Datové sklady a využití datové struktury typu hvězda pro prostorová data

Jiří Horák[1] , Bronislava Horáková[2]

[1]Institut geoinformatiky, HGF, VŠB-TU Ostrava, 17.listopadu 15,
70833, Ostrava-Poruba, ČR
jiri.horak@vsb.cz

[2] Institut geoinformatiky, HGF, VŠB-TU Ostrava, 17.listopadu 15,
70833, Ostrava-Poruba, ČR
bronislava.horakova@vsb.cz

Abstrakt. V práci byly rozlišeny 3 typy datových skladů v pojetí jednotného a integrovaného úložiště dat, analytického datového skladu s multidimenzionální strukturou a datového skladu s transakčními daty v multidimenzionální struktuře. U datového skladu 2.typu je poskytnut úvod do teorie multidimenzionálních datových skladů. Použití datových struktur typu hvězda je demonstrováno v aplikační oblasti prostorových socioekonomických dat a hydrologických dat. Jsou dokumentovány možnosti, výhody a slabiny takových modelů.

Klíčová slova: datové sklady, metadata, prostorová data

1 Datový sklad

Pojem datový sklad je v poslední době značně frekventovaný a používá se v různých souvislostech, někdy pouze intuitivně pro označení rozsáhlého úložiště dat.

Pokusme se nejdříve vymezit varianty chápání datového skladu:

1. datový sklad 1.typu - datový sklad ve smyslu organizovaného, jednotného a integrovaného úložiště dat.
2. datový sklad 2 .typu – datový sklad typu data warehouse, který bychom mohli označit přívlastkem „analytický“ datový sklad, protože k jeho základním vlastnostem patří vedle integrace dat z transakčních databází, jejich agregace a uložení v multidimenzionálních strukturách, právě optimalizace z hlediska dotazování a analýzy dat. Předpokládá se následná aplikace systému OLAP.
3. datový sklad 3.typu – datový sklad používaný pro ukládání originálních dat v primární podobě ve formě multidimenzionální struktury především s cílem vytvoření centrálního úložiště s přesným popisem originálních dat. Jde tedy o transakční systém s multidimenzionální strukturou zpravidla typu hvězda.

Pojem dimenzionalita, resp. multidimenzionální struktury tu není chápán jako prostorový a jako dimenze se nepoužívají prostorové souřadnicové osy, ale popis faktorů, které data ovlivňují a které nás zajímají z hlediska analýzy.

Naším cílem není přitom jen vyjasnění terminologie a vztahů jednotlivých případů, ale především snaha ukázat jiné pojetí datových skladů, které může být pro některé aplikační oblasti a způsoby řešení inspirující a přínosné.

2 Datový sklad 1.typu

Cílem takového datového skladu je provádět transformaci originálních dat z původních míst uložení (ze samostatných informačních systémů), odstínit rozdíly v datech, provést jejich sjednocení, „standardizaci“, uložení a zpravidla i zpřístupnění v jednotném prostředí (resp. prostředí s jednotným, standardním přístupem). Zpravidla jde o centralizované řešení (může být ovšem i distribuované, avšak s jednotnou logickou a metodickou částí). Tento typ datového skladu neprovádí agregaci dat (ve smyslu ukládání pouze sumarizovaných či jinak agregovaných prostorových dat).

Vhodným příkladem jsou datové sklady GIS veřejné správy.

Příklady:

- datový sklad Vodstvo ČR (VUV T.G.Masaryka v Brně http://www2.vuv.cz/),
- datový sklad CENIA (Bukáček, 2006) (obr.1)

Obr.1. Datový sklad obsahující data integrovaná z různých zdrojů (Bukáček 2006)

- datový sklad Informačního systému o silniční a dálniční síti České republiky (ISSDS ČR), obsahující veškeré informace vztahující se k síti sledovaných pozemních komunikací, tj. k dálnicím, silnicím I., II. a III. třídy (Kružík, 2004),

- datový sklad IDC ÚHÚL Brandýs nad Labem (Mansfeld, 2003), obsahující vedle tématických dat pro lesní hospodářství (oblastní plány rozvoje lesů, lesní hospodářské plány a osnovy, informace o inventarizaci lesů) také různá podkladová data
- datový sklad GIS krajského úřadu dle návrhu základní architektury GIS (Maršík, 2004)
- datové sklady GIS na úrovni města (viz např. http://gis.plzen-city.cz/ogis/tech.htm)

Takto koncipované datové sklady jsou zaměřeny na centrální skladování a poskytování dat sjednocených z původních zdrojů.

Např. v případě datového skladu IDC ÚHÚL se používají importy dat ZABAGED pro konstrukci DMT (Mlčoušek 2004). Rovněž se uvádí použití tzv. informačního standardu lesního hospodářství jako metody, pomocí které se zajišťuje uložení dat do tvaru pokud možno nezávislého na technologii.

Podobně studie metainformačního systému MEDIS ŽP také hovoří o datovém skladu s konzolidovanými, agregovanými daty (Notes, 2003).

Z firemních projektů je možné uvést jako příklad datový sklad firmy GEODIS Brno s.r.o. Ten díky zpracování různých orginálních dat obsahuje ortofotomapy, přesný digitální model terénu a digitální model povrchu, digitální trojrozměrnou vektorovou bázi ekvivalentní mapám 1:5000 a trojrozměrné modely budov a jiných objektů (Plšek, 2004), přitom poskytuje několik úrovní informace (Plšek, 2003) v závislosti na měřítku pozorování či aplikace.

V některých případech se pojem datový sklad používá pro databázová řešení, kde jsou v relační či objektové databázi uložena i geometrická data jako součást popisu prostorových objektů (především jde o prostorová rozšíření relačních SŘBD). Z popisu však často není zřejmé, zda se ukládají originální nebo upravená, integrovaná data.

Příkladem může být popis architektury Geostore (http://www.geostore.cz/index.asp), kde se hovoří o datovém skladu geografických dat realizovaném v prostředí relační databáze.

Podobným způsobem se používá datový sklad i v projektu Jednotné technické mapy Zlínska. Podobně Orlík et al. (2005) specifikují uložená data v PostGIS jako datový sklad atd.

V některých případech je transformace chápána jako jednorázová a ne jako proces zachovávající originální data a systémy. Následně slouží vytvořený datový sklad přímo pro transakční operace.

Příkladem takového přístupu může být zpracování původních CAD souborů pro Deutsche Bahn AG a vytvoření datového skladu jako centralizovaného úložiště dat se sjednoceným způsobem uložení v prostředí Oracle, nad kterým se již vedou veškeré další operace (Corbley, 1999).

3 Datový sklad 2.typu

Zde se používá pojem datový sklad v klasickém pojetí známém z prostředí databázových a informačních systémů.

Datový sklad (data warehouse) představuje architekturu, obvykle založenou na relačním SŘBD, která se používá pro údržbu historických dat získaných z databází operativních dat, která byla transformována, sjednocena a zkontrolována před jejich použitím v databázi datového skladu (Strange in Pokorný 2004). Datový sklad tedy nepoužívá přímo data získaná při běžných transakcích (systémy typu OLTP on-line transaction processing), ale provádí jejich selekci, homogenizaci (např. odstranění sémantických rozdílů v jednotlivých zdrojových transakčních databázích) a především agregaci podle všech kritérií, která jsou považována za významná z hlediska možného dotazování.

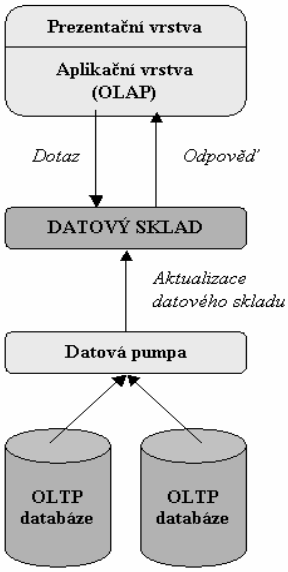
OLTP (On Line Transaction Processing) popisuje zpracování dat v operativní databázi. Zaměřuje se na ukládání a zpracování záznamů o jednotlivých uskutečněných transakcích a typicky zajišťují aktuální stav jisté evidence. Vedle běžné ekonomické agendy (vyřizování objednávek, sledování aktuálních kontaktů na zákazníky, dodavatele, nabízení aktuálních výrobků apod.) zde můžeme řadit i např. informační systém o území (aktualizace geometrie a vlastností silniční sítě, aktuální názvy ulic, poslední evidence obyvatel, aktuální užití a pokryv krajiny, zemědělská a lesnická produkce), dopravní systém (aktuální stav komunikační sítě, řídících prvků, aktivních agentů), vodohospodářský systém (množství a rozmístění aktuálních srážek, předpověď, aktuální průtok, stav regulačních prvků). Realizace se dnes zajišťuje zpravidla pomocí relační databázové technologie. Dominují aktualizační transakce, zaměřené na jeden či několik málo záznamů (tj. v dané chvíli upravujeme údaje pouze k jednomu geoprvku), používají se jednoduché dotazy (vyber objekty dané třídy, poskytni informace o daném místě apod.), je požadováno zajištění současného přístupu více uživatelů (s editačními právy), přitom systém musí zvládnout velké množství požadavků.

Z hlediska časového jsou v systému OLTP významná „aktuální“ data a data historická (tj. ta která pozbyla aktuálnosti) jsou archivována mimo OLTP systém, aby nedocházelo k jeho zbytečnému zatěžování.

Z hlediska datového skladu slouží OLTP systémy jako zdroje dat, která jsou pravidelně čerpána pomocí datové pumpy (obr.2)

Základním cílem takového datového skladu je tedy uložit sjednocená a integrovaná data pro zefektivnění následující analytické a statistické práce s daty. Využívají se přitom data historická, v některých případech i externí.

V některých případech se používá datový sklad i pro uložení primárních dat, což představuje přechod k 3.typu datového skladu. V takovém případě hovoříme o dvou vrstvách, kdy „nultá“ vrstva obsahuje neupravená originální data, která jsou v určitých intervalech importována z různých zdrojů. Další vrstvu již představují data konzolidovaná (sjednocená, integrovaná, verifikovaná apod.) uložená vhodným způsobem a určená pro provádění analýz a tvorbu výstupů.



Obr. 2. Vztah mezi OLTP, DW a OLAP (upraveno dle Vítek 2002)

Výsledkem tvorby datového skladu je poskytnutí sjednocených a agregovaných (odvozených) dat pro náročnější dotazování, potřebné pro získání koncentrované informace obsažené v původních datech. Tato informace může sloužit pro podporu rozhodování, ať již s využitím specifických technik dolování dat (typický nástroj datových skladů) nebo v navazujících systémech typu OLAP (či dříve EIS), které jsou vybaveny vhodným grafickým rozhraním pro koncové uživatele a dovolují využít mechanismů agregace dat, sledování trendů, porovnání vývoje apod.

