VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Vizualizácia dát

Ing. Ladislav Ruttkay

17.12.2007

Anotácia

Hlavným predmetom práce je vizualizácia dát. Vo svojom úvode však oboznamuje čitateľa so základnými pojmami samotnej reprezentácie dát, dátových modelov a používaných techník. Pre celkový obraz nevynecháva základné popisy a definície. V závere pojednáva o trendoch v danej oblasti v budúcnosti.

Obsah

1. Úvod

Súčasnosť niekedy nazývame ako *informačnú revolúciu*. Je to pomenovanie však na mieste? Nejedná sa skôr o *dátovú explóziu!?* Je totiž zásadný rozdiel medzi množstvom nahromadených dát, a množstvom informácií, ktoré z toho môžeme získať. Ideál, ktorý sa však snažíme dosiahnuť je mať správnu informáciu na správnom mieste a v správny čas. Práve technológie, ktoré sa k tomu približujú, tak získavajú široké praktické uplatnenie.

V tejto oblasti sa stretávame s pojmami ako OLAP, dolovanie dát, získavanie znalostí z databáz, ... a pod. Tieto technológie nám pomáhajú určovať trendy, výsledky, štatistiky,... uplatniteľné v iných odvetviach a tým ich ekonomický dopad na správnosť rozhodnutí je len ťažko oceniteľný.

Stále však podstatné a strategické rozhodnutie ostáva na užívateľovi, ktorý na základe informačnej podpory učiní rozhodnutie. A práve tu sa dostávame k častej téme komunikácie človeka s počítačom. Ako dokážeme prehľadne zobraziť predaj 100 produktov v 50 krajinách sveta, ak chceme získať prehľad za rok, mesiac, týždeň,...? Ako efektívne komunikovať s počítačom v takomto množstve dát. Rozborom týchto metód zobrazenia a interakcie, taktiež plánmi a očakávaniami sa bude zaoberať práve táto práca.

1.1 Business inteligence

Každá organizácia dnes disponuje obrovským množstvom dát, ktorých zdrojom bolo buď ručné, alebo automatické získanie. Tieto dáta môžu v prenesenom význame tvoriť určitú „pamäť“ organizácie. Podobne, ako každý inteligentný jedinec má schopnosť na základe zážitkov optimalizovať svoje chovanie a poučiť sa z chýb, tak aj organizácie sa snažia využiť práve *Business Inteligence*.

Riešenia Business inteligencie transformujú dáta do použiteľných informácií, podstatných pre rozhodovanie. Umožňujú získanie prehľadu o trendoch, lojalite zákazníkov, dodávateľov, znižujú finančné náklady, umožňujú nachádzanie nových možností predaja,... atď. Prínosom je, že sa dokážu pozerať na informácie z viacerých pohľadov a dimenzií. Riešenie Business Inteligencie odpovedá na otázku „Čo sa stane, ak ... ?“ a nie na otázku „Čo sa stalo ... ?“ [1]

Zložky „pamäte“ organizácie :

- **Pamäť relatívne dlhodobá** - zo zákona je nutné niektoré transakcie archivovať niekoľko rokov
- **Pamäť ťažko dostupná** – OLTP systémy sú optimalizované na spracovanie transakcií, ale prístup k historickým dátam je veľmi náročný
- **Pamäť nekonzistentná** – je typické že behom „života“ IS dochádza k zmenám spôsobu vyjadrenie identických skutočností v OLTP systéme, napr. výmenou technológie.
- **Pamäť s exponenciálne rastúcim objemom uložených dát** – odhaduje sa, že celosvetovo sa objem OLTP dát zdvojnásobí každých 18 až 24 mesiacov.

1.2 On-Line Analytical Processing

On-Line Analytical Processing (OLAP) je softvérová technológia, ktorá umožňuje analytikom a manažérom, získanie rýchleho, konzistentného, interaktívneho pohľadu na informácie, ktoré boli transformáciami získané z dát, aby zobrazovali skutočne dimenzie pochopiteľné užívateľom.[3]

Cieľom OLAP nástrojov je poskytnúť multidimenzionálne náhľady na dáta ktorá sa za nimi nachádzajú. Aby bol tento cieľ dosiahnutý, tak tieto nástroje používajú multidimenzionálne modely na ukladanie a zobrazenie dát. Dáta sú uložené v kockách, alebo hyperkockách, ktoré sú definované v multidimenzionálnom priestore. Každá dimenzia v sebe zahrnuje istú množinu agregačných úrovní. Typické OLAP operácie v sebe obsahujú agregáciu a deagregáciu informácií (*roll-up* a *drill-down*) naprieč dimenziou. Taktiež medzi používané operácie patrí vybranie špecifických častí kocky, zmenu orientácie multidimenzionálneho pohľadu na dáta na obrazovke (*pivoting*).

V posledných rokoch práve OLAP a dátové sklady sa stali hlavným predmetom výskumu v databázovej oblasti. [4] Jednou zo základných problematík, s ktorou sa stretáva vývoj a výskum OLAP, je modelovanie dát. Niekoľko modelov bolo adaptovaných a použitých v praxi a vo výskume boli študované ďalšie. Všetky však zdieľajú niekoľko základných konceptov, ako napr. dimenzie a dimenzionálne hierarchie. Napriek tomu ale stále nie je formálne definovaný a široko akceptovaný, logický alebo konceptuálny jednotný model dát.

1.2.1 Pravidlá OLAP

Existuje 12 základných pravidiel OLAP, ktoré sformuloval Dr. E. F. Codd [6]. 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 *roll-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.

1.3 Získavane znalostí z databáz

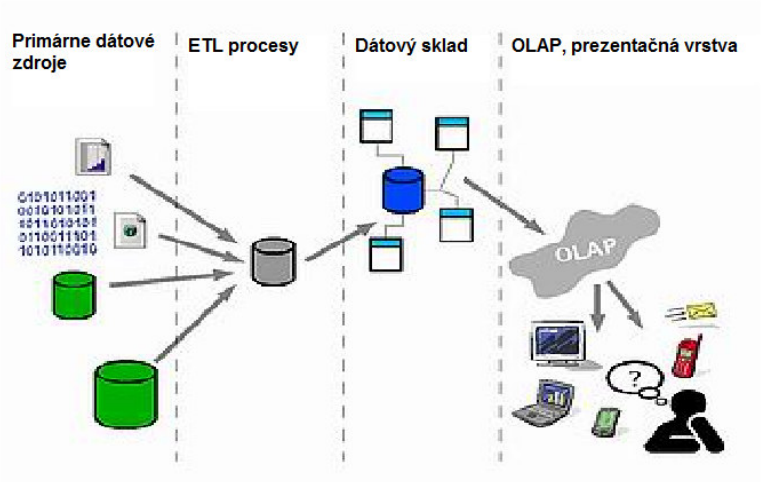
Ak môžeme OLTP systém prirovnať k nervovému systému organizácie, dátový sklad potom predstavuje jej „pamäť“. Aj geniálny OLTP systém a obrovská „pamäť“ však nestačia organizácií na to, aby bola schopná optimalizovať svoje procesy a „poučiť“ sa zo znalostí. Potrebujeme mať také nástroje, ktoré by umožnili objaviť v dátach informácie, ktoré pomôžu pri rozhodovaní pracovníkom na rôznych úrovniach riadenia. Hovoríme o *získavaní znalostí z databáz (Knowledge Discovery in Databases - KDD).*

KDD je medziodborová disciplína, ktorá využíva štatistické metód, vizualizáciu dát, expertné systémy, strojové učenie a taktiež nástroje pre On-Line Analytical Processing (OLAP). KDD je možné chápať ako proces netriviálneho objavovaní implicitných, dopredu neznámych a potenciálne použiteľných informácií v dátach. [2]

Celý proces KDD delíme na niekoľko časti v ktorých dochádza k používaní rôznych technológií.

