ORGANIZACE DAT PRO JEJICH ANALÝZU[1]

Dušan HÚSEK a Hana ŘEZANKOVÁ
ÚIVT AV ČR, VŠE KSTP

Abstract: In this paper problems of data storage and exploration in DataWarehouse are discussed. Basic variations of relational data model are described and comparison with multidimensional model are shown. Following schemes based on relational models are mentioned: Star Schema and Snowflakes Schema.

Резюме: В этой работе дискутированы проблемы хранения данных и их использование в складе данных. Описаны основные варианты релационной модели данных и показано сравнение с мультидимензиональной моделей. Упомянуты следующие схемы основаные на релационной модели: Star scheme и Snowflake scheme. Собременно превосходит подход комбинирующий релационную модель данных вместе с мультидимензиональной моделей.

Klíčová slova: Sběr dat, těžba dat, organizace dat, datový sklad, podnikový informační sklad, statistické databáze, OLAP, ROLAP

1 Úvod

Nové technologie v oblasti počítačových věd (Computer Science) rozšiřují pojem „statistické výpočty“ (Statistical Computing) i na oblasti

- sběru dat a jejich šíření (elektronická výměna dat, sítě),
- organizace dat a jejich správy (statistické systémy řízení báze dat),
- interpretace výsledků a vytváření výstupních sestav (metody umělé inteligence).

Vynořují se též nové perspektivy pro samotnou analýzu dat, založené na využití počítačové grafiky a neuronových sítí (viz [8]).

Všechny tyto oblasti lze zahrnout do statistického výpočetního prostředí. Na jedné straně tedy organizace dat a jejich správa je součástí statistické analýzy, na straně druhé bývá statistická analýza často zařazována jako součást aplikací vytvářených v rámci systémů řízení báze dat (SŘBD). Jedním typem takovýchto aplikací jsou podnikové informační systémy (Enterprise Information Systems - EIS).

EIS mohou být rozčleněny do tří kategorií.

[1] Tato práce vychází z výsledků grantu č.102/94/0728 GA ČR

I. EIS založené na relačním SŘBD

Univerzálním rozhraním k tomuto SŘBD je jazyk SQL (Structured Query Language), který dominoval databázové technologii téměř 15 let. Konkrétním typem aplikací v tomto prostředí jsou administrativní a provozní informační systémy OLTP (On-line Transaction Processing), jež jsou určeny pro správu obchodních dat (účtování), lze sem zařadit i systémy pro rezervaci jízdenek a letenek apod. Součástí těchto systémů tedy ještě nemusí být statistické zpracování, i když mohou být počítány souhrnné charakteristiky.

V tomto případě je cílem výpočtů získat odpovědi na otázky typu: „Kolik jsme dnes prodali výrobků?“

II. EIS založené na multidimenzionálním modelu

Cílem aplikací je vytvářet multidimenzionální pohled na data. Tento typ aplikací je označován jako OLAP (On-line Analytical Processing). Z hlediska organizace dat rozlišujeme

- ROLAP (Relational OLAP), kdy multidimenzionalita nespočívá v uložení dat, ale až v jejich prezentaci (pro uložení je používána relační databáze) a
- MOLAP (Multidimensional OLAP), kdy je princip multidimenzionality uplatňován i při organizaci dat.

V obou případech jsou hlavním výstupem různé typy tabulek a grafů z oblasti popisné statistiky.

Systémy, které jsou založeny na principu OLAP, jsou označovány jako EIS, přičemž tato zkratka v daném případě znamená Executive (manažerský) IS.

Obecně mohou tyto systémy zahrnovat i prostředky pro rozsáhlejší statistickou analýzu dat, proto je lze označit jako systémy pro podporu rozhodování (DSS - Decision Suport Systems).

Typickým dotazem v takovémto systému je otázka typu: „Kteří jsou nejlepší zákazníci pro produkt X v regionu Y?“

III. EIS založené na „těžbě dat“ či „dolování dat“ (Data Mining)

Lépe by bylo charakterizovat činnost těchto systémů jako „těžba informací z dat“. Jde v podstatě o důkladnou statistickou analýzu zjišťující závislosti v datech, trendy v časových řadách apod. Dalším charakteristickým rysem v procesu „Data Mining“ je využívání neuronových sítí.