Zpravidla se musí použít určitá vhodná rozšíření jazyka SQL nebo jiný způsob dotazování, protože jazyk SQL není vždy vhodný pro provádění požadovaných analytických operací (Groff, Weinberg 2005). Uveďme pro ilustraci pouze 1 situaci, kdy není využtí jazyka SQL otpimální. Pokud potřebujeme nalézt 1 nejlepší případ dle zadaných kritérií, provádí se nejdříve dle jazyka SQL množinové (resp. relační) operace a až na konci se provede setřídění a vybere se 1 záznam, což není příliš efektivní postup.

Navazující systém OLAP (On Line Analytical Processing) lze charakterizovat jako technologii pro zpracování dat z databáze datového skladu s využitím velkého množství kladených dotazů.

K typickým OLAP analytickým operacím patří:

- Drill-down (postup po hierarchii dolů, získávání většího detailu),
- Roll - Up (postup po hierarchii nahoru, získávání více agregovaných dat)
- Drill-Across (spojení několika faktorových tabulek se stejnou granularitou)
- Slice-and-Dice (dělení dat)
- Pivot (záměna dimenzí u vytvářeného pohledu)

Zaměřme se nyní blíže na koncepci datového skladu. Definice Billa Inmona (Humpries 2002) říká:

„Datový sklad je podnikově strukturovaný depozitář subjektově orientovaných, integrovaných, časově proměnných, historických dat použitých na získávání informací a podporu rozhodování. V datovém skladu jsou uložena atomická a sumární data."

Subjektovou orientací se rozumí orientace na takový subjekt, podle kterého jsou data v datovém skladu kategorizována. Subjektem může být zákazník, zaměstnanec, výrobek, územní jednotka a podobně. Požadavek integrace zahrnuje např. zavedení jednotné terminologie či jednotných veličin. Ukládané údaje by měly být konzistentní a důvěryhodné. Časovou variabilitu můžeme chápat jako uložení série snímků, z nichž každý reprezentuje určitý časový úsek.

Jak již bylo vysvětleno na obr. 2, datový sklad je zpravidla fyzicky i logicky oddělen od provozních systémů. Data z provozních systémů se převádějí do datového skladu, kde se po transformaci ukládají způsobem, který vyhovuje analytickému a prezentačnímu zpracování výstupů. Je potřebné si uvědomit, že je třeba optimalizovat strukturu s ohledem na dotazy, které budou nad daty prováděny.

Tabulka 1. Přehledné srovnání vlastností OLTP a OLAP (Groff, Weinberg 2005)

Databázová charakteristika	Databáze OLTP	Databáze datového skladování
obsah dat	aktuální data	historická data
datová struktura	tabulky organizované v souladu s transakční strukturou	tabulky organizované pro snadné chápaní a dotazování
velikost tabulky	tisíce řádků	miliony řádků
vzorec přístupu	předem určený pro každý typ transakce, která má být zpracována	rozmanitý, podle rozhodnutí, jež je třeba učinit
řádky prohledávané podle jedné žádosti	desítky	tisíce až miliony
rychlost přístupu	mnoho obchodní transakcí za sekundu nebo minutu	mnoho minut nebo hodin na jeden dotaz
typ přístupu	čtení, vložení, aktualizace	téměř 100% čtení
zaměření výkonu	kapacita transakcí	čas pro dokončení dotazu

K realizaci klasických datových skladů se používají relační databázové modely s odpovídající multidimenzionální strukturou (systémy ROLAP) nebo speciální multidimenzionální databáze, které multidimenzionální struktury podporují nativně (místo uložení dat v relačních tabulkách aplikují zpravidla vícerozměrná pole, známá z programovacích technik) a mají k dispozici specializovaný multidimenzionální SŘBD (Pokorný 2004). Druhé prostředí se zdá být obecně vhodnější, i když dochází k řídkému obsazení vzniklé multidimenzionální kostky.

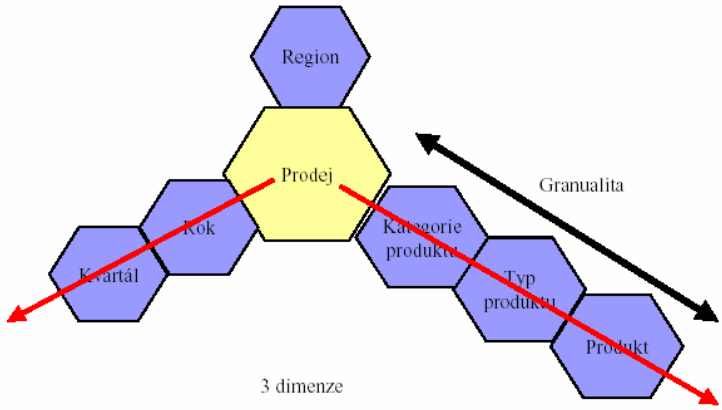
V multidimenzionální struktuře se používají termíny dimenze a fakta.

Dimenze reprezentují jednotlivé aspekty, podle kterých jsou organizována data (přesněji fakta) a podle kterých došlo k agregaci dat. Na jejich základě se provádí analýza agregovaných dat. Dimenze mají zpravidla hierarchickou strukturu. Jednotlivé prvky hierarchie se pak používají k seskupování dat (faktů). V klasických, ekonomických aplikacích je vždy přítomna ekonomická dimenze a čas jako 2 povinné dimenze (Dohnal, Pour, 1997); další dimenze se již navrhují s ohledem na preference uživatelů. Může to být dimenze výrobků, zaměstnanců, zákazníků, ale také geografická dimenze.

Například dimenze Čas a Geografie je možné definovat jako typické víceúrovňové hierarchie s úrovněmi dny, týdny, měsíce, čtvrtletí, roky; resp. obce, okresy, kraje, státy. Potom se konkrétní jev sleduje za určité období a za daný územní celek.
Většina dimenzí se mění jen pomalu a často mají charakter číselníků.

Fakta představují (agregované) hodnoty, které jsou zajímavé pro rozhodování. Představují takzvaná souhrnná data, která jsou obvykle numerická, měřitelná a získávají se opakovaně, což vytváří vztahy M:N mezi dimenzemi. V souvislosti s popisem faktů se uvádějí 2 vlastnosti, které je potřebné sledovat (Pokorný 2004):

- Granularita – vlastnost, která určuje úroveň podrobnosti faktů. Závisí přitom na úrovni podrobnosti dimenzí. Vysoká granularita znamená uložení dat v nízkém stupni agregace (např. údaje za sčítací obvody), tedy s velkým detailem dat.
- Aditivita faktů - vlastnost určující, zda je možné fakta sumarizovat podle dimenzí. Atributy, do kterých se ukládají fakta, můžeme rozlišit na atributy aditivní, které lze agregovat podle všech dimenzí, semiaditivní, které lze agregovat jen podle některých dimenzí, a neaditivní atributy.



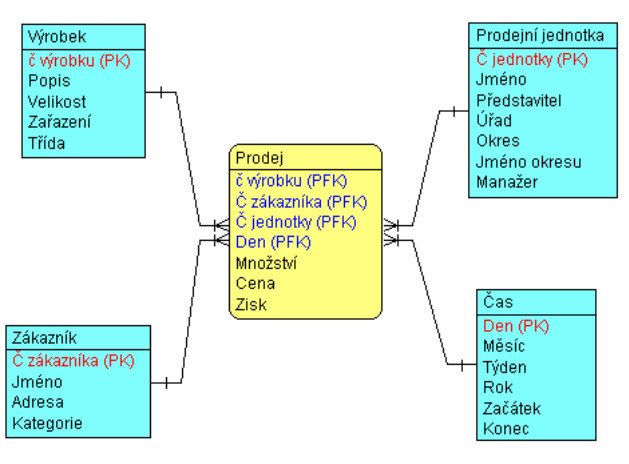
Obr. 3. Koncepce multidimenzionality (Pirkl, 2004)

Existují 2 základní varianty multidimenzionální datové struktury, používané pro datové sklady. Jde o:

1. hvězdy a souhvězdí
2. sněhové vločky

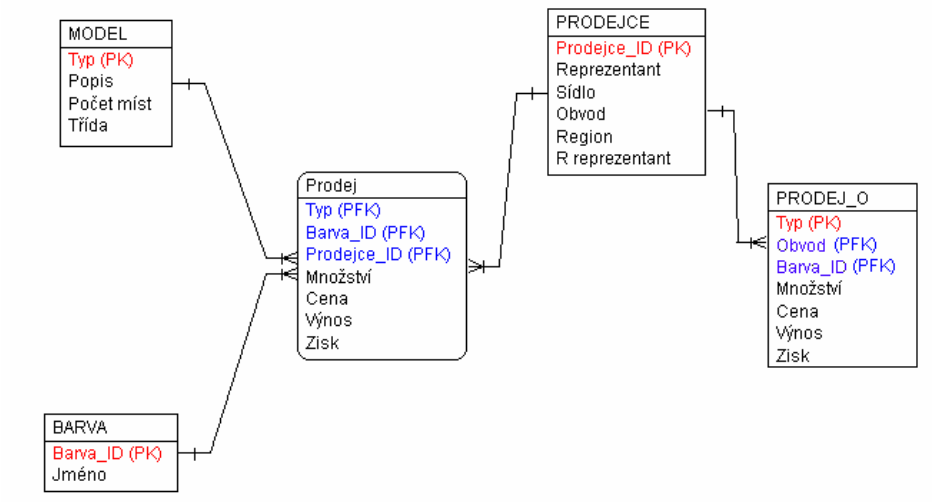
Někteří autoři hovoří samostatně o kostce (hyperkostce) či krychli.

Struktura typu *hvězda* (tzv. hvězdicové schéma) obsahuje nejméně 1 hvězdu, v jejímž centru je tabulka faktů, ve které jsou uložena fakta spolu s klíčem tvořeným kombinací identifikátorů všech dimenzí (tedy kombinací cizích klíčů). Na ní jsou napojeny tabulky dimenzí, které popisují vlastnosti vždy 1 dimenze. Typicky se předpokládá, že mezi jednotlivými dimenzemi nejsou žádné závislosti.