- Výber dát
- Príprava dát – čistenie, transformácie,..
- Objavovane znalostí
- Prezentácia znalostí – podstatná práve vizualizácie, ktorej sa budeme hlbšie venovať.

1.4 Architektúra systému



Obrázok 1 Architektúra Business Inteligence systému

Architektúru systému Business inteligence, je možné rozdeliť na niekoľko častí ako vidíme na Obrázku 1. Svoju podstatnú rolu zohráva subsystém ETL *(extraction, transformation, loading)*, ktorý slúži k extrakcií dát z primárnych systémov, k ich úpravám, kontrole a zisteniu kvality. ETL systém následne tieto dáta prevedie do špecializovaného úložiska, optimalizovaného pre efektívne analýzy a rýchle poskytnutie dát vizualizačnej vrstve. Tá potom zaistí zobrazenie dát a interakciu s užívateľom.

Takto navrhnutý systém môže byť technicky realizovaný rôznymi spôsobmi a technológiami. Pre malé riešenia sú všetky subsystémy implementovaná ako jedna aplikácia, určená k inštalácii na jednom počítači (Desktop OLAP). Výhoda je ich jednoduchosť inštalácie a nekomplikovaná implementácia elementárnych analýz. Problémom však je zaistenie ďalšieho rozvoja systému, rastu dát, pridávaní analýz a pod. Preto danú technológiu vo veľkých projektoch nenájdeme. Väčšie systémy, využívajú výhradne viac vrstvú architektúru s dátovým skladom, ako centrálnym úložiskom dát.

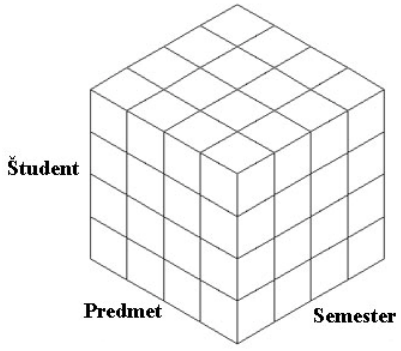
2. Multidimenzionálny databázový model

V reprezentácií dát sa stretávame často s problémom, nie neznámym taktiež v iných vedných kategóriách. Jedná sa o praktické nasadenie, ktoré môže mať iné podmienky od akademického výskumného prostredia. Z užívateľského hľadiska je zrejme, že nie sme ochotný čakať na odozvu programu a vyhodnotenie niekoľko minút a takáto aplikácia preto často nenachádza reálne uplatnenie. Z tohto dôvodu musia byť dáta ktoré chceme vizualizovať, čo najkompletnejšie podporované samotným dátovým modelom, ktorý umožní ich rýchle načítanie. Z tohto dôvodu ak chceme zobrazovať OLAP dáta multidimenzionálne tak je nemysliteľné aby to bolo vybudované nad relačným databázovým modelom, ale používa sa multidiemnzionálny databázový model. Preto sa základom tohto modelu musíme venovať aby bolo zrejme aké možnosti ponúka, ktoré môžu byť následne vizualizované.

Multidimenzionálna databáza je teda typ databázy, ktorý je optimalizovaný pre dátové sklady (*Data Warehouse)* a OLAP 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.

2.1 Hyperkocka

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 (nie len 3 ale aj niekoľko desiatok). Príkladom trojdimenzionálneho modelu môže byť hyperkocka s dimenziami *Študent, Predmet, Semester:*



Obrázok 2 Hyperkocka

2.2 Základné pojmy multidimenzionálneho databázového modelu

2.2.1 Bázová dimenzia

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

2.2.2 Ortogonalita dimenzií

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.

2.2.3 Merateľná dimenzia

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.

2.2.4 Odvodené meradlo

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

2.2.5 Časová dimenzia

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.

2.2.6 Subdimenzia

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

2.2.7 Agregovaná dimenzia

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

2.2.8 Konsolidačná dimenzia

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

2.2.9 Multidimenzionálna databáza

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 ($m>>n$).

2.3 Fakty a Dimenzia

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 údajov:

- **Fakty** - numerické merné jednotky obchodovania. Prvotné fakty sa môžu kombinovať alebo vypočítať pomocou iných faktov a vytvoriť tak merné jednotky.
- **Tabuľka 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 členmi *(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 hierarchickú štruktúru, napr.:

Čas: rok, kvartál, mesiac, týždeň, deň,...

Škola: vysoká škola, fakulta, odbor, ročník, skupina,...

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

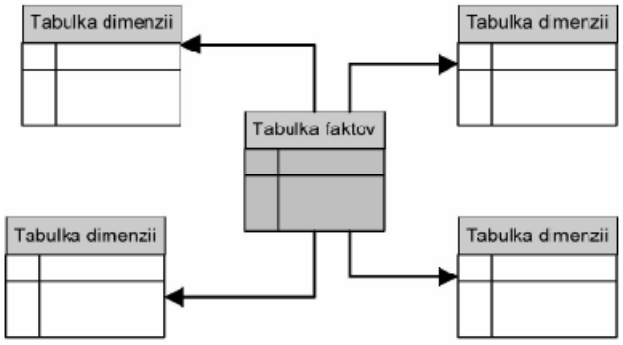
2.4 Schémy tabuliek dimenzií

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

- Hviezdicové schéma – *star schema*
- Schéma snehovej vločky – *snowflake schema*

2.4.1 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.

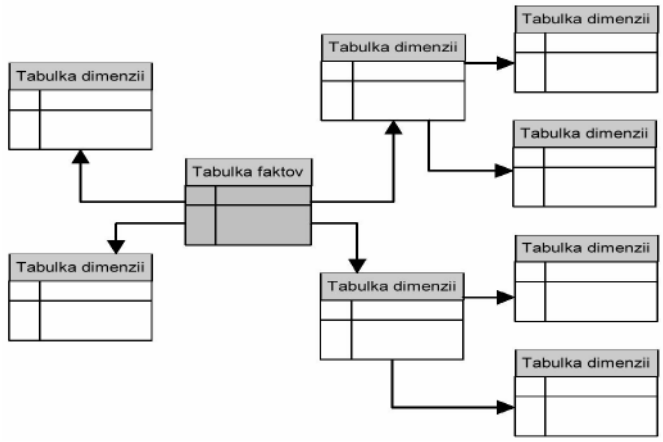


Obrázok 3 Hviezdicová schéma

2.4.2 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 normalizova-

ných tabuliek, ale má nižší dotazovací výkon, lebo obsahuje veľké množstvo spojených tabuliek.



Obrázok 4 Schéma snehovej vločky

2.5 Porovnanie relačného a multidimenzionálneho modelu

2.5.1 Relačný model

Entitno-relačný model ako prvý navrhol v roku 1970 E.F.Codd pracujúci pre IBM. Táto technika sa stala veľmi populárna a osvedčená pri uchovávaní obrovského množstva organizovaných dát s veľkou transakčnou odozvou. Prístup k dátam ale môže vyžadovať zložité spojenia (*join*) viacerých tabuliek a agregácie, čo však je netriviálne pre koncového užívateľa, ktorý často musí požiadať o odbornú pomoc.

