FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
UNIVERZITY KOMENSKÉHO V BRATISLAVE

Multidimenzionálny databázový model a OLAP

DIPLOMOVÁ PRÁCA

Juraj Fehér **2007**

Multidimenzionálny databázový model a OLAP

DIPLOMOVÁ PRÁCA

Juraj Fehér

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky
Katedra informatiky

Informatika

Školiteľ diplomovej práce
RNDr. Ján Šturc, CSc.

BRATISLAVA 2007

Čestne prehlasujem, že túto diplomovú prácu som vypracoval samostatne a použil som iba literatúru a elektronické dokumenty uvedené v zozname na konci práce.

V Bratislave, Máj 2007

.....

Ďakujem svojmu diplomovému vedúcemu RNDr. Ján Šturcovi, CSc. za cenné rady, a pomoc pri písaní diplomovej práce. Taktiež chcem poďakovať svojim blízkym za ich podporu počas písania tejto práce.

Abstrakt

Hlavnou témou diplomovej práce je multidimenzionálny databázový model a OLAP analýza. Diplomová práca sa zaoberá transformáciou dát a znalostí do dát štruktúrovaných použitím sady konceptov, rôznych metód, multidimenzionálnych štruktúr a multidimenzionálnych databáz. Taktiež sa zaoberá OLAP analýzou a jej možnosťami.

Mojim cieľom je predviesť štruktúru multidimenzionálnych databáz, ich použitie, ukázať funkčnosť a formálne zadefinovanie. Porovnať s fungovaním transakčných databáz, ich účel a vhodnosť použitia. Zadefinovať OLAP, porovnať varianty OLAP, analyzovať nástroje OLAP, ich architektúru, funkcionalitu, použitie a ich zdokonalenie vzhľadom na predchádzajúce verzie. Porovnať silu a podporu rozhodovania na databázových systémoch.

Kľúčové slová: Business Intelligence, data warehouse, on-line analytical processing, OLAP, hypercube.

Obsah

Zoznam obrázkov

Zoznam tabuliek

Zoznam skratiek

AMO	Analysis Management Objects
AS	Analysis Services
AW	Analytic Workspace
DB	Database
DBMS	Database Managment System, Systém riadenia bázy dát
DSV	Data Source Views
DWH	Data Warehouse, Dátový sklad
ETL	Extraction, Transformation, Loading
ETT	Extraction, Transformation, Transport
HOLAP	Hybridný OLAP
KPI	Key Performance Indicator, kľúčové indikátory
MDBMS	Multidimensional Database Management System
MDX	Multidimensional Expressions
MOLAP	Multidimenzionálny OLAP
ODS	Operational Data Store, Sklad prevádzkových dát
OLAP	On-Line Analytical Processing
OLTP	On-Line Transaction Processing, Transakčné databázy
BI	Business Intelligence
OWB	Oracle Warehouse Builder
RDBMS	Relačné databázové systémy
RDL	Report Definition Language
ROLAP	Relačný OLAP
SSIS	SQL Server Integration Services
UDM	Unified Dimensional Model
XMLA	XML for Analysis

Úvod

Popri procese zhromažďovania veľkého množstva údajov, prichádza potreba tieto údaje analyzovať, získavať z nich informácie v požadovanom čase a v správnej forme, dostupnej nielen pre odborných analytikov a ľudí s informatickým vzdelaním, ale aj pre užívateľov z oblasti riadenia a manažmentu. Získanie pravdivých a relevantných informácii v správnom čase sa stáva nevyhnutnou požiadavkou pri strategickom rozhodovaní a podpore rozhodovania. Pod názvom OLAP sú zahrnuté technológie, metódy a prostriedky, ktoré umožňujú ad-hoc analýzu údajov multidimenzionálneho charakteru. OLAP umožňuje užívateľovi pracovať s údajmi veľmi flexibilne a analyzuje dáta z mnohých hľadísk. OLAP je rozšírenou súčasťou Business Intelligence a oblasť aplikácie nájde v obchodnom prostredí, manažérskom rozhodovaní, finančnom sektore, pri zostavovaní rozpočtu, odhadovaní a analýzy trendov rôznych veličín a podobných oblastí. Databáza určená pre služby OLAP využíva multidimenzionálny databázový model, ktorý svojou štruktúrou dovoľuje komplexné analýzy a ad-hoc dotazovanie s veľmi rýchlou odozvou. Existuje veľa OLAP nástrojov slúžiacich na tvorbu zostáv, analýzu, reportovanie. Podpora pre tieto nástroje je v súčasnosti plne implementovaná.

Táto práca mapuje súčasný stav problematiky, implementáciu v praxi a porovnáva multidimenzionálny model a OLAP nástroje z rôznych aspektov. V prvých kapitolách prinášam ucelený súbor definícii a prehľad o problematike. Mapujem rozdiely v oblasti využitia transakčných databáz a dátových skladov. Definujem logický multidimenzionálny databázový model a porovnávam ho s relačným. Taktiež definujem OLAP, oblasti použitia a porovnávam implementačné varianty OLAP.

V druhej časti práce sa sústredím na OLAP produkty a nástroje, konkrétne ich podporu v databázových systémoch Microsoft a Oracle. Porovnávam ich architektúru, funkcionalitu, využitie. Taktiež rozoberám rozdiely vzhľadom na predchádzajúce verzie a podporu oblasti Business Intelligence. Dávam dôraz na sadu samostatných OLAP nástrojov ako Cognos a venujem sa aj výkonu databázových systémov vzhľadom na podporu rozhodovania a výkon analytického spracovania.

1 Business Intelligence

Termín Business Intelligence (BI) prvýkrát v roku 1989 definoval Howard Dresner zo spoločnosti Gartner Group ako „množinu konceptov a metodík, ktoré zlepšujú rozhodovací proces za použitia metrík alebo systémov založených na metrikách.“. Aj keď táto oblasť IT nie je príliš stará, využitie nástrojov BI sa rapídne rozvíja. Za posledné roky zaznamenala značný architektonický, ale aj vecný rozvoj.

Business Intelligence je stratégia práce s informáciami, proces získavania, ukladania, analýzy, správy dát. Je to proces transformácie údajov na informácie a prevod týchto informácií na poznatky. Účelom Business Intelligence je teda konvertovať veľké objemy údajov na poznatky, ktoré sú potrebné pre koncového užívateľa. Tieto poznatky môžeme potom efektívne využiť napríklad v procese rozhodovania.

Poznatky (informácie) takto získané nie sú len konkrétne záznamy, alebo množiny záznamov. Môžu to byť aj poznatky získané z pozorovania nejakej veličiny z manažérskeho prostredia, alebo poznatky určujúce závislosti medzi údajmi. Preto moderné databázové servery obsahujú rozsiahlu podporu pre budovanie dátových skladov (data warehouse), analýzy OLAP (Online Analytical Processing) a data mining (dolovanie, odkrývanie dát).

Systémy BI poskytujú hlboké zákaznícke analýzy nad veľkými objemami dát. Pomocou dataminingových modelov vieme napríklad predpovedať z histórie chovania klientov, ktoré máme uložené vo forme dát, chovanie konkrétneho klienta v danej situácii. „Napríklad je možné identifikovať, kedy klient zvažuje odísť ku konkurencii, odhaliť pokus o bankový podvod, či ohraničiť cieľovú skupinu pre ponuku konkrétneho produktu alebo služieb“ Petr Železník z firmy SAP[1].

1.1 Hierarchia informačných úrovní

Základom všetkého sú údaje. Údaje obsahujú len jednoduché fakty, ale pritom vieme, že niekde vo vnútri množine údajov sú ukryté určité informácie. Tieto informácie však odhalíme až vtedy, keď pridáme k dátam súvislosti. Keď navyše do hry vstupuje okrem informácií aj tvorivá inteligencia, tak získavame znalosti. Keď znalosti

zovšeobecníme, získavame “múdrosť”, čo znamená schopnosť presného zhodnotenia znalostí a ich následne uplatnenie v reálnej praxi.[2]

Obr. 1 Hierarchia informačných úrovní

1.2 Základné zložky Business Intelligence

Medzi základné komponenty Business Intelligence patrí:

- Zdrojové systémy, transakčné systémy, kontrolné systémy - tvoria primárne alebo produkčné databázy - sú zdrojom dát pre BI, presné rozdelenie a porovnanie si ukážeme neskôr.
- Prostriedky výberu, transformácie a prenosu dát z produkčných databáz do multidimenzionálnych štruktúr - dátové pumpy.
- Integrácia zdrojových systémov - prenosy dát medzi zdrojmi a multidimenzionálnymi databázami v reálnom čase. Je nutné čerpať a pracovať s aktuálnymi dátami ihneď.
- Uloženie netransformovaných dát, zníženie dopadu prevádzky BI na výkon zdrojových systémov, optimalizácia.

1.3 Nástroje Business Intelligence

Nástroje, ktoré využíva Business Intelligence, sú softwarové aplikácie navrhnuté pre podporu procesu BI. Väčšinou to sú hlavne nástroje, ktoré pomáhajú pri analýze a prezentácii dát. Niektoré nástroje však obsahujú ETL funkcionalitu (Extract, Transform, Load čo v preklade znamená výber, transformácia a prenos dát).

Existujú nasledovné typy nástrojov BI:

- Online Analytical Processing, známe ako OLAP (vrátane HOLAP, ROLAP, MOLAP). Sústredia sa na možnosti manažmentu a podpory rozhodovania. Sú to výkonné informačné systémy, ktoré podporujú analýzu veľkého množstva dát s použitím rôznych perspektív.[3]
- Reporting Software, nástroje na prácu so spracovanými dátami, na definovanie zostáv a podoby výstupov pre užívateľa.
- Data Mining, analytická metóda získavania netriviálnych skrytých a potenciálne užitočných informácii z dát. Niekedy sa chápe ako analytická súčasť dobývania znalostí z databázy.[4]
- Business Performance Management je novou generáciou Business Intelligence. Pomáha efektívne využívať finančné, ľudské, materiálne a iné zdroje. Je to množina procesov pomáhajúcich k optimalizácii výkonnosti rozhodovania. Predstavuje sústavu, rámec pre organizovanie, automatizáciu a analýzu metodológii, metrík, procesov a systémov, ktoré sa zúčastňujú na procese rozhodovania. [5]

1.4 Produkty Business Intelligence

Medzi Open Source produkty patria: *Openl, Pentaho, YALE*. Vývojom komerčných BI produktov sa zaoberajú napríklad: *ACE COMM, Actuate, Applix, Business Objects, Cognos, ComArch, Dimensional Insight, Hyperion Solutions Corporation, MaxQ Technologies, MicroStrategy, Oco, Oracle Corporation, OutlookSoft, Panorama Software, Pentaho, PROPHIX, Pilot Software, Inc., QlikView, SAP Business Information Warehouse, SAS Institute, Siebel Systems, SPSS, Teradata.*

2 Dátové sklady (Data Warehouse)

Pojem Data Warehouse môžeme definovať ako stratégiu prístupu k informáciám, ktoré sú určené pre rozsiahle analýzy. Jedná sa o integrované, subjektovo orientované, stále a časovo rozšíriteľné údaje usporiadané pre podporu potrieb manažmentu. Údaje sa v dátových skladoch nevytvárajú, ale prenášajú z jednotlivých aplikácií. Hovoríme o historických, agregovaných a priebežne rozširovaných údajoch.

Od vybudovania dátových skladov sa očakáva výrazne vyšší výkon spracovania dát a informácii na všetkých stupňoch riadenia. Zmyslom vybudovania dátového skladu je vytvoriť v organizácii jednotnú, homogénnu a konzistentnú dátovú základňu, nad ktorou sa potom dajú efektívne používať nástroje a systémy na podporu rozhodovania.

Data Warehouse (DWH) je súčasťou väčšieho celku analytického spracovania údajov – OLAP (On-Line Analytical Processing), ktorého časťou sú mechanizmy umožňujúce vyhľadávanie a analýzu informácií bez vedomostí špeciálnych postupov (SQL dotazy) – EIS (Executive Information System), DSS (Decission Support System) a MIS (Management Information System).

2.1 Klasifikácia informačných potrieb

Informačné potreby užívateľov informačných systémov môžeme vo všeobecnosti rozdeliť do dvoch úrovní.[6]

2.1.1 Operatívna úroveň

Tiež niekedy nazývaná ako prevádzková úroveň. Na tejto úrovni sa nachádzajú všetky potreby každodennej prevádzky systému ako napríklad príjem objednávok, vedenie skladových zásob, výroba, účtovníctvo, personálne a mzdové agendy atď. Tieto aplikácie pracujú v princípe pomocou ad hoc transakcií a preto spracovanie na tejto úrovni tiež nazývame OLTP (On-Line Transaction Processing). Databázy, v ktorých sa realizuje táto úroveň nazývame operatívne, transakčné, alebo prevádzkové databázy.

2.1.2 Dispozičná úroveň

Dispozičná úroveň je nadradená úrovni operatívnej úrovni hlavne z hľadiska dlhodobého charakteru riadenia organizácie. K tomu je potrebné mať databázu obsahujúcu istý časový horizont a rôzne kumulácie a agregácie. Rozhodovanie na tejto úrovni je vykonávané využitím rôznych modelov, analýz premieňaním dát na informácie, na základe ktorých je možné vykonať správne rozhodnutia. Spracovanie na tejto úrovni nazývame OLAP (On-Line Analytical Processing). Dôležitým znakom tejto úrovne je charakter prístupu užívateľov – iba čítanie.

2.2 Klasifikácia databázových systémov

Nasleduje stručný prehľad od transakčných systémov cez operačno-taktické pre podporu rozhodovania, až po informačné systémy pre podporu rozhodovania vrcholového riadenia. Systémy vyššej úrovni v sebe zahrňujú aj systémy nižšej úrovni.[2]

2.2.1 Systémy OLTP (Online Transaction Processing)

Sú to relačné databázy, v ktorých vykonávame veľké množstvo transakcií v reálnom čase. Nazývame ich tiež operatívne, produkčné alebo transakčné databázy. Transakčné databázy, do ktorých sa ukladajú aktuálne operatívne údaje, sú organizované ako relačné, čo znamená, že údaje sú uložené v databázových tabuľkách, medzi ktorými sú relačné vzťahy vyplývajúce z aplikačnej logiky.

Primárnym cieľom transakčných databázových systémov je umožniť užívateľom databázového serveru vykonávanie veľkého množstva transakcií online (napríklad obchodných, bankových a pod.) Cieľom transakčných databázových systémov je automatizácia činností, ktoré sa opakujú. Parí sem zautomatizovanie bežných úloh ako napríklad vedenie účtovníctva, spracovanie miezd a platov, evidenčné systémy atď. Typickou vlastnosťou týchto systémov je, že veľká časť celkového spracovania je vykonávaná už pri vkladaní dát resp. tesne po ňom. K zdroju údajov v rovnakom čase pristupujú užívatelia, ktorí údaje z databázy čítajú, zapisujú, prípadne vykonávajú jednoduchšie analýzy. Údaje v transakčných databázach by mali byť uložené v normalizovaných tabuľkách, ktoré vyhovujú aspoň podmienkam 2NF alebo 3NF. To

má za dôsledok veľa atomických, relačne zviazaných tabuliek. Práve preto je analýza veľkého množstva takto uložených údajov často neefektívna, pomalá a tiež ťažšie vytvoriteľná.

V transakčných systémoch je využívaný princíp relačných databáz. Medzi hlavné obmedzenia relačných databáz patrí absencia komplexných analytických nástrojov a potenciálne obmedzenie údajov, ku ktorým je možné v rozumnom čase pristupovať.

Aj keď sa v praxi často zamieňajú pojmy MIS, DSS, EIS, existujú jemné odchýlky medzi nimi. Do Business Intelligence zahrňujeme MIS, DSS, OLAP aj EIS systémy.

2.2.2 Systémy MIS (Management Information Systems)

Systémy pre podporu riadenia a rozhodovania. Nasadzujú sa na operatívno-taktickej úrovni. Hlavnou úlohou je poskytovanie kvalitných informácii pre riadiacich pracovníkov vo forme komplexných prehľadov a agregovaných zostáv podľa rôznych časových, priestorových a iných hľadísk. Ide napríklad o vyhodnocovanie a analýzu minulých období, ako aj prognózovanie budúceho vývoja, pričom informácie spĺňajú požiadavky na rýchlosť, kvalitu a orientáciu na trh. Pôvodné MIS požiadavky na zostavu sa odosielajú vývojovému teamu, ktorý zostavu vytvorí a poskytne, ale až po určitom čase. V tomto období už však zostava nemusí byť aktuálna a použiteľná.

2.2.3 Systémy DSS (Decision-Support Systems)

Ide o systémy pre podporu rozhodovania. Sú nadstavbou MIS a na rozhraní taktického a strategického rozhodovania. Poskytujú výsledky rôznorodých zložitých analýz. Využitie majú v procese rozhodovania riadiacich pracovníkov.

2.2.4 Systémy EIS (Executive Information Systems)

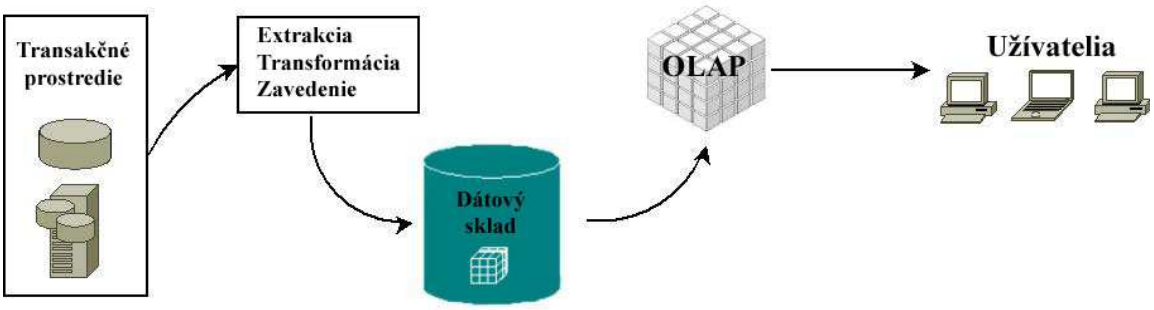
Informačné systémy pre vrcholové riadenie na strategickej úrovni. Zastrešujú všetky predošlé. Ich úlohou je získavanie podkladov pre strategické rozhodovanie vo finančnej a personálnej oblasti. Pojem EIS postupne nahradzuje pojem DSS.

2.2.5 OLAP (Online Analytical Processing)

Ide o informačné systémy pre analýzu veľkého množstva údajov. Výsledkom analýzy sú súhrny a reporty slúžiace ako podklad pre rozhodovanie a riadenie procesov. OLAP systémom sa budem viac venovať v ďalších kapitolách.

2.3 Dátový sklad (Data Warehouse)

Data Warehouse možno označiť ako centrálny podporný systém, ktorý obsahuje údaje z rôznych interných a externých zdrojov, zhromažďuje ich, vytvára medzi nimi vzťahy, a tým pôsobí ako databanka pre ostatné systémy riadenia. DWH môže poskytnúť len také informácie, ktoré získal zo svojich zdrojov a závisí od kvality jednotlivých údajov a ich zdrojov, a nie od použitých prostriedkov. Vytvára sa individuálne na báze existujúcich informačných systémov a dôležitých informácií.



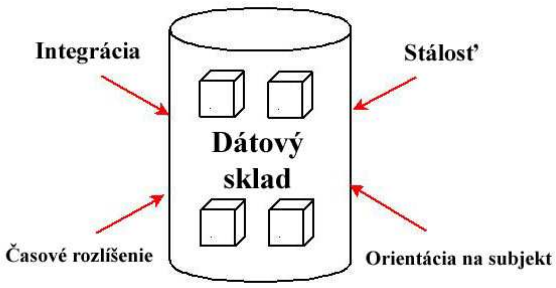
Obr. 2 Dátový sklad

Údaje sa získavajú a ukladajú do transakčných (operatívnych) databáz, ktoré môžu byť v rôznych oddeleniach firiem, pripadne rozličných lokalitách. Tieto údaje sa v pravidelných intervaloch zozbierajú, predspracujú a zavedú do dátového skladu.

2.3.1 Definícia dátového skladu

Koncom 80-tych rokov William Inmon (považovaný za otca dátových skladov) definoval dátový sklad nasledovne:

Definícia: Dátový sklad je subjektovo orientovaná, integrovaná, časovo rozlíšená a stála zbierka dát určená na podporu procesu manažérskeho rozhodovania. [6]



Obr. 3 Definícia dátového skladu

2.3.1.1 Orientácia na subjekt (Subject-oriented)

Pri orientácii na subjekt sú dáta kategorizované podľa subjektu, ktorým môže byť napr. zákazník, dodávateľ, zamestnanec, výrobok a podobne. Inak povedané, všetky dáta v databáze sú organizované tak, že elementy súvisiace s rovnakou udalosťou, subjektom, sú spojene dohromady. Subjekty sú ohraničené organizačnými a procesnými hranicami a vyžadujú informácie z viacerých zdrojov.