Obr. 4. Schéma typu hvězda (podle Pokorný 2004)

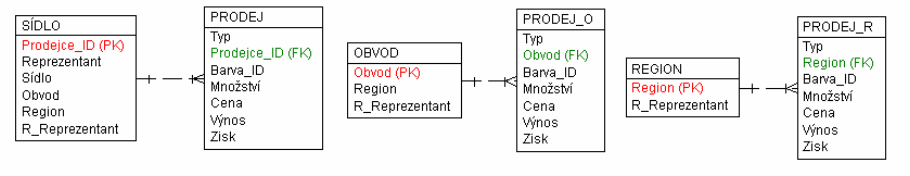
Souhvězdí (galaxie) pak představuje schéma s více hvězdami, které sdílejí některé dimenzionální tabulky.



Obr. 5. Schéma typu souhvězdí (podle Pokorný 2004)

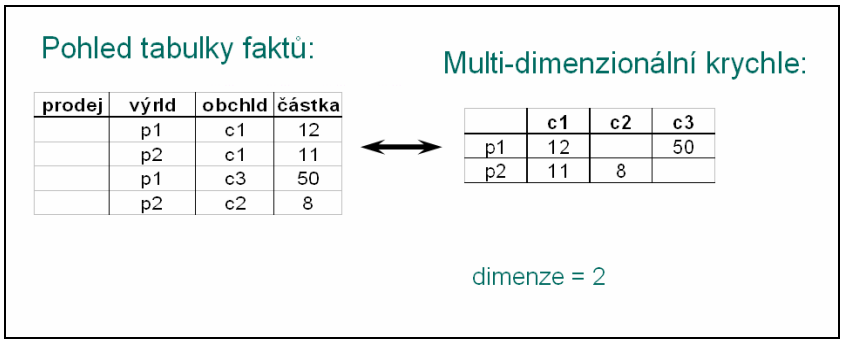
Z hlediska realizace hierarchie v dimenzi je vhodné rozlišovat dimenze s implicitní hierarchií a explicitní hierarchií. Implicitní hierarchie je vyjádřená zařazením potřebných atributů přímo v tabulce dimenze, která není normalizovaná. Explicitní hierarchie vytváří pro danou dimenzi hierarchický řetěz tabulek (např. obec – okres – kraj – oblast – stát). Oba přístupy mají své výhody.

Struktura typu *sněhová vločka* používá dílčích agregovaných tabulek faktů, spojených jen s 1 dimenzí.

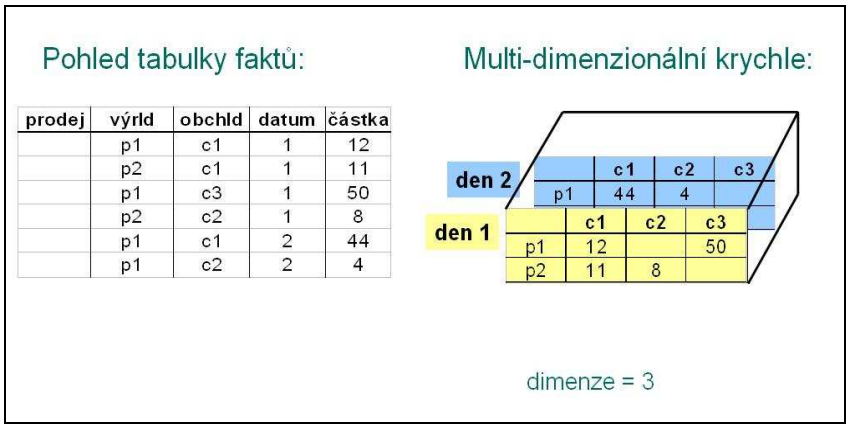


Obr. 6. Část schématu sněhových vloček pro dimenzi prodejce (podle Pokorný 2004)

Pojetí multidimenzionální datové struktury typu hyperkostka nám nejlépe přiblíží obrázky 7 a 8, kde jsou zobrazeny výchozí tabulky faktů a jim odpovídající multidimenzionální kostky. V prvním případě (obr .7) vzniká matice, přidáním časové dimenze pak krychle (obr. 8).



Obr. 7. Dvoudimenzionální kostka (in Kunz, 2006)



Obr. 8. Multidimenzionální krychle (in Kunz, 2006)

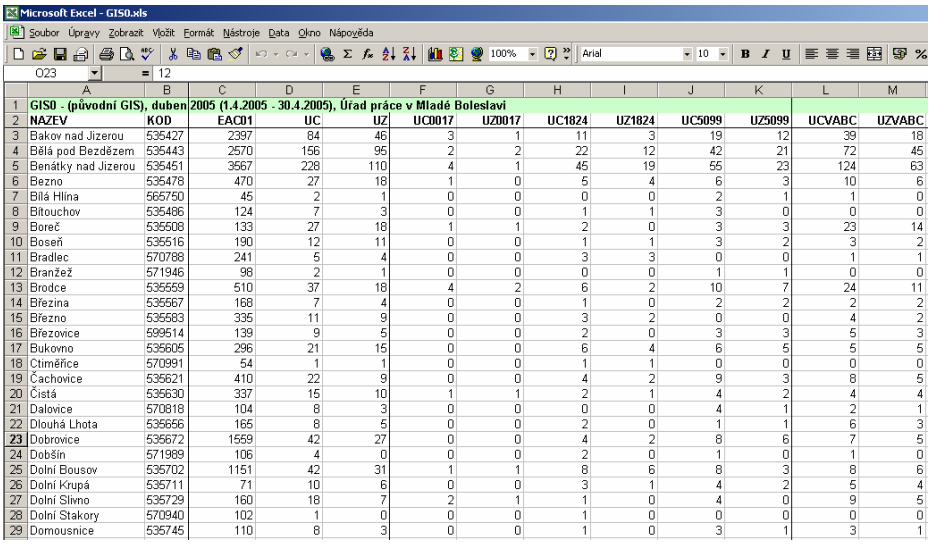
Protože tvorba jednoho rozsáhlého datového skladu pro organizaci se zohledněním různých aplikačních potřeb je obtížná, vznikají často tzv. *datová tržiště* (data mart) pro extrakci dílčí části dat.

Při budování datového skladu je nutné zajistit i jeho počáteční naplnění a pozdější doplňování. Můžeme hovořit o aktualizaci datového skladu, ale ve smyslu přidávání záznamů.

4 Návrh datového skladu 2.typu pro vybraná socioekonomická prostorová data

Jedním z výsledků projektu „Implementace nástrojů prostorové analýzy trhu práce v činnosti úřadů práce“, realizovaného ve spolupráci s MPSV ČR a úřady práce byl návrh agregace vybraných primárních údajů a ukazatelů z operativní databáze úřadů práce (Horák 2001 nebo Horák 2002). Firma OKsystém s.r.o. realizovala pro tento účel funkci, generující z informačního systému OKpráce statistiku označovanou jako GIS statistika, protože jejím základním účelem bylo poskytovat údaje pro konstrukci kartogramů a kartodiagramů mapující situaci na trhu práce v území příslušného úřadu práce. Tato GIS statistika se pravidelně připravuje počátkem měsíce a obsahuje údaje k poslednímu dni předchozího měsíce. Exportovaná data jsou uložena v souborech XLS. V původní verzi se exportoval pouze jeden soubor, obsahující celkem 34 primárních údajů (počet osob v dané kategorii, počet volných míst) za každou obec v okrese a 34 ukazatelů, tj. údajů vypočtených z primárních hodnot a charakterizujících situaci na trhu práce.

Od roku 2005 jsou k dispozici další 3 exporty, podstatně rozšiřující množinu dostupných údajů (dnes již 142 primárních údajů a zhruba stejný počet ukazatelů).



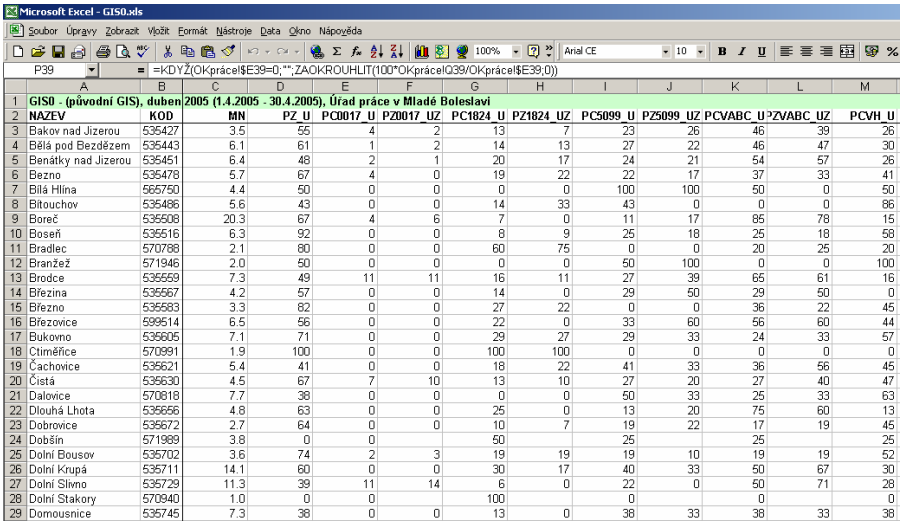
Microsoft Excel - GIS0.xls

Soubor Úpravy Zobrazit Vložit Formát Nástroje Data Okno Nápověda

O23 = 12

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	GIS0 - (původní GIS), duben		2005 (1.4.2005 - 30.4.2005), Úřad práce v Mladé Boleslavi										
2	NAZEV	KOD	EACI01	UC	UZ	UC0017	UZ0017	UC1824	UZ1824	UC5099	UZ5099	UCVABC	UZVABC
3	Bakov nad Jizerou	535427	2397	84	46	3	1	11	3	19	12	39	18
4	Bělá pod Bezdězem	535443	2570	156	95	2	2	22	12	42	21	72	45
5	Benátky nad Jizerou	535451	3567	228	110	4	1	45	19	55	23	124	63
6	Bezno	535478	470	27	18	1	0	5	4	6	3	10	6
7	Bílá Hlína	565750	45	2	1	0	0	0	0	2	1	1	0
8	Bitouchov	535486	124	7	3	0	0	1	1	3	0	0	0
9	Boreč	535508	133	27	18	1	1	2	0	3	3	23	14
10	Boseň	535516	190	12	11	0	0	1	1	3	2	3	2
11	Bradlec	570788	241	5	4	0	0	3	3	0	0	1	1
12	Branžež	571946	98	2	1	0	0	0	0	1	1	0	0
13	Brodce	535559	510	37	18	4	2	6	2	10	7	24	11
14	Březina	535567	168	7	4	0	0	1	0	2	2	2	2
15	Březno	535583	335	11	9	0	0	3	2	0	0	4	2
16	Březovice	599514	139	9	5	0	0	2	0	3	3	5	3
17	Bukovno	535605	296	21	15	0	0	6	4	6	5	5	5
18	Čtiměřice	570991	54	1	1	0	0	1	1	0	0	0	0
19	Čachovice	535621	410	22	9	0	0	4	2	9	3	8	5
20	Čistá	535630	337	15	10	1	1	2	1	4	2	4	4
21	Dalovice	570818	104	8	3	0	0	0	0	4	1	2	1
22	Dlouhá Lhota	535656	165	8	5	0	0	2	0	1	1	6	3
23	Dobrovice	535672	1559	42	27	0	0	4	2	8	6	7	5
24	Dobšín	571989	106	4	0	0	0	2	0	1	0	1	0
25	Dolní Bousov	535702	1151	42	31	1	1	8	6	8	3	8	6
26	Dolní Krupá	535711	71	10	6	0	0	3	1	4	2	5	4
27	Dolní Slivno	535729	160	18	7	2	1	1	0	4	0	9	5
28	Dolní Stakory	570940	102	1	0	0	0	1	0	0	0	0	0
29	Domousnice	535745	110	8	3	0	0	1	0	3	1	3	1