Podstatné je si uvedomiť, že nie všetky relačné databázy musí byť vhodné konvertovať na multidimenzionálne. Ak sa jedná napr. o databázu *Meno/Priezvisko/Dátum narodenia*, tak takáto multidimenzionálna databáza by obsahovala riedke záznamy. V relačnej schéme by boli nutné len 3 porovnania ale v multidimenzionálnej až 9 (3x3). Multidimenzionálne databázy sú navrhnuté za účelom manipulácie a analýzy komplexných databázových štruktúr s veľkým množstvom dát a relácií.

2.5.1.1 Výhody relačných databáz

- 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

2.5.1.2 Nevýhody relačných databáz

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

2.5.2 Multidimenzionálny model

Medzi hlavné plusy tohto modlu patrí hlavne zlepšená prezentácia a navigácia v dátach. Taktiež jednoduchá údržba, keďže dáta sú uložené rovnakým spôsobom ako sú prezentované a nie sú vyžadované ďalšie výpočty a zložité dotazy.

2.5.2.1 Výhody multidimenzionálnych databáz

- rýchly a 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

2.5.2.2 Nevýhody multidimenzionálnych databáz

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

2.5.3 Porovnanie charakteristík

	Relačné databázy	**Multidimenzionálne databázy**
Množina	Entita	Bázová dimenzia
Popis	Atribúty	Agregovaná dimenzia
Dimenzionalita	Dvojdimenzionálna tabuľka	Viacrozmerná štruktúra
Čas	Náročná manipulácia	Elementárna manipulácia
Modelovanie	Normalizované ERD, DFD	*Star, Snowflake* schéma
Prístup k dátum	SQL	MDX, variácie SQL

Tabuľka 1 Porovnanie charakteristík databáz

3. Vizualizácia

Každý deň sa stretávame s vizualizovanými dátami v rôznych podobách v novinách, časopisoch, televíznych správach, predpovedí počasí,...a pod. Vidíme stĺpcové grafy, porovnávajúce rast HDP, čiarové grafy ako ukazujú vývoj teploty za posledné sledované obdobie, a takto by sme mohli uviesť veľké množstvo ďalších príkladov. Položme si základnú otázku: Prečo používame vizualizáciu? Vizualizácia pomáha pochopeniu dát, je bližšie k ľudskému vyjadrovaniu a poskytuje nám ucelený obraz na dáta. Je dokázané, že človek najviac vníma a zapamätá si obrazové dáta. Niekoľko tabuliek čistých dát dokážeme vyjadriť pomocou často len jedného grafu, z ktorého môžeme zreteľne vidieť určité informácie bez zbytočných detailov, ktoré nás v tom momente nemusia zaujímať.

Vizualizácia ma svoje nesporné miesto v *Business Inteligece*, pretože dokáže zobraziť dáta a ich prezentovať tak, že je možné z nich určiť predtým neobjavené závislosti, predpovedať vývoj a určiť ďalšie možnosti a smerovanie firmy.

Ak hovoríme o vizualizácií, tak musíme rozlišovať dva základné od seba odlišné pojmi, ktorými sú: *Vizualizácie dát* a *Vizuálne dolovanie dát.*

3.1 Vizualizácia dát

Dátové vizualizačné nástroje a postupy nám pomáhajú vytvárať dvoj a trojrozmerné obrazy dát, z ktorých môžu byť jednoduchšie interpretované znalosti vyplývajúce z týchto dát. Pomocou skúmania zobrazených dát, môžeme odhaliť zaujímavé, netriviálne, pôvodné nevidené a potenciálne použiteľné informácie z dát.

3.1.1 Dátové typy

Ak hovoríme o dátach obecne, tak ich môžem zaradiť do určitej *domény*. Doména predstavuje abstraktnú kategóriu pre dátovú jednotku. Ak budem napr. uvažovať nad údajom: *Ján Novák 3.12.1983,* tak *Ján* je v doméne *meno, Novák* v doméne *priezvisko* a *3.12.1983* je v doméne *dátum narodenia.* Je zrejme, že práve domény budú určovať názov stĺpcov v tabuľkách v ktorých sa budú nachádzať dáta. Predpokladáme že v jednej doméne sa budú nachádzať dáta významovo podobné, teda nevložíme *Novák* do domény *dátum narodenia*. Z tohto môžeme odvodiť, že dáta v doméne budú rovnakého typu, hovoríme, že doména je istého typu. Pod týmto pojmom rozumieme typy napr.: *reťazec, čas, celé číslo, ...*

Typy domén delíme na *diskrétne* a *spojité*, podľa toho, aké hodnoty zaznamenávajú. Domény s diskrétnym typom sú také, ktoré môžu obsahovať iba konečné množstvo rôznych

hodnôt. Sú to samozrejme domény označujúce napr. deň v týždni alebo pohlavie, o ktorých môžeme povedať že sú v podstate výčtového typu. Taktiež diskrétne sú ale typy *celé čísla* a *reťazce*, aj keď ich potenciálny počet je nekonečný, prípadne ich dĺžka je nekonečná. Je nutné však uviesť, že ich hodnoty sú navzájom nespojité a medzi dvoma po sebe nasledujúcim hodnotami sa už žiadne ďalšie nenachádzajú. Práve táto skutočnosť ich radí medzi diskrétne typy.

Domény spojitého typu obsahujú potenciálne nekonečné množstvo spojitých hodnôt. Medzi dvoma hodnotami stále dokážeme nájsť inú hodnotu , a takto by sme mohli pokračovať až do nekonečna. Ak uvedieme matematicky príklad, tak napr. pre reálne čísla platí, že medzi hodnotami *0* a*1* sa nachádza nekonečné množstvo reálnych čísel. Medzi spojité typy patria *reálne čísla, čas, teplota, vlhkosť,* a pod.

3.1.2 Vizualizačné a dátové dimenzie

Aby sme správne chápali vizuálizáciu a zobrazenia dát, tak musíme korektne rozlišovať dátové a a vizualizačné dimenzie.

Vizualizačné dimenzie predstavujú koordináty a vlastnosti zobrazovacieho priestoru. Ak budeme zobrazovať dvoj rozmerne, tak uvažujeme o osi *x* a *y,* teda dve vizualizačné dimenzie. V 3D priestore to bude analogicky s osou *z*. Medzi vlastnosti zobrazovacieho priestoru môže patriť taktiež zobrazovanie *farieb*, *jasu*, *veľkosti* zobrazovaných objektov. Týmto spôsobom dokážeme zobraziť aj viac ako 2, prípadne 3 dimenzie. Je však nutné si uvedomiť isté obmedzenia, ktorými sú množstvo okom rozlišovaných farieb, jasov, veľkostí. Ak by sme napr. farbou chceli zobraziť určitú hodnotu, tak uvedenie jednej skutočnosti v prípadnej legende by už komplikovalo prehľadnosť zobrazenia a taktiež by sme mohli použiť iba jednoznačne okom rozlíšiteľný počet farieb.