2.3.1.2 Integrácia (Integrated)

Dáta v dátovom sklade musia byť jednotné a integrovateľné. To znamená, že údaje týkajúce sa konkrétneho predmetu sa do dátového skladu ukladajú len raz. Medzi prostriedky ako docieliť integráciu patrí konzistentné názvoslovie, konzistentné jednotky meraných veličín, konzistentné kódovacie štruktúry, konzistentné domény a formáty atribútov. Vo fáze prípravy a zavedenia musia byť dáta upravené, vyčistené a zjednotené. Pre dátového analytika sú nedôveryhodné a nekonzistentné dáta neprípustné. Dáta v dátových skladoch sú teda konzistentné a kvalitné skôr, než sú sprístupnené užívateľom.

2.3.1.3 Časové rozlíšenie (Time-variant)

Údaje v dátovom sklade reprezentujú určitý časový úsek a sú platné v istom časovom momente. Táto charakteristika je rozdielna od operatívnych dát (z transakčných databáz), ktoré musia byť platné v momente prístupu. Časový horizont údajov v dátovom sklade je dlhodobejší (roky) na rozdiel od operatívnej úrovne (dni, mesiace). Štruktúra kľúčových atribútov obsahuje aj element času, ktorý v operatívnej databáze nemusí byť

uvádzaný. Toto dáva možnosť analyzovať v správach (reportoch, zostavách) zmeny údajov za určité časové obdobie. Po tom, ako je v dátovom sklade zaznamenaná konkrétna snímka dát z operatívnej databázy, zvyčajne sa tieto dáta v dátovom sklade už nemodifikujú ani nemažú, ale pridávajú sa iba nové snímky.

2.3.1.4 Stálosť (Non-volatile)

Na operatívnej úrovni sú do databázy údaje pravidelne vkladané, mazané a modifikované, a to záznam po zázname. Pri dátových skladoch sa ale nemenia ani neodstraňujú, iba sa v pravidelných intervaloch pridávajú nové údaje. Neexistujú teda žiadne operácie zmeny dát počas normálneho spracovania. Z toho vyplýva, že väčšina metód pre optimalizáciu údajov a transakčný prístup k údajom je v dátovom sklade nepotrebná.

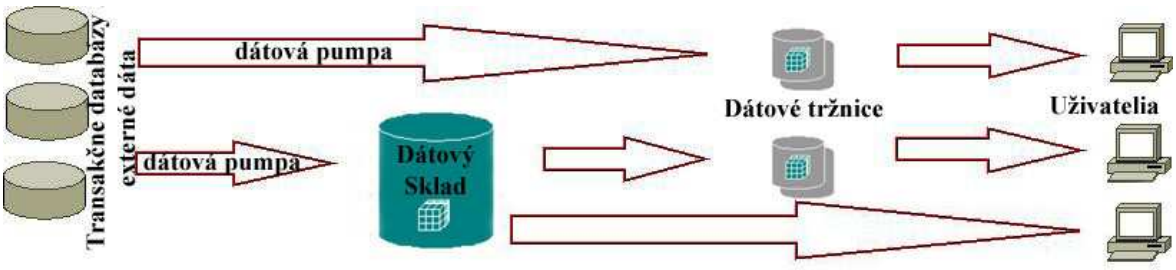
2.3.2 Dátové Tržnice (Data mart)

Definícia: Dátová tržnica, DM, Data mart je špecifický, subjektovo orientovaný sklad dát, navrhnutý na základe špecifických požiadaviek menšej množiny užívateľov.[8]

Sú to teda určité presne špecifikované oddelené podmnožiny dátového skladu, určené pre istú skupinu ľudí (menšie organizačné zložky firmy). Dáta pre dátové tržnice sú vyberané s cieľom vyhovieť špecifickým požiadavkám častí organizácie. Dátové tržnice sú často preferované podnikom ako prvý krok k vybudovaniu dátového skladu a môžu byť použité ako dôkaz správnosti koncepcie dátového skladu.

Dátová tržnica je špeciálna verzia dátového skladu, ktorá tiež obsahuje snímok operatívnych dát. Základný rozdiel je v tom, že pri vytváraní dátovej tržnice sa zameriava na špecifické, preddefinované potreby konkrétnych užívateľov a konfiguráciu dát. Dáta uložené v dátových tržniciach je možné používať predovšetkým ako podklad pre cielené analýzy.

Výhoda tohto usporiadania na viac samostatných dátových skladíšť je jednoduchšia a rýchlejšia implementácia a z toho vyplývajúce rýchlejšie prínosy pre užívateľa. Na druhú stranu, medzi nevýhody patrí, že môže dochádzať k nekonzistencii medzi jednotlivými tržnicami. Táto schéma je tiež náročnejšia na údržbu.



Obr. 4 Dátové tržnice

2.3.3 Sklad prevádzkových dát (Operational Data Store)

Definícia: Sklad prevádzkových dát je architektonická budova, kde sú hromadne uložené zjednotené dáta.[12]

Toto uloženie môže byť tiež definované ako množina zjednotených databáz navrhnutých k podpore prevádzkového sledovania. Na rozdiel od databáz pre aplikácie OLTP, ktoré sú orientované transakčne s dôrazom na funkcionalitu, ODS obsahuje predmetovo orientované celopodnikové dáta. Dáta v ODS sú nestále, aktuálne a podrobné. Poskytujú zjednotený pohľad na dáta v prevádzkových systémoch a aktuálny pohľad na prevádzku. Dáta sú transformované a integrované do konzistentného zjednoteného celku z pôvodných a ďalších prevádzkových systémov. Dáta v ODS sú pravidelne, v reálnom čase a nepretržite obnovované, takže výsledný obraz zodpovedá poslednému aktuálnemu stavu. Dátové sklady na rozdiel od ODS obsahujú historické snímky dát pre porovnanie v rámci rôznych časových období. [13]

	DWH	**ODS**
Účel	Podpora manažérskeho rozhodovania	Aktuálny pohľad na prevádzku
Podobnosti	Zjednotené, predmetovo orientované dáta	Zjednotené, predmetovo orientované dáta
Rozdiely	Statické dáta Historické dáta Súhrnné dáta	Nestále dáta Súčasné dáta Podporné dáta

Tab. 1 Porovnanie DWH vs ODS

2.3.4 Dátová pumpa

Informácie sú do DWH prenášané z transakčných systémov pomocou softwarových komponentov – **dátových púmp**. Úlohy ktoré dátové pumpy plnia sú:

- selekcia a extrakcia dát z produkčného systému
- transformácia extrahovaných dát
- reštrukturalizácia podľa potrieb užívateľov DWH
- agregácia dát podľa vybraných kritérií
- konsolidácia dát z rôznych dátových zdrojov
- vytváranie časových radov

Dátová pumpa je proces napĺňania dátového skladu dátami z dátových zdrojov. Napĺňanie z transakčných systémov väčšinou prebieha v čase, keď je predpoklad nízkeho zaťaženia, aby sa nepredlžovala doba odozvy pre užívateľov týchto systémov (nočné hodiny, víkendy).

Tento proces môžeme rozdeliť do troch fáz: výber (výber dát prostredníctvom rôznych metód), transformácia (overenie, čistenie, integrovanie a časové označenie dát) a prenos dát (premiestnenie dát do dátového skladu). Tieto nástroje a postupy sa tiež označujú ETL (Extraction, Transformation, Loading) alebo ETT (Extraction, Transformation, Transport).[9]

2.3.4.1 Výber

V tejto fáze dochádza k napájaniu na rôzne dátové zdroje a získavaniu požadovaných informácií na ďalšie spracovanie. Týmito zdrojmi sa chápe rôzne nehomogénne operatívne prostredie na rôznych hardwarových platformách (PC, iMAC a pod.) operačných systémoch (Windows, Unix, Linux, Sun, Solaris atď.), databázových systémoch (MS SQL Server, Oracle, Informix, IBM DB2 a pod.), rôzne archívne systémy, podnikové systémy a taktiež rôzne formáty súborov v súborovom systéme.

Pre samotný výber dát sa používajú rôzne nástroje a prístupy. Môžeme použiť nástroje ktoré generujú kód pre výber a majú univerzálne použitie. Alebo vlastné aplikácie a externé programy vo vyšších procedurálnych programovacích jazykoch C++, C#. Výhodou vlastných externých programov je výkonnosť, natívny prístup k DB. Pre

menšie množstvo údajov je výhodné vytvoriť prístupovú bránu (gateway). Táto metóda má však pri väčších objemoch dát nízky výkon a zaťažuje sieť.

2.3.4.2 Transformácia

Táto fáza zahŕňa procesy slúžiace k premene extrahovaných dát do podoby prijateľnej pre navrhnutý dátový sklad. Ide o operácie:

- **Validácia** - Overovanie správnosti dát z dátového zdroja. Je neprípustné mať nekvalitné, zle štruktúrované dáta. Použitie nekvalitných dát vedie k chybným alebo minimálne nepresným zostavám.
- **Prečisťovanie** - Korekcie, prípadne odstránenie nesprávnych dát, ktoré sú nepostačujúce. Táto fáza sa tiež zvykne nazývať cleasing, scrubbing. Niekedy vzhľadom na nízky prínos a vzhľadom k nákladom a náročnosti nemá zmysel čistiť údaje.
- **Integrácia –** zjednotenie dát z viacerých dátových zdrojov do konzistentného formátu (dátové typy, formáty). Sem patrí napríklad problém nejednoznačnosti údajov, keď jeden typ údaju môže byť uložený v rôznych zdrojoch v rozličnom formáte (napríklad pohlavie vo formáte Male/Female, M/F, man/woman a podobne). Medzi ďalšie problémy patrí formát čísel a textových reťazcov.
- **Derivácia –** výpočet odvodených dát, problém s chýbajúcimi údajmi, duplicitné dáta, zjednotenie hodnôt dát.
- **Denormalizácia –** združovanie dát z viacerých normalizovaných tabuliek do jednej denormalizovanej tabuľky (faktov) za účelom zníženia počtu spojovania tabuliek.
- **Agregácia –** vytvorenie, výpočet požadovaných súhrnov z detailných dát.

2.3.4.3 Prenos dát

Zavŕšením etapy ETL je prenos, ukladanie, alebo zavedenie do dátového skladu. Prenos spočíva v presune údajov a ich uložení do databázových tabuliek. Je to netriviálny proces a musí byť plánovaný a automatizovaný v najvyššej možnej miere. Po zavedení spravidla prebieha indexovanie, aby bol prístup k dátam optimalizovaný.

Pri prvotnom naplnení dátového skladu môže ísť o veľké množstvo údajov. Následne sa dáta obnovujú v pravidelných cykloch spúšťania dátovej pumpy. Existujú tri základne scenáre pre ukladanie dát do dátového skladu:

- **Celková obnova** – celý dátový sklad bude nahradený novými dátami.
- **Prírastková obnova** – ukladajú sa len pridané dáta väčšinou za nejaké časové obdobie.
- **Synchronizácia** – ukladajú sa len zmenené záznamy.

2.3.4.4 Synchronizácia

Súhrn techník synchronizácie pri ukladaní dát do dátového skladu poskytuje CDC – Change Data Caprure (v preklade snímanie zmenených dát). Definuje dva prístupy k synchronizácii. [10]

Statické snímanie dát – nezaznamenávajú sa zmeny dát prevedené medzi dvoma prevodmi, snímkami.

- Jednoduché statické snímanie – vytvorí sa periodický snímok zdrojových dát a tieto sa načítajú do dátového skladu.
- Snímanie podľa časových pečiatok – s využitím časovej pečiatky zmeny na detekciu zmeny záznamu. Výhodou je zníženie počtu prenášaných dát. Nevýhodou je nutnosť evidencie časových pečiatok na strane zdrojového systému.
- Snímanie porovnávaním súborov – zmeny sa identifikujú porovnávaním rozdielov pred a po snímaní, porovnávaním celých záznamov.

Inkrementálne snímanie dát. **–** zaznamenáva sa každá zmena v dátach

- Snímanie s prispením aplikácie – operatívny systém zapisuje všetky zmenené záznamy do osobitného súboru, tabuľky.
- Snímanie založené na triggroch – menené záznamy sú uschovávané do vyhradenej tabuľky pomocou databázových triggrov.
- Snímanie transakčného žurnálu – zmeny záznamov sú zaznamenávané do transakčného žurnálu príslušného RDBMS.

2.3.5 Metadáta

Definícia: Metadáta sú dáta o dátach. Sú formou abstrakcie, ktorá popisuje štruktúru a obsah dátového skladu.

Celý proces načítania dát do skladu riadia **metadáta** (dátový slovník, dáta o dátach), zhotovené administrátorom alebo importované z iných zdrojov. Popisujú zdroje dátových štruktúr, obsah dátového skladu, pravidlá pre transformáciu dát a ostatné elementy.

Proces uloženia do DWH môže byť automatizovaný pomocou špeciálnych postupov a špecifických utilít, závislých na použitom dotazovacom nástroji.

2.3.5.1 Využitie metadát

- Metadáta stanovujú obsah dát v sklade – pomáhajú administrátorom a užívateľom určiť a pochopiť dátové položky.
- Uľahčujú vykonávanie analýz – rýchla lokalizácia požadovaných dát, správna interpretácia, informácie o dátových formátoch, definícia dát.
- Sú formou auditu transformácie dát – popisujú transformáciu zdrojových dát do dátového skladu, informácie o pôvode dát, generovanie extrakčných a transformačných skriptov.
- Zvyšujú a udržujú kvalitu dát – definovanie prípustných hodnôt pre položky v dátovom sklade, popis pravidiel pre opravu chýb.

2.3.5.2 Štyri vrstvy metadát

- metadáta analýzy a návrhu – vytvárané pri analýze a návrhu databázového systému. Patria medzi ne diagramy dátového modelu, popis entít a atribútov, kľúčov a vzťahov, názvoslovie, popis domén.
- metadáta databázy – katalóg, systémové tabuľky príslušnej databázy na danom databázovom systéme. Zahrňujú definície tabuliek, pohľadov, stĺpcov, indexov, primárnych a cudzích kľúčov a podobne.
- vrstva prevodu dát – definuje vzťah medzi produkčným systémom (systémami) a samotným DWH. Obsahuje modely obidvoch systémov a vzťahy alebo transformačné procesy medzi nimi, informácie o mapovaní, verifikácii.

- koncová vrstva (aplikačná) – určená pre pohľad na dátové štruktúry zo strany užívateľa, katalóg výstupných zostáv a dotazov, prístupové práva, pomocné informácie.

2.4 Porovnanie transakčných databáz a dátových skladov.

Hlavným rozdielom medzi transakčnými databázami a dátovými skladmi je, že transakčné databázy (OLTP) sú určené na ukladanie operatívnych údajov a dátový sklad je navrhnutý a optimalizovaný na rozsiahle analýzy.

Výsledkom dotazov OLTP sú tabuľky, zostavy a súhrny získané agregačnými funkciami. OLTP sú kvôli jednoduchému dotazovaniu a vďaka vylúčeniu prebytočnosti spravidla normalizované a teda operatívne údaje sú komplexné a vysoko štruktúrované. Dosahujú vysokých výkonov skôr pri transakciách on-line ako pri zložitých analýzach. Sú write-optimized, teda optimalizované na zápis.

Dátový sklad je databáza, ktorá je navrhnutá ako prostriedok na dotazovanie a analýzu. Obsahuje read-only dáta, ktoré sú vyhodnocované a analyzované omnoho efektívnejšie ako regulárne OLTP transakčné databázy. Sú teda optimalizované na čítanie, read-optimized. Vytvorenie DWH vedie priamo k zvýšeniu kvality analýzy, štruktúra tabuliek je jednoduchšia (udržuje iba potrebné informácie v jednoduchších tabuľkách), je štandardizovaná (používa dobre zdokumentované tabuľkové štruktúry) a je denormalizovaná, (znižuje viazanosť tabuliek medzi sebou a odozvu pri zložitých dotazoch).

Decentralizovanosť systémov OLTP je ďalšou prekážkou pri použití pre analýzy. OLTP nemajú k dispozícii integrovaný zdroj údajov zo všetkých operačných systémov. Údaje na základe ktorých sa tvorí analýza sú roztrúsené v rôznych, spravidla heterogénnych, systémoch OLTP. To sťažuje tvorbu komplexných analýz, keďže sa tieto údaje musia integrovať skôr, ako je možné z nich získať potrebné informácie.

Problémy pri použití transakčných databáz pre analýzy nastávajú napríklad pri integrácii dát z jednotlivých systémov. Integrácia sa nemusí podariť, časová náročnosť

prípadných analýz (aj u nie príliš zložitej analýzy) je vysoká alebo u veľkého objemu nahromadených údajov je vyťaženosť databáz príliš veľká.

2.4.1 Porovnanie OLTP a OLAP

Systémy OLTP a systémy OLAP sa dajú porovnať podľa troch hlavných kritérií: účelu, koncepčnej schémy a technologických rozdielov.

2.4.1.1 Porovnanie podľa účelu

	OLTP Systémy	**OLAP Systémy**
Hlavná funkcia	Automatizácia operácii alebo procesov, dennodenné operácie	Poskytovanie optimálnych informácií pre rozhodovanie
Orientácia	Customer-oriented	Market-oriented
Užívatelia	Úradník, IT profesionál	Znalostný analytik
Účel dát	Kontrola a chod základných komerčných úloh	Napomáhať s plánovaním, riešením problémov a podpory rozhodovania

Tab. 2 Porovnanie OLTP a. OLAP podľa účelu

2.4.1.2 Porovnanie podľa koncepčnej schémy

	OLTP Systémy	**OLAP Systémy**
Hlavná funkcia	Vkladá dáta do systému	Získava informácie zo systému
Vkladanie a aktualizácia údajov	Krátke a rýchle zadávanie, zmena, rušenie, čítanie dát iniciované užívateľom	Užívatelia majú možnosť iba čítať. Periodické obnovovanie, dopĺňanie dát.
Zdroj dát	Operatívne dáta: OLTP systémy sú vlastné zdroje dát	Konsolidované dáta: zdroje dát pochádzajú z externého prostredia, z viacerých OLTP databáz
Dotazy	Relatívne štandardizované a jednoduché dotazy. Návratová hodnota je niekoľko záznamov.	Komplexné dotazy umocňované agregáciami
Čo dáta reprezentujú	Snímok aktuálnych dát, obmedzené historické dáta, izolované	Historické dáta, konsolidované, rozdielna organizácia, viacrozmerný pohľad

Pohľad na dáta	Detailný, plocho relačný	Sumarizovaný, viacrozmerný
Použitie	Štruktúrované, opakované	Ad hoc
Činnosti	Automatizácia rutinných činností. Procesy	Možnosť kreativity užívateľov pri práci s dátami, analýzy, prezentácie. Analýza
Aktivity	Podporujú každodenné firemné aktivity	Podporujú dlhodobé stratégie firmy
Počet užívateľov	Tisíce	Stovky
Veľkosť Databázy	100 MB-GB	100 GB-TB

Tab. 3 Porovnanie OLTP vs OLAP podľa koncepčnej schémy

2.4.1.3 Porovnanie podľa technologických rozdielov

	OLTP Systémy	**OLAP Systémy**
Prístupová rýchlosť	Typicky veľmi rýchla. Zlomky sekúnd až sekundy	Záleží na množstve požadovaných dát, aktualizácia a komplexné dotazy môžu trvať hodiny, rýchlosť môže byť optimalizovaná vytvorením indexov. Sekundy až hodiny
Databázový design	ER dátový model, vysoko normalizovaný s viacerými tabuľkami	Typicky de-normalizovaný s menej tabuľkami, Star alebo Snowflake dátový model
Indexovanie	Nadmerné používanie indexov spomaľuje operácie modifikácie	Indexovacie techniky môžu byť použité v neobmedzenom rozsahu
Integrita dát	Sa zaisťuje dátovým modelom a sčasti na aplikačnej úrovni	Sa zaisťuje pri zavádzaní dát (transformácia dát)
Orientácia	Application-oriented design, orientácia na aplikáciu	Subject-oriented design, orientácia na predmet záujmu a čas
Prístupový model	Krátke, atomické transakcie	Väčšinou read-only Operácie a historické dáta
Počet transakcií	Spracovávajú veľké množstvo transakcií	Spracovávajú malý počet zložitých dotazov
Úložný priestor	Môže byť relatívne malý	Väčší, zapríčinený existenciou aggregačných štruktúr a historických dát, potrebuje viac indexov ako OLTP

Backup a Recovery	Pravidelné zálohovanie, operatívne dáta sú kritické pre chod systému a strata znamená výraznú škodu	Niektoré systémy majú možnosť jednoduchého obnovenia údajov z OLTP dát v prípade straty, alebo poškodenia

Tab. 4 Porovnanie OLTP vs OLAP podľa technologických rozdielov

2.4.2 Prečo použiť Data Warehouse

Prof. Sham Navathe z Georgia Institute of Technology rozdelil dôvody na dve skupiny podľa výkonu a funkčnosti.[11]

Výkon

- Transakčné databázy sú navrhnuté a optimalizované na zápis (write-optimized).
- Komplexné OLAP dotazy majú veľkú výkonnosť pri použití DWH a nezablokujú systém.
- Špeciálna organizácia dát, prístupu a implementačných metód pre viacrozmerné dotazovanie.