Obr. 9. Ukázka obsahu listu OKpráce pro základní šablonu



Microsoft Excel - GIS0.xls

Soubor Úpravy Zobrazit Vložit Formát Nástroje Data Okno Nápověda

P39 = =KDYŽ(OKpráce!$E39=0;"";ZAOKROUHLIT(100*OKpráce!Q39/OKpráce!$E39;0))

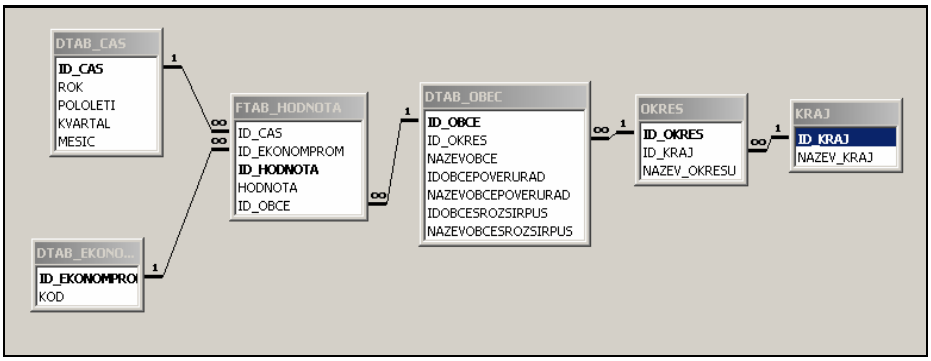
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	GIS0 - (původní GIS), duben		2005 (1.4.2005 - 30.4.2005), Úřad práce v Mladé Boleslavi										
2	NAZEV	KOD	MN	PZ_U	PC0017_U	PZ0017_UZ	PC1824_U	PZ1824_UZ	PC5099_U	PZ5099_UZ	PCVABC_U	PZVABC_UZ	PCVH_U
3	Bakov nad Jizerou	535427	3.5	55	4	2	13	7	23	26	46	39	26
4	Bělá pod Bezdězem	535443	6.1	61	1	2	14	13	27	22	46	47	30
5	Benátky nad Jizerou	535451	6.4	48	2	1	20	17	24	21	54	57	26
6	Bezno	535478	5.7	67	4	0	19	22	22	17	37	33	41
7	Bílá Hlína	565750	4.4	50	0	0	0	0	100	100	50	0	50
8	Bitouchov	535486	5.6	43	0	0	14	33	43	0	0	0	86
9	Boreč	535508	20.3	67	4	6	7	0	11	17	85	78	15
10	Boseň	535516	6.3	92	0	0	8	9	25	18	25	18	58
11	Bradlec	570788	2.1	80	0	0	60	75	0	0	20	25	20
12	Branžež	571946	2.0	50	0	0	0	0	50	100	0	0	100
13	Brodce	535559	7.3	49	11	11	16	11	27	39	65	61	16
14	Březina	535567	4.2	57	0	0	14	0	29	50	29	50	0
15	Březno	535583	3.3	82	0	0	27	22	0	0	36	22	45
16	Březovice	599514	6.5	56	0	0	22	0	33	60	56	60	44
17	Bukovno	535605	7.1	71	0	0	29	27	29	33	24	33	57
18	Čtiměřice	570991	1.9	100	0	0	100	100	0	0	0	0	0
19	Čachovice	535621	5.4	41	0	0	18	22	41	33	36	56	45
20	Čistá	535630	4.5	67	7	10	13	10	27	20	27	40	47
21	Dalovice	570818	7.7	38	0	0	0	0	50	33	25	33	63
22	Dlouhá Lhota	535656	4.8	63	0	0	25	0	13	20	75	60	13
23	Dobrovice	535672	2.7	64	0	0	10	7	19	22	17	19	45
24	Dobšín	571989	3.8	0	0		50		25		25		25
25	Dolní Bousov	535702	3.6	74	2	3	19	19	19	10	19	19	52
26	Dolní Krupá	535711	14.1	60	0	0	30	17	40	33	50	67	30
27	Dolní Slivno	535729	11.3	39	11	14	6	0	22	0	50	71	28
28	Dolní Stakory	570940	1.0	0	0		100		0		0		0
29	Domousnice	535745	7.3	38	0	0	13	0	38	33	38	33	38

Obr. 10. Ukázka obsahu listu Ukazatelé pro základní šablonu

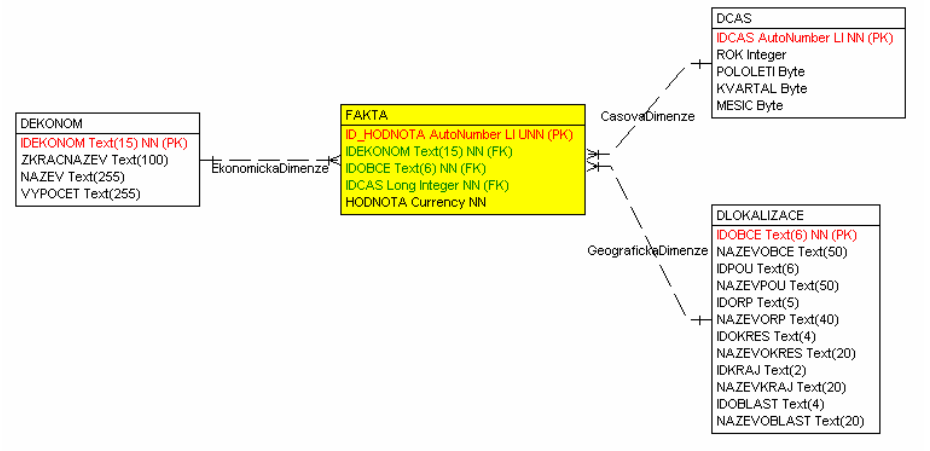
S rostoucím počtem údajů je zřejmé, že pro řadu analýz takové souborově orientované uložení dat z GIS statistik není vhodné. Např. není jednoduché propojovat data napříč tabulkami v jednotlivých souborech při požadavku sledování časového vývoje daného ukazatele.

Pro nové uložení dat je možné navrhnout multidimenzionální datovou strukturu typu hvězda s částečnou explicitní hierarchií (obr. 11) nebo z důvodu požadavku na zachycení vývoje administrativního uspořádání území využít pouze implicitní

hierarchie (obr.12). Fakta (tedy vlastní číselné hodnoty) jsou uložena v tabulce FAKTA (obr. 13).



Obr. 11. Struktura typu hvězda s částečnou externí hierarchií pro dimenzi administrativního uspořádání (Kunz 2006)



Obr. 12. Struktura typu hvězda s implicitní hierarchií pro dimenzi administrativního uspořádání

FTAB_HODNOTA : Tabulka

ID_CAS	ID_EKONOMPI	HODNOTA	ID_OBCE
26	40	29.00	551694
27	40	29.00	551694
28	40	29.00	551694
29	40	17.00	551694
30	40	17.00	551694
31	40	17.00	551694
32	40	20.00	551694
33	40	14.00	551694
34	40	14.00	551694
35	40	13.00	551694
36	40	11.00	551694
1	41	13.00	551694
2	41	9.00	551694
3	41	10.00	551694
4	41	9.00	551694
5	41	8.00	551694
6	41	11.00	551694
7	41	11.00	551694
8	41	10.00	551694
9	41	0.00	551694
10	41	0.00	551694
11	41	0.00	551694
12	41	0.00	551694
13	41	0.00	551694
14	41	0.00	551694

Obr. 13. Obsah tabulky FAKTA

Definice ekonomické dimenze je uložena v tabulce DEKONOM. Odpovídá popisu ekonomických proměnných v souborech GIS statistiky (struktura viz obr. 12, obsah obr. 14). Tabulka obsahuje seznam celkem 315 vstupních dat a ukazatelů, které mohou být v tabulce faktů evidovány.