Dátové dimenzie predstavujú dátové domény, o ktorých sme už pojednávali, prípadne zložené dátové domény. Ak by sme totiž používali napr. doménu *meno* a *priezvisko,* tak by sme mohli tieto dve domény reprezentovať ako dimenziu celé *meno.* Je zrejme že dátové dimenzie, podobne ako domény, môžu byť spojité a diskrétne. Počet dátových dimenzií nie je nijak obmedzený. Vo vyhodnocovaní dát sa stretávame s interakciu domén z rôznych tabuliek, kde sa zameriavame za získanie nových vypočítaných, prípadne agregovaných hodnôt. Napr. výsledkom interakcie domény *pobočka firmy* a domény *produkty* by mohli byť tržby za konkrétny produkt v konkrétnej pobočke, ale taktiež aj počet predaných kusov. V chápaní dátových dimenzií nie je problém zaviesť interakciu s treťou a ďalšou dimenziou. V našom

príklade môžeme uvažovať napr. doménu *čas*. Takže získané výsledky by určovali počet konkrétnych predaných produktov v konkrétnej pobočke firmy za konkrétne časové obdobie.

Dátové dimenzie sú matematicky zrejme, a môžeme uvažovať o interakciách veľkého množstva domén, ktorých výsledkom budú reálne použiteľné dáta. Problémom však ostáva ich interpretácia, pretože dátové domény je nutne interpretovať prostredníctvom domén vizualizačných, ktorých počet je obmedzení ľudským chápaní. Tento problém možno označiť ako jeden zo základných problémom vizualizácie a množstvo vedeckých skupín, ale aj komerčných organizácií sa zaoberá práve možnosťami zobrazenia databázových domén do domén vizuálnych. Táto sféra výskumu je stále otvorené a ponúka nám nové a zaujímavé riešenia. O niektorých budeme hovoriť aj v tejto práci.

3.1.3 Nástroje vizualizácie dát

Vizualizačné nástroje sú určené k dvoj alebo trojdimenzionálnemu zobrazeniu dát. Niektoré techniky sa používajú už storočia, iné sa postupne vyvíjajú. Existujú aj nástroje, ktoré dokážu dynamicky odpovedať na interakciu a animovať aj väčšie množstvo dimenzií.

Nástroje vizualizácie dát môžeme klasifikovať do dvoch základných tried:

- Multidimenzionálne vizualizácie
- Hierarchické a mapové vizualizácie

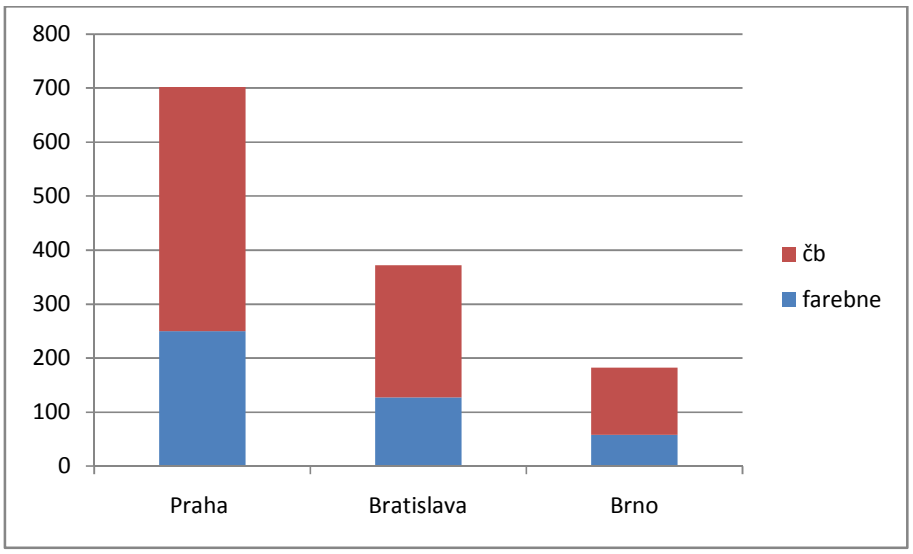
3.1.3.1 Multidimenzionálne vizualizačné nástroje

Multidimenzionálne vizualizačné nástroje umožňujú užívateľom vizuálne porovnať dátové dimenzie medzi sebou za použitia koordinátového systému. Taktiež umožňujú odhalenie závislosti medzi dvoma dátovými doménami, ak sa ich zobrazenia podobajú. Medzi tieto vizualizačné nástroje patria rôzne typy tabuliek, ich variácií a grafov, ako napr.: stlpcové, koláčové, histogramy, bodové, čiarové,...

Stĺpcový graf

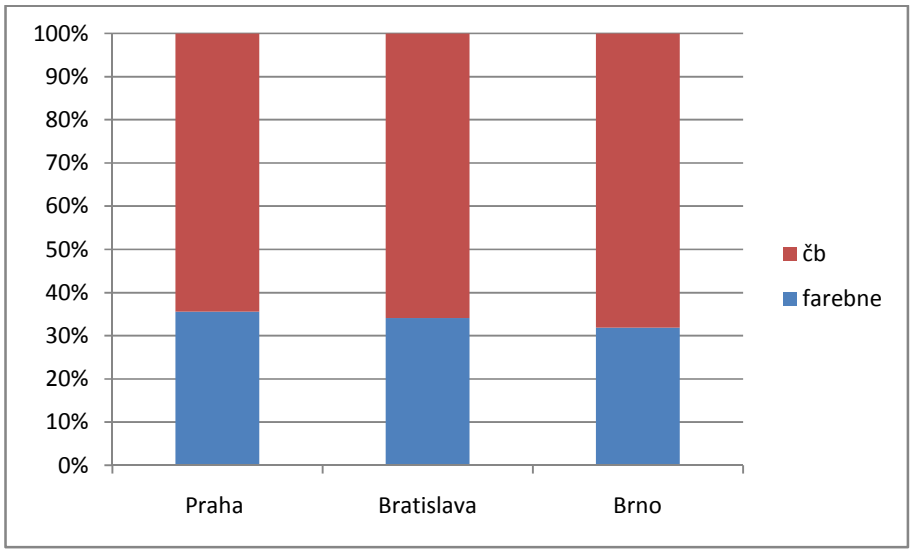
Tento typ grafu porovnáva spojitú dátovú dimenziu krížom cez diskrétnu dátovú dimenziu v dvoj rozmernom súradnom systéme. Stĺpcový graf svojim účelom je podobný spojnicovému grafu, ale ten môže porovnávať dve spojité dimenzie. Veľkosť daného stĺpca svojim obsahom vyjadruje veľkosť zobrazovanej hodnoty. Ak vyjadrenie dimenzií osami vymeníme, tak získame *graf pruhový*. Grafické vyjadrenie môže tiež nadobúdať rôzny vzor alebo farbu stĺpcov. Existujú variácie stĺpcových alebo riadkových grafov, ktoré dokážu zobraziť skutočnosť, že dáta v stĺpci sú delené do viacerých kategórií. Uveďme ako príklad stĺpcový

graf, ktorý zobrazuje počet vytlačených strán pre jednotlivé oddelenia firmy. Jednoduchý stĺpcový graf zobrazí iba počet strán ale ak vytvoríme *skladaný graf*, tak môžeme zobraziť kategórie: *farebne, čb.*