Typickými dotazy v takovémto systému jsou takové, jejichž cílem je získat odpovědi na otázky typu: „Na jaký druh zákazníků bychom se měli zaměřit?“ „Jaký druh událostí způsobil koupi produktu s těmito zákazníky?“

Vstupem pro DSS jsou data získávaná často z různých zdrojů (z různých operačních systémů a z různých formátů dat). Obvykle jde o rozsáhlé datové

soubory.

Manipulace s daty, která zahrnuje jejich získávání, modelování a zpracování do základních informací důležitých pro řízení, je označována jako „Data Warehousing“. Prostředí, které tyto činnosti umožňuje, je nazýváno datovým skladem (Data Warehouse). Jako prvky datového skladu jsou uváděny i OLAP a EIS.

V literatuře se ve výše uvedených pojmech vyskytuje značná nejednotnost, a to i v rámci jedné firmy. Například firma SAS dříve zahrnovala veškerou možnou statistickou analýzu jako součást datového skladu. Také „Data Mining“ bylo uváděno jako prvek datového skladu, a to jednak při transformaci dat, jednak při jejich využití. Tento pojem byl spojen především s využitím neuronových sítí. Nyní je „Data Mining“ chápán jako komplexní proces, který využívá jak neuronové sítě, tak klasické statistické metody. Datový sklad tvoří vstup do tohoto procesu. Je tedy určen pouze ke získávání dat z různých zdrojů a k jejich převedení do formy využitelné v procesu „těžby dat“.

2 Data Warehouse (DW)

Datový sklad je tedy určen především k ukládání a integraci dat mimo provozní databázi a zajištění manipulace s těmito daty. DW může využívat i několik modelů pro ukládání dat současně (fyzických modelů). Mimo fyzického modelu dat v DW se používá ještě logický model dat na úrovni konceptuální a model transformací.

Rozlišujeme tedy

- logické modelování dat,
- fyzické modelování dat a
- modelování transformací.

Cílem modelování dat v DW je určit strukturu a obsah DW a definovat, jak převést data do DW. Definice modelu je ukládána jako metadata v metabázi DW. Logický model definuje entity, které budou vyžadovány v DW.

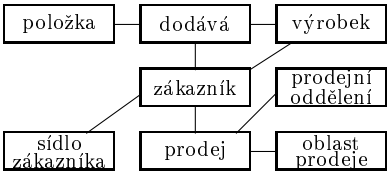
Tento logický model má být převeden do fyzického modelu dat, který definuje architekturu DW. Na úrovni fyzického modelu by měly být zohledněny i požadavky vyplývající z funkční analýzy na výběry či sumarizace dat a podobně.

2.1 Fyzický model dat

V předešlém bylo řečeno, že na fyzické úrovni se může vyskytovat několik modelů dat. Příkladem DW, který podporuje takovou strategii, je DW od společnosti SAS.

Relační model dat

Základním elementem v tomto modelu je relace. Na úrovni uživatele pak je relace repezentována tabulkou. Pro manipulaci s daty existují neprocedurální jazyky, které však vyžadují, k zajištění integrity dat a jejich neredundance, aby data byla ukládána v tzv. normalizovaném stavu. Standardem pro manipulaci s daty v tomto modelu se stává jazyk SQL (Structured Query Language). SŘBD pracující s tímto modelem jsou optimalizovány pro on-line transakční zpracování, na poměrně jednoduché dotazy a aktualizační příkazy. Na konceptuální úrovni lze k modelování dat použít Entity-Relationship (E-R) model.