DEKONOM : Tabulka

IDEKONOM	ZKRACNAZEV	NAZEV	VYPOCET
UZKZAM9	Počet uch. žen požadujících KZAM9	Počet evidovaných uchazečů o zaměstnání - požadující primá	Počet evidovaných uchazečů o zaměstnání : požadující primárně zam
UCABS	Počet uchazečů - absolventů	Počet evidovaných uchazečů o zaměstnání - absolventi	Počet evidovaných uchazečů o zaměstnání : absolvent
UZABS	Počet uchazečů - ženy absolventky	Počet evidovaných uchazečů o zaměstnání - absolventi - žen	Počet evidovaných uchazečů o zaměstnání : žena a současně absol
UCMLA	Počet uchazečů mladistvých	Počet evidovaných uchazečů o zaměstnání - mladiství	Počet evidovaných uchazečů o zaměstnání : (věk 17 let a méně) a sc
UZMLA	Počet uchazečů žen mladistvých	Počet evidovaných uchazečů o zaměstnání - mladiství - ženy	Počet evidovaných uchazečů o zaměstnání : (věk 17 let a méně) a sc
UCABSE6	Počet uchazečů - absolventů nad 6m.	Počet evidovaných uchazečů o zaměstnání - absolventi v evid	Počet evidovaných uchazečů o zaměstnání : absolvent a současně v
UZABSE6	Počet uchazečů - ženy absolv. nad 6m.	Počet evidovaných uchazečů o zaměstnání - absolventi v evid	Počet evidovaných uchazečů o zaměstnání : absolvent a současně v
UCMLAE6	Počet uchazečů mladistvých nad 6 m.	Počet evidovaných uchazečů o zaměstnání - mladiství v evide	Počet evidovaných uchazečů o zaměstnání : (věk 17 let a méně) a sc
UZMLAE6	Počet uch. žen mladistvých nad 6 m.	Počet evidovaných uchazečů o zaměstnání - mladiství v evide	Počet evidovaných uchazečů o zaměstnání : (věk 17 let a méně) a sc
VMC	Počet hlášených volných míst	Počet hlášených volných pracovních míst	Počet hlášených volných pracovních míst podle fyzického místa prac
UCE0	Počet nově evidovaných uchazečů	Počet nově evidovaných uchazečů v daném měsíci	Počet nově evidovaných uchazečů v daném měsíci
UCEX	Počet uchazečů, kteří odešli z evidence	Počet uchazečů, kteří odešli z evidence v daném měsíci („vyř	Počet uchazečů, kteří odešli z evidence v daném měsíci („vyřazení ce
EAZ01	Počet ekonom. aktivních obyvatel - žen	Počet ekonomicky aktivních obyvatel - ženy	
UCDOS	Počet dosažitelných uchazečů	Počet dosažitelných uchazečů	
UZDOS	Počet dosažitelných uchazečů - žen	Počet dosažitelných uchazečů - žen	
UCDOSMN	Míra nezaměstnanosti z dosažitelných uchaze	Míra nezaměstnanosti z dosažitelných uchazečů	100*UCDOS/EAC01
UZDOSMN	Míra nezaměstnanosti z dosažitelných uchaze	Míra nezaměstnanosti z dosažitelných uchazečů - žen	100*UZDOS/EAZ01
MN	Míra nezaměstnanosti [%]	Míra nezaměstnanosti [%]	100*UC/EAC01
PZ_U	Podíl žen [%]	Podíl žen na celkovém počtu uchazečů o zaměstnání [%]	100*UZ/UC
PC0017_U	Podíl věku 17 let a méně [%]	Podíl věkové skupiny 17 let a méně [%]	100*UC0017/UC
PZ0017_UZ	Podíl žen věku 17 let a méně [%]	Podíl žen věkové skupiny 17 let a méně [%]	100*UZ0017/UZ
PC1824_U	Podíl věku 18 - 24 let [%]	Podíl věkové skupiny 18 - 24 let [%]	100*UC1824/UC
PZ1824_UZ	Podíl žen věku 18 - 24 let [%]	Podíl žen věkové skupiny 18 - 24 let [%]	100*UZ1824/UZ
PC5099_U	Podíl věku 50 let a více [%]	Podíl věkové skupiny 50 let a více [%]	100*UC5099/UC
PZ5099_UZ	Podíl žen věku 50 let a více [%]	Podíl žen věkové skupiny 50 let a více [%]	100*UZ5099/UZ
PCVABC_U	Podíl uch. se základ. stupněm vzd. [%]	Podíl uchazečů se základním stupněm vzdělání [%]	100*UCVABC/UC
PZVABC_UZ	Podíl žen se zákl. stupněm vzd. [%]	Podíl žen se základním stupněm vzdělání [%]	100*UZVABC/UZ
PCVH_U	Podíl vyučených [%]	Podíl vyučených [%]	100*UCVH/UC

Záznam: 1 z 315

Obr. 14. Obsah tabulky DEKONOM

Časová dimenze je realizována pomocí tabulky DCAS (viz obr. 12). Hierarchie byla vyjádřena pouze pro úroveň měsíc, kvartál, pololetí a rok.

Z hlediska geoinformatiky je samozřejmě nejdůležitější realizace geografické dimenze (případně dimenzí).

Nejvíce problematické je konzistentní řešení změn v geograficky vymezení území, ke kterým dochází. Obecně bychom tedy měli řešit problém napojení na „stavovou" topologii geografické databáze (Rapant 2005).

Nejdříve si položme otázku, zda je potřebné registrovat změny průběhu hranic a další drobné geometrické úpravy. Pokud budeme předpokládat, že drobné geometrické změny hranic zřejmě neovlivní socioekonomickou situaci, můžeme navrhnout realizaci systému, kde budeme popisovat změny pouze na základě příslušnosti malých územních jednotek do sledovaných celků (popsat územní změnu změnou příslušnosti katastrálního území, sčítacího obvodu či základní sídelní jednotky do daného celku např. zde obce).

S tím je spojena otázka volby vhodné granularity uložených údajů.

Pokud není možné ukládat data pro menší územní jednotky, je nutné zajistit alespoň externě jejich přepočty na standardizované jednotky např. OkrBruntál9604 (tj. území okresu Bruntál, platné mezi 1.1.1996 a 31.12.2004, tj. po odloučení Zlatých Hor a odloučením 3 obcí k 1.1.2005)

Vzniká tak jednodušší a robustnější systém, kdy k prostorové lokalizaci budeme používat pouze geokódy.

Pro geografickou dimenzi byla nejdříve připravena struktura s částečně explicitní hierarchií dle požadavku normalizace dat. Vzhledem k charakteru datového skladu ale bylo rozhodnuto připravit pouze implicitní hierarchii, tedy všechny hierarchické úrovně popsat přímo v tabulce DLOKALIZACE (obr. 12). Toto nenormalizované řešení nám umožní řešit různou příslušnost obcí do jednotlivých celků v průběhu času. Abychom odlišili jednotlivé situace pro 1 obec, rozšířili jsme primární klíč této dimenzionální tabulky o číslo verze. Doplňujícím atributem se stává datum platnosti, které vyjadřuje, od kdy je daná situace platná.

Příklad v tabulce 2 ukazuje zápis situace s rozdílným zařazením 3 obcí (Huzová, Moravský Beroun a Norberčany) do okresu a kraje před a po 1.1.2005.

Tabulka 2. Tabulka DLOKALIZACE včetně verze a platnosti dat

Kod	Obec	Okres	Kraj	**verze**	platnost
597414	Huzová	Bruntál	MSK	**1**	
597678	Moravský Beroun	Bruntál	MSK	**1**	
597686	Norberčany	Bruntál	MSK	**1**	
597414	Huzová	Olomouc	OLK	**2**	1.1.2005
597678	Moravský Beroun	Olomouc	OLK	**2**	1.1.2005
597686	Norberčany	Olomouc	OLK	**2**	1.1.2005

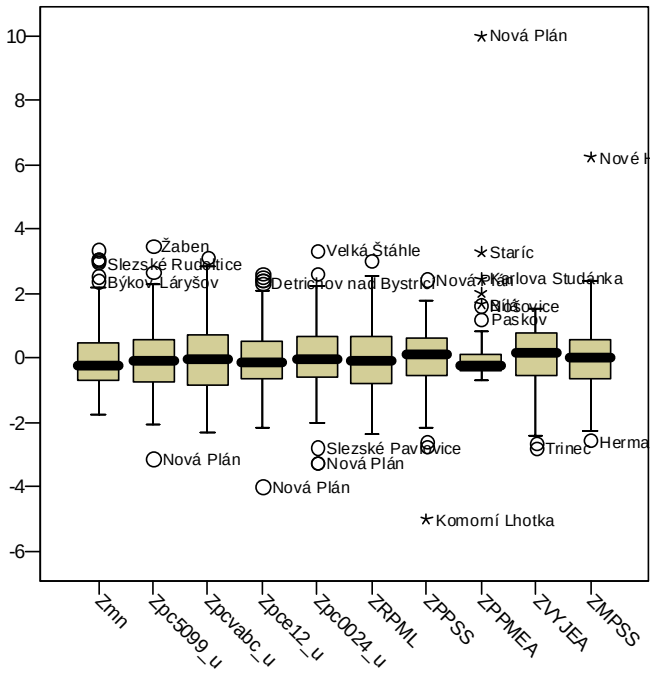
V takovém systému lze provádět samostatné agregace dle jednotlivých úrovní hierarchie s ohledem na verzi (resp. období platnosti). Pro agregaci na elementární úrovni potřebujeme externí přepočty, jestliže si verze dané jednotky neodpovídají. Je ovšem vhodné vždy kontrolovat počet členů agregace.

Zpracování dat uložených ve struktuře typu hvězda může být prováděno přímo ve vytvořené speciální databázové aplikaci s využitím řady agregačních SQL dotazů nebo lze využít některého již vytvořeného programového produktu s ohledem na

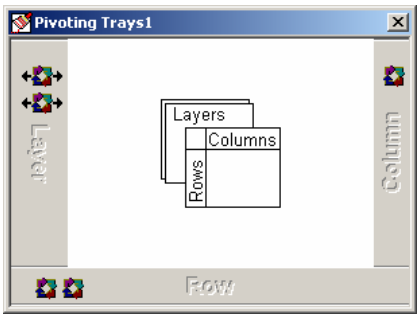
požadavky, kladené na aplikaci. Jedním z takových produktů, který je orientován na statistické vyhodnocení dat, je program SPSS. Základní informace lze nalézt na http:\\www.spss.cz.

Způsob práce v SPSS přímo zvýhodňuje uložení sledovaných fakt do 1 sloupce, protože je tak orientována i tvorba výběrů či hromadné generování statistik pro daný atribut s rozlišením dle definovaných skupin. Obr. 15 ukazuje výsledek vytvoření takové jednoduché statistiky pro vybrané socioekonomické ukazatele (míra nezaměstnanosti, podíl mladých a starších na počtu nezaměstnaných, podíl vyjíždějících atd.) v obcích Moravskoslezského kraje jako podklad pro vyhodnocení situace v území. Např. obec Nová Pláň dosahuje v řadě ukazatelů extrémních hodnot a značně se tedy odlišuje.

Dále má program SPSS dobře propracovaný nástroj pro Pivoting výsledné statistiky (tedy záměnu dimenzí v pohledu) (obr.16), kde přetahováním symbolů pro agregační skupiny mezi pozicí řádek-sloupec-vrstva lze snadno měnit agregační kritéria pro zpracování dat.