Funkčnosť

- Chýbajúce dáta: Podpora rozhodovania vyžaduje historické dáta, ktoré transakčné databázy štandardne neuschovávajú, alebo uschovávajú iba v obmedzenom množstve.
- Zjednotenie dát: vyžaduje sa zjednotenie a centralizácia dát z viacerých heterogénnych zdrojov, transakčných databáz a ďalších externých zdrojov.
- Kvalita dát: pri použití viacerých zdrojov dát dochádza k nekonzistencii dátovej reprezentácie, kódu, formátu.

2.4.3 Ďalšie nevýhody a problémy transakčných databáz pre analytické účely

Ďalšie nevýhody a problémy pri použití transakčných databáz pre analytické účely možno sformovať do nasledujúcich bodov.[2]

2.4.3.1 Zložité odladenie problémov, hľadanie závislostí jednotlivých veličín

Nie je jednoduché nájsť príčiny a vysvetlenie problémov ktoré nastanú. Je problém nájsť konfliktné hodnoty sledovaných veličín, hlavne keď závisia na viacerých faktoroch. U zložitejších javov nie je jasné na čom a v akej miere je sledovaná veličina závislá. Vynechanie dôležitej závislosti môže mať za následok čiastočné, alebo úplne skreslenie sledovaných údajov.

Pri použití štandardných zostáv sú výstupy pri širšom zadaní parametrov analýzy z produkčných systémov príliš rozsiahle. Analýza je poskytovaná pre širšiu skupinu užívateľov. Líši sa v detailnosti vrstiev. Je potrebné používať súčty, minimá a maximá sledovaných hodnôt v agregovaných zostavách.

2.4.3.2 Náročnosť na výkon hardwaru a operačného systému

Dlhý čas výpočtu a neúmerné vyťaženie hardwarových prostriedkov a operačného systému. Dôsledkom je predĺženie času odozvy pre užívateľa, ktorý vykonáva transakciu a tiež aj pre užívateľa, pre ktorého je výsledok analýzy určený. Operácia môže trvať desiatky minút až hodiny. Pri vytváraní agregovanej zostavy sa počítajú rádovo milióny čísel. Pri preťažení sa transakčný systém stáva nepoužiteľný, pretože prestáva plniť svoje primárne funkcie. Výpočet potrebných agregácií, predpovedí a podobne trvá v transakčných systémoch veľmi dlho, a preto môže mať za následok preťaženie siete a neúmerne zaťaženie databázového stroja.

Riešenie zníženia zaťaženia databázového servera:

- Je možné využiť čas, keď je predpoklad, že bude systém najmenej vyťažený (napríklad v noci). Treba však zvážiť dôsledok dlhého časového intervalu medzi zadaním požiadavky na vytvorenie zostavy a vytvorením zostavy.
- Ďalšia možnosť je využitie vopred sumarizovaných agregácií. Prakticky ide o vytvorenie pomocnej tabuľky, do ktorej budú uložené agregované hodnoty a pri spracovaní dotazu sa bude vyhľadávať v tejto tabuľke. Výhoda je ľahká implementácia, nevýhoda je obmedzenie dotazov iba na agregované hodnoty.
- K zníženiu zaťaženia DB servera pri spracovaní dotazov prispeje aj prenesenie dát, ktoré potrebujeme pre výpočet agregovaných zostáv na separátny server.

Prenesená časť databázy nebude určená na manipuláciu s dátami, ale iba na vytváranie zostáv pre analýzu ukazovateľov. Dáta sú týmto ohraničením špecificky uložené za účelom optimalizácie pri čítaní veľkého množstva dát na úkor zápisu. Aby nedochádzalo k zbytočnému prenášaniu celého obsahu databáz je potrebné využiť transformačné procedúry na prenos iba nových alebo zmenených záznamov.

2.4.3.3 Historické údaje

Typickou vlastnosťou transakčných databázových systémov je to, že neexistuje možnosť v dostatočnom rozsahu uchovávať historické údaje potrebné ku komplexnej analýze. Dôvodom je disková kapacita, ktorá je často nepostačujúca. Veľa údajov sa v niektorých prípadoch neuchováva vôbec.

V niektorých prípadoch (účtovné databázy, poisťovne, banky, rezervačné systémy, burzy), keď je aj v transakčných databázach potrebné uchovávať historické údaje, sa tieto údaje uchovávajú rozdelením databázy na operatívnu a historickú časť, využitím temporálnych databáz. Temporálne databázy využívajú časové pečiatky na zavedenie pojmu času do databázy. Tieto zaznamenávajú k položkám čas transakcie a čas platnosti (rozšírenie položky o štart/stop). Následná aktualizácia dátového skladu z temporálnej databázy je len rutinná záležitosť.

2.4.3.4 Nehomogénna štruktúra údajov

Údaje na základe ktorých sa tvorí analýza sú roztrúsené v rôznych spravidla heterogénnych systémoch OLTP. Identické údaje v decentralizovaných systémoch môžu byť uložené v rôznych tvaroch a formátoch. Napríklad číselný kód alebo rodné číslo (vďaka oddeľovaču) môže byť uložený ako string, alebo ako integer.

2.4.3.5 Časová náročnosť

Postupné pripojenie k nehomogénnym systémom a analýza viacerých údajov naraz môže byť z technického hľadiska problém. Často je tiež potrebná zmena štruktúry niektorej z transakčných databáz.

3 Multidimenzionálny databázový model

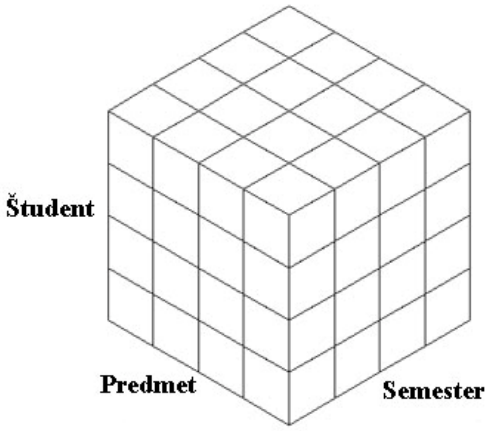
Multidimenzionálna databáza je typ databázy, ktorý je optimalizovaný pre data warehouse a OLAL aplikácie. Multidimenzionálne databázy často využívajú zdroj z existujúcich relačných databáz. Používajú princíp dátovej kocky resp. hyperkocky, tiež nazývanej kocka (data cube, hypercube) na reprezentáciu dimenzií dát dostupných pre užívateľa.

V tejto kapitole sa budem venovať logickému multidimenzionálnemu modelu (ktorý využíva Microsoft SQL Server, Oracle a ďalšie).

Príklad hyperkocky:

Prevažná väčšina údajov je organizovaná v relačnej databáze v dvojrozmerných relačných tabuľkách. Výsledkom agregácie a analýzy býva obvykle multidimenzionálna dátová štruktúra – hyperkocka.

Multidimenzionálny databázový model si môžeme najjednoduchšie predstaviť ako priestorovú kocku. Každá kocka môže mať niekoľko dimenzií (aj niekoľko desiatok). Príkladom trojdimenzionálneho modelu môže byť hyperkocka s dimenziami Študent, Predmet, Semester:



Obr. 5 Hyperkocka

Údaje sa nachádzajú v prienikoch jednotlivých dimenzií. Môžeme napríklad analyzovať údaje len za určité časové obdobie - semester. Alebo aby sme vyhodnotili výsledky konkrétneho študenta ktoré dosahoval počas piatich semestrov v predmete Programovanie.

3.1 Definícia multidimenzionálneho databázového modelu

Formálna definícia dimenzie a dimenzionálnej databázy.[9]

Definícia: Bázová dimenzia je konečná diskrétna množina D s mohutnosťou > 1, prvkami ktorej sú atomické hodnoty nejakej veličiny ekonomického charakteru, dôležité z pohľadu užívateľa. Prvky množiny D sa nazývajú členovia dimenzie.

Definícia: Majme dimenzie D_1 a D_2, množinu E, $E \subseteq D_1 \times D_2$. Potom platí:

$$D_1 \perp D_2 \Leftrightarrow D_1 \times D_2 \equiv E \wedge \forall d_2 \in D_2; \neg\exists f : d_2 = f(d_{11}, d_{12}, ..., d_{1|D1|}), d_{1i} \in D_1, i \in \{1,...,|D_1|\}$$

Dimenziu nazývame ortogonálna k inej dimenzii práve vtedy keď, pre každý člen jednej dimenzie môžu v reálnom svete existovať všetky členy druhej dimenzie a medzi členmi týchto dimenzií neexistuje funkčná závislosť. V multidimenzionálnej databáze sú všetky bázové dimenzie vzájomne ortogonálne. Každý člen bázovej dimenzie je atomický, t.j. nesmie byť pre účely modelu rozložiteľný do podčastí.

Definícia: Merateľná dimenzia je špeciálny typ bázovej dimenzie, ktorej členmi sú premenné, o hodnoty ktorých sa užívateľ zaujíma. Členy merateľnej dimenzie nazývame meradlá. Multidimenzionálna databáza musí obsahovať páve jednu merateľnú dimenziu.

Definícia: Odvodené meradlo je také meradlo, ktoré môžeme vyjadriť ako funkciu jedného alebo viacerých meradiel. Základné meradlo je každé meradlo, ktoré nie je odvodené.

Definícia: Časová dimenzia je bázová dimenzia, ktorej členmi sú časové obdobia. Multidimenzionálna databáza môže obsahovať najviac jednu časovú dimenziu. Časové obdobia časovej dimenzie by mali byť vzájomne súvislé.

Definícia: Subdimenzia je množina disjunktných podmnožín členov bázovej dimenzie, ktoré majú nejakú spoločnú vlastnosť.

Definícia: Agregovaná dimenzia je subdimenziou, v ktorej zjednotenie členov je izomorfné s príslušnou bázovou dimenziou.

Definícia: Konsolidačná dimenzia je zoskupenie prvkov bázovej dimenzie, ktoré zhŕňa do jedného meradla alebo množiny meradiel pre bázovú dimenziu. Hierarchie môžu obsahovať niekoľko úrovní.

Definícia: Multidimenzionálna databáza je n-rozmerný priestor bázových dimenzií, z ktorých jedna musí byť merateľná dimenzia, a nad ktorým môžeme existovať m-rozmerný priestor agregovaných dimenzií (m>>n).

3.1.1 Fakty a Dimenzie

Do multidimenzionálnych databáz sa ukladajú upravené dáta, ktoré sú podkladom pre získanie sumarizovaných a agregovaných údajov. Na rozdiel od relačných databáz sa používajú prevažne nenormalizované tabuľky, ktoré rozdeľujeme na dva druhy: na tabuľky faktov a tabuľky dimenzií. Každá kocka OLAP je teda vytvorená na základe týchto dvoch údajov:

- Fakty – numerické merné jednotky obchodovania. Prvotné fakty sa môžu kombinovať alebo vypočítať pomocou iných faktov a vytvoriť tak merné jednotky.
- Tabuľky faktov – je hlavná tabuľka, na ktorú sú viazané tabuľky dimenzií. Uchováva veľké množstvo dát. Dáta sa nemenia často. Spravidla je len jedna

tabuľka faktov pre jednu kocku. Tabuľka faktov býva najväčšia tabuľka v databáze. Môže vytvárať rôzne schémy.

- Dimenzie – obsahujú logicky alebo organizačne hierarchicky usporiadané údaje. Možno povedať, že to sú textové popisy obchodovania, teda že charakterizujú dáta. Elementy sú členovia(members) niektorej dimenzie.
- Tabuľky dimenzií – obsahujú usporiadané údaje. Sú naviazané na tabuľku faktov, alebo na inú tabuľku dimenzií. Sú spravidla menšie ako tabuľky faktov a dáta sa v nich nemenia tak často. Veľmi často sa používajú časové, produktové a geografické dimenzie. Obsahujú atribúty popisujúce fakty.

Tabuľky dimenzií obvykle používajú stromovú (hierarchickú) štruktúru napr.:

- Čas: ◦ rok, ◦◦ kvartál, ◦◦◦ mesiac
- Škola: ◦ vysoká škola ◦◦ fakulta ◦◦◦ odbor ◦◦◦◦ ročník ◦◦◦◦◦ krúžok

Máme možnosť zjemňovať - *drill-down* a zovšeobecňovať - *roll-up* hierarchickú úroveň dimenzie na nižší, alebo vyššiu úroveň (momentálna pozícia v hierarchii).

Priestor pre celú kocku je dopredu určený. Jednotlivé záznamy sa nachádzajú na priesečníkoch dimenzií. S rastúcim počtom rozmerov multidimenzionálnej databázy veľmi rýchlo rastú aj požiadavky na úložnú kapacitu. V praxi sa používajú rôzne technológie na kompresiu objemu použitého diskového priestoru. V prípade, že sa na všetkých priesečníkoch dimenzií nenachádzajú údaje, kocku nazývame aj riedka kocka.

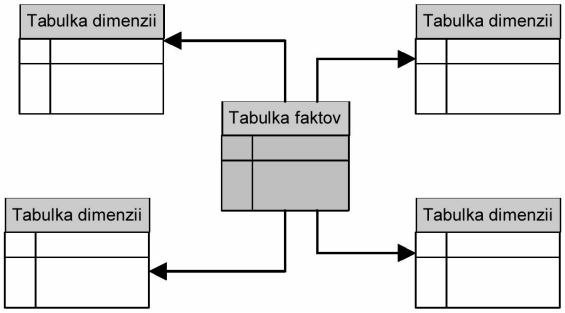
3.1.2 Schémy tabuliek dimenzií

Kocku vytvárame na základe dimenzionálneho modelu, ktorý má určité topologické usporiadanie - schému. Môže byť dvoch typov:

- hviezdicová schéma (star scheme)
- schéma snehovej vločky (snowflake scheme)

Hviezdicová schéma

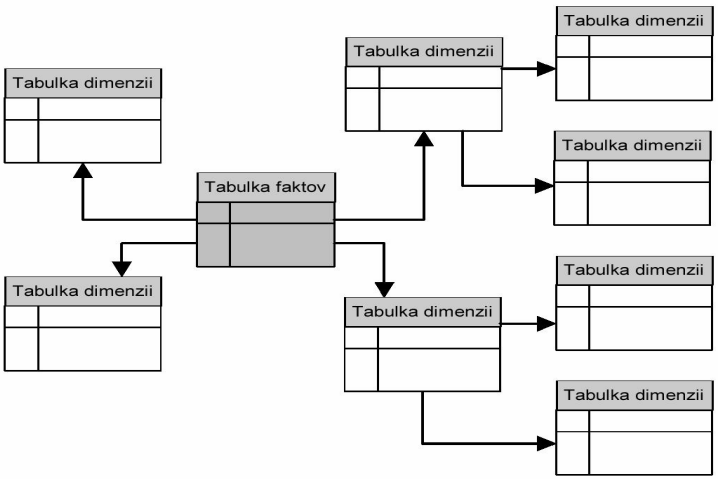
Tento model sa používa najčastejšie na modelovanie multidimenzionálnych dát pomocou relačného modelu. Hviezdicová schéma sa skladá z tabuľky faktov. Tabuľka faktov je denormalizovaná a obsahuje cudzie kľúče, ktoré sa vzťahujú k primárnym kľúčom v tabuľkách dimenzií. Každá dimenzia môže obsahovať úrovne hierarchie, ktoré sú reprezentované jednotlivými stĺpcami v príslušnej tabuľke Dimenzie. Hviezdicová schéma nemá normalizované dimenzie ani relačné prepojenia medzi tabuľkami dimenzií. Dôsledok toho je, že je ľahko pochopiteľná ale vďaka nenormalizovaným dimenziám je vytvorenie modelu relatívne pomalé. Model poskytuje vysoký dotazovací výkon, keďže všetky údaje sa získajú naraz a nemusia sa skladať z relačných tabuliek.



Obr. 6 Hviezdicová schéma

Schéma snehovej vločky

V niektorých prípadoch, sa používa táto modifikácia hviezdicovej schémy. Rozdiel oproti hviezdicovej schéme je v tom, že má normalizované tabuľky dimenzií. Schéma snehovej vločky oproti hviezdicovej schému obsahuje niektoré dimenzie zložené z viacerých relačne spojených tabuliek. Tento model umožňuje rýchlejšie zavádzanie údajov do normalizovaných tabuliek, ale má nižší dotazovací výkon, lebo obsahuje veľké množstvo spojených tabuliek.



Obr. 7 Schéma snehovej vločky

3.2 Porovnanie relačného a multidimenzionálneho modelu

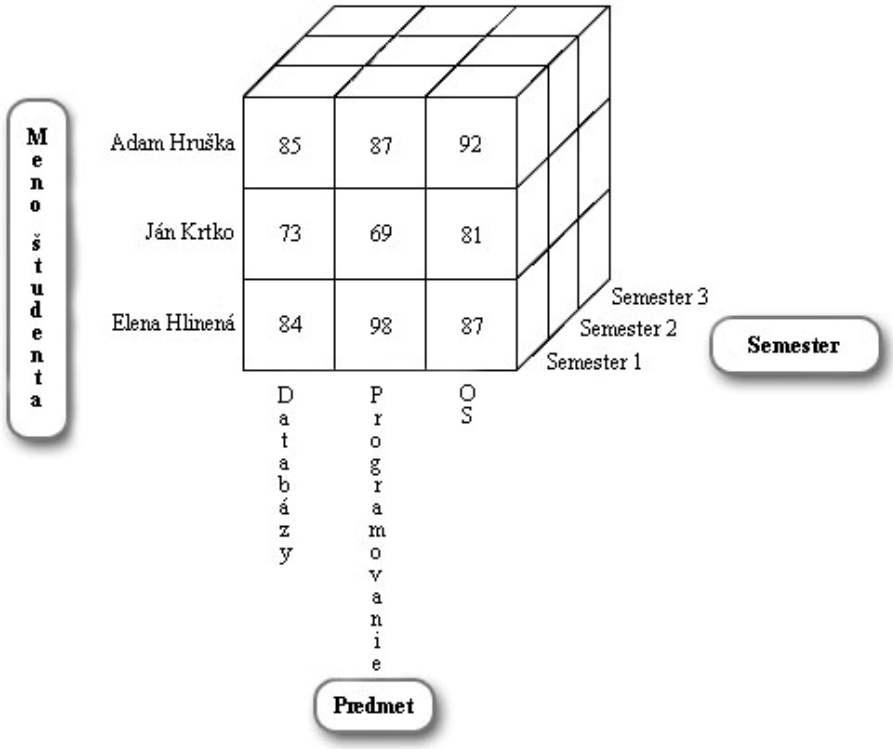
Medzi hlavné obmedzenia relačných databáz patrí absencia komplexných analytických nástrojov a potenciálne obmedzenie údajov, ku ktorým je možné v rozumnom čase pristupovať. Východiskom je organizácia údajov do multidimenzionálnych štruktúr. Multidimenzionálne štruktúry reprezentujú vyššiu úroveň organizácie ako relačné tabuľky. Štruktúra je sama o sebe reprezentovaná viac intuitívne, pretože perspektíva dát je uložená priamo podľa dimenzií ako ich vnímame.

Majme napríklad časť jednoduchej relačnej tabuľky študentov a ich body zo skúšok z predmetov, ktoré trvajú viac semestrov (predpokladajme, že každý trvá napríklad 8 semestrov).

id	Meno študenta	Predmet	Body	Semester
11	Adam Hruška	Databázy	85	1
21	Adam Hruška	Programovanie	87	1
24	Adam Hruška	OS	92	1
28	Ján Krtko	Databázy	73	1
42	Ján Krtko	Programovanie	69	1
58	Ján Krtko	OS	81	1
55	Elena Hlinená	Databázy	84	1
57	Elena Hlinená	Programovanie	98	1
60	Elena Hlinená	OS	87	1
..	...	...	...	...

Tab. 5 Príklad 1

Príklad hyperkocky s troma dimenziami: Študent, Predmet, Semester.



Obr. 8 Príklad hyperkocky

Je vidieť, že body nemusíme určiť ako dimenziu, hodnoty bodov sú obsiahnuté v príslušnej bunke. Ďalšia výhoda je odstránenie duplikácií, keďže pre každú skúšku v semestri nemusí byť opakované meno študenta. Podobne to je aj s názvami predmetov. Výhodou je rýchla manipulácia s dátami, keď môžeme analyzovať výsledky študenta v jednom semestri, prípadne výsledky študentov jedného predmetu.