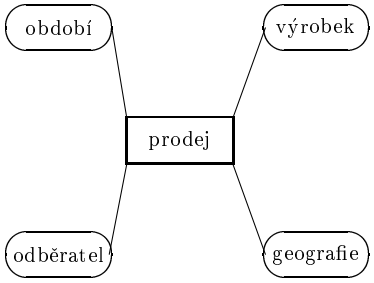
Obr.: E-R diagram

Protože v rámci DW je databáze využívána uživateli pouze pro dotazování, je možno uložit data do databáze v nenormalizované podobě. Důsledkem je však enormní nárůst diskového prostoru. Pomoci má následující řešení.

Model hvězda (Star schema)

Model hvězda má být odpovědí na problémy použití relačního modelu v oblasti řízení, a tedy i na úrovni agregovaných dat. Je založen na fyzickém aparátu relačních SŘBD, ale data nejsou ukládána v normalizované podobě. V rámci tohoto modelu jsou oddělena fakta - většinou kvantitativní údaje - od kategoriálních dat neboli dimenzí. Fakta jsou uložena v centrální nenormalizovaných tabulkách. Takto je redukován počet tabulek a přitom je omezena potřeba jejich spojování (Join) v průběhu práce s databází ve DW. Tímto způsobem je zlepšena odezva systému. Denormalizace má za následek redundanci dat a tudíž i větší paměťové nároky, integritu dat je nutné zajišťovat procedurálními prostředky.

Časová dimenze		*údaje o prodeji*		*dimenze produktu*
čas (klíč)	$\longrightarrow$	čas (klíč)		
den		produkt (klíč)	$\longrightarrow$	produkt (klíč)
měsíc		prodané množství		popis
čtvrtletí		cena v korunách		značka
rok		ostatní fakta		kategorie
				oddělení
				dodavatel (údaje)
				atd.

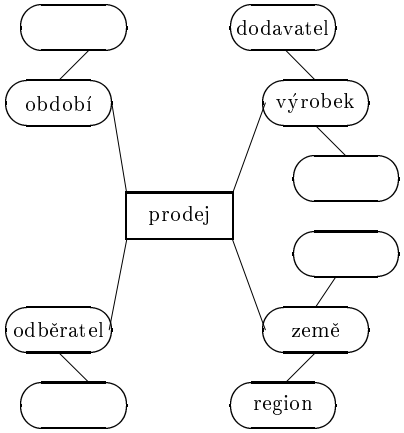


Obr.: Model hvězda

Model sněhová vločka (Snowflake schema)

Sleduje stejné cíle jako model hvězda, ale vychází se zde z předpokladu, že tabulky dimenzí potřebují alespoň částečnou normalizaci, protože denormalizace dimenzí může mít za následek stále vysokou redundanci údajů v databázi (a tudíž i problémy s ukládáním dat a údržbou databáze).

Časová dimenze		*údaje o prodeji*		*dimenze produktu*
čas (klíč)	$\longrightarrow$	čas (klíč)		
den		produkt (klíč)	$\longrightarrow$	produkt (klíč)
měsíc		prodané množství		popis
čtvrtletí		cena v dolarech		značka
rok		ostatní fakta		kategorie
				oddělení
				dodavatel (klíč) $\longrightarrow$
				atd.



Obr.: Model sněhová vločka

Multidimenzionální model dat (MDMD)

Protože stále roste používání nástrojů pro zpracování dat technologií OLAP, kde základním modelem pro pohled na data je multidimenzionální kostka, je pravděpodobné, že tento datový model se prosadí i na nejnižší fyzické úrovni (teď se na situaci díváme z pohledu dodavatelů relačních SŘBD).