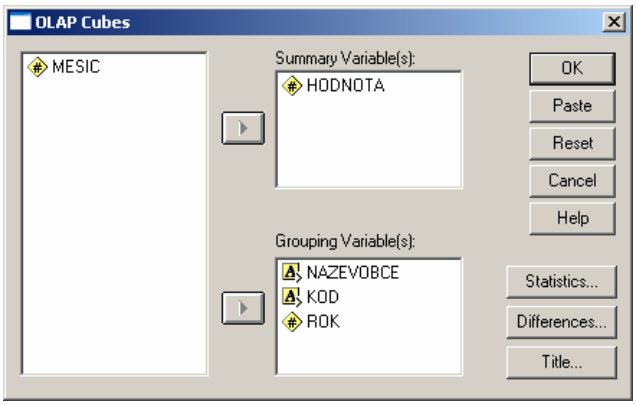


Obr. 15. Krabicový graf pro část ekonomických ukazatelů v obcích MSK v prostředí SPSS

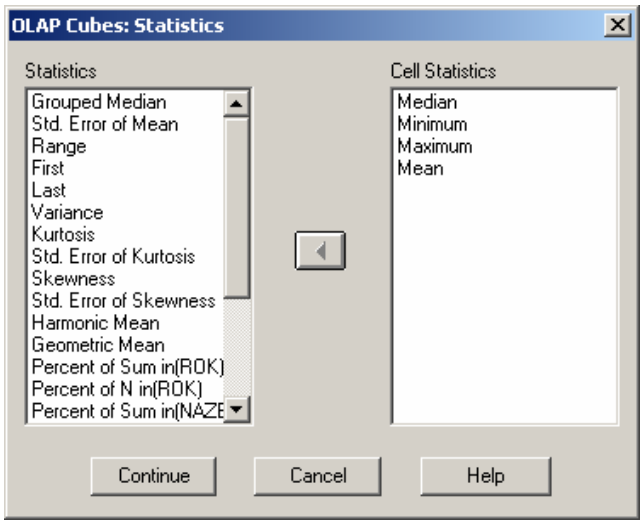


Obr. 16. Pivoting tabulkového výstupu v prostředí SPSS

Navíc program obsahuje přímo možnost definovat OLAP kostky a generovat potřebné statistiky (obr. 17, obr. 18). Vybrané statistické ukazatele je možné považovat za realizaci další dimenze pohledu na fakta (medián, průměr, minimum či maximum v podstatě představují hodnoty další, statistické dimenze).



Obr. 17. Definice OLAP kostky v prostředí SPSS pro daný příklad (Kunz 2006)



Obr. 18. Výběr statistik pro výpočet hodnot v buňce v prostředí SPSS (Kunz 2006)

Vytvořený report je interaktivní a snadno umožňuje jednak vybírat zájmové administrativní jednotky či ekonomické proměnné, tak rovněž pomocí pivotingu dosáhnout změny uspořádání výstupu a také způsobu agregace výsledných statistických ukazatelů. Obr. 19 ukazuje agregaci údajů o míře nezaměstnanosti pro město Bruntál dle vybraných roků, obr. 20 agregaci podle měsíců.

OLAP Cubes

NAZEVOBCE	Bruntál
KOD	MN

	ROK	Median	Minimum	Maximum	Mean
HODNOTA	1995	5.5000	4.90	6.30	5.5250
	1996	5.6000	4.90	6.40	5.6750
	1997	5.9000	5.30	7.00	6.0667
	Total	5.7500	4.90	7.00	5.7556

Obr. 19. Výsledná agregace dat podle roků v prostředí SPSS (Kunz 2006)

OLAP Cubes

NAZEVOBCE	Bruntál
KOD	MN

	MESIC	Median	Minimum	Maximum	Mean	N
HODNOTA	1	6.3000	6.20	6.90	6.4667	3
	2	6.4000	6.10	6.80	6.4333	3
	3	6.4000	5.90	6.40	6.2333	3
	4	5.7000	5.20	5.80	5.5667	3
	5	5.1000	4.90	5.50	5.1667	3
	6	4.9000	4.90	5.50	5.1000	3
	7	5.3000	5.10	5.30	5.2333	3
	8	5.5000	5.10	5.80	5.4667	3
	9	5.7000	5.50	6.00	5.7333	3
	10	5.5000	5.10	5.80	5.4667	3
	11	5.9000	5.50	6.00	5.8000	3
	12	6.3000	5.90	7.00	6.4000	3
	Total	5.7500	4.90	7.00	5.7556	36

Obr. 20. Výsledná agregace dat podle měsíců v prostředí SPSS (Kunz 2006)

Tento příklad dokumentoval možnosti tvorby a využití datového skladu 2.typu, ovšem konkrétní aplikace je zatím ve stádiu návrhu.

5 Datový sklad 3.typu

Hlavní motivací pro vytváření multidimenzionálních datových struktur pro ukládání originálních, transakčních dat je schopnost dobře vystihnout faktory evidence a významu ukládaných dat.

Tento význam dobře charakterizuje definice datového skladu dle Kimballa (2003): „Dimenzionální modelování začíná rozdělením světa do měřených dat a kontextu. Měřená data jsou bvykle numerická a jsou získávána opakovaně. Fakta jsou vždy obklopena kontextem většinou textového charakteru, který je pravdivý ve chvíli, kdy jsou fakta zaznamenána. Jestliže jsou fakta skutečně opakovaně měřená data, zjistíte, že tabulka faktů vždy vytváří charakteristické M:N vztahy mezi dimenzemi. Tento přístup vytváří koncept multidimenzionální struktury skládající se z tabulky faktů, obklopené dimenzionálními tabulkami“.

V datovém skladu 3.typu se tedy nepokoušíme unifikovat data při jejich sběru, ale naopak se snažíme, aby při pořizování byla data zapsána v primární, co nejméně změněné podobě a současně aby se zapisovala řada metadat, která umožňují následné správné využití dat.

Současně můžeme pozorovat i určitý pozitivní vedlejší efekt, spočívající v úspoře místa pro řídce obsazené struktury. Neukládají se totiž prázdné hodnoty, ale pouze známé hodnoty.

Charakteristický problém u popisu kontextu měřených dat představuje zachycení skutečnosti změny sémantického obsahu popisovaného atributu, do kterého se data ukládají.

Definice atributu se totiž obecně mění s:

- Časem
- Územím (rozlohou)
- Národními specifiky

Předpokládejme, že budeme ukládat míru nezaměstnanosti v takovém skladu. Problém spočívá v tom, že ačkoliv je míra nezaměstnanosti celkem jednoznačně definována (zjednodušeně řečeno počet nezaměstnaných ku počtu ekonomicky aktivních obyvatel), praktická implementace definice se vyvíjí a mění jak v čase, tak i mezi jednotlivými státy a dokonce i s ohledem na různou úroveň územní jednotky. Jde o to, jak se zjišťuje počet nezaměstnaných (např. na úrovni České republiky či kraje se stanovuje z Výběrového šetření pracovních sil, kdežto na úrovni obce se použije pouze počet uchazečů o zaměstnání s bydlištěm v dané obci evidovaných na územně příslušném úřadu práce), jak se stanovuje aktuální počet ekonomicky aktivních (které skupiny osob do nich zařadíme - ženy na mateřské dovolené, osoby v pracovní neschopnosti, vojáci, studenti v souběhu s praxí atd. - a jak jejich počet zjistíme) i jak se uplatní další úpravy (např. aplikace klouzavého průměru např. pro výpočty na úrovni okresů). Je evidentní, že se míra nezaměstnanosti stanovená odlišnými způsoby liší a bohužel to významně ovlivňuje výsledky. Některé poklesy míry nezaměstnanosti v naší nedávné historii tak lze jednoznačně připsat na vrub změněné metodice a ne v dané chvíli mediálně a politicky oslavované změně hospodářské situace.

Přitom nelze zpravidla provést homogenizaci těchto dat na jakýsi univerzální atribut s jasnou sémantikou. Pokud pracujeme na stejné úrovni hierarchii (např. srovnáváme údaje obcí mezi sebou) a sledujeme dané území v krátkém časovém úseku (aby se nezměnila metodika přípravy dat), je potřebné použít oficiální publikované údaje

Proto je potřebné uložit sémantický popis (alespoň zkráceně, ve formě odkazu na metodu stanovení) explicitně, tj. jako další atribut, resp. novou dimenzi v multidimenzionálním světě. Jiné řešení, např. paralelní evidence několika atributů s odlišným významem (MN1, MN2, MN3 dle metodiky 1, 2 a 3) stěží povede k cíli, navíc vznikají problémy s navazujícím zpracováním.

Problémem řešení rozdílné sémantiky dat v tomto kontextu se zabývají nejenom teoretické práce, ale i i několik evropských projektů (např. HarmonIT http://www.harmonit.org/ nebo HarmoniRiB - Moore, Tindall and Bech, 2003).

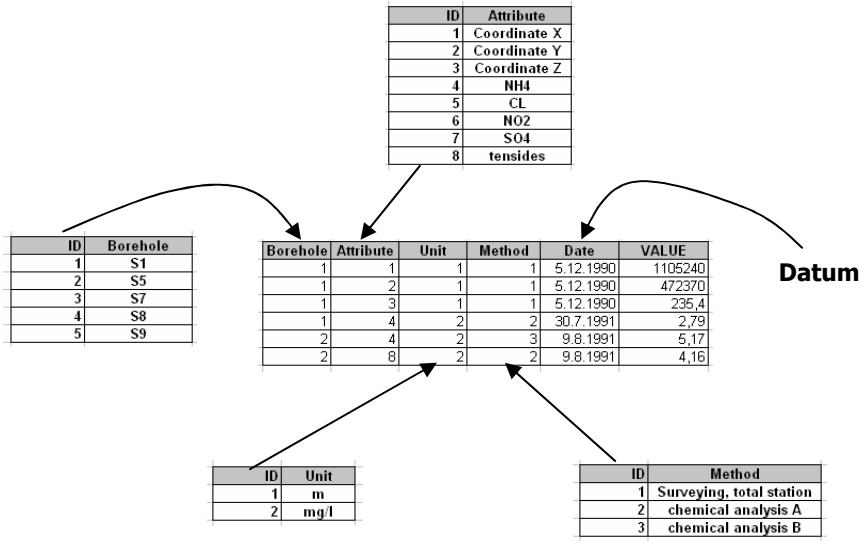
Jako jiný příklad můžeme uvést měření množství atmosférických srážek. Projevuje se zde hlavně vliv:

- Času – pokud jde o historická data, byla pravděpodobně změřena, pokud jde o data pro budoucnost, jde o předpověď (i když ve formě „změřených“ údajů z modelu)
- Metody, které můžeme obecně rozdělit na

- metody pro měřené hodnoty – množství mohlo být stanoveno na srážkoměrné stanici (a jaký byl její typ), nebo radarovým odhadem nebo byl použit adjustovaný radarový odhad
- metody pro predikované hodnoty – např. využití modelu ALADIN, modelu GFS.