Porovnanie na základe niektorých kritérií:

- **Veľkosť.** Veľkosť multidimenzionálnych databáz je obmedzená. Aj napriek trendu zvyšovania limitu sú pre rozsiahle dátové sklady lepšou variantou relačné nástroje a výkonné a škálovateľné relačné databázové platformy.

- **Nestálosť zdrojových dát.** Veľmi nestále dáta je možné lepšie spracovať pomocou relačných technológií. Pravidelné plnenie hyperkociek a aktualizácia vyžaduje veľa času. Práve nedostatok času pre neustále vykonávanie týchto činností môže byť prekážka.
- **Agregácie.** Hyperkocky lepšie podporujú agregácie, avšak pridaním podpory navigácie v rámci agregácií priamo v relačných databázových platformách sa dajú rozdiely minimalizovať. Časovo náročnejšie v relačných systémoch však zostáva zjemňovanie a zovšeobecňovanie.
- **Dodatočné Investície a ľudský potenciál**. Pokračovanie vo využívaní relačných technológii, nástrojov a odborných znalostí zaisťuje ďalšie zužitkovanie vynaložených investícii a znižuje technické riziká. Multidimenzionálne databázové platformy pridávajú do systémovej architektúry ďalšiu vrstvu, ktorá vyžaduje vyhradenie nových zdrojov pre administráciu a správu. Pokročilí užívatelia uprednostňujú rozsah funkčnosti dostupnej v multidimenzionálnych nástrojoch OLAP. Naopak užívatelia, ktorí vyžadujú priamy prístup k dátovému skladu a k dátam, lepšie využijú relačné nástroje OLAP.

3.2.1 Výhody a nevýhody relačnej databázy

Relačný databázový model používa na uschovávanie dát dvojrozmernú štruktúru riadkov a stĺpcov v tabuľkách, pričom tabuľky môžu byť napojené cez kľúčové hodnoty, t.j. na základe entitno-relačného modelu. Tento model ako prvý navrhol v roku 1970 E.F. Codd pracujúci pre IBM a svojou prácou predbehol dobu a možnosti vtedajšej výkonnosti počítačov vzhľadom na jeho databázový systém.

Technika Entitno-relačného modelu a štruktúrovanie dát v normalizovaných tabuľkách sa stala veľmi populárna a osvedčená pri uschovávaní obrovského množstva organizovaných dát s veľkou transakčnou odozvou. Prístup k dátam však môže vyžadovať zložité spojenia (joiny) viacerých tabuliek a agregácie, čo je však netriviálne pre koncového užívateľa, ktorý je nútený pri svojich dotazoch využiť služby IT odborníkov.

Nie všetky databázy má zmysel konvertovať do multidimenzionálnych štruktúr. Relačná databáza je napríklad výhodnejšia, keď uvažujeme kocku s dimenziami meno, priezvisko a hodnotou vek. Nemá zmysel ukladať do multidimenzionálnych štruktúr dáta, kde je predpoklad výrazne riedkych záznamov. Efektivita prístupu je v takýchto prípadoch vyššia v relačných databázach, napríklad pre tabuľku s tromi záznamami meno/priezvisko/vek budú potrebné 3 porovnania, pričom pri multidimenzionálnej štruktúre 9 (3x3). Multidimenzionálne databázy sú navrhnuté za účelom manipulácie a analýzy komplexných databázových štruktúr pracujúcich s veľkým množstvom dát a relácií.

Výhody relačnej databázy:

- dostatočný potenciál ľudských zdrojov
- dostatočný potenciál softwaru a nástrojov pre vývoj, ladenie a generovanie zostáv
- použiteľnosť v transakčných databázach a aj v dátových skladoch
- prístup k dátam v reálnom čase
- jednotný sklad dát
- pružná schéma

Nevýhody relačnej databázy:

- absencia komplexných analytických nástrojov
- potenciálne obmedzenie objemov údajov, ku ktorým je možné v danom čase pristupovať

3.2.2 Výhody a nevýhody multidimenzionálnej databázy

Niektoré výhody multidimenzionálnej databázy som prezentoval na príklade aj v úvode tejto kapitoly. Medzi hlavné plusy patrí zlepšená prezentácia a navigácia v dátach. Taktiež ľahká údržba, keďže dáta sú uložené rovnakým spôsobom, ako sú prezentované a nie sú vyžadované výpočty, dotazy. V relačných databázach je možnosť optimalizácie, ale nemôžeme optimalizovať databázu pre každý dotaz zvlášť. Zvýšený výkon umožňuje efektívne využitie OLAP aplikácii.

Výhody multidimenzionálnej databázy:

- rýchly komplexný prístup k veľkému objemu údajov a navigácia v nich
- prístup k multidimenzionálnym a relačným dátovým štruktúram
- možnosť efektívnej a komplexnej analýzy dát
- lepšia prezentácia dát
- možnosť modelovania situácii a vytvárania prognóz, orientácia na užívateľa
- ľahká údržba
- vysoký výkon
- zložité analýzy

Nevýhody multidimenzionálnej databázy:

- vyššie nároky na kapacitu diskového priestoru
- problémy pri zmene dimenzií bez prispôsobenia časovej dimenzie

3.2.3 Porovnanie charakteristík

	Relačné databázy	**Multidimenzionálne databázy**
Množina	Entita	Bázová dimenzia
Popis	Atribúty	Agregovaná dimenzia
Dimenzionalita	Dvojrozmerná tabuľka	Viacrozmerná štruktúra
Čas	Manipulácia s časovými hodnotami ťažšia	Práca s časovými hodnotami je elementárna
Modelovanie	Normalizované ERD, DFD	Hviezdicová schéma, snehová vločka
Prístup k dátam	SQL	MDX, úpravy SQL

Tab. 6 Relačné databázy vs Multidimenzionálne databázy

4 OLAP (On-Line Analytical Processing)

Pod názvom OLAP sú zahrnuté technológie, metódy a prostriedky, ktoré umožňujú ad-hoc analýzu multidimenzionálnych informácií. OLAP dovoľuje užívateľovi pracovať s údajmi veľmi flexibilne a analyzuje dáta podľa mnohých hľadísk.

4.1 Definicia OLAP

OLAP (On-Line Analytical Processing) je druh softwarovej technológie ktorá slúži ku spracovaniu údajov (ich transformácii) uložených v dátovom sklade do podoby pre koncových užívateľov, teda manažérov a analytikov. Umožňuje konzistentný a interaktívny prístup k širokému spektru možných pohľadov na informácie.

Ďalšia z definícii: OLAP je voľne definovaná množina princípov, ktoré poskytujú dimenzionálny rámec pre podporu rozhodovania.

4.1.1 Funkcionalita OLAP

Funkcionalitu OLAP charakterizuje dynamická multidimenzionálna analýza dát za analytickej a navigačnej podpory pre užívateľa. Implementácia OLAP je v prostredí klient/server, za poskytovania sústavnej rýchlej odozvy na dotazy, bez ohľadu na veľkosť databázy a jej zložitosť.

Funkcionalita je najčastejšie implementovaná pomocou osobitného OLAP serveru. OLAP server má buď vlastnú multidimenzionálnu databázu alebo v reálnom čase plní dátové štruktúry z inej databázy (väčšinou relačnej).

Funkcionalita umožňuje:

- výpočty a modelovanie naprieč dimenziami, skrz hierarchie, naprieč členmi
- analýza trendov v rozličných časových periódach
- rozdeľovanie podmnožín pre zobrazovanie
- zostup a vzostup do nižších a vyšších úrovní konsolidácie (drill-down/ drill-up)

- prienik do príslušnej detailnej úrovne dát
- rotácie pre porovnania v nových dimenziách príslušnej oblasti
- sústavne rýchlu odozvu na dotazy, bez ohľadu na veľkosť databázy a jej zložitosť

4.1.2 Pravidlá pre OLAP

Existuje 12 základných pravidiel OLAP, ktoré sformuloval Dr. E. F. Codd [14]. Tieto pravidlá boli napísané pre architektúru produktu dodávateľa Arbor Software (Hyperion Solutions).

1. Multidimenzionálny konceptuálny model: OLAP by mal poskytovať užívateľovi multidimenzionálny model tak, aby zodpovedal jeho potrebám a aby tento model mohol využívať pre analýzu zhromaždených údajov.

2. Transparentnosť: To, aby užívateľ mohol naplno využívať svoju produktivitu, odbornosť a prostredie docielime tým, že technológia systému OLAP, jej databáza a architektúra výpočtu bude transparentná. Dôležitá je heterogénnosť vstupných dát, ktorú zaistíme v procese ETL.

3. Dostupnosť: Systém OLAP by mal pristupovať len k údajom, ktoré sú potrebné pre analýzu. Systém by mal navyše byť schopný pristupovať ku všetkým takýmto údajom, nezávisle na tom, z ktorého heterogénneho podnikového zdroja pochádzajú a ako často sú obnovované.

4. Stabilná výkonnosť: Užívateľ nesmie pocítiť žiadne podstatné zníženie výkonu, aj keď veľkosť databáz postupom času rastie.

5. Architektúra klient/server: Systém OLAP musí fungovať na základe architektúry klient-server. Dôležitá je cena, výkon, flexibilita, interoperatibilita.

6. Generická dimenzionalita: Každá dimenzia údajov musí byť ekvivalentná v štruktúre aj operačných schopnostiach.

7. Dynamická manipulácia s riedkymi maticami: Systém OLAP musí byť schopný prispôsobiť svoju fyzickú schému na konkrétny analytický model, ktorý optimálne ošetrí riedke matice za udržania požadovanej úrovne výkonu.

8. Podpora viacerých užívateľov: Systém OLAP musí byť schopný podporovať viac užívateľov alebo skupiny užívateľov pracujúcich súčasne na konkrétnom modeli.

9. Neobmedzené operácie naprieč dimenziami: Systém OLAP musí rozoznať dimenzionálne hierarchie a automaticky vykonávať výpočty v rámci dimenzií a medzi dimenziami.

10. Intuitívna manipulácia s dátami: Užívateľské rozhranie musí umožňovať všetky manipulácie s údajmi v pre neho prístupnom (user-friendly) prostredí. Napríklad pre operácie ako drill down a drill up.

11. Flexibilné výstupy: Schopnosť usporiadať riadky, stĺpce a bunky spôsobom, ktorý umožní analýzu a intuitívnu prezentáciu analytických zostáv.

12. Neobmedzené dimenzie a úrovne agregácií: V závislosti na požiadavkách podnikania môže mať analytický model viac dimenzií, pričom každá z nich môže mať viacnásobné hierarchie. Analytický model by nemal byť umelo obmedzovaný počtom dimenzií alebo úrovňou agregácií.

Tieto pravidlá sú však často kritizované. Ako nedostatok sa uvádza nekompletnosť, miešanie technických, užívateľských a obchodných otázok, vynútenie architektúry klient/server, zahrnutie niektorých triviálnych pravidiel.

4.2 Implementačné varianty OLAP

Z hľadiska vlastného uloženia dát môžeme databázové systémy DWH rozdeliť do troch skupín, na základe implementačných variantov, MOPAL, ROLAP a HOLAP.

4.2.1 MOLAP (Multidimenzionálny OLAP, MDBMS, MS-OLAP)

Multidimenzionálny OLAP je technológia, ktorá na implementáciu multidimenzionálneho modelu využíva pre tento účel špeciálne vyvinutý OLAP server s vnútornou architektúrou databázy optimalizovanou pre multidimenzionálne dáta. Využíva dvojvrstvovú architektúru klient/server. Je to model, v ktorom prebieha

spracovanie všetkých funkcií aplikácií na dvoch komponentoch – klient a databázový server. Údaje sú ukladané do špecializovanej multidimenzionálnej databázy, do n–rozmerného priestoru. Počet dimenzií (n) zodpovedá počtu pohľadov na údaje. Uloženie údajov je závislé na ich predkompilácii a obmedzených možnostiach dynamicky pridávať nové pohľady. Samozrejmosťou je alokácia diskového priestoru v závislosti na počte dimenzií.

Výhodou je veľmi vysoká rýchlosť spracovania údajov s priamym prístupom užívateľov (aplikačná vrstva je vo vnútri MDBMS). Technológia je vhodná pre menšie aplikácie s obmedzenou multidimenzionalitou. Databáza musí byť periodicky kompilovaná. Existuje ešte riešenie firmy Speedware, ktoré umožňuje ukladať údaje do multimaticového systému, čo umožní prakticky neobmedzenú veľkosť MDBMS a rieši problém riedkeho zaplnenia dát v spojitosti s lepším využitím diskového priestoru.

Dáta sa získavajú buď z dátového skladu, alebo operatívnych zdrojov. Údaje ukladáme vo vlastných dátových štruktúrach.

4.2.1.1 Charakteristika MOLAP

- dvojvrstvová architektúra klient/server
- dáta ukladané do MDBMS v n–rozmernom priestore
- pred uložením dát na disk potreba alokácie priestoru
- veľká rýchlosť spracovania dotazov
- potrebná stála rekompilácia

4.2.1.2 Úložná kapacita MOLAP

Hlavnou nevýhodou týchto systémov je náročnosť na úložnú kapacitu. Dáta v multidimenzionálnom priestore sú veľmi rozptýlené, takže vzniká riedka matica. Štatisticky veľkosť S multidimenzionálnej databázy je približne [7]:

$$S = 8 \times \prod_{i=1}^{n} |Di| \qquad \text{[kB]},$$

kde n je počet bázových dimenzií a $|D_i|$ je mohutnosť bázovej dimenzie i.

4.2.1.3 Porovnanie MOLAP a ROLAP

MOLAP sa odlišuje hlavne tým, že potrebuje predpočítané dáta a ich uloženie v kocke. Ukladá ich optimalizované vo viacrozmernej polovej úložnej štruktúre a nie do relačnej databázy. Produkty využívajúce MOLAP sú napríklad *Microsoft Analysis Services*, *Essbase*, *TM1*.

Výhody MOLAP

- rýchle vyhodnotenie dotazov, vďaka optimalizovanému uloženiu, viacrozmernému indexovaniu a caching
- vyžaduje menší úložný priestor (oproti relačnému modelu) vďaka kompresívnym technikám
- automatický výpočet agregovaných dát
- umožnenie použitia indexovacích techník na poli (ako natural indexing)

Nevýhody MOLAP

- načítavanie dát je zdĺhavé, hlavne pri veľkých množstvách, využíva sa inkrementálny prístup, prípadne načítavanie len zmenených dát
- zložitejšie dotazovacie modely pri dimenziách s veľkou mohutnosťou

4.2.2 ROLAP (Relačný OLAP)

Vznikli snahou prispôsobiť relačný model DWH modelu. Je založený na relačných databázach a trojvrstvovej architektúre klient/server. Vykonávanie aplikačného programu je rozdelené medzi tri komponenty: klient (prezentačná funkcia), aplikačný server (obchodná logika) a databázový server (vlastná manipulácia s údajmi), čím je dosiahnutá väčšia flexibilita v prípade zmien.

Nevýhodou sú vyššie náklady na zavedenie a zabezpečenie systému. Výhodou je otvorenosť prostredia, využitie SQL. Ako ďalšia výhoda sa uvádza, že sa nevyžaduje duplicita dát (transakčné dáta sú uložené v relačnom systéme), avšak z dôvodu zvýšenia výkonu, agregácie a časových rezov je potrebné dáta aj tak duplikovať.

Popis architektúry je zjednodušený. Stupeň kompilácie dát závisí na voľbe administrátora. Pre modelovanie štruktúry dátového skladu nad relačnou databázou sa používajú logické schémy snow-flake (snehová vločka). U databáz dochádza k redundanciám, ktoré sú potrebné na niekoľkonásobné skrátenie doby odozvy. Relačný

model je dvojrozmerný. Viacrozmerné pohľady sa musia vyriešiť pomocou dôkladnej indexácie a duplikácie tabuliek.

Najjednoduchší model vytvorený pomocou dimenzionálneho modelovania sa skladá z tzv. faktovej tabuľky, ktorej primárny kľúč je zložený z rôznych dimenzií. Najčastejšie používaný dimenzionálny model sa nazýva hviezdicová schéma – každá dimenzia faktovej tabuľky je nahradená cudzím kľúčom, zodpovedajúcim dimenzii tabuľky. Vrstva metadát organizuje údaje podľa tematických okruhov.

ROLAP je flexibilná technológia pri práci nad veľkým rozsahom dát a väčším množstvom multidimenzionálnych pohľadov. Databáza sa nemusí neustále rekompilovať a neexistuje problém riedkeho zaplnenia priestorových dát. Rýchlosť spracovania je však oveľa nižšia oproti multidimenzionálnej tabuľke.

4.2.2.1 Charakteristika ROLAP

- dôsledok úspechu relačných databáz
- snaha o prispôsobenie relačnej DB pre DWH
- trojvrstvová architektúra klient/server
- pre modelovanie štruktúry DWH nad relačnou DB sa využíva schéma snehovej vločky
- databáza nie je normalizovaná
- viacrozmerný pohľad riešený indexáciou a duplikáciou tabuliek
- čas vedený len ako pevný dátum
- existuje možnosť použitia kombinácie MOLAP/ROLAP

4.2.2.2 Porovnanie ROLAP a MOLAP

OLAP Report, najväčší nezávislý pozorovateľ OLAP produktov na základe svojho prieskumu tvrdí, že na základe prieskumu za roky 2001-2005 spoločnosti, ktoré používali ROLAP vykazovali nižší výkon ako tie s MOLAP [15].[1]Produkty využívajúce ROLAP sú napríklad *Microsoft Analysis Services, Microstrategy, Business Objects, Oracle BI, Mondrian* **.**

Výhody ROLAP

- ROLAP je považovaný za lepšie škálovateľný, hlavne pri modeloch s dimenziami s veľkou mohutnosťou (rádovo miliónmi členov)
- Načítavanie dát je rýchlejšie vďaka rozmanitosti nástrojov a možnosti prispôsobenia dátového modelu.
- Dáta sú uložené v štandardnej relačnej databáze a môžu byť prístupne aj pre SQL reportovacie nástroje.
- ROLAP nástroje efektívnejšie pracujú s neagregovanými údajmi (textovými údajmi) ako MOLAP.

Nevýhody ROLAP

- ROLAP nástroje vykazujú nižšiu výkonnosť ako MOLAP nástroje.
- Načítavanie agregačných tabuliek musí byť riadené pomocou samostatného ETL kódu.
- Veľa ROLAP produktov vynecháva vytváranie agregačných tabuliek. Vyhodnocovanie dotazov sa týmto spomaľuje, kvôli pristupovaniu k väčším a detailnejším tabuľkám.
- Niektoré špeciálne techniky MOLAP nie sú dostupné, napríklad ako hierarchické indexovanie. ROLAP využíva najnovšie vylepšenia jazyka SQL, operácii ako CUBE, ROLLUP.

4.2.3 HOLAP (Hybridný OLAP)

Hybridný OLAP kombinuje výhody MOLAP a ROLAP. Princíp činnosti je v možnosti voľby užívateľa, ktorá časť údajov ostane v relačnej forme (informácie s vyššou mierou detailov) a ktorá bude agregovaná do multidimenzionálnej databázy (informácie s vyššou mierou agregácie). Základnou podmienkou je transparentné použitie MOLAP pre dáta s vyšším stupňom agregácie a ROLAP pre prácu s dátami na detailnejšej úrovni. Komerčne dostupný produkt tohto typu obsahuje v sebe kategórie EIS. Produkty využívajúce HOLAP sú napríklad *Microsoft Analysis Services*, *MicroStrategy*.

5 MS SQL Server 2005

Microsoft SQL Server je DBMS vyvinutý spoločnosťou Microsoft. Je najčastejšie používaný pre malé a stredné databázy, ale za posledných 5 rokov sa ich využitie rozšírilo aj na veľké podnikové databázy. V tejto kapitole som čerpal niektoré informácie z [32][16].

Implementácia Business Intelligence na platforme MS SQL Serveru 2005 sa rozdeľuje na tri časti podľa zamerania a funkcionality:

- **Integration Services** – získavanie dát z externých aj interných nehomogénnych zdrojov, ich transformácia, integrácia, syntéza (ETL funkcionalita).
- **Analysis Services –** analýza dát, obohatenie dát, hierarchizácia dát, hľadanie závislostí. Umožňuje organizovať dáta do intuitívnych štruktúr pre podporu preddefinovaných aj jednorazových dotazov, ktoré dokážu identifikovať pravidlá, vzťahy a trendy.
- **Reporting Services –** prezentácia a distribúcia dát, forma a rozsah výstupov. Platforma pre generovanie zostáv umožňuje vytvárať zostavy (reporty) v reálnom čase aj podľa definovaných časových plánov. Zostavy môžu byť prístupné z webového prehliadača, kancelárskych aplikácií, špecializovaných obchodných nástrojov.