Co stojí ukládání dat v multidimenzionálním modelu dat

Celkovou velikost multidimenzionální databáze je dána vzorcem

$$8 * \prod_{i=1}^{n} D_i,$$

kde n je počet základních dimenzí a D_i je kardinalita i-té základní dimenze. V reálných aplikacích tak můžeme dospět k neuvěřitelně velkým hodnotám potřebného diskového prostoru. Tak například jsou-li u společnosti sbírány po dobu pěti let měsíčně údaje o 5000 zákaznících, 2000 výrobcích v pěti továrnách, distribuovaných přes tři distribuční místa, přičemž chceme sledovat (a tudíž i ukládat do Data Warehouse) 10 ukazatelů ve třech různých formách - plánované, aktuální a predikované, pak velikost potřebného diskového prostoru bude

$$8 \times (5 \times 12) \times 5000 \times 2000 \times 5 \times 3 \times 10 \times 3 = 2,16\,\mathrm{PB} \ \text{(petabajtu)}$$

V realitě pak mnoho z těchto kombinací nebude existovat. Většina zákazníků bude nakupovat jenom některé výrobky a jen občas, takže se bude jednat

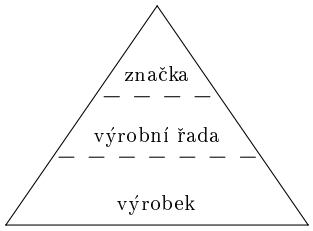
o řídkou matici. V jednom konkrétním případě z praxe velmi blízkému výše uvedenému příkladu bylo pro uložení dat potřeba pouhých 800 MB.

Multidimenzionální model dat se vyznačuje flexibilitou, která umožňuje uchovávat velké objemy „řídkých“ dat, tzn. že prázdné položky se neukládají.

V současné době se vedou spory o to, zda je vůbec nezbytné, při výkonnosti současné výpočetní techniky využívat specializované modely dat pro jejich ukládání v DW a následné zpracování v rámci OLAP.

Kombinovaný model dat

Nicméně v přechodné době se zdá, že bude nejvhodnější použít kombinaci těchto dvou technologií (Oracle), tzn. podrobná data mít uložena v relační databázi a agregovaná neboli souhrnná data na multidimenzionálním neboli OLAP serveru. Tento model vyniká flexibilitou, která umožňuje uchovávat velké objemy „řídkých“ dat a na druhé straně zase dovoluje, aby manipulace s údaji v rámci složitých ekonomických modelů, jako je hierarchie výrobků, nebo organizací probíhala v intuitivnějším multidimenzionálním prostředí.



Obr.: Kombinace relačního a multidimenzionálního modelu

Aby byla možná koexistence těchto dvou modelů, v DW musí existovat možnost těsné vazby jazyka SQL na MDMD. Zjednodušeně řečeno je nezbytné, aby jazyk pro manipulaci s daty na úrovni RMD měl těsnou vazbu na MDMD. V současnosti se u některých firemních výrobků objevují rozšíření jazyka SQL, která si viditelně kladou za cíl umožnit získávat-předávat agregované údaje vyšší- vrstvě pracující s multidimenzionálním modelem dat (viz rozšíření jazyka SQL od společnosti Microsoft o příkazy CUBE a ROLLUP).

Podle posledních výzkumů organizace pro průzkum trhu Gartner Group bude klíčovým trendem na trhu OLAP snaha o těsné připojení multidimenzionálních serverů relačním DW.

2.2 Požadavky na manipulace s daty

Hlavní požadavkem na manipulace s daty v DW je jejich intuitivnost v tom smyslu, že se používá pojmový aparát známý spíše z oblasti vyhodnocování dat (než z oblasti jejich ukládání):

- drill-down - zobrazení detailních dat, která byla použita pro agregace,
- roll-up - zobrazení agregovaných hodnot na nejbližší vyšší hierarchické úrovni,
- drill-everywhere - rozšíření aktuální analýzy v libovolném směru,
- dicing - rotace nebo převrácení multidimenzionální tabulky (hyperkostky),
- slicing - vyčlenění jedné vrstvy z multidimenzionální tabulky,
- pivoting - transpozice dvourozměrné tabulky (záměna sloupců za řádky a naopak),
- surfing - navigování v datech pomocí „drag and drop“ operací,
- data mining - prosévání velkých objemů dat s cílem nalezení (skrytých) vztahů mezi nimi,
- join - spojování podle odpovídajících si dimenzí.

3 ROLAP, OLAP a EIS

Jako příklad multidimenzionálního uložení dat si uveďme převod dat z relačního modelu v systému Media EIS pomocí speciálního programovacího jazyka.