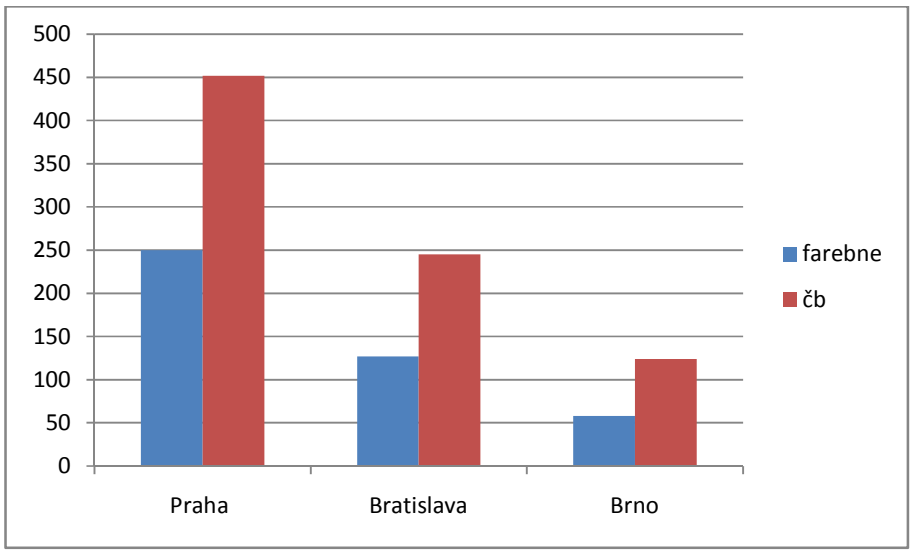
Obrázok 5 Skladaný graf

Ďalšou alternatívou v skladanom grafe je zobrazenie percentuálneho významu, kde bude prebiehať celou osou *y*, čo bude značiť 100%. Percentuálny pomer zložiek bude reprezentovaný pomerom obsahov zobrazených časti jedného stĺpca. Z tohto grafu nedokážeme určiť počet reprezentovaných jednotiek, ale iba porovnať ich percentuálne rozloženie.



Obrázok 6 100% Skladaný stĺpcový

Posledným, častou používaným variantom, je graf s*tĺpcový skupinový*, kedy hodnoty zložiek nie sú nad sebou v jednom stĺpci ale nachádzajú sa vedľa seba, v prípade 3D zobrazenia môžu byť za sebou.

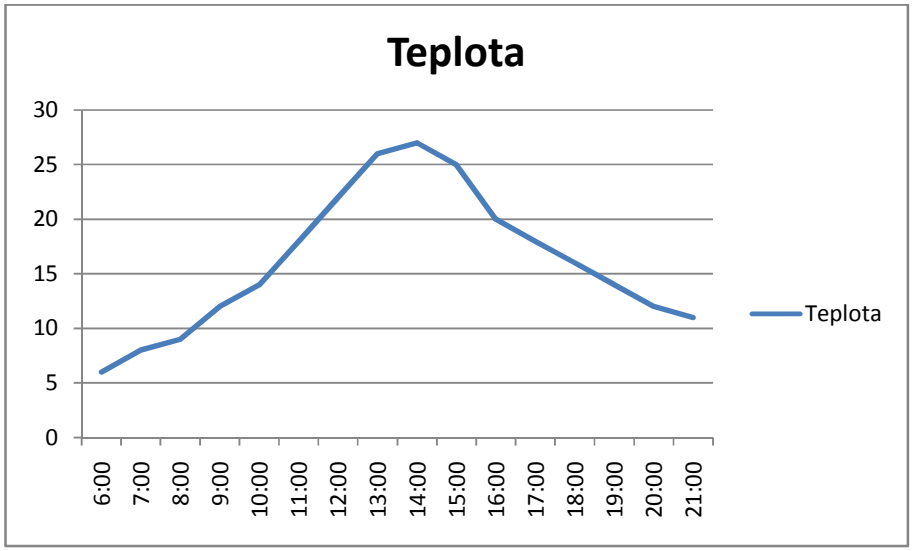


Obrázok 7 Skupinový stĺpcový

Zaujímavým použitím obyčajných stĺpcových grafov je práve vizualizácia dátového rozloženia. V tomto prípade hovoríme o tzv. *histograme.*

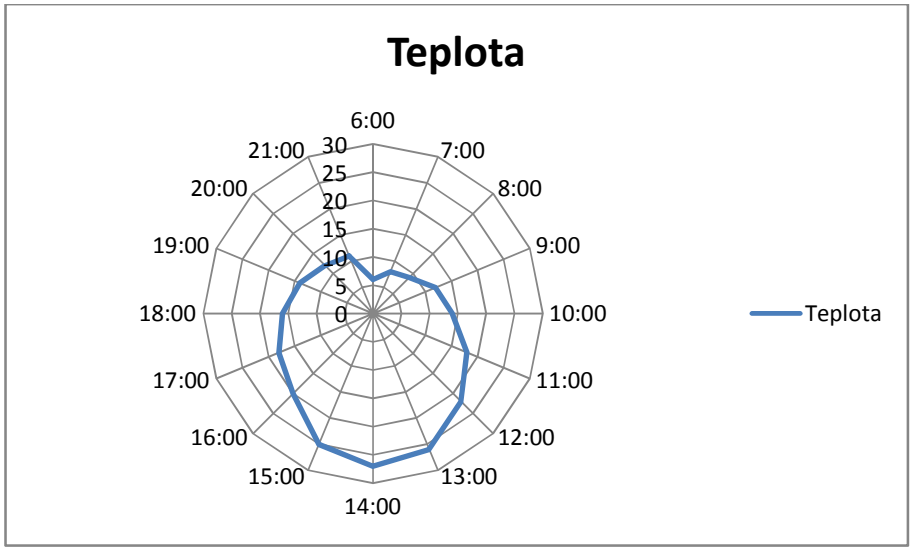
Spojnicový graf

Vo svojej najjednoduchšej forme je *spojnicový graf* skupinou dát zobrazených ako body na *x, y* súradnom systéme, podľa možností, spojených čiarami. Tento graf dokáže zobraziť interakciu medzi dvoma spojitými aj diskrétnymi dátovými doménami aj ich kombináciou. Často sa práve používa na zobrazenie zmeny istej domény v priebehu času, ale môžeme použiť aj dve nečasové dimenzie a zobraziť medzi nimi súvislosť. V podstate by sa jednalo o graf funkcie. Niektoré varianty zobrazujú spojnicu ako lomenú čiaru, ďalšie sa snažia zobraziť krivku, iné zvýrazňujú nosné body.



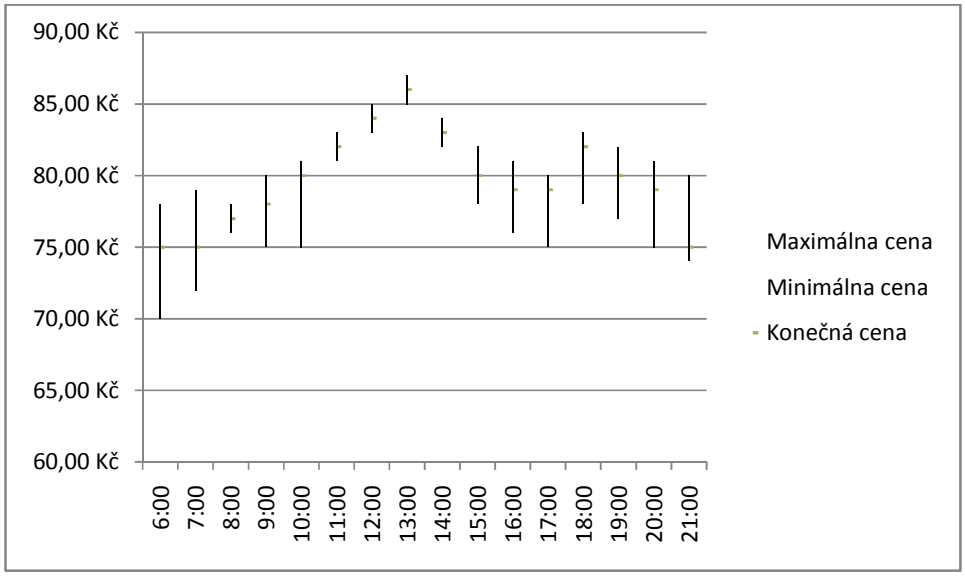
Obrázok 8 Spojnicový graf

Zaujímavú variantu tvorí paprskový, alebo *radiálny graf*, keď závislosť *x* a *y* nezobrazujeme v pravouhlom súradnom systéme, ale v 360 stupňovom systéme.