5.1 Architektúra

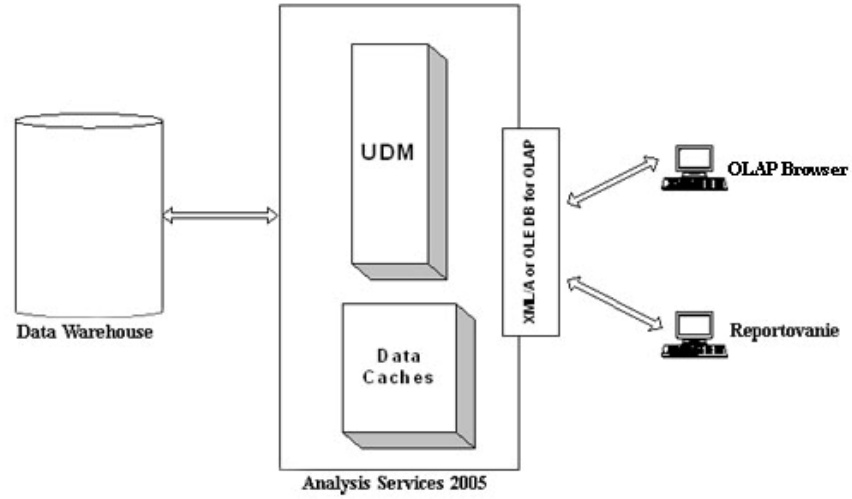
Zdrojové súbory obsahujúce OLAP definície objektov sa ukladajú do skladiska (Repository) v súboroch a formáte XML. Prístup k nim zabezpečuje MS Analysis Management Objects (AMO). Pre správu analytických služieb, analytických objektov a databáz slúži aplikácia MS SQL Management Studio. Na návrh a vytváranie objektov sa používa nástroj MS SQL Server BI Development Studio alebo priamo MS Visual Studio 2005.

AMO posiela požiadavky na vrstvu analytických služieb, ktoré obsahujú aj modul MDX Engine (má na starosti kalkulácie). Týmto sa znížili nároky výpočtový výkon

klienta a taktiež sa odľahčila sieťová komunikácia medzi klientom a analytickými službami. Došlo k zjednoteniu rozhrania XMLA, ktoré sa stalo jediným systémovým rozhraním. Stále však ostáva možnosť pripojenia cez rozhranie ODBO (OLE DB for OLAP) vďaka ovládaču na strane klienta, ktorý transformuje volania do natívneho rozhrania XMLA.

Vrstvy Business Intelligence sú zjednotené do jedného modelu UDM – Unified Dimensional Model. Jeden dimenzionálny model stačí pre generovanie reportov (zostáv) a aj OLAP hyperkociek.

Vo verzii Enterprise je možné využiť proaktívne cacheovanie. Je to využitie vyrovnávacej MOLAP cache pamäte na ukladanie najčastejšie používaných alebo očakávaných výsledkov agregácií. Vo vzťahu k užívateľovi je cache pamäť neviditeľná. Umožňuje aj synchronizáciu multidimenzionálnych dát v pamäti cache vzhľadom k primárnemu dátovému zdroju. Dáta zostávajú v relačných databázach a spočítané agregácie sa ukladajú do multidimenzionálnych štruktúr. Princíp fungovania spočíva v sledovaní zmien a aktualizácii multidimenzionálnych dát v cache pamäti (vymazaním a vypočítaním nových agregácií).



Obr. 9 UDM - Unified Dimensional Model

Jadro data mining API tvorí DMX jazyk, ktorý je popísaný v špecifikácii OLE DB for Data Mining specification, plus language enhancements for SQL Server 2005 Data Mining [18] od Microsoftu.

Multidimensional Expressions (MDX) je dotazovací jazyk, ktorý nám umožňuje pracovať a dotazovať multidimenzionálne dáta. Je založený na XMLA špecifikácii s rozšírením pre SQL Server 2005 Analysis Services. MDX využíva výrazy zložené z identifikátorov, hodnôt, príkazov, funkcií, a operátorov ktoré vyhodnocuje Analysis Services a poskytuje návratovú hodnotu objektu (napríklad množina členov) alebo skalárnu hodnotu (string, číslo). MDX má použitie pre dotazovanie dát z klientskej aplikácie do kocky v SQL Server 2005 Analysis Services a na formátovanie dotazu (výsledku). Ďalej na vykonávanie úloh vrátane definícií vypočítaných členov, KPIs, administratívnych úloh vrátane zabezpečení dimenzií a buniek. MDX je syntakticky podobný jazyku SQL, ktorý sa používa v relačných databázach, ale nie je rozšírením jazyka SQL. Jazyk je popísaný v Multidimensional Expressions (MDX) Reference [19].

5.1.1 Porovnanie architektúry s MS SQL Server 2000

Jadrom architektúry MS SQL Serveru 2000 je objektový model pre správu analytických služieb DSO (Decision Support Objects). Definičné súbory vznikajú pri návrhu OLAP objektov a ukladajú sa do skladiska (Repository), čo je databáza pod správou sql servera, alebo kancelárska databáza MS Access. Pre správu, konfiguráciu analytických služieb a pre prácu s analytickými objektmi slúži nástroj Analysis Manager. Dáta, hyperkocky a dimenzie sú uložené v súborovom systéme a prístup k nim zabezpečuje vrstva OLE DB pre OLAP. O prístup z klientskych aplikácii sa stará služba Pivot Table Services (PTS). K dispozícii je aj rozhranie XMLA SDK využívajúce nový štandard XMLA (toto rozhranie bolo pridané až neskôr), pričom je potrebné použitie MS Internet Information Server a aplikácie ISAPI.

Vrstvy databázových, analytických a reportovacích služieb sú oddelené. Existuje možnosť analyzovať dáta a generovať reporty z relačných databáz aj analytických databáz. Každá hyperkocka môže byť tvorená len z jedného dátového zdroja. Pomocou MDX môžeme vyšpecifikovať podmnožinu dát a vypísať ju do dvojrozmernej tabuľky. Nevýhoda oddelených vrstiev je redundancia dát, keďže tie isté dáta sú často uložené v relačnej a aj v multidimenzionálnej databáze. Naproti tomu MS SQL Server 2005 umožňuje využitie výhod oboch typov databáz, pri využití UDM (Unified Dimensional Model).

Z užívateľského hľadiska nastáva integrácia nástrojov Query Analyzer, Enterprise Manager, DTS do Visual Studia 2005 IDE BI Workbench a SQL Workbench, ktoré preberajú ich funkcionalitu.

Z architektonického hľadiska AS 2005 podporujú výpočty a caching iba na strane servera. V oblasti configuration management je možné dátové zdroje a definície kociek meniť iba manuálne v závislosti na novom prostredí, AS 2005 zabezpečuje niekoľko obslužných programov. V AS 2005 vrstva Data Source Views (DSV) prináša organizáciu vyššej úrovne, keďže umožňuje organizáciu a kontrolu nad rámec predchádzajúcich možností (napríklad výber podmnožiny objektov v DS, premenovávanie stĺpcov, pridávanie vypočítaných virtuálnych stĺpcov). Taktiež DSV dovoľuje prácu na kockách a štruktúre aj bez priameho pripojenia na dátový zdroj.

Čo sa týka programátorských zmien, tak v AS 2005 je model a MDX syntax zjednodušený. Poskytuje možnosť písať MDX výrazy ako procedurálne skripty s usporiadanou postupnosťou príkazov, čo znižuje riziko nekonečnej rekurzie a umožňuje odstraňovanie chýb (debugovanie) krok za krokom. Skriptom môžeme obmedziť ich prístup a upravovať ich výkonnosť. Medzi ďalšie programátorské zmeny patrí XMLA, Server Trace Events, Analysis Management Objects (umožňuje vytvárať BI objekty programovateľne). Ostatné zmeny sa týkajú Data Miningu, Proactive Caching, KPI framework a bezpečnosti.

Komponent	MS SQL Server 2000	SQL Server 2005
Integrácia, transformácia a presun dát	Data Transformation Services (DTS)	SQL Server 2005 Integration Services
Relačný data warehouse	SQL Server 2000 relačná databáza	SQL Server 2005 relačná databáza
Multidimenzionálna databáza	SQL Server 2000 Analysis Services	SQL Server 2005 Analysis Services
Data mining	SQL Server 2000 Analysis Services	SQL Server 2005 Analysis Services
Riadený reporting	SQL Server 2000 Reporting Services	SQL Server 2005 Reporting Services
Ad hoc reporting	neaplikovateľné	SQL Server 2005 Reporting Services
Ad hoc dotazy a analýzy	Microsoft Office produkty (Excel, Office Web Components, Data	Microsoft Office produkty (Excel, Office Web Components, Data

	Analyzer, SharePoint Portal Server)	Analyzer, SharePoint Portal Server)
Vývojové nástroje pre databázový server	SQL Server 2000 Enterprise Manager, Analysis Manager, Query Analyzer, rôzne ďalšie nástroje	SQL Server 2005 Business Intelligence Development Studio
Nástroje pre správu databázového servera	Enterprise Manager, Analysis Manager	SQL Server Management Studio

Tab. 7 Komponenty BI a integrované nástroje v MS SQL Server 2000 a 2005

5.1.2 XMLA

XML for Analysis je novým štandardom pre komunikáciu s analytickými službami. Vychádza z existujúcich štandardov webových služieb (XML, SOAP) a na klientsky počítač nie je potrebné nič inštalovať. Je nezávislý na platforme aj operačnom systéme. Technologicky je nahradením rozhrania OLEDB for OLAP a datamining. Je spravovaný XMLA Councilom, ktorý bol založený firmami Microsoft, Hyperion a SAS.

Podpora v analytických službách MS SQL Servera 2005 je natívna a je jediným rozhraním pre prístup k analytických službám. Pri zostavovaní príkazov a interpretácii XMLA odpovedí sa využívajú knižnice ADOMD.NET pre získavanie dát a AMO pre správu serveru. Základné rozhranie disponuje metódami:

```
Discover
(
[in] RequestType As EnumString,     // typ požiadavku
[in] Restrictions As Restrictions,  // obmedzenia pre dáta a informácie, pre výstup
[in] Properties As Properties,      // parametre pre metódu
[out] Result As Rowset              // výstup vo forme Rowsetu
)
Execute
(
[in] Command As Command,            // príkaz
[in] Properties As Properties,      // parametre pre metódu
[out] Result As Resultset           //výstup vo forme Rowsetu
```

)

V BI Development Studio, Deplopment Wizard, Management Studio, je možné s kódom XMLA vykonávať operácie Save (uloženie definícii do formátu XML), Build (kompilácia a kontrola konzistentnosti), Deploy (zbalenie XML dokumentu do SOAP-XMLA a odoslanie na server), Process (naplnenie dimenzií a hyperkociek dátami).

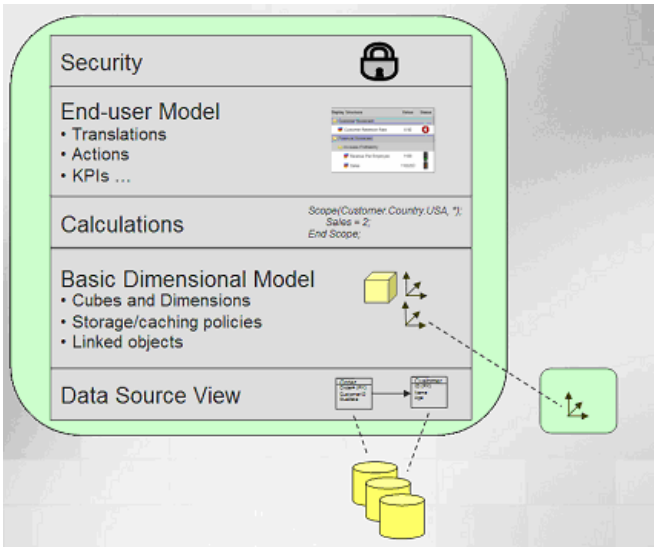
5.2 Nástroje pre prácu s Analytickými službami

Analytické služby sú koncipované ako služby na pozadí (podobne ako databázový server). MS SQL Server 2005 ponúka zjednotené užívateľské rozhranie pre nástroje slúžiace na administráciu, dotazovanie, návrh štruktúr a modelov. Sú to:

- MS SQL Server Management Studio
- Business Intelligence Development Studio

5.2.1 Dev Studio

MS Development Studio je vlastne podmnožina MS Visual Studia 2005, obmedzená pre vývoj BI aplikácii. Architektúra BI nadstavby nad databázou s využitím UDM je rozdelená do štyroch vrstiev. Najnižšia, prvá vrstva je Data Source View, slúžiaca na špecifikáciu tabuliek, pohľadov, atribútov, relačných väzieb a taktiež zjednocuje dáta z heterogénnych zdrojov. Na základe vytvorenej relačnej schémy, sa vytvárajú dimenzie a hyperkocky. Ďalšou úrovňou je úroveň základného dimenzionálneho modelu, zaoberajúca sa spôsobom ukladania, výpočtom a cacheovaním multidimenzionálnych dát. Nasledujúca, tretia vrstva je End-user Model, definujúca manipuláciu s dátami a výsledky súhrnu na najvyššej úrovni. Posledná vrstva sa zaoberá bezpečnosťou.



Obr. 10 UDM – Unified Dimensional Model [17]

5.2.2 MS SQL Management Studio

Vychádza z vývojového prostredia MS Visual Studio 2005 a je vybudované na základe MS Development Studia. MS SQL Management Studio je komplexne integrované prostredie pre správu databázového servera. V porovnaní s predchodcom zlučuje nástroje Enterprice Manager (administrácia servera), Query Analyser (konzolová aplikácia pre prácu s SQL dotazmi) a Analysis Manager (nástroj pre manipuláciu s multidimenzionálnimi databázami a modelmi).

Dotazovací mód Management Studia okrem klasických SQL a T-SQL dotazov je možné prepnúť do rôznych režimov:

- MDX Query (multidimenzionálne výrazy)
- DMX Query (výrazy pre dataminingové modely)
- XMLA Query (XML pre analytické služby)
- SQL Mobile Query (dotazy pre SQL Server CE)

5.3 Integračné služby

Integračné služby majú za účel získavanie dát z externých aj interných heterogénnych zdrojov, ich transformáciu, prečistenie a integráciu. Uplatnenie je teda pri

zbere dát a reorganizácii štruktúre dát. Taktiež sa starajú o aktualizáciu dát v dátových skladoch.

Hlavnou vrstvou je Data Transformation Pipeline (DTP), ktorá podporuje zdrojový a cieľový adaptér. Na nižších vrstvách sa vykonáva extrakcia a transformácia a patrí sem aj DTR API.

Nástroj DTS Designer, ktorý je súčasťou vývojového prostredia BI Development Studio, slúži pre návrh, virtuálne modelovanie, ladenie projektov, prevod transformácii a integráciu dát. Má dva oddelené grafické editory, pre návrh dátových tokov a pre návrh riadiacich tokov.

5.4 Reportovacie služby

Reportovanie služby si môžeme predstaviť ako bezstavový server, ktorý je súčasťou MS SQL Serveru 2005, spravujúceho metadáta, definície objektov a podobne. Tieto dáta majú vo vlastnej databáze, ktorá je tiež vo správe MS SQL Serveru 2005. Sú implementované ako služba operačného systému MS Windows, rovnako ako napríklad webový server.

Reportovacie služby plnia funkciu podpory rozhodovania, slúžia na návrh reportu a starajú sa o generovanie výstupu v elektronickej alebo papierovej forme. Reporty môžu byť statické alebo interaktívne (pomocou rôznych ovládacích prvkov môžeme report prispôsobovať).

Reportovacie služby sa skladajú z týchto častí:

- **Report Server** – webové, reportovacie služby.
- **Report Model Designer** – modelov reportov
- **Report Designer** – vizuálny nástroj pre vytváranie reportov
- **Report Manager** – správa, prehliadanie reportov
- **Report Builder** – vytváranie reportov

5.4.1 Architektúra

Základná vrstva je tvorená MS SQL Server Catalog, pod správou MS SQL Serveru. Je využívaná Report Serverom na ukladanie metadát, časových snímkov,

definícií reportov a zabezpečenia. Nad touto vrstvou je vrstva primárnych aplikačných rozhraní URL, WMI (správa reportovacích služieb) a webové služby. Vrstva zdieľaných služieb slúži na vygenerovanie reportu v danom formáte a taktiež na nastavenie bezpečnostných parametrov. Nasleduje medzivrstva pre klientske aplikácie, zobrazovanie dát s využitím webovej služby IIS / ASP.NET. Taktiež sú tu komunikačné rozhrania SOAP pre ostatné nástroje, WMI rozhranie pre prístup z administrátorskej konzoly.

5.4.2 Životný cyklus reportu

Životný cyklus reportu má tieto fázy: návrh, správa a doručenie. V prvej etape prebieha návrh reportov, ktoré sú potom publikované Report Serverom. Reporty sú definované vo vizuálnom prostredí BI Development Studia a výsledkom je XML dokument s obsahom zdrojového kódu v jazyku RDL (Report Definition Language). XML obsahuje koreňové elementy, URI namespace, základné parametre reportu (rozmery formulárov, okrajov). Reporty sa navrhujú pomocou Report Buildera, pričom je potrebné poznať štruktúru dát. Využíva .NET Framework 2.0.

V druhej etape riadime návrhy reportov, adresárov, zdrojov a pod.. Je možné využiť API webové služby. Report Manager je nástroj pre správu reportov. Možnosti správy sa rozdeľujú do kategórií General (popis a ostatné informácie o reporte), Parameters (nastavenie parametrov pre reporty), Data Sources (správa a zabezpečenie dát), Execution (využívanie cacheovania, časový rozvrh generovania reportu), History (ukladanie časových snímkov), Security (zabezpečenie dát).

V poslednej etape určujeme spôsob a formu doručenia. Reporty môžu byť vygenerované na požiadanie, alebo na základe časových rozvrhov. Doručenie prebieha od klienta k reportu (napr. cez URL adresu stránky reportu) a od reportu ku klientovi (napr. elektronickou poštou). Dostupné formáty výstupu sú webové formáty (HTML 4, HTML 3.2, HTML w/OWC), formáty tlače (TIFF, RTF, PDF) a dátové formáty (Excel, XML, CSV).

5.4.3 Jazyk RDL

Výsledkom návrhu reportu je XML dokument s obsahom zdrojového kódu v jazyku RDL (Report Definition Language). Sú to metadáta o dátach, návrhových schémach a vlastnostiach jednotlivých prvkov. Výrazy v jazyku RDL používajú syntax programovacieho jazyku VisualBasic.NET. Interaktívny parametrický report v MS SQL Server 2005 umožňuje meniť prezentovanú množinu dát zmenou parametrov.

Medzi grafické prvky návrhu patria Textbox (textové pole), Line (čiara), Table (tabuľka), Matrix (maticová tabuľka), Rectangle (obdĺžnik), List (zoznam), Image (obrázok), Subreport (vnorený report), Chart (graf). Medzi prvky zobrazovania dát patria List (voľne uloženie dát), Table (dáta organizované do stĺpcov) a Matrix (dynamická organizácia dát, pivot table). Agregačné funkcie operujú nad množinou záznamov. Medzi ne patria AVG, Count, CountDistinct, First, Last, Max, Min, RowNumber, RunningValue, StDev, StDevP, Sum, Var, VarP.

5.5 Analytické služby

Pri vytváraní kocky je k dispozícii Cube Wizard, aj s voľbou auto build. Možnosť práce s kockou v prostredí BI Development Studia rozširuje nástroj Cube Builder s nasledujúcimi možnosťami:

- Cube structure – prehliadanie a úprava modelu OLAP kocky. Relačné vzťahy medzi tabuľkami faktov a dimenzií.
- Dimension usage – definícia použitia dimenzií v OLAP kockách, editácia vzťahov medzi tabuľkami faktov a dimenzií.
- Calculations – návrh zdrojového kódu definície kalkulácie.
- KPIs – kľúčové indikátory (Parameter Key Performance Indicator), kľúčová metrika vyžadujúca cieľ, ktorý by sa mal splniť. Určia sa hodnoty, ktoré prevyšujú doporučenú toleranciu a ktoré sa blížia očakávanej hodnote. KPI je definovaný pomocou MDX výrazu: hodnota, cieľová hodnota, stav, trend, typ vizualizácie.
- Actions

- Partitions – partície, môžeme kocku rozdeliť podľa rôznych kritérií na jednotlivé partície. Existuje možnosť meniť parametre cacheovania.
- Perspectives – perspektívy, zjednodušený pohľad na kocku, logická vrstva.
- Translations – viacjazyčné popisy atribútov dimenzií, viacjazyčné konštanty.
- Browser – prehliadnutie a kontrola výsledkov analýzy.

5.5.1 MS Office ako klient analytických služieb

Pre najširší okruh bežných užívateľov je najvhodnejšia klientska aplikácia, ktorá je súčasťou kancelárskeho balíka alebo intranetový, internetový prehliadač. Kancelársky balík Microsoft Office poskytuje veľkú podporu analytických služieb na rôznych úrovniach a umožňuje pracovať v pripojenom alebo odpojenom režime. Môžeme vytvárať OLAP kocky priamo na klientskej úrovni aj bez analytického serveru, pomocou MS Excelu. S analytickými dátami dokáže pracovať aj databázový program MS Access.