Společné uložení všech variant měření srážek v jednom atributu (s odlišením metody) má několik pozitivních efektů. Např. při mapování srážek v území můžeme všechny varianty v 1 „vrstvě“ (při shodném čase) kombinovat s cílem získání více lokalizovaných hodnot v území pro lepší interpolaci. Pak je ovšem vhodné interpolaci provádět s ohledem na odlišnou míru neurčitosti těchto dat.
Dokonce v dané vrstvě můžeme použít i data odvozená např. z časové řady při výpadku daného měření na stanici, pokud by toto řešení bylo vhodnější než plošná interpolace. Důvodem může být např. nízká kontinuita hodnot v dané části pole, tedy nízká plošná korelace s okolními stanicemi.

Data jsou tedy uložena v datovém skladu s multidimenzionální strukturou. To umožňuje zapisovat měřená data s plnými metadaty, popisujícími kdy, kde, co a jak bylo měřeno (obr.21).

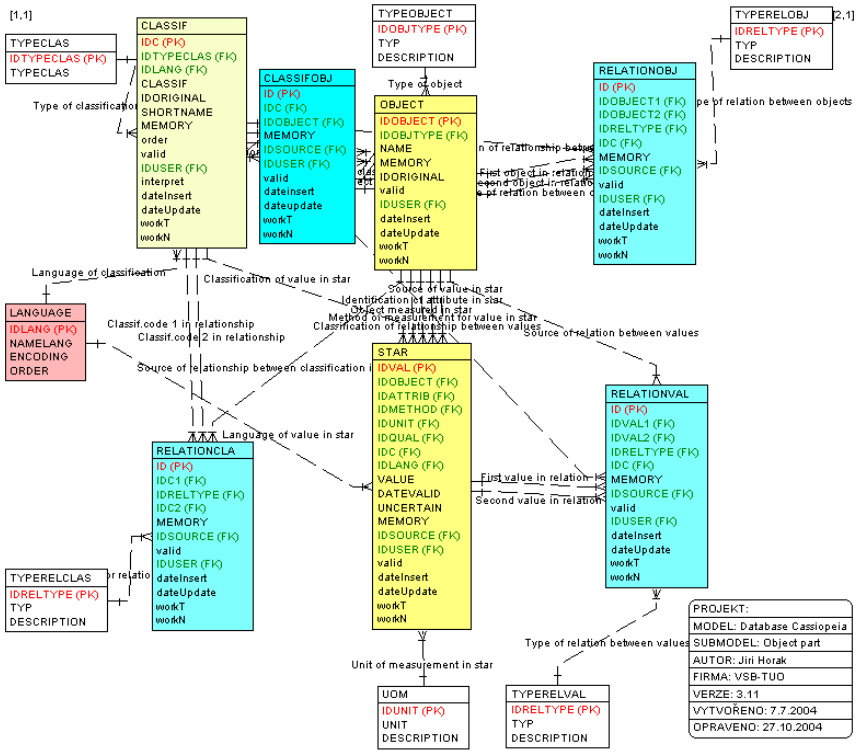


Obr. 21. Princip ukládání transakčních dat v multidimenzionální struktuře

Takového postupu a datové struktury bylo využito v projektu TRANSCAT pro databázi Cassoipeia (Horák et al., 2006). Plná sestava dimenzí zahrnuje:

- Objekt (kde se měří)
- atribut (co se měří)
- metoda (např. identifikace analytické metody)
- jednotka měření
- kvalifikátor (=, >,<, stopy apod.)

- čas (datum a čas)
- hodnocení neurčitosti
- datová sada (potřebné kvůli vazbě na plná metadata, copyright)
- klasifikační schéma (v případě ukládání kódů hornin, typů vrstev, půdy, vegetace, apod.)
- zdroj
- jazyk



Obr. 22. Struktura objektové části databáze Cassiopeia (Horák et al., 2005)

Vzniklá struktura (obr. 22) je sice relativně málo srozumitelná pro běžného uživatele, ale má řadu výhod při přístupu aplikace k datům a především jsou data zachována v podobě, která by neměla omezovat budoucí využití.

Pohled na data uložená v tabulce faktů je na obr. 23, fakta jsou uložena ve sloupci VALUE, většina dalších sloupců představuje realizaci identifikátorů dalších dimenzí, jako je objekt měření (IDOBJECT), identifikátor atributu který se měřil (IDATTRIB), identifikátor způsobu měření (IDMETHOD), identifikátor jednotky měření (UOM) atd.

STAR : Tabulka

IDVAL	IDOBJECT	IDATTRIB	IDMETHOD	IDUNIT	IDQUAL	IDC	IDLANG	VALUE	DATEVALID	UNCERT	MEMO	IDSOURCE	IDUSER
VSX453097	VSB005401	VSX452133	VSX452135	26	1	0	EN	0.13	22.7.2005 7:00:00			VSB000238	ADM000003
VSX453096	VSX451886	VSX452133	VSX452135	26	1	0	EN	0.13	22.7.2005 6:00:00			VSB000238	ADM000003
VSX453095	VSX451885	VSX452133	VSX452135	26	1	0	EN	0.13	22.7.2005 6:00:00			VSB000238	ADM000003
VSX453094	VSB005401	VSX452133	VSX452135	26	1	0	EN	0.13	22.7.2005 6:00:00			VSB000238	ADM000003
VSX453093	VSX451886	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 5:00:00			VSB000238	ADM000003
VSX453092	VSX451885	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 5:00:00			VSB000238	ADM000003
VSX453091	VSB005401	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 5:00:00			VSB000238	ADM000003
VSX453090	VSX451886	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 4:00:00			VSB000238	ADM000003
VSX453089	VSX451885	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 4:00:00			VSB000238	ADM000003
VSX453088	VSB005401	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 4:00:00			VSB000238	ADM000003
VSX453087	VSX451886	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 3:00:00			VSB000238	ADM000003
VSX453086	VSX451885	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 3:00:00			VSB000238	ADM000003
VSX453085	VSB005401	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 3:00:00			VSB000238	ADM000003
VSX453084	VSX451886	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 2:00:00			VSB000238	ADM000003
VSX453083	VSX451885	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 2:00:00			VSB000238	ADM000003
VSX453082	VSB005401	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 2:00:00			VSB000238	ADM000003
VSX453081	VSX451886	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 1:00:00			VSB000238	ADM000003
VSX453080	VSX451885	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 1:00:00			VSB000238	ADM000003
VSX453079	VSB005401	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005 1:00:00			VSB000238	ADM000003
VSX453078	VSX451886	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005			VSB000238	ADM000003
VSX453077	VSX451885	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005			VSB000238	ADM000003
VSX453076	VSB005401	VSX452133	VSX452135	26	1	0	EN	0.36	22.7.2005			VSB000238	ADM000003
VSX453075	VSX451886	EKF000023	VSX452134	26	1	0	EN	0	20.7.2005 13:00:00			VSB000238	ADM000003
VSX453074	VSX451885	EKF000023	VSX452134	26	1	0	EN	0	20.7.2005 13:00:00			VSB000238	ADM000003
VSX453073	VSB005401	EKF000023	VSX452134	26	1	0	EN	0	20.7.2005 13:00:00			VSB000238	ADM000003
VSX453072	VSX451886	EKF000023	VSX452134	26	1	0	EN	0	20.7.2005 14:00:00			VSB000238	ADM000003
VSX453071	VSX451885	EKF000023	VSX452134	26	1	0	EN	0	20.7.2005 14:00:00			VSB000238	ADM000003
VSX453070	VSB005401	EKF000023	VSX452134	26	1	0	EN	0	20.7.2005 14:00:00			VSB000238	ADM000003
VSX453069	VSX451886	EKF000023	VSX452134	26	1	0	EN	0	20.7.2005 15:00:00			VSB000238	ADM000003

Obr. 23. Pohled na tabulku STAR obsahující evidované údaje

Pokud bychom měli shrnout hlavní charakteristiky datového skladu 3.typu, je možné konstatovat, že tento typ skladu zpravidla:

- neobsahuje integrovaná a sjednocená data
- neobsahuje agregovaná data
- neslouží přímo pro analýzy, ale pro uložení dat
- díky předchozím bodům nemusí obsahovat pouze číselná data , ale je možné ukládat i např. text

Z pohledu možnosti ukládat text lze konstatovat, že takové řešení přináší určité výhody. Může realizovat originální zápis čísel - např. zápis „2,00“ určuje přesnost na 2 desetinná místa, nebo zápis „10-15“ jako výsledek stanovení. Námitka, že pak není možné uplatnit integritních omezení je lichá, protože ta již stejně nelze snadno realizovat díky použití 1 sloupce pro uložení hodnot různých atributů. Dále je možné strukturu použít pro ukládání jazykových variant textů (včetně kódování) - viz obr 24, kde je ve sloupci IDLANG uveden identifikátor příslušného jazyka, platného pro text ve sloupci VALUE. Ukládání jazykových verzí musí být samozřejmě řešeno správně i s ohledem na používanou znakovou sadu.