Obrázok 9 Radiálny spojnicový graf

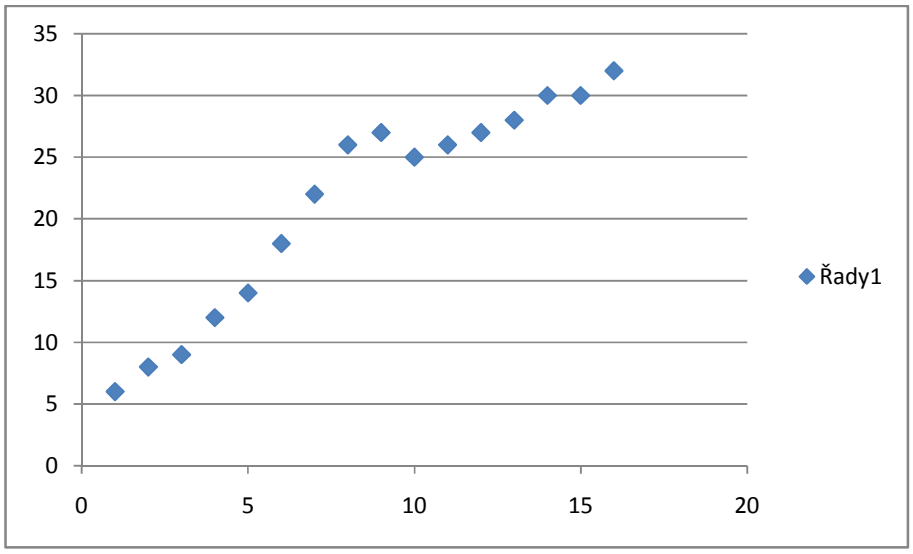
Novodobejšiu formu spojnicového grafu predstavujú *burzové grafy*, ktoré zobrazujú v časovom priebehu niekoľko skutočností, ako je napr.: maximálna cena, minimálna cena, výsledná cena. Takto je možné zobraziť vývoj akcií na burze a rozpätie hodnôt v ktorých sa daná cena nachádzala. Ich použitie sa však nezužuje iba na burzy.



Obrázok 10 Burzový graf

Bodový graf

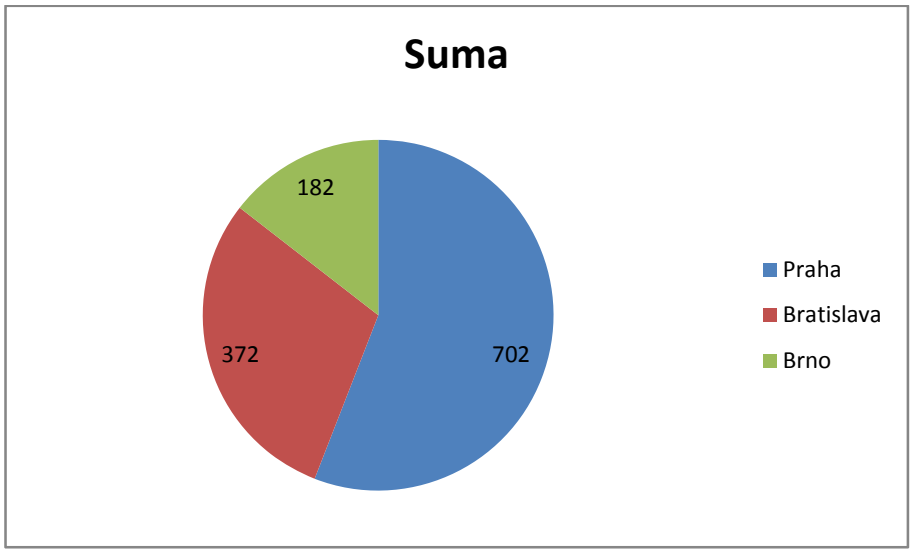
Bodové grafy sú typicky používané na porovnanie dvojíc hodnôt. Jedna hodnota je prezentovaná na *x*-ovej osy a druhá hodnota je reprezentovaná na *y*-ovej osy, ich priesečník je v grafe označený ako bod, prípadné iná zvolená značka. Je možné zaviesť aj zobrazenie tretej dátovej domény použitím osi *z*. Bodové grafy sa od spojnicových líšia tým, že priesečníky hodnôt nie sú spojené čiarou.



Obrázok 11 Bodový graf

Výsekový graf

Výsekové grafy zobrazujú príspevok častí interpretovaných dát do celku. Časti výsekového grafu môžu zobrazovať jednotkový podiel na celkovej sume ale taktiež percentuálny. V podstate na vizualizácií veľkostí jednotlivých časti nedôjde k zmene, zmenia sa len popisy hodnôt.



Obrázok 12 Výsekový graf

Tieto uvedené grafy predstavujú len základnú vzorku všetkých možností. Väčšie množstvo variácií a podrobnejší popis je možné nájsť v referencii [7].

Tabuľky

Je nutné spomenúť, že základným vizualizačným nástrojom je aj zobrazenie samotnej tabuľky. Síce obsahuje len dáta, grafickou úpravou formátu je možné získať zlepšenie jej prehľadnosti. Z jednoduchých tabuliek, ktoré zobrazujú len dve dimenzie, boli odvodené aj *kontingenčné tabuľky*, ktoré dokážu zobraziť väčšie množstvo dimenzií v 2D. Ich prehľadnosť a tým aj použiteľnosť je však otázna, a často neprináša očakávané, požadované, prehľadné výsledky.

3.1.3.2 Hierarchické a mapové vizualizácie

Hierarchické, mapové a iná špeciálne vizualizačné nástroje sa od klasických multidimenzionálnych odlišujú tým, že zdôrazňujú a používajú štruktúru modelu dát samotných. Ako príklad môžeme uviesť organizačné grafy alebo rodokmeň. Niektoré dáta obsahujú v sebe dedičnú hierarchickú štruktúru. V takomto prípade môže byť vizualizácia pomocou stromov

práve vhodnou možnosťou. Ďalšie dáta môžu byť úzko späté s krajinami a mapami a sú zobrazované geografickými mapovými grafmi.

Stromové grafy

Stromové grafy prezentujú dáta vo forme stromovej štruktúry. Každá úroveň grafu sa skladá s niekoľkých paralelných vetiev. Každý *uzol* v strome môže reprezentovať celý podstrom, ktorého je koreňom. Stromový graf zobrazuje kvantitatívne a hierarchické vzťahy medzi zobrazovanými dátami. Každý uzol môže obsahovať informácie vo forme čísel, stĺpcov, farieb a pod., o agregovaných dátových hodnotách, ako napr.: suma, priemer, počet. Čiary nazývané *hrany* spájajú uzly a ukazujú vzťahy medzi dátam.