Pri pripojení k OLAP serveru pomocou klientskej aplikácie sa ťažisko výkonu prenesie na server, ktorý klientskej aplikácii len odovzdá hotové výsledky. Sumarizované hodnoty vypočíta OLAP server a MS Excel dáta iba zobrazuje, prípadne ďalej spracuje a ukladá ako pohľady. Taktiež umožňuje export do formátov dokumentu, napríklad PDF.

MS Office využíva kontingenčné tabuľky (Pivot Table), ktoré umožňujú výmenu riadkov, stĺpcov, pričom riadky a stĺpce môžu mať hierarchickú štruktúru. Na návrh kontingenčnej tabuľky MS Excel 2003 a 2007 máme možnosť využiť vizuálne prostredie a šablóny.

6 Oracle 10*g*

Oracle je DBMS, moderný multiplatformový databázový systém s možnosťami spracovávania dát, vysokým výkonom a jednoduchou škálovateľnosťou. Databázový systém Oracle je vyvíjaný firmou Oracle Corporation. Aktuálna verzia je Oracle Database 10g. V nasledujúcej kapitole som niektoré informácie čerpal z [31].

Multidimenzionálne dáta sú uložené v analytic workspaces, kde sa k ním pristupuje pomocou OLAP rozhrania databázy Oracle. Samotné analytic workspaces sú uložené v tabuľkách v relačnej schéme a je možné s nimi zaobchádzať ako s inými relačnými tabuľkami. Sú k nim pridelené ID vlastníka (užívateľa) a ostatných užívateľov ktorým bol pridelený prístup k nim. Analytic workspaces boli navrhnuté práve na vysporiadanie sa s fyzickou organizáciou multidimenzionálnych dát. Sú založené na modeli indexového viacrozmerného poľa, ktoré zabezpečuje priamy prístup k položkám. Ako rozhranie na vyváranie a správu analytic workspaces slúži Analytic Workspace Manager. Umožňuje vytvoriť logický dimenzionálny model dát a jeho uloženie do XML súboru, taktiež mapu do relačných dátových zdrojov, načítanie a agregáciu dát. Analytic Workspace Manager spravuje životný cyklus analytic wokrspaces. Obsahuje aj nástroje pre vylepšenie z predchádzajúcej verzie Oracle9i a Oracle Express Server.

Ako natívny jazyk pre analytic workspaces sa používa OLAP DML. Je to dotazovací a definičný jazyk dát pre analytic workspaces. Definuje zásobníky dát a umožňuje manipuláciu s nimi. Všetky operačné úrovne (GUIs, Java, SQL) komunikujú s OLAP DML. Viac o DML uvádza Oracle OLAP DML Reference [20].

Infraštruktúru pre analýzu dát tvoria tri vrstvy:

- Oracle Database 10g
- Oracle Business Intelligence 10g
- Oracle Business Intelligence Tools 10g (Developer Suite)

6.1 Oracle Database 10*g*

Platforma umožňuje uloženie dát vo všetkých potrebných formách pre následnú analýzu ako napríklad prevádzkové systémy, dátový sklad, dátové tržnice. Taktiež zaisťuje manipuláciu s dátami ako ETL proces, multidimenzionálny sklad, beh dataminingových algoritmov. Pri požiadavkách na prenosy v reálnom čase sa využíva pre príjem, spracovanie a odoslanie správy vlastnosť databázy Oracle Advanced Queing. Databáza zaisťuje aj plánované spustenie jednotlivých úloh a krokov ETL a EAI procesov.

Databáza podporuje analýzu dát priamo v databáze, dotazovanie nad veľkými objemami dát, správa sumarizácii – Materialized Views, analytické funkcie. Tieto vlastnosti následne využívajú nástroje pre reportovanie, dotazovanie a pokročilé analýzy.

Pokročilé OLAP analýzy zahrňujú modelovanie výpočtov analytických funkcií. Oracle 10g zaisťuje multidimenzionálny sklad a kalkulačný engine, ktoré sú v dispozícii ako rozšírenie relačnej databázy OLAP Option. K dátam v MOLAP sklade je možné pristupovať ľubovoľným reportovacím, alebo analytickým nástrojom, ktorý využíva jazyk SQL, alebo natívne pomocou OLAP API. Data-miningové algoritmy sú implementované priamo v databáze a sú dostupné ako rozšírenie databázy Data Mining Option. V dispozícií je aj rozhranie pre integráciu data miningu do ľubovoľných aplikácií.

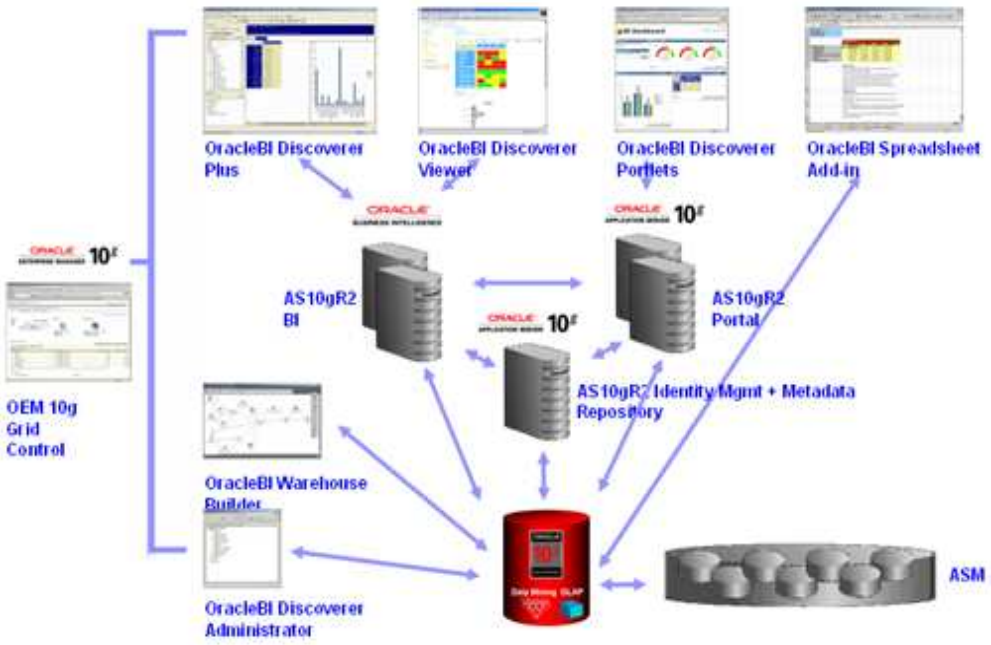
6.2 Oracle Business Intelligence 10*g*

Uvedené v Decembri 2005. Oracle Business Intelligence 10g je kompletné, samostatné riešenie pre potreby základných analýz ako reportovanie, ad-hoc dotazovanie, pokročilých MOLAP aj ROLAP analýz. Prinášajú možnosť vývoja vlastných aplikácií, správy metadát riešenia, návrh, tvorbu a monitorovanie dátových skladov. Charakteristická architektúra je trojvrstvová, škálovateľná, centrálny sklad.

Oracle Business Intelligence 10g je možné prepojiť s produktom Oracle Portal, ktorý podľa potreby umožňuje užívateľom kombinovať výstupy analýz dát s informáciami z iných externých zdrojov.

Súčasťou sú nástroje pre distribúciu pripravených reportov a dotazovania. Umožňujú prístup k metadátam dátového skladu, ETL procesu a monitorovanie priebehu ETL procesu (nástroj Warehouse Builder). Pre základné ale aj pokročilé MOLAP analýzy a dotazovanie je určený nástroj Oracle Discoverer Plus s pripojením k relačnému alebo multidimenzionálnemu skladu. Vďaka spomínanej priamej podpore Oracle Databázy pre analýzu dát, pokrývajú tieto nástroje veľký počet požiadaviek na analýzu. Je možné použiť rozšírenie funkcionality MS Excel: Oracle Spreadsheet Add-In, taktiež Discover Viewer alebo Report Services pre distribúciu multidimenzionálnych alebo jednoduchých analýz a predpripravených reportov užívateľom.

- **OracleBI Discover** – dotazovanie, reportovanie, analýza
- **OracleBI Spreadsheet Add-in** – Microsoft Excel Integrácia, priamy prístup
- **Discoverer Portlets** – nastavovateľné súhrny
- **OracleBI Warehouse Builder –** posilňuje kvalitu dát, ETL funkcionalitu
- **Oracle Business Intelligence Beans –** vývoj BI aplikácii



Obr. 11 Architektúra Oracle BI 10g

6.2.1 OracleBI Discover

Je to nástroj pre ad-hoc dotazovanie a taktiež analytický a reportovací nástroj. Doručovanie reportov uskutočňuje cez Web Prehliadač. Poskytuje prístup do relačných, aj OLAP dát. Skladá sa z dvoch vrstiev. Prvá je Discoverer J2EE vrstva a druhá Discoverer Plus & Viewer.

Oracle BI Discover Plus – Nástroj pre tvorbu základných a pokročilých analýz. Nie je potrebná znalosť SQL. Je to kompletné prepracovanie z Discoverer Plus 9iAS / 10gR1. Zabezpečuje funkcionalitu ako vytváranie a ukladanie pracovných zošitov (workbooks), drill to detail, rotacie, drill to related. Umožňuje prácu s grafmi, určovanie podmienok, triedenie, kalkulácie a parametrizovanie, dolovanie dát z OLTP do OLAP. Oracle BI Discover Plus existuje v dvoch verziách Discoverer Plus (náhrada Discoverer Desktop) a Discoverer Plus OLAP (náhrada Oracle Sales Analyzer).

Rozdiely	**Discoverer Plus**	**Discoverer Plus OLAP**
Zdroj dát	Relačné dáta (Oracle, ODBC)	OLAP dáta (Oracle Relational, AW)
Uloženie Metadát	End User Layer	Discoverer Catalog
Typ User Interface	Single Document Interface	Multiple Document Interface
Podmienky a kalkulácie	Edit Worksheet Single Dialog	BI Beans Query a Calculation Beans
Sumarizácie	Materialized Views	Materialized Views, AWs
Administratívny nástroj	Discoverer Admin, OWB	ASControl, AWM2, OWB “Paris” (future)

Tab. 8 Discoverer Plus a Discoverer Plus OLAP rozdiely

Discoverer Viewer – Prehliadanie už pripravených reportov, analýz. Jedná sa o tenký klient, DHTML prehliadanie workbookov. Pokrýva OLTP aj OLAP, má možnosť ukladať vykonané zmeny. Poskytuje funkcie ako pivoting, rezy, triedenia, zmena hodnôt atribútov, formátovanie vzhľadu reportov. Tlačenie reportov je možné vo výstupe pdf, xls, html, csv, rtf. Umožňuje publikáciu reportov pomocou mailu.

Discoverer Portlets – Publikovanie vytvorených reportov a analýz. Stará sa o doručovanie reportov pomocou Oracle Portal. Vyžaduje identitu vrstvy manažmentu skladu metadát a vrstvy Oracle Portal. Umožňuje prácu s grafmi, tabuľkami.

6.2.1.1 Porovnanie OracleBI Discover a Discoverer Desktop

Discoverer Plus 9iAS/10g je nástupcom Discoverer Desktop a prináša veľa zlepšení. Medzi užívateľské nedostatky Discoverer Desktop patrí napríklad nedostatok užívateľskej voľnosti pri vyberateľných bunkách (dragging, dropping). Formátovanie je možné iba cez Worksheet Wizard. Tlačenie je možné iba cez web prehliadač. Taktiež neumožňuje uloženie do súborového systému. Medzi výhody Discoverer Desktop patrí to, že je založené na webovom rozhraní a preto nie je potrebná žiadna inštalácia klienta.

Uvedené nedostatky rieši Discoverer 10.1.2. Prináša zlepšenie vzhľadom na grafické rozhranie a škálovateľnosť. Rieši a umožňuje formátovanie úrovní buniek, reorganizáciu pomocou drag and drop, dopĺňa Analysis Bar, umožňuje drill to related item, drill to detail. Taktiež pridáva podporu tlačenia zostáv a ukladanie do súborového systému. Oproti Discover Desktopu pridáva riadiaci panel, skvalitnenie grafov, integráciu s OLAP dátami, doručovanie cez web (nie je potrebná inštalácia klienta).

6.2.2 OracleBI Spreadsheet Add-in

Jedná sa o dodatok pre Microsoft Excel 2000, XP, 2003 s možnosťou aj samostatnej inštalácie, ktorý je prístupný z Excel Menu bar**.** Je možné ho označiť ako kľúčový komponent, orientovaný na koncového užívateľa a dáva mu možnosť dotazovať, zobrazovať a riadiť Oracle OLAP dáta priamo z MS Excel. Umožňuje prístup na čítanie do Oracle OLAP dát (relačných alebo AW).

6.2.3 Oracle Business Intelligence Beans

Vyžaduje JDeveloper 10.1.2. Slúži pre dotazovanie tenkého klienta. Zlepšuje integráciu Jdeveloper rozšírením JSP Tag knižnice, rozšírenou UIX podporou, podporuje vizuálne návrhové nástroje. Úroveň objektovej bezpečnosti zabezpečuje BI Beans Catalog.

6.3 Oracle Business Intelligence Tools 10*g*

Patria sem nástroje pokrývajúce všetky potreby vývoja a správy systémov pre analýzu dát. Nástroje sú úzko zviazané s databázou, Business Intelligence serverom a nástrojmi, ktoré ich využívajú. Nástroje sú integrované a zdieľajú metadáta v štandarde CWM (Common Warehouse Model) vznikajúce pri návrhu a správe BI a DWH riešení. Pre ich uloženie používajú databázu Oracle.

Súčasťou balíku je nástroj pre návrh a správu dátových skladov a procesov pre jeho plnenie (ETL) Warehouse Builder, v ktorom je zahrnuté aj prepojenie na SAP (SAP Integrator) a mainframe (Pure Extract) a podpora čistenia zákazníckych dát (Pure Integrate). Obsahuje aj nástroj pre návrh reportov Reports Builder a administrátorskú časť nástroja pre dotazovanie Discover Administration Edition. Využívajú špeciálne vlastnosti databázy pre reporting a dotazovanie.

Pre vývoj analytických aplikácii nad multidimenzionálnym skladom je k dispozícii vývojové prostredie pre Java JDeveloper spoločne s knižnicami Oracle Business Intelligence Beans. Pomocou JDevelopera môžeme vyvíjať aj aplikácie využívajúce data miningové algoritmy implementované v databáze.

7 Cognos

Spoločnosť Cognos patrí medzi najväčších svetových producentov firemných informačných systémov (BI) a plánovacieho softvéru pre veľké firmy. Je to produktové riešenie nezávislé na DBMS, nadstavbový manažérsky informačný systém. Je kompatibilné s viacerými DBMS vrátane MS SQL Serveru, Oracle, alebo IBM DB2 a ďalšími. Najnovším predstaviteľom produktovej generácie Cognosu je Cognos 8 Business Intelligence. Jedná sa o nástupcu Cognos Series 7. V nasledujúcej kapitole som niektoré informácie čerpal z [33],[34].

7.1Architektúra Cognos 8 BI

Cognos 8 BI je postavený na jednotnej univerzálnej a otvorenej Service Oriented Architecture (SOA) architektúre. Cognos 8 BI nadviazal na nástroj ReportNet, ktorý je taktiež postavený na SOA architektúre s charakteristickými vlastnosťami ako škálovateľnosť, podpora, UNICODE, viacjazyčnosť. Umožňuje využitie existujúcej IT infraštruktúry ako napríklad bezpečnostných systémov, RDBMS, aplikačných serverov, webových aplikácií, sieti a metadát. Je integrovateľný s web portálmi IBM WebSphere, SAP Enterprise Portal, Plumtree Portal. Je multiplatformový, podporuje operačné systémy Windows, Unix, Linux. Taktiež podporuje rôzne zmiešané aplikačné prostredie.

Cognos 8 BI je založený na webovom užívateľskom a administrátorskom rozhraní. Architektúra webových služieb (web services) minimalizuje požadované zdroje pre vývoj, prácu a údržbu. Využíva otvorené štandardy ako XML, SOAP a WSDL. Výhody využitia webového rozhrania sú, že nie je potrebná inštalácia klientskej aplikácie, pluginov, alebo appletov. Princíp „ukáž a klikni" rozhranie zvyšuje adaptabilitu pre užívateľov, znižuje závislosť na administrátoroch a odbornosť. Administrácia je centralizovaná a distribuovaná priamo k objektu záujmu. Umožňuje vzdialený prístup do systému. Prípadné vykonávané zmeny sa zobrazia všetkým užívateľom súčasne.

Uvediem porovnanie s ReportNetom z ktorého Cognos 8 BI vychádza. Webový portál bol rozšírený o Metric Studio (nahradzujúce Metrics Manager), Event Studio (miesto komponentu NoticeCast), Analysis Studio, ktoré umožňuje OLAP analýzu dát (nahradzuje PowerPlay Web). Model vo Framowork Manageri je doplnený o možnosť dimenzionálneho modelovania relačných dát, s využitím dimenzií, faktov. Dovoľuje definovať hierarchie, ukazovatele a ich alokáciu priamo v modeli metadát. Spoločný dotazovací stroj umožňuje vykonávať multidimenzionálnu analýzu v Analysis Studio a reportovanie v Report Studiu aj nad relačnými aj multidimenzionálnymi dátami.

Architektúra webových služieb je trojvrstvová, skladá sa z prezentačnej vrstvy, aplikačnej vrstvy a dátovej vrstvy.

7.1.1 Prezentačná vrstva

Obsahuje užívateľské a administrátorské rozhranie dostupné cez webové rozhranie. Nachádzajú sa tu aj nástroje pre tvorbu komplexných zostáv alebo analýzu dát. V tejto vrstve je taktiež aj rozhranie pre MS Excel: Cognos Office connections a Window Client Add-Ins. Štruktúrou prezentačnej vrstvy sa dosahuje distribuovateľnosť a jednoduchosť.

7.1.2 Aplikačná vrstva

Jadrom aplikačnej vrstvy je modul BI Dispatcher implementovaný ako Java servlet. Prijíma požiadavky od klientskych aplikácií prezentačnej vrstvy a rozdeľuje ich medzi lokálne služby ako repoty service, presentation service, metrics service, event service delivery service, monitoring service, content manager service, schedunling service, batch report service, alebo ich nasmeruje na iný dispatcher.

Integrácia je v spoločnom dotazovacom stroji query engine, ktorý spracúva požiadavky na dáta pre reportovanie a analýzu. Umožňuje pristupovať k relačným dátam pomocou SQL a multidimenzionálnym dátam pomocou MDX, BAPI. Výsledkom je zvýšená konzistentnosť.

7.1.3 Dátová vrstva

Cognos podporuje veľký počet štandardov a dátových zdrojov bez nutnosti dodatočného kódovania alebo migrácie dát. Stratégia otvoreného prístupu k dátam umožňuje spracovávať rôzne dátové zdroje do spoločného modelu metadát.

Spoločný model metadát sa vytvára pomocou Framework Managera. Ako metadátové zdroje je možné použiť Oracle, MS SQL Server, XML, JDBC. Definície zostáv, analýz, metadáta, výstupy sú uložené v RDBMS. Model metadát obsahuje podporu pre viacjazyčné aplikácie a formátov meny, dátumu.

7.2 Nástroje Cognos 8 BI

Cognos 8 BI poskytuje nástroje slúžiace na analýzu dát, ad-hoc dotazovanie, reportovanie, vytváranie komplexných reportov nezávisle od zdroja dát relačných alebo multidimenzionálnych dát.

7.2.1 Analysis Studio

Analysis Studio je webové prostredie Cognos 8 BI slúžiace na analýzu dát. Funkcionalitou a rozhraním vychádza z PowerPlay Web. Nastalo niekoľko zmien v rozhraní návrhu analýzy. Medzi ne patrí nová sumárna lišta, možnosť vložiť do riadkov a stĺpcov viac položiek z rôznych dimenzií. Medzi nové analytické možnosti patrí rozšírené triedenie a zoraďovanie, funkcia Top X, možnosť vytvárania filtrov na kategórie a kombinovania filtrov, rozšírené možnosti načítavania kategórií. Funkcionalitu reportovania nad OLAP dátami, prezeranie dát prebrali nástroje Cognos Viewer, Report Studio, Query Studio. Analysis Studio je teda určené len výhradne pre analýzu dát.

7.2.2 Query Studio

Query Studio je nástroj na ad-hoc dotazovanie a jednoduché reportovanie. Na vytvárané dotazy je možnosť použiť štandardnú šablónu, predpripravenú v Report Studiu. Ďalšie možnosti sú podmienené formátovanie hodnôt v stĺpcoch dotazu, reťazenie filtrov

pomocou logických operátorov AND, OR, vnáranie a vynáranie pre zobrazenie detailnejších dát nižšej alebo vyššej hierarchickej úrovni (drill down, drill up).