Mějme číselné proměnné Náklady a Tržby a kategoriální proměnné Typ výrobku, Země, Město, Rok. V systémech EIS jsou používány pojmy

- ukazatelé, resp. proměnné, pro sledované ukazatele (Náklady, Tržby),
- kategorie, resp. dimenze, pro kategoriální proměnné (Typ výrobku, Oblast, Období) a

- prvky kategorií, resp. hodnoty dimenzí pro hodnoty kategoriálních proměnných, které mohou být hierarchicky uspořádány, např. pro kategorii Oblast můžeme uvažovat prvky
  Česká republika
    Praha
    Brno
    Ostrava
  Slovenská republika
    Bratislava
    Košice...

Pokud jsou v relačním modelu uloženy např. denní údaje, je možné v EIS ukládat pouze sumarizované údaje za čtvrtletí, příp. rok apod. Výhodou tohoto multidimenzionálního uložení je rychlejší přístup k datům, který se výrazně projeví právě v případě sumarizace údajů.

Příklad převodu:

```
Data:
93,10,1,1,Od,13.08163278,0.9863551116,0.3846000037,0.1896836753
93,10,1,1,Kl,428.6858174,42.91145032,11.31730558,14.7896607
93,10,1,1,St,81.62091786,0.45707714,3.65661712,2.938353043
93,10,1,1,Pr,1.559124015,7.4370215515e-002,2.2295473414e-002,
1.1225692908e-002
93,10,1,1,Br,47.44030662,5.949014451,0.5123553115,0.7969971513
93,10,1,1,Ko,2.139941585,4.7934691513e-002,1.5407579415e-002,
9.244547649e-002
93,10,1,1,Ru,11.89853022,0.5354338597,0.385512379,0.1070867719
93,10,1,1,Ja,55.07167615,1.872436989,0.4956450853,
7.7100346605e-002
93,10,1,2,Od,1.449892224,0.1035223048,3.0012769047e-002,
5.8720635091e-002
```

Popis převodu:

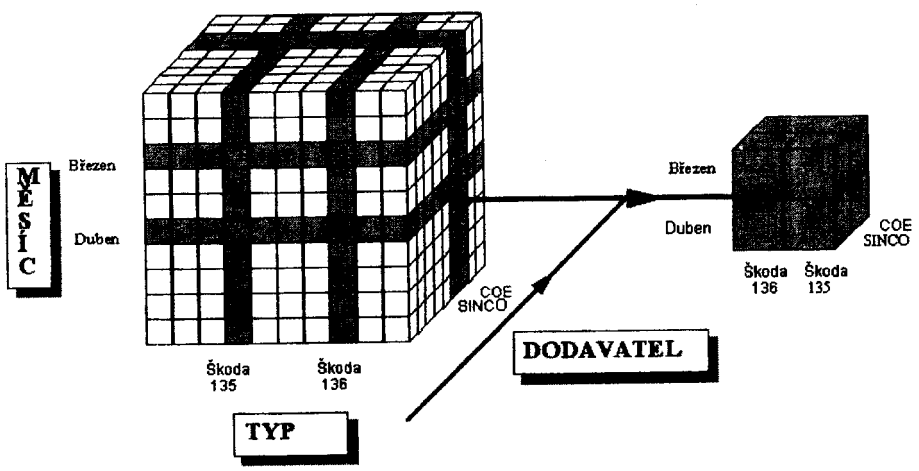
```
LOADER Pro_Doba
INPUT ASCII COMMA FILE = \uv{Pro_doba.txt}
 rok kvartal pracov sektory papirny prodCas Cas1 Cas2 Cas3
SELECT
 Rok = rok
 Ctvrtleti = kvartal
[
 _1Ct. = 01 - 03
 _2Ct. = 04 - 06
```

```
 _3Ct. = 07 - 09
 _4Ct. = 10 - 12
 ]
 Zdroje = pracovnici
[
 Delnici = 01
 Technici = 02
 Management = 03
 ]
 Sektory = sektory
[
 Prejimka = 01
 Vyroba = 02
 Obchodnici = 03
 Administrativa = 04
 ]
 Papirny = papirny
[
 Klatovy = "Kl"
 Strakonice = "St"
 Prachatice = "Pr"
 Rumburk = "Ru"
 Jablonec = "Ja"
 Bruntal = "Br"
 Odry = "Od"
 Kosice = "Ko"
 ]
OUTPUT
 ProduktivniCas = TOTAL  (prodCas) ERASE
 NeproduktivniCas OF _Skoleni = Cas1
 NeproduktivniCas OF Dovolena = Cas2
 NeproduktivniCas OF NahradniPrace = Cas3
```