Mapové vizualizácie

Tieto grafy sú práve vhodné na vizualizáciu prísne teritoriálne vymedzených dát, vzťahujúcich sa na istý geografický región. Príslušná doména dát je zobrazená na mape vo vymedzených regiónoch. Hodnoty pre dané oblasti sú zobrazené prostredníctvom čísel a farieb, alebo sú kombinované so stĺpcovými a inými grafy, čím je možné pre každú oblasť zobraziť väčšie množstvo dát, prípadne ich závislosť.

3.2 Vizuálne dolovanie dát

Nástroje vizuálneho dolovania dát, zobrazujú vizuálne model získavania znalostí dát z databáze a umožňujú náhľad do získavania informácií takéhoto algoritmu. Je podstatné, že týmto spôsobom je možné lepšie daný algoritmus nie len pochopiť, ale aj overiť, či informácie sú získavané zo správnych zdrojov. Taktiež tieto nástroje pomáhajú pochopiť a ohodnotiť dôležitosť vydolovaných dát a ich vnútornú podstatu.

Vstupom do nástroja pre vizualizáciu dát je model, ktorý je chápaný ako kolekcia generalizácií alebo vzorov, nachádzajúcich sa v dátach a predstavujúcich istú formu abstrakcie. Ľudia sa bežne dokážu poučiť zo svojich skúseností. Vyspelejšie nástroje získavania znalostí, taktiež môžu obsahovať rozhodovanie mechanizmy, ktoré môžu napodobňovať učenie. Výsledné získané informácie od takýchto nástrojov však je náročné overiť, prečo nástroj dospel k daným záverom. Práve vizualizácia však dokáže proces zobraziť a priblížiť, prečo došlo k danému rozhodnutiu. Túto možnosť však musí aj samotný nástroj pre dolovanie dát podporiť a uviesť jasné dôvody prečo došlo k danému rozhodnutiu. Niektoré nástroje totiž fungujú ako *black box* a dôvody svojich rozhodnutí neuvádzajú, takže ich nemôžeme ani vizualizovať.

Medzi vhodné príklady patrí vizualizácia *rozhodovacieho stromu* aplikácie pre dolovanie dát. Niektoré nástroje však používajú zložitejšie mechanizmy ako sú napr.: neurónové siete. Tie obsahujú veľké množstvo spojení jednoduchých procesných jednotiek, segmentovaných do vstupných, skrytých a výstupných úrovní. Zobrazenie takéhoto rozsiahleho modelu siete by bolo značne náročné a neprehľadné a je stále predmetom skúmania.

4. OLAP Vizualizačné modely

Vizualizácia patrí ku kardinálnym témam v oblasti databázového výskumu. OLAP, ako technológia pre podporu rozhodovania, je úzko spojený s oblasťou výskumu vizualizácie. V kontexte OLAP technológií, predstavuje vizualizácia dát technicky a postupy pre prezentáciu OLAP špecifických informácií koncovým užívateľom. Aj keď databázová komunita očakáva, že vizualizácia bude maz značnú dôležitosť v danej oblasti, a výskum dokázal vyvinúť techniky zobrazenia veľkého množstva dát, tak stále OLAP vizualizácie neboli zahrnuté do bežne používaných pokročilých vizualizačných techník.[8]

Napriek výskumu OLAP, boli snahy o obrazovú reprezentáciu len niekoľké, budeme hovoriť o troch. [9,10,11] Prvou snahou z komerčného oblasti, bolo zverejnenie štandardu *OLE DB* multidimenziálnych databáz firmou Microsoft.[9] V tomto štandarde stránka reprezentácie dát ale zohrávala jednu s hlavných úloh. V danom prístupe je používaný silný dotazovací jazyk na poskytnutie komplexných reportov, vytvorených z niekoľkých OLAP kociek dát, resp. ich podmnožinami. Druhým bol akademický prístup nazvaný *páskový model (Tape model)*, založený na tzv. *páskach(Tapes)*. Posledným OLAP vizualizačným postupom o ktorom budeme pojednávať je *Cube Presentation model – CPM*.

4.1 OLEDB

OLEDB (*Object Linking and Embedding, Database*) je aplikačným rozhraním navrhnutým firmou Microsoft pre prístup do rozdielnych úložísk dát rovnakým spôsobom. Je to množina rozhraní implementovaných použitím *Component Object Model(COM)*. OLEDB bolo navrhnuté ako vyššia úroveň náhrady a následník pre *Open Database Connectivity(ODBC)*, za podpory väčšieho množstva rozdielnych nerelačných databáz, akými sú objektové databázy, alebo tabuľkové procesory, ktoré nie nutne podporujú SQL. OLEDB rozdeľuje úložisko dát od aplikácie, ktorá potrebuje pristupovať k dátam cez množinu abstrakcií. Je konceptuálne rozdelené na *poskytovateľa(Provider)* a *zákazníka(Customer)*. Zákazníkom sú aplikácie ktoré pristupujú k dátam, a poskytovateľ je softwarový komponent ktorý implementuje rozhranie a týmto poskytuje dáta zákazníkovi. OLEDB je súčasťou technológie *Microsoft Database Access Components (MDAC)*. MDAC predstavuje skupinu Microsoft technológií, ktoré spolu spolupracujú a vytvárajú rámec(*framework*), ktorý dovoľuje programátorom jednotný a úplný prístup k vývoju aplikácií a prístup k takmer akýmkoľvek dátam. Takto môže byť zabezpečený prístup od jednoduchých textových súborov až po komplexné databázové systémy. Taktiež môže umožniť prístup k hierarchickým úložiskám dát, ako napr. emailové

systémy. OLEDB poskytuje štandárd jazyka *MultiDimensional eXpression (MDX*), pre výpočet a prezentáciu OLAP kociek. OLEDB prináša aj niekoľko nevýhod, medzi ktoré patria, chýbajúce teoretické pozadie, ktoré neobsahuje žiadnu definíciu schémy multikocky, a kombinuje prezentačné s viualizačnými záležitosťami. Výsledkom je príliš komplexný a niekedy ťažko použiteľný, ale silný dotazovací jazyk.[9]

4.2 Páskový model

Páskový model *(Tape model)* sa skladá zo štrukturalizovanej hierarchie pások, ktoré významovo odpovedajú dimenziám. Každá páska sa skladá zo záznamových stôp *(Tracks)*, ktoré odpovedajú úrovniam dát. Priesečníky záznamových stôp definujú multidimenzionálnu maticu. V tomto modeli definujeme základné operácie: [12]

- Vkladanie a mazanie záznamových stôp
- Menenie poradia stôp, alebo ich radenie
- Postupné prechádzanie stôp
- *Roll-up*
- *Drill down*

4.3 Cube Presentation Model

Cube Presentation Model (CPM), sa skladá z dvoch vrstiev. Prvá je *logická*, ktorá obsahuje formuláciu kociek a *prezentačná* vrstva, ktorá obsahuje vizuálnu prezentáciu týchto kociek na 2D obraze. Rozdelenie kopíruje základnú myšlienku celého modelu ktorá spočíva v rozdelení na logické získavanie dát, ktorá sa nachádza v logickej vrstve modelu CPM, a na dátovú prezentáciu, ktorá sa nachádza v prezentačnej vrstve modelu CPM. Uvedené rozdelenie umožňuje prípadne rozdelenie implementácie alebo ich oddelenú výmenu.