7.2.3 Report Studio

Slúži na vytváranie komplexných reportov nad dátami z relačných databáz, multidimenzionálnymi dátami a kockami PowerCube. Cognos Viewer ponúka dynamické reporty. Report Studio je aj dashboardingovým nástrojom. Dostupné sú nové typy grafov, diagramov. Ponúka možnosť vytvárať krížové tabuľky s uzlami, definovanie ľubovoľných lokálnych väzieb a prepájanie dotazov. Existuje možnosť využiť aj existujúce, vytvorené analýzy v Analysis Studiu, aj dotazy z Query Studia.

7.2.4 Data Manager

Jedná sa o vylepšenie produktu DecisionStream 7. Komponent Data Manager umožňuje vytvoriť aplikáciu manažérskeho informačného systému. Obsahuje ETL nástroj pre spracovávanie dát z rôznych dátových zdrojov do jednej spoločnej databázy pozostávajúcej z predmetne orientovaných dátových tržníc, a spoločnej dimenzionálnej základni (Dimensional Framework). Dimenzionálna základňa zabezpečuje konzistentnosť dát v dátových tržniciach dátového skladu. Ako grafické návrhové rozhranie pre tvorbu, úpravu ETL procesov slúži Data Manager Designer. Výkon pre RTL procesy poskytuje Data Manager Engine a sieťové služby Data Manager Network Services. Na prístup k dátovým zdrojom ľubovoľných typov využíva spoločné dotazovacie rozhranie Cognos Querying Framework.

Data Manager 8 podporuje kódovanie UNICODE, čo umožňuje spracovávať dáta z rôznych systémov (ich kódových stránok) a konvertovať ich do UNICODE pre uloženie do dátových tržníc. Umožňuje návrh efektívnejších ETL procesov pomocou odvodených dimenzií, filtrov na úrovni DataStream a Delivery, particionovanie tabuliek, automatizované riadiace atribúty faktových tabuliek, rozvinutie hierarchií, podporu nových dátových typov ako BLOB.

7.2.5 Metric Studio

Metric Studio je Cognos Scorecarding. Pomáha orientovať tímy a upravovať taktiku na plnenie strategických cieľov a porovnávať ich pomocou metrík. Predstavuje nástupcu pre Metrics Manager v.2. Metric Studio poskytuje spoločné rozhranie na sledovanie výkonnosti a správu obsahu. Umožňuje sprístupniť stav najdôležitejších metrík na portál Cognos Connection (taktiež správa metrík), prípadne použiť informovanie prostredníctvom e-mailu.

7.2.6 Event Studio

Event Studio je určené na identifikovanie individuálnych problémov alebo javov a následné využitie týchto udalostí. Predstavuje nástupcu pre Cognos NoticeCast 7. Event Studio je nástroj pre definovanie a správu udalostí užívateľmi. Užívateľ môže definovať podmienku na základe položiek, širokej škály dostupných funkcií a výpočtov. Tiež umožňuje definíciu činností v prípade splnenia podmienky a časový rozvrh kontroly dát agentom. Užívatelia majú možnosť nechať sa upozorniť o konkrétnej operatívnej udalosti, udalosti s záväznosťou na metriky (performance related event). Taktiež je následne možné automatizovať činnosti nasledujúce po vzniku udalosti ako oznámenie, spustenie reportu, spustenie agenta, zapísať alebo zmeniť hodnotu databázy, spustiť webovú službu.

8. Porovnanie OLAP nástrojov a databázových systémov

V predchádzajúcom texte som sa snažil architektonicky porovnať balíky OLAP nástrojov. Taktiež som porovnával zlepšenie a zmeny oproti predchádzajúcim verziám. V nasledujúcich kapitolách porovnávam nástroje navzájom.

8.1 Porovnanie OLAP nástrojov

Microsoft aj Oracle sú považované za OLAP lídrov a ich BI platforma je porovnateľná s ostatnými ako Hyperion, IBM. Oracle BI riešenia však nedosahujú významný trhový podiel, čo môže byť zapríčinené tým, že ich nástroje u užívateľov nie sú zaužívané, prípadne vyššia cena za balíky produktov. Všetky tri balíky produktov sú všetky viac ako reportovacie a návrhové nástroje, všetky ukladajú definície prvkov reportu, zabezpečujú flexibilnú bezpečnosť reportu a rozsiahle distribučné možnosti reportu, každý má svoju vlastnú množinu podporovaných metadát, osobitnú inštaláciu a rozdielne výstupné užívateľské a návrhové rozhranie.

8.1.1 Porovnanie vzhľadom na prostredie

Reporting Services Microsoftu je licencovaný komponent SQL Serveru 2005, ktorý je poskytnutý pri kúpe SQL Servera. Nie je možné ho použiť s inými DBMS. Podobne aj Oracle Business Intelligence nie je možné použiť s inými DBMS, avšak je poskytovaný len v Enterprise Edition (Oracle Database 10*g*), ale je dostupný tiež ako samostatný produkt. Naproti tomu Cognos 8 BI je nezávisle licencovaný a kompatibilný s viacerými DBMS ako MS-SQL Server, IBM-DB2, Oracle. Tu môže zavážiť cenový faktor, keďže spomenutá BI podpora je v MS poskytnutá s SQL Serverom a nie je potrebná dodatočná licencia na vyššej úrovni. Teda zviazanosť nástrojov s SQL Serverom, alebo zviazanosť nástrojov s Oracle môže priniesť výhody v podobe lepšej natívnej podpory. Na druhej strane práve nezávislosť riešenia OLAP nástrojov Cognos

prináša faktor univerzálnosti, avšak sa sústredí a je určený iba ako kompletný súbor BI nástrojov.

Práve vďaka zviazanosti nástrojov je nutné sa zaoberať aj výberom správneho DBMS, aby vyhovoval aj iným architektonickým, užívateľským požiadavkám a použiteľnosti vzhľadom na zavádzané prostredie. MS SQL Server poskytuje podporu pre ukladanie, indexovanie a dotazovanie XML dokumentov vo vnútri relačnej databázy vrátane XML. U Oracle túto funkcionalitu zabezpečuje Oracle XML DB. Ďalší dôležitý faktor môže byť natívna podpora .NET v MS SQL Serveri a natívna podpora Javy v Oracle. Výhoda MS SQL Server 2005 je integrované návrhové a správcovské prostredie s MS Visual Studio 2005 a vôbec integrácia s produktmi Microsoftu. Oracle v rámci Oracle Business Intelligence Tools poskytuje vývojové prostredie Java Jdeveloper. Záleží na tom, ktorý prístup v rámci vývoju aplikácií podporujeme a využívame.

Všetky riešenia zhodne prinášajú podporu kancelárskeho balíka Microsoft Office, pri použití ako klientskej aplikácie. U Microsoftu je to automatická podpora, u Oracle je to Oracle BI Spreadsheer Add-in komponent slúžiaci na dotazovanie, zobrazovanie, riadenie Oracle OLAP dát priamo z MS Excel. Cognos taktiež obsahuje podporu pre pripojenia cez MS Excel.

Spomínané riešenia zhodne poskytujú možnosť využitia webového rozhrania ako klientskej aplikácie pre prezeranie, manipuláciu s reportami a administrátorské rozhranie. Výhody tohto rozhrania som už rozvinul, patrí sem napríklad nepotrebnosť inštalácie klientskej aplikácie, pluginov, zvýšená adaptabilita užívateľov, znížená potreba administrácie a nižšie požiadavky na odbornosť.

MS SQL Server je postavený na podpore operačného systému MS Windows (MS Windows 2000 Server SP4, MS Windows Server 2003), naproti tomu Oracle je multiplatformový (Unix, Linux, FreeBSD, MS Windows, Novell Netware, IBM OS/390 or MVS and VMS) a Cognos je tiež multiplatformový (MS Windows, Sun Solaris, HP-UX, IBM AiX). Práve toto môže byť rozhodujúci faktor, prekážka pri rozhodovaní, keďže riešenie Microsoftu je obmedzené iba na operačný systém Windows.

8.1.2 Dotazovanie multidimenzionálnych dát

Spomenuté riešenia Microsoft, Oracle aj Cognos podporujú obe, aj relačné a aj multidimenzionálne dátové sklady. U Oracle dochádza k určitému rozdeleniu a duplikácii funkcionality u relačnej databázy a u Oracle OLAP. Relačné databázy obsahujú príkazy dimension, rollup, cube, model a Oracle Olap podporuje tie isté funkcie. Relačné databázy podporujú riadenie agregácií pomocou materializovaných náhľadov (materialized views), Oracle OLAP ich podporuje použitím agregačných máp. Relačné databázy majú veľa BI funkcií ako lead, lag, rank, ratio. Oracle OLAP má vlastnú množinu matematických, štatistických funkcií.

Ako dotazovací jazyk ktorý umožňuje spracovať a dotazovať multidimenzionálne dáta má Microsoft aj Oracle vlastné riešenie. Microsoft využíva SQL Server MDX založený na XMLA špecifikácii a poskytuje DDL (Data Definition Language) syntax pre manipuláciu s dátovými štruktúrami. Ako natívny jazyk pre prístup k multidimenzionálnym dátam, dotazovanie, navigáciu a analytickým funkciám využíva Oracle OLAP DML. Ďalšie rozhrania, ktoré Microsoft poskytuje, sú AMO (Analysis Management Objects) pre prístup k zdrojovým súborom OLAP definícií uložených vo formáte XML. V SQL Serveri je zjednotené systémové rozhranie XMLA štandardom pre komunikáciu s analytickými službami, využívajúc knižnice ADOMD.NET. Naproti tomu Oracle využíva Java OLAP API pre aplikácie a umožňuje pripojenie, navigáciu, selekciu a niektoré analytické funkcie, prípadne spustenie OLAP DML príkazu. Taktiež u Oracle pomocou preddefinovaných PL/SQL balíkov je možné pristupovať k OLAP príkazom, alebo OLAP multidimenzionálnym náhľadom priamo. Cognos poskytuje integrovaný dotazovací stroj – query engine, ktorý umožňuje pristupovať ku všetkým zdrojom dát, relačným pomocou SQL aj multidimenzionálnym pomocou MDX, BAPI.

8.1.3 Porovnanie na základe ETL.

Oracle ponúka nástroj Oracle Warehouse Builder (OWB), Microsoft rieši ETL pomocou integračných služieb a ako ETL nástroj Cognosu slúži Data Manager. OWB je grafický nástroj na návrh, vytvorenie a načítanie Oracle dátového skladu. Je to návrhové prostredie, ktoré pracuje s metadátami a nie priamo na fyzických objektoch (tie sa vytvárajú z metadáta katalógov). Poskytuje nám plnohodnotnú množinu ETL

transformácií, mapovanie, change management, možnosť porovnávania verzií, časovanie zmien. Nevýhoda je, že všetky zmeny dátového skladu by mali byť vykonávané pomocou OWB, pretože zmeny urobené z vonkajšieho prostredia sa neodrazia v zmenách metadát. Ďalší problémom sú funkcie ETL spracovávania, ktoré môžu byť programované iba v závislosti na tabuľky, o ktorých má OWB metadáta.

Microsoft SSIS môže byť využité aplikáciami dátového skladu a tak isto aj ostatnými databázovými aplikáciami. Podporuje pohyb a transformáciu z hetegorénnych dátových zdrojov (oboma smermi), vrátane zdrojov RSS a cieľom SQL Server Reporting Services. SSIS tak isto podporuje veľkú množinu ETL transformácii, mapovanie a poskytuje množinu predpripravených funkcií (fuzzy matching), ktoré však podobne ako u Oracle môžu byť dodatočne dopĺňané.

Cognos Data Manager taktiež umožňuje spracovať dáta z mnohých dátových zdrojov do jednej databázy (predmetne orientované dátové tržnice). Využívajú dimenzionálnu základňu (Dimensional Framework), ktorá zabezpečuje konzistentnosť dát. Poskytuje grafické návrhové rozhranie na tvorbu a úpravy ETL procesov. Výkon zabezpečuje Data Manager Engine, ktorého procesy môžu byť spustiteľné aj z vonkajšieho prostredia, samostatne bez prítomnosti celého katalógu Data Managera. Taktiež poskytuje veľkú množinu ETL transformácii, mapovanie, filtre.

8.1.4 Porovnanie Metadát

V Oracle každý produkt má svoju vlastnú množinu metadát, uloženú v SQL tabuľkách. Správa týchto katalógov pri zmene štruktúry nie je automatická, ale aplikácie spravujú štruktúry metadát pomocou PL/SQL volaní. Je možné urobiť zmenu objektu, ktorá sa nezohľadní v týchto katalógoch, avšak sa to môže odraziť v nesprávnom fungovaní aplikácie.

Na druhej strane Microsoft využíva jednotný, integrovaný prístup k metadátam. Pomocou Data Sources Views zadefinujeme schému zdrojových dát, Report Builder aplikuje schému na obsah a umožní navigáciu. UDM pridáva business náhľad, podporuje metriky a KPIs. Ľubovoľný nástroj môže využívať metadáta na svoj účel.

Cognos využíva spoločný model metadátového modelu, do ktorého spracováva ľubovoľnú kombináciu dátových zdrojov. Spoločný metadátový model vytvára

Framework Manager z metadátových zdrojov ako Oracle, Microsoft, SQL, Server, XML, JDBC. Sú uložené v RDBMS.

8.1.5 Porovnanie reportovacích možností

Oracle reportovací nástroj je Oracle Reports. Umožňuje prístup k rôznym zdrojom ako SQL tabuľky, OLAP, XML súbory, textové súbory. Poskytuje výstup reportu vo formátoch PDF, XML, HTML, RTF, MS Excel. Na doručovanie reportov môžeme využiť tlačenie, e-mail a OracleAS Portal. Taktiež máme možnosť formátovania reportu využitím tabuliek, matíc, mailových návestí a ďalšie.

Podobne aj Reporting Services od Microsoftu poskytuje vstupy SQL tabuľky, Analysis Services kocky, XML, textové súbory. Výstupy podporuje tie isté PDF, MS Excel, XML, HTML. Možnosti doručovania a formátovania reportu sú tiež porovnateľné.

Pri zdroji dát Cognos môže využiť najväčšiu variabilitu. Ako relačné zdroje Oracle, SQL, IBM, Teradata, Sybase, ODBC. Ako OLAP zdroje Cognos OLAP, SAP BW, Microsoft SSAS, Essbase, Oracle 10G, IBM DB2 CubeViews. Ďalej napríklad XML, Java beans, JDBX, LDAP, WSDL, textové súbory, Excel súbory , Access súbory. Variabilita možných vstupov je teda vyššia. Formáty výstupu a možnosti doručovania sú porovnateľné s SQL Server RS a Olap Reports. Možnosti formátovania sú väčšie, keď môžeme využiť dynamické reporty, grafy, schémy a veľa ďalších prvkov.

8.1.6 Ďalšie možné Business Intelligence riešenia

Medzi ďalšie dostupné multiplatformové riešenia v oblasti Business Intelligence patrí Hyperion System 9. Jedná sa o business performance management system. V rámci neho BI platforma zameriavajúca sa na všetky typy analýz a reportov je Hyperion System 9 BI+. Ďalšou zložkou je Applications+, čo je súbor integrovaných aplikácií pre manažérov, analytikov pre sledovanie výkonnosti, presností predpovedí a podmienok trhu. Data Management Services je skupina produktov pre riadenie finančných a operatívnych dát, metadát a kvality dát. Viac o ich produkte je možné nájsť v zdroji [21].

Ďalšie nezávislé riešenie prináša Business Objects so skupinou produktov BusinessObjects XI. Poskytujú nástroje na reportovanie, dotazovanie, analýzu, performance management, information management. Viac o ich produkte [22].
Iné riešenia Business Intelligence riešenia prináša: MicroStrategy [23], SAP [24], Cartesis [25], Applix [26], Infor [27].

8.1.7 Analýza podielu na trhu

Vychádzam z analýzy reálneho podielu OLAP produktov na trhu vykonanej nezávislou výskumnou organizáciou OLAP Report [28].

Trh s OLAP produktmi rastie viac ako sa predpokladalo, za posledný rok bol zaznamenaný nárast o 16.4%. Podiel Microsoftu za posledný rok akceleroval pravdepodobne vďaka uvedeniu MS SQL Serveru 2005. Za posledných 5 rokov nenastáva žiadna zmena u prvých 6 miest. Najväčší predajcovia ako Microsoft, Oracle, SAP, Business Objects poskytujú OLAP nástroje ako súčasť väčšieho balíka, napríklad MS Analysis Services a OLAP nástroje SAP BW nie sú predávané samostatne ale v sade produktov. Oracle a SAP ktoré majú tendenciu dominovať v iných sektoroch sú relatívne slabé na trhu OLAP technológii. V nasledujúcich rokoch splynutím Hyperionu s Oracle, sa však dostane Oracle na druhé miesto za Microsoft.

Pri rozhodovaní o nasadení OLAP produktu sa ukazujú rozdiely aj vzhľadom na geografický región, alebo veľkosť spoločnosti. Napríklad MS Analysis Services sú typicky preferované menšími spoločnosťami, na druhej strane SAP BW, MicroStrategy väčšími. Podobne, Business Objects a SAP sú relatívne silnejšie v Európe, a v Severnej Amerike sú to naopak MicroStrategy a Hyperion. Veľké spoločnosti, ktoré sa nešpecializujú iba na OLAP ako Microsoft, Oracle majú silnú pozíciu u malých BI špecialistov. Čo sa týka zamerania, MicroStrategy má silnú pozíciu v maloobchode, Applix vo finančnom a poisťovacom sektore a Microsoft a Business Objects v IT priemysle.

Vendor	2006		2005		2004		2003	
	Pozícia	Podiel	Pozícia	Podiel	Pozícia	Podiel	Pozícia	Podiel
Microsoft	1	31.6 %	1	27.9 %	1	27.3 %	1	26.1 %
Hyperion Solutions	2	18.9 %	2	19.2 %	2	20.6 %	2	21.9 %
Cognos	3	12.9 %	6	14.0 %	3	14.1 %	3	14.1 %
Business Objects	4	7.3 %	4	7.4 %	4	7.2 %	4	7.7 %
MicroStrategy	5	7.3 %	5	7.2 %	5	7.1 %	5	6.2 %
SAP	6	5.8 %	6	5.9 %	6	6.0 %	6	5.8 %
Cartesis	7	3.7 %	8	3.9 %	10	3.1 %	10	3.1 %
Applix	8	3.6 %	10	3.3 %	9	3.2 %	9	3.1 %
Infor	9	3.5 %	7	5.1 %	7	4.9 %	7	4.9 %
Oracle	10	2.8 %	9	3.4 %	8	3.7 %	8	4.0 %

Tab. 9 Analýza podielu na trhu [32]

8.2 Porovnanie databázových systémov

V nasledujúcej časti porovnám databázové systémy, vzhľadom na podporu rozhodovania a výkonu analytického spracovania.

Použijem sadu porovnávacích testov (benchmarkov) TPC-H, neziskovej organizácie Transaction Processing Performance Council [29], zameranej na definovanie porovnávacích testov v oblasti transakčného spracovania a databáz. Zastúpenie a členstvo v nej majú Compaq, Data General, Dell, EMC, HP, IBM, Informix, Microsoft, NCR, Oracle, Sequent, SGI, Sun, Sybase, Unisys.

8.2.1 TPC-H

TPC-H je množina porovnávacích testov zameraných na podporu rozhodovania a výkonu dátového skladu. Obsahuje skupinu obchodne orientovaných ad-hoc dotazov a zhodných modifikácii dát. Dotazy a dáta v databáze sú vybrané zmysluplne. Je zameraný na systém na podporu rozhodovania, ktoré majú za účel analyzovať obrovské množstvo dát, vykonávať dotazy s veľkou zložitosťou a odpovedať rozhodujúce otázky ohľadne obchodu. Špecifikácia verzie TPC-H 2.6.0 je publikovaná a dostupná [30].

Metrika výkonnosti vychádzajúca z TPC-H sa nazýva Composite Query-per-Hour Performance Metric (QphH@Size), a odzrkadľuje niekoľko aspektov schopnosti systému vykonávať dotazy. Tieto aspekty zohľadňujú napríklad veľkosť databázy, silu vykonania dotazu, pričom dotaz je zadaný v súvislom toku, priechodnosť dotazov, keď sú zadané viacerými užívateľmi súčasne. TPC-H používa tiež metriku ceny vzhľadom na výkon (TPC-H Price/Performance metric), ktorú vyjadruje ako $/QphH@Size a je to pomer medzi celkovou cenou systému a jeho výkonu.

TPC-H umožňuje užívateľom porovnať možné riešenia vzhľadom na preddefinované prostredie skladajúce sa z: databázového systému, operačného systému, serverovej architektúry a úložných polí. Revidované výsledky testov umožňujú analyzovať silné a slabé stránky, možnosti špecifickej kombinácii komponentov.