4 Závěr

K ukládání dat v rámci technologie Data Warehouse jsou v součastnosti používány varianty relačního modelu dat a model multidimezionální. Posledně jmenovaný model je bližší koncovému uživateli, umožňuje efektivně ukládat a pracovat s agregovanými daty. Na druhé straně zdrojem dat pro data Warehouse jsou operativní data uložena především v relačních databázích. Podle posledních výzkumů organizace pro průzkum trhu Gartner Group bude klíčovým trendem na trhu OLAP snaha o těsné připojení multidimenzionálních serverů relačním DW.

Údaje o prodeji



References

[1] Anderson,S., Walker,E.: Building a SAS Data Warehouse. A SAS Institute White Paper, SAS Institute Inc., 1995.

[2] Bannister,F.E.: OLAP - A Question of Definition. DATASEM'96, Brno 20.-22.10.1996, CS-Compex, a.s., str. 211-222.

[3] Codd,E.F., Codd,S.B., Salley,C.T.: Providing OLAP (On-line Analytical Processing) to User-Analyst. E. F. Codd & Date, Inc., 1993.

[4] Emmerich,T., Walker,E.: SAS Institute's Rapid Warehousing Methodology. A SAS Institute White Paper, SAS Institute Inc., 1996.

[5] Greenberg,H.: OLAP, nebo ROLAP? ComputerWorld, VII(1996), č.30, str.16.

[6] Held,G., Neville,P.: Data Mining with the SAS System: From Data to Business Advantage. A SAS Institute White Paper, SAS Institute Inc., 1996.

[7] Kozák,J.: Použitá data nevyhazovat! ComputerWorld, VII(1996), č.30, str.16.

[8] Lauro,C.: Computational Statistics or Statistical Computing, is that the question? Daily Bulletin of COMPSTAT'96, 29th August 1996, Barcelona.

[9] Lhoták,M., Benešovský,M.: Data Warehousing aneb nic nového pod sluncem? DATASEM'96, Brno 20.-22.10.1996, CS-Compex, a.s., str.203-210.

[10] Linthicum,D.S.: Spolupráce clienta se serverem v praxi. PC Magazine Czech Edition, ročník IV (1996), číslo 5, str.61-75.

[11] Media EIS. Dokumentace k programovému systému, Speedware, 1995.

[12] Mena,J.: Vytěžte vlastní data. ComputerWorld, VII(1996), č.36, str.14.

[13] Morton,S., Anderson,A.: Requirements-driven Data Modelling for the SAS Data Warehouse. A SAS Institute White Paper, SAS Institute Inc., 1996.

[14] Oracle Warehouse. Firemní materiál, Oracle, 1995.

[15] SYBASE - Interactive Warehouse. PC Magazine Czech Edition, ročník IV (1996), číslo 9, str.71-74 (Speciální inzertní sekce).

[16] Štverka,I., Havlena,V.: Změňte vaše data v konkurenční výhodu (1.). ComputerWorld, VII (1996), č.38, str.15-16.

[17] Štverka,I., Havlena,V.: Změňte vaše data v konkurenční výhodu (2.). ComputerWorld, VII (1996), č.39, str.16-18.

[18] Weber,J.: Data Warehouse, ano? Proč? A jak? DATASEM'96, Brno 20.-22.10.1996, CS-Compex, a.s., str.195-202.