety, prírodovedné, humanitné...). Aký vplyv má typ absolvovanej strednej školy na dosiahnuté výsledky na vysokej škole, prípadne na jej neabsolvovanie. Umožní zistiť najvhodnejšie regióny a typy škôl pre vedenie náboru na získanie svojich budúcich študentov a podobne.

Dátový sklad by mal fakulte poskytnúť nový pohľad na dáta, ktoré vlastní a na ich základe aj nový pohľad na jej budúcich, aktuálnych i minulých študentov.

Tento príspevok bol podporovaný **grantovou agentúrou VEGA v rámci projektu číslo 1/4078/07**, za čo všetci autori vyslovujú poďakovanie.

Zoznam bibliografických odkazov:

[1] INMON, W. H. *Building the Data Warehouse*. Wiley Computer Publishing, 2002. ISBN 0-471-08130-2
[2] LACKO, L. *Datové sklady, analýza OLAP a dolování dat s příklady SQL Servera Oracle*. Computer Press, 2003. ISBN 80-7226-969-0
[3] LEGERSKÝ, M. *Realizácia dátovej pumpy pre dátový sklad fakulty*. Diplomová práca. Trnava MtF STU, 2005.