STAR : Tabulka

IDVAL	IDOBJECT	IDATTRIB	IDMETHOD	IDUNIT	IDQUA	IDC	IDLANG	VALUE
VSB002699	VSB000207	VSB002309	ADM000000	12	1	0	CZ	Správa Chráněné krajinné oblasti Jeseníky
VSB002700	VSB000218	VSB002309	ADM000000	12	1	0	GR	Ένωση Γεωργικών Συνεταιρισμών Δράμας
VSB002701	VSB000208	VSB002309	ADM000000	12	1	0	CZ	Agentura ochrany přírody a krajiny, středisko Olomc
VSB002711	VSB000219	VSB002309	ADM000000	12	1	0	GR	Περιφέρεια Α.Μ.Θ.
VSB002712	VSB000220	VSB002309	ADM000000	12	1	0	GR	ΔΕΥΑ Καβάλας
VSB002713	VSB000221	VSB002309	ADM000000	12	1	0	GR	Νομαρχιακή Αυτοδιοίκηση Δράμας
VSB002714	VSB000222	VSB002309	ADM000000	12	1	0	GR	Δημοτική Επιχείρηση Ύδρευσης-Αποχέτευσης Ξάνθ
VSB002731	VSB000088	VSB002309	ADM000000	12	1	0	GE	Bayerisches Landesamt für Umweltschutz
VSB002732	VSB000089	VSB002309	ADM000000	12	1	0	GE	Bayerisches Landesamt für Wasserwirtschaft
VSB002733	VSB000090	VSB002309	ADM000000	12	1	0	GE	Bayerisches Geologisches Landesamt
VSB002734	VSB000091	VSB002309	ADM000000	12	1	0	GE	Wasserwirtschaftsamt Amberg
VSB002735	VSB000092	VSB002309	ADM000000	12	1	0	GE	Wasserwirtschaftsamt Deggendorf
VSB002736	VSB000324	VSB002309	ADM000000	12	1	0	CZ	Obecní úřad Běšiny
VSB002737	VSB000325	VSB002309	ADM000000	12	1	0	SP	Instituto Nacional de Meteorología
VSB002738	VSB000326	VSB002309	ADM000000	12	1	0	GE	Wasserwirtschaftsamt Regensburg
VSB002739	VSB000327	VSB002309	ADM000000	12	1	0	GE	Wasserwirtschaftsamt Weiden
VSB002740	VSB000328	VSB002309	ADM000000	12	1	0	GE	Bund Naturschutz Bayern e.V. Kreisgruppe Cham
VSB002747	VSB000335	VSB002309	ADM000000	12	1	0	GE	EUROPARC Föderation
VSB002748	VSB000336	VSB002309	ADM000000	12	1	0	GE	Umweltbüro Regen
VSB002749	VSB000376	VSB002309	ADM000000	12	1	0	PT	Direcção Regional do Ambiente e Ordenamento do
VSB002750	VSB000379	VSB002309	ADM000000	12	1	0	SP	Instituto Geológico y Minero de España
VSB002751	VSB000380	VSB002309	ADM000000	12	1	0	SP	Consejo Nacional del Agua (CNA)
VSB002752	VSB000381	VSB002309	ADM000000	12	1	0	SP	Consejo del Agua de la Cuenca
VSB002753	VSB000083	VSB002309	ADM000000	12	1	0	GE	BUND Bayern, Kreisgruppe Freyung/Grafenau
VSB002754	VSB000085	VSB002309	ADM000000	12	1	0	EN	IFB Eigenschenk GmbH
VSB002755	ADM000013	VSB002309	ADM000000	12	1	0	PL	VŠB - Techniczny universytet
VSB002756	VSB000099	VSB002309	ADM000000	12	1	0	PL	Starostwo Powiatowe Nysa; Wydział Rolnictwa i Oc
VSB002757	VSB000100	VSB002309	ADM000000	12	1	0	PL	Urząd Miejski w Nysie
VSB002758	VSB000102	VSB002309	ADM000000	12	1	0	PL	Opolski Urząd Wojewódzki, Wydział Środwiska i R
VSB002759	VSB000103	VSB002309	ADM000000	12	1	0	PL	Regionalny Zarząd Gospodarki Wodnej we Wrocław
VSB002760	VSB000104	VSB002309	ADM000000	12	1	0	PL	Wojewódzki Zarząd Melioracji i Urządzeń Wodnych

Obr. 23. Pohled na tabulku Star obsahující jazykové mutace názvů organizací

Bohužel se ukazuje, že uživatel není nijak motivován zapisovat celou řadu metadat a tak se snaží tuto práci zjednodušit. V tomto směru musíme konstatovat, že plné využívání těchto možností přinese pravděpodobně až podstatné rozšíření automatizovaného záznamu, ukládání a přenosu metadat s každými pořizovanými či manipulovanými daty.

Závěrem je možné konstatovat, že na základě praktických zkušeností multidimenzionální databáze operativních dat je značně těžkopádná pro analýzy a dotazy. Pro tyto účely doporučujeme generovat sekundární tabulky s odvozenými informacemi.

6 Závěr

Frekvetnovaný pojem datový sklad se dnes používá v různém kontextu. Principielně je možné rozlišit několik variant chápání datového skladu, 3 z nich zahrnují:

- datové sklady ve smyslu organizovaného, jednotného a integrovaného úložiště dat,
- datové sklady analytické, typu data warehouse, který integrovaná data ukládá do multidimenzionální struktury, optimalizované pro dotazování a analýzy,
- datové sklady pro ukládání neupravených, původních dat s plnými metadaty v multidimenzionální struktuře.

Datový sklad 1.typu je značně frekventován v budovaných velkých GIS projektech ať již zaměřených na podporu veřejné správy či na činnost jednotlivých organizací.

Filosofie datového skladu typu data warehouse je postavena na multidimnezionálním konceptu, kdy jsou integrovaná a homogenizovaná data ukládána do datové struktury typu hvězda, sněhová vločka, případně hyperkostka. V takové struktuře je nutné navrhnout adekvátní dimenzionální tabulky, správně zvolit úrovně hierarchiea granulity ukládaných fakt, vybrat explicitní či implicitné realizaci hierarchie, samozřejmě i vhodně organizovat tabulyk faktů.

Návrh datového skladu pro vybraná socioekonomická data z oblasti trhu práce dokumentuje stávající situaci a popisuje návrh řešení se 3 dimenzemi a implicitní hierarchií geografické dimenze. Je doporučeno řešení, kdy je k lokalizaci použito pouze geokódů a časový vývoj se zaznamenává pomocí skladebnosti menších územních jednotek a vyznačení čísla verze a data platnosti.

Základní analytické možnosti OLAP jsou dokumentovány na jednoduchém příkladu v prostředí SPSS.

Datový sklad 3.typu je určen pro ukládání původních, transakčních dat do multidimenzionální datové struktury. Motivací k takovému jednání je snaha postihnout plný kontext (metadata) vázaný k jednotlivým, cenným údajům. Sémantické problémy s definicí atributu mohou být obecně spojeny s faktorem času, geografickými aspekty (rozloha, národní specifika), resp. metodami pořízení či získání dat, což je dokumentováno na příkladu.

Praktická realizace takové struktry byla provedena v databázi Cassiopeia pro projekt TRANSCAT. Teoretické výhody takové struktury jsou v praxi zastíněny neochotou pořizovat neautomatizovaně velké množství metadat a možným geometrickým nárůstem objemu dat.

Literatura

1. Bukáček R.: Plánované síťové geoinformační služby. *Konference Geoinformatika ve veřejné správě.* Brno 2006.
2. Corbley, K.: V Evropě nejrozsáhlejší GIS pro užití v evidenci nemovitostí: Německá dráha vede prostorová a atribuční data v jediném systému. VÚGTK, 1999. http://www.vugtk.cz/nzk/c4-99/corbley.htm
3. Dohnal, J., Pour, J. *Architektury informačních systémů*. Ekopress, 1997, 301 s., ISBN 80-86119-02-5.
4. Grof J., Weinberg P. *SQL kompletní průvodce.* ComputerPress 2004. ISBN 80-251-0369-2.
5. Horák, J. Projekt Implementace nástrojů prostorové analýzy trhu práce v činnosti úřadů práce. *Geoinfo ročník 8, 2001, číslo 4.* ISSN 1212-4311.
6. Horák, J. Implementace geoinformačních nástrojů v činnosti úřadů práce. *Konference GIS Ostrava 2002.* Ostrava, 2002, ISSN 1213-2454. <http://gis.vsb.cz/Publikace/Sborniky/GIS_Ova/GIS_Ova_2002/Sbornik/Referaty/horak2.htm>

7. Horak J., Unucka J., Stromsky J., Marsik V., Orlik A. TRANSCAT DSS architecture and modelling services. *Control & Cybernetics, vol 35, No.1 (Geographic Information Systems and Decision Support: New Approaches and Applications).* Varšava, Polsko 2006.
8. Horák J., Unucka J., Stromský J., Orlík A.. Příspěvek projektu TRANSCAT pro integrovaný management povodí v pohraničních oblastech. *Konference Hydrologické dni 2005.* Bratislava, Slovensko, 2005.
9. Humphries M. a kol. *Data warehousing – návrh a implementace.* Computer Press, 2002. ISBN 80-7226-560-1.
10. Kimball, R. Fact Tables and Dimension Tables. 2003. http://www.intelligententerprise.com//030101/602warehouse1_1.jhtml
11. Kouba Z. Datové sklady. Dobývání znalostí z databází 2000. *Sborník přednášek, FIS VŠE Praha*. Praha 2000.
12. Kružík, P. Liniový referenční systém v Geodatabase. *13.konference uživatelů GIS ESRI a Leica Geosystems.* Praha, 2004.
13. Kunz J. *Tvorba multidimenzionální databáze.* Semestrální práce. Ostrava 2006.
14. Mansfeld V. Datový sklad IDC ÚHÚL v informačním systému LH. *Lesnická práce č. 09.* 2003. http://lesprace.silvarium.cz/content/view/478/
15. Maršík V. Využití standardů OpenGIS při návrhu architektury GIS Krajského úřadu. *Konference GIS Ostrava 2004.* Ostrava 2004.
16. Mlčoušek M., Fryml J., Šach F. Datový sklad IDC ÚHÚL Brandýs nad Labem - poskytování dat o lese prostřednictvím internetových a mobilních technologií. *Konference GIS Ostrava 2004.* Ostrava 2004.
17. Moore, Tindall, Bech. The HARMONIRIB database functional specification. *Requirements report.* Dánsko, 2003. http://www.harmonirib.com.
18. Notes CZ. http://www.notes.cz/NotesWebsite.nsf/($OpenByAlias)/MetainformacniSystemPripadovaStudie?OpenDocument&Hierarchy=Sluzby. 2003
19. Orlík A., Růžička J., Stromský J., Děrgel P., Kamler J. Správa časoprostorových dat v prostředí PostgreSQL. http://gisak.vsb.cz/gportal/modules.php?name=News&file=article&sid=19 . 2005
20. Pirkl D. Tvorba datových skladů – pohled zevnitř. *Konference Datakon.* Brno 2004. http://www.datakon.cz/datakon04/d04_it_pirkl.pdf
21. Plšek V. Víceúrovňový datový sklad a jeho využití v GIS. *12.konference uživatelů GIS ESRI a Leica Geosystems v ČR.* Praha 2003.
22. Plšek V. 3D data pro 3D GIS, komplexní datový sklad a možnosti jeho využití ve státní správě i v soukromých organizacích. *13.konference uživatelů GIS ESRI a Leica Geosystems v ČR.* Praha 2004.
23. Pokorný, J. Konstrukce databázových systémů. Vysokoškolská skripta ČVUT, Praha 2004.
24. Rapant P. *Geoinformační technologie.* Vysokoškolská skripta. VŠB – TU Ostrava. Ostrava, 2005. http://gis.vsb.cz/publikace/Skripta_sylaby/u_git/GIT2005.pdf.
25. Vítek: *Datové sklady a OLAP*. 2002. http://datamining.xf.cz/view.php?cisloclanku=2002102808.

Annotation:

Data warehouses and an application of star data structure for spatial data

This paper distinguishes three kinds of data stores – integrated data store, data warehouses with a multidimensional data structure, and data warehouse with transaction data in the multidimensional data structure. Introdution to the theory of data warehouses. Examples of an application of star data structure for spatial socioeconomic data as well as hydrological data. Advantages, disadvantages and possibilities of the models are discussed.