Pri dosiahnutom výkone sa musí brať na zreteľ počet procesorov, uzlov, serverová architektúra. Dôležité je rýchle zdieľanie výsledkov a načítavanie naprieč CPU uzlov, vysoko rýchlostná úložná infraštruktúra prenosu dát medzi uzlami a poľami, fyzické vrstvovanie dát medzi viacerými diskami, médiami, segmentácia veľkých tabuliek na zníženie množstva dát požadovaného pri dotazovaní. Nasledujúce porovnania budú zoradené do kategórií na základe veľkosti databázy, budem brať do úvahy výsledky dosiahnuté len za posledné 2 roky.

8.2.2 100 GB

V Tejto kategórii sa nenachádzajú porovnania s databázou Oracle. Môže to byť aj kvôli orientácii Oracle skôr na veľké databázy, naopak Microsoft sa sústredí na využívanie v menších.

Najlepší výkon 19,323 QphH bol dosiahnutý na systéme HP Proliant DL585G2 4P, OS MS Windows Server 2003 EE x64 SP1, za cenu 10.67 USD/QphH pri databáze MS SQL Server 2005 x64 EE SP1, s využitím procesora AMD 8220SE 2.8 GHz. Na rovnakej zostave s použitím procesora Intel DC 7140 3.4GHz, bol dosiahnutý výkon 7,120 QphH za cenu 7.91 USD/QphH.

Celkovo, prvých 8 miest vzhľadom na výkon bolo dosiahnutých na databáze MS SQL Server 2005.

Identifikácia testu: 106092501; 25.9.2006; TPC-H v. 2.3		
Systém: HP Proliant DL585G2 4P ; CPU: AMD 8220SE 2.8 GHz; #Proc.: 4; #Cores: 8; #Threads:8; Cluster: Nie; Load Time: 1.11h; Database Size Ratio: 31.68		
Operačný systém: Microsoft Windows Server 2003 Enterprise x64 Edition SP1		
Databáza: Microsoft SQL Server 2005 x64 Enterprise Edt. SP1		
Celková cena systému: 205,988 USD	**Výkon:** 19,323 QphH	**Cena:** 10.67 USD/QphH
http://www.tpc.org/results/FDR/tpch/HP_DL585G2_100GB_4P_2.8GHzDC_FDR.pdf		

Najnižšia cena 4.61USD/QphH bola dosiahnutá na databáze Sun Sybase IQ 12.6 Single, systéme SunFire X4100, OS Sun Solaris 10, avšak bol dosiahnutý výkon 4,132 QphH. Druhé miesto s výkonom 14,923 QphH za cenu 6,04USD bolo na databáze MS SQL Server 2005 x64 EE na systéme PowerEdge 6950/2.8GHz/2MB.

Identifikácia testu: 106062602; 23.6.2006; TPC-H v. 2.3		
Systém: SunFire X4100; CPU: AMD Opteron - 3.0 GHz; #Proc.: 2; #Cores: 2; #Threads: 2; Cluster: Nie; Load Time: 1.65h; Database Size Ratio: 2.72		
Operačný systém: Sun Solaris 10		
Databáza: Sun Sybase IQ 12.6 Single		
Celková cena systému: 19,057 USD	**Výkon:** 4,132 QphH	**Cena:** 4.61 USD/QphH
http://www.tpc.org/results/FDR/tpch/sunfire.x4100.3.0_tpch.sybaseIQ.100gb.fdr.pdf		

8.2.3 300 GB

Najlepšie výsledky výkonu dosahuje databáza Oracle 10g. Na systéme HP BladeSystem ProLiant BL480c Cluster 16P DC, OS Red Hat Enterprise Linux 4 dosahuje Oracle 10g R2 EE pri použití clustrovania výkon až 40,411 QphH za cenu 18.67 USD/QphH. Pre porovnanie rovnaká zostava ale s 8 procesormi Intel X5355 2.66Ghz dosahuje výkon 30,765 QphH za cenu 22.90 USD/ QphH.

Databáza MS SQL Server 2005 EE x64 dosahuje na systéme HP ProLiant DL585 G2, OS MS Windows Server 2003 EE x64 výkon 18,298 QphH za ceny 13.67 USD/QphH. Najnovší test na databáze IBM DB2 UDB 8,2 a systéme IBM System x3650, ukazuje výkon 10,165 QphH za ceny 15.40 USD/ QphH.

Identifikácia testu: 106121901; 18.12.2006; TPC-H v. 2.5		
Systém: HP BladeSystem ProLiant BL480c Cluster 16P DC ; CPU: Intel 5160 Xeon 3.0GHz; #Proc.: 16; #Cores: 32; #Threads: 32; Cluster: Ano; Load Time: 1.22h; Database Size Ratio: 55.68		
Operačný systém: Red Hat Enterprise Linux 4		
Databáza: Oracle Database 10g release2 Enterprise Edt		
Celková cena systému: 754,232 USD	**Výkon:** 40,411 QphH	**Cena:** 18.67 USD/QphH
http://www.tpc.org/results/FDR/tpch/HP%20ProLiant_BL480c_16PDC_061218_fdr.pdf		

Identifikácia testu: 106103101; 31.10.2006; TPC-H v. 2.3		
Systém: HP ProLiant DL585 G2; CPU: AMD 8220SE 2.8GHz; #Proc.: 4; #Cores: 8; #Threads: 8; Cluster: Nie; Load Time: 3.28h; Database Size Ratio: 24.96		
Operačný systém: Microsoft Windows Server 2003 Enterprise x64 Edition		
Databáza: Microsoft SQL Server 2005 Enterprise Edt (x64)		
Celková cena systému: 250,057 USD	**Výkon:** 18,299 QphH	**Cena:** 13.67 USD/QphH
http://www.tpc.org/results/FDR/tpch/HP_DL585G2_300GB_4P_2.8GHzDC_FDR.pdf		

Identifikácia testu: 106100602 ; 6.10.2006; TPC-H v. 2.3		
Systém: IBM System x3650; CPU: Intel Xeon Processor 5160 3.0 GHz; #Proc.: 2; #Cores: 4; #Threads: 4; Cluster: Nie; Load Time: 3.14h; Database Size Ratio: 12.80		
Operačný systém: SUSE LINUX Enterprise Server 9		
Databáza: IBM DB2 UDB 8.2		
Celková cena systému: 156,535 USD	**Výkon:** 10,165 QphH	**Cena:** 15.40 USD/QphH
http://www.tpc.org/results/FDR/tpch/ibm.x3650-Linux.h.300GB.fdr.100606.pdf		

Najnižšia cena 6.29 USD/QphH bola dosiahnutá na databáze Sybase IQ 12.6 Single, systéme SunFire X4200, OS Sun Solaris 10, avšak za výkonu 4,936 QphH.

Identifikácia testu: 106062601; 23.6.2006; TPC-H v. 2.3		
Systém: SunFire X4200; CPU: AMD Opteron - 3.0 GHz; #Proc.: 2; #Cores: 2; #Threads: 2; Cluster: Nie; Load Time: 5.04h; Database Size Ratio: 2.95		
Operačný systém: Sun Solaris 10		
Databáza: Sybase IQ 12.6 Single		
Celková cena systému: 31,033 USD	**Výkon:** 4,936 QphH	**Cena:** 6.29 USD/QphH
http://www.tpc.org/results/FDR/tpch/sunfire.x4200.3.0_tpch.sybaseIQ.300gb.fdr.pdf		

8.2.4 1000 GB

V tejto kategórii veľkosti databázy dominuje s dosiahnutým výkonom 68,100 QphH za ceny 59.00 USD/QphH databáza Oracle 10g R2 EE na systéme HP Integrity Superdome – Itanium2/1.6 GHz-64p/64c a OS HP UX 11.i V2 64 bit. Najlepší výkon s databázy IBM DB2 UDB 8.2 má hodnotu 53,451 QphH a cenu 32.80 USD/QphH bol dosiahnutý na systéme IBM eServer xSeries 346 a OS SUSE LINUX Enterprise Server 9. Šiesty najlepší výkon v poradí bol dosiahnutý s databázou MS SQL Server 2005 E. Itanium E. a má hodnotu 33,488 QphH za cenu 27.00 USD/QphH na systéme HP Integrity rx8640 – Itanium2/1.6 GHz-16p/32c a OS MS Windows Server 2003 Datacenter Itanium Ed SP1.

Identifikácia testu: 105080801; 8.8.2005; TPC-H v. 2.1		
Systém: HP Integrity Superdome – Itanium2/1.6 GHz-64p/64c ; CPU: Intel Itanium2 - 1.6 GHz; #Proc.: 64; #Cores: 64; #Threads: 64; Cluster: Nie; Load Time: 1.10h; Database Size Ratio: 41.62		
Operačný systém: HP UX 11.i V2 64 bit		
Databáza: Oracle Database 10g R2 Enterprise Edt w/Partitioning		
Celková cena systému: 4,008,065 USD	**Výkon:** 68,101 QphH	**Cena:** 59.00 USD/QphH
http://www.tpc.org/results/FDR/tpch/hp_tpch_sd_1TB_050808_fdr.pdf		

Identifikácia testu: 105021401; 14.2.2005; TPC-H v. 2.1		
Systém: IBM eServer xSeries 346; CPU: Intel Xeon - 3.6 GHz; #Proc.: 64; #Cores: 64; #Threads: 128; Cluster: Ano; Load Time: 0.82h; Database Size Ratio: 26.25		
Operačný systém: SUSE LINUX Enterprise Server 9		
Databáza: IBM DB2 UDB 8.2		
Celková cena systému: 1,753,144 USD	**Výkon:** 53,451 QphH	**Cena:** 32.80 USD/QphH
http://www.tpc.org/results/FDR/tpch/x346.linux.1TB.051219.fdr.pdf		

Identifikácia testu: 106071801; 18.7.2006; TPC-H v. 2.3		
Systém: HP Integrity rx8640 – Itanium2/1.6 GHz-16p/32c ; CPU: Intel DC Itanium2 Processor 9050 - 1.6 GHz; #Proc.: 16; #Cores: 32; #Threads: 32; Cluster: Nie; Load Time: 11.68h; Database Size Ratio: 11.16		
Operačný systém: Microsoft Windows Server 2003 Datacenter Itanium Ed SP1		
Databáza: Microsoft SQL Server 2005 Enterprise Itanium Ed.		
Celková cena systému: 903,908 USD	**Výkon:** 33,488 QphH	**Cena:** 27.00 USD/QphH
http://www.tpc.org/results/FDR/tpch%5CHP-rx8640-1000G-SQL-07-18-2006-FDR.pdf		

Najnižšia cena 12.56 USD/QphH pri výkone 12,087 QphH bola dosiahnutá na databáze MS SQL Server 2005 E. IA64 SP1 a systéme NovaScale 3045, OS MS Windows Server 2003 E IA64 SP1.

Identifikácia testu: 107031202; 6.3.2007; TPC-H v. 2.5		
Systém: NovaScale 3045; CPU: Intel Dual-Core Itanium2 1.6 GHz; #Proc.: 4; #Cores: 8; #Threads: 16; Cluster: No; Load Time: 27.97h; Database Size Ratio: 5.71		
Operačný systém: Microsoft Windows Server 2003 Enterprise IA64 Edt. SP1		
Databáza: Microsoft SQL Server 2005 Enterprise IA64 Edt SP1		
Celková cena systému: 151,870 USD	**Výkon:** 12,087 QphH	**Cena:** 12.56 USD/QphH
http://www.tpc.org/results/FDR/tpch/Bull_NS3045_03062007_TPCH_MSSQL2005_64-bit_FDR.pdf		

8.2.5 3000 GB

Najlepší výkon v kategórii je 114,713 QphH za cenu 36.68 USD/QphH bola dosiahnutá na databáze Oracle 10g R2 EE, na systéme Sun Fire[TM] E25K server a OS Sun Solaris 10. Súčasne je to aj 2. najnižšia dosiahnutá cena.

Identifikácia testu: 107040901; 9.4.2007; TPC-H v. 2.6		
Systém: Sun Fire[TM] E25K server; CPU: Sun UltraSPARC[TM] IIIi Cu 1800 MHz; #Proc.: 72; #Cores: 144; #Threads: 144; Cluster: Nie; Load Time: 4.87h; Database Size Ratio: 21.60		
Operačný systém: Sun Solaris 10		
Databáza: Oracle Database 10g R2 Enterprise Edt w/Partitioning		
Celková cena systému: 4,207,126 USD	**Výkon:** 114,714 QphH	**Cena:** 36.68 USD/QphH
http://www.tpc.org/results/FDR/tpch/sun_fire_e25k_3tb_040907_fdr.pdf		

Najnižšia dosiahnutá cena je 32.34 USD/QphH, ale iba pri výkone 54,465 QphH. Bola dosiahnutá na databáze IBM DB2 UDB 8.2 a systéme IBM eServer xSeries 346, OS Suse Linux Enterprise Server 9.

Identifikácia testu: 105051901; 18.5.2005; TPC-H v. 2.1		
Systém: IBM eServer xSeries 346; CPU: Intel Xeon - 3.6 GHz; #Proc.: 64; #Cores: 64; #Threads: 128; Cluster: Ano; Load Time: 2.39h; Database Size Ratio: 8.75		
Operačný systém: Suse Linux Enterprise Server 9		
Databáza: IBM DB2 UDB 8.2		
Celková cena systému: 1,761,686 USD	**Výkon:** 54,466 QphH	**Cena:** 32.34 USD/QphH

http://www.tpc.org/results/FDR/tpch/x346.linux.3TB.051219.fdr.pdf

8.2.6 10000 GB

V tejto kategórii sú zaujímavé 2 výkony. Najlepší je 180,108 QphH dosiahnutý na databáze IBM DB2 UDB 8.2 s clustrovaním za cenu 47.00 USD/QphH s použitím systému IBM System p5 575 with DB2 UDB 8.2 a OS IBM AIX 5L V5.3.

Najnižšia dosiahnutá cena 32.91 USD/QphH s výkonom 171,380 QphH má databáza Oracle 10g R2 EE pri použití systému HP Integrity Superdome-DC Itanium2/1.6GHz/64p/128c a OS HP-UX 11i v3 64 bit. Ukazuje sa orientácia a využiteľnosť databázy IBM DB2 vzhľadom na 10000 GB veľkosť databázy. Microsoft zastúpenie nemá.

Identifikácia testu: 106071701; 14.7.2006; TPC-H v. 2.3		
Systém: IBM System p5 575 with DB2 UDB 8.2; CPU: IBM POWER5+ - 2.2 GHz; #Proc.: 128; #Cores: 128; #Threads: 256; Cluster: Ano; Load Time: 2.74h; Database Size Ratio: 11.78		
Operačný systém: IBM AIX 5L V5.3		
Databáza: IBM DB2 UDB 8.2		
Celková cena systému: 8,467,124 USD	**Výkon:** 180,108 QphH	**Cena:** 47.00 USD/QphH
http://www.tpc.org/results/FDR/tpch/IBM_575_10TB_20060714_FDR.pdf		

Identifikácia testu: 106120401; 30.11.2006; TPC-H v. 2.3		
Systém: CPU: Intel Dual Core Itanium 2 9040 1.6 GHz; #Proc.: 64; #Cores: 128; #Threads: 128; Cluster: Nie; Load Time: 5.86h; Database Size Ratio: 11.07		
Operačný systém: HP-UX 11i v3 64 bit		
Databáza: Oracle Database 10g R2 Enterprise Edt w/Partitioning		
Celková cena systému: 5,640,390 USD	**Výkon:** 171,380 QphH	**Cena:** 32.91 USD/QphH
http://www.tpc.org/results/FDR/tpch/hp_tpch_sd_10TB_fdr_022707.pdf		

9 Záver

Oblasť OLAP analýzy sa za posledných niekoľko rokov dostala do popredia záujmu, vďaka požiadavkám na podporu analytického spracovania údajov. S ňou úzko súvisí aj multidimenzionálny databázový model, z ktorého vychádza. U najväčších poskytovateľov OLAP nástrojov ako Microsoft, Oracle nie je možné analyzovať a porovnať OLAP nástroje samostatne, pretože ich riešenia sú úzko prepojené a poskytované ako súčasť väčšieho produktu.

Cieľom práce bola formálna definícia logického modelu, ilustrácia oblasti využitia a analýza možností a využitia modelu. OLAP možnosti ponúkajú veľa implementačných variantov a vo svojej práci sa snažím poskytnúť analýzu ich využitia vzhľadom na existujúce prostredie.

Prínosom práce je analýza problematiky multidimenzionálnych databáz, porovnanie analytického prostredia, nástrojov, možností a techník, architektúry a vylepšení oproti predchádzajúcim verziám, ktoré ponúkajú databázové systémy MS SQL Server 2005, Oracle 10*g* a taktiež samostatné produktové riešenie Cognos.

Prípadné rozšírenia tejto práce môžu byť nasledujúce:

- Implementácia fyzickej úrovne multidimenzionálneho dátového modelu
- Porovnanie ďalších poskytovateľov OLAP nástrojov
- Porovnanie dotazovacieho jazyka pre OLAP databázy (napríklad MDX)
- Indexovanie v multidimenzionálnych štruktúrach
- Multidimenzionálne clusterovanie

10 Zoznam použitej literatúry

[1] Pavel Louda, Business intelligence v novém hávu, 23.11.2006, http://www.computerworld.cz

[2] Ľuboslav Lacko, 2003. Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL Serveru a Oracle. Computer Press, 2003. 488 s. ISBN 80-7226-969-0.

[3] James O'Brien and George Marakas, 2005. Management Information Systems, 7th ed. McGraw-Hill, 2005, 592 s. ISBN-10**:** 007293588X.

[4] Berka, Petr, 2003. Dobývání znalostí z databází. Praha: Academia, 2003. 366 s. ISBN 80-200-1062-9.

[5] Wade, David and Ronald Recardo, 2001. *Corporate Performance Management*. Butterworth-Heinemann, 2001, ISBN 0-87719-386-X.

[6] Zborník prednášok konferencie DATASEM ’96, Podľa: [9].

[7] Data Warehousing Knowledge Center www.datawarehousing.org, DATASEM ’96, Podľa: [9].

[8] Data Mart vs Data Warehouse - The Great Debate, 14.3.2006, http://opensourceanalytics.com

[9] Karol Matiaško, Luboš Vnuk, Katarína Ševčíková, 2001. Dátové sklady ako informačný zdroj pre podporu rozhodovania. Fakulta riadenia a informatiky, Žilinská univerzita, Bulletin SIS. - č. 2, 2001.

[10] Ardent Software, Warehouse Watch – http://ardentsoftware.com, Podľa: [9].

[11] Sham Navathe,1999. Decision Support, Data Warehousing, and OLAP. Georgia Institute of Technology, 1999.

[12] W.H. Inmon, C. Imhoffa, G. Battase, 1995. Building the Operational Data Store. John Wiley & Sons, Inc. 1995. 276 s. ISBN:0-471-12822-8.

[13] Mark Humphries, M. W. Hawkins, M. C. Dy, 2001. Data warehousing - návrh a implementace, Computer Press, 2001. 256 s. ISBN 80-7226-560-1.

[14] Codd E.F. Codd S.B. & Salley C.T. 1998, Providing OLAP to User-Analysts: An IT Mandate, E.F.Codd & Associates 1998. 24 s. 131 9811.

[15] Nigel Pendse, The OLAP Survey 6, 21.2.2007, http://www.olapreport.com/survey.htm

[16] Ľuboslav Lacko, 2006. Business Intelligence v SQL Serveru 2005, Computer Press a. s. 2006. 391 s. ISBN 80-251-1110-5.

[17] Ľuboslav Lacko, Datový sklad v Yukonu podruhé, 6.7.2005, www.dbsvet.cz

[18] Data Mining Extensions (DMX) Reference, http://msdn2.microsoft.com/en-us/library/ms132058.aspx

[19] Multidimensional Expressions (MDX) Reference, http://msdn2.microsoft.com/en-us/library/ms145506.aspx

[20] Oracle OLAP DML Reference , http://www.oracle.com/technology/products/bi/olap/OLAP_DML_10.2.zip

[21] Hyperion Solutions, http://www.hyperion.com

[22] Business Objects, http://www.businessobjects.com/

[23] MicroStrategy, http://www.microstrategy.com/

[24] SAP, http://www.sap.com/

[25] Cartesis, http://www.cartesis.com

[26] Applix, http://www.applix.com/

[27] Infor, http://www.infor.com/

[28] OLAP Report, http://www.olapreport.com/market.htm

[29] Transaction Processing Performance Council, http://www.tpc.org/

[30] Špecifikácia TPC-H, Verzia 2.6.0, http://www.tpc.org/tpch/spec/tpch2.6.0.pdf

[31] Oracle, http://www.oracle.com

[32] Microsoft, http://www.microsoft.com

[33] Cogent, http://www.cogent.sk

[34] Cognos, http://www.cognos.com