4.3.1 Logická vrstva

Logická vrstva obsahuje[11]:

- **Dimenzie:** definované ako mriežku úrovní
- **Funkciu prechodu**: mapujúcu hodnoty medzi súvisejúcimi úrovňami v dimenzií
- **Detailne dátové množiny:** ktoré modelujú tabuľky faktov na najnižšej informačnej granulite pre všetky v nich obsiahnuté dátové dimenzie
- **Kocky:** definované ako agregácie nad detailnými dátovými množinami.

4.3.2 Prezentačná vrstva

4.3.2.1 Entity prezentačnej vrstvy

Prezentačná vrstva využíva k zobrazovaniu na obrazovke niekoľko entít:

- **Bod** (*Point*): Bod na osách odpovedá klasickému ponímaniu bodu v matematike. Je charakterizovaný prislúchajúcou rovnosťou výberu podmienok nad určitou úrovňou. Je ale možné niekoľko dátových dimenzií zobraziť na prezentačnej ose, preto definícia bodu nevylučuje prípad, že bod reprezentuje niekoľko dát s rôznych dátových dimenzií..
- **Osa** (*Axe*): Osa môže byť chápaná ako množina bodov. V CPM uvažujeme o dvoch druhoch: *Neviditeľná* osa, *Obsahová* osa. Neviditeľná osa určuje miesto pre úrovne množín dát, ktoré majú byť zobrazené. Obsahová osa, má zložitejšiu rolu, určuje umiestnenie pre obsah multikociek, určených nad agregovanými detailnými dátami.
- **Multikocka** (*Multicube*): Multikocka je definovaná na miltidimenzionálnym priestorom, zahrnujúcim skupinu ôs v jednom pohľade. Z druhého hľadiska je definovaná nad spodnými dátami, poskytujúca všetky dáta, ktoré budú filtrované a agregované pred samotnou prezentáciu užívateľovi. Multikocky sa taktiež objavujú v modeli CPM v mapovaní medzi multidimenzionálnym priestorom a spodnými dátami, ktoré poukazuje na výpočet multidimenzionálneho obsahu.
- **2D-plát** (*Slice*) – 2D plát je 2D vrstvou dát, ktoré môžu byť zobrazené na obrazovke. Nech existuje miltikocka s *k* osami, potom 2D-Slice môžeme definovať nie len nad ľubovoľnou dvojicou z uvedených osí, ako by sa na prvý pohľad logicky mohlo zdať, ale podľa definície bodu, ktorá hovorí že on sám môže zobrazovať niekoľko dát s rôznych dátových dimenzií, tak 2D-Slice môžeme definovať, ako množinu bodov skladajúcich sa z hodnôt s maximálne z *k-2* dimenzií.
- **Páska** (*Tape*) – Páska je stĺpec alebo riadok cez 2D-Slice dát, konštruovaný paralelne na jednu z ôs. Opäť ak uvažujeme 2D-Slice, vytvorený nad multikockou s *k* osami, tak páska môže byť analogicky definovaná ako množina bodov, kde každý môže reprezentovať niekoľko hodnôt s *k-1* dimenzií.

- **Kríženie**(*Cross-join*): Ak budeme uväzovať o jednej páske kolmej na jednu z osí a o druhej páske kolmej na inú z osi, tak ich prienik je bunka, v obecnom prípade aj skupina buniek, ktoré označujeme ako kríženie (*cross-join*). Takže kríženie môžeme definovať aj ako prienik dvoch neparalelných pások.
- **Obsahová funkcia** (*Content Function*): Na úrovní schémy, predpokladáme funkciu priradzujúcu miery na obsahové osy multikocky, spolu s poradím alebo obmedzeniami.

4.3.3 Príklad CPM

Pre jednoduchosť uvedieme príklad, podľa [13], v ktorom budeme uvažovať o prehľade predaja a produktov za určitý čas. Zavedieme multikocku *SalesCube*, nad dimenziami *Products*, *Salesman*, *Time*, *Geography*. Niektoré z dimenzií obsahujú niekoľko úrovní agregácie. Dimenziu *Time* obmedzíme len na rok 1991, a dimenziu *Product* budeme ignorovat v svojich rozkladových triedach tým, že budeme uvažovať stále o všetkých produktoch.

Zakaždým, keď chceme zobraziť na 2D obrazovku viac ako dve dimenzie, tak potrebujeme získať kríženie daných dimenzií, v dvoch osách. Pre ukážku použijeme dimenziu *Salesman* (obmedzenú pre ukážku len pre dvoch predajcov) a *Geography* na osách stĺpcov, a *Time* necháme na riadkových osách. Je nutné si uvedomiť, že dimenzia *Geography* obsahuje niekoľko úrovní agregácie, ako napr. *Štát (Country)*, *Oblasť (Region)* a *Mesto (City)*. Podobné platí aj pre dimenziu *Time*, kde uvažujeme o *Štvrťroku (Quarter)* a *Mesiacoch (Months)*.

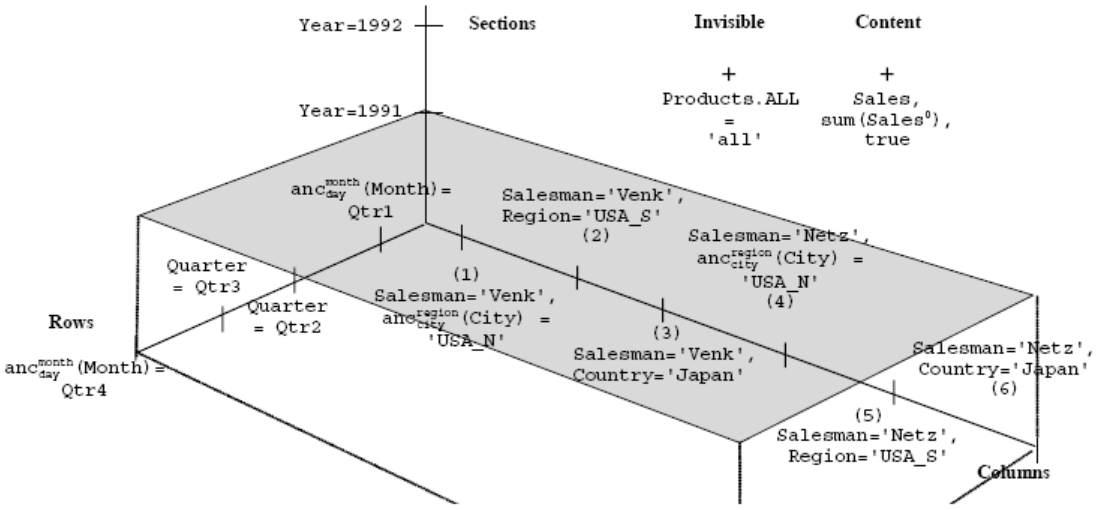
Uvažujme nasledujúci dotaz:

```
SELECT CROSSJOIN({Venk,Netz},{USA_N.Children,USA_S,Japan}) ON COLUMNS
{Qtr1.CHILDREN,Qtr2,Qtr3,Qtr4.CHILDREN} ON ROWS
FROM SalesCube
WHERE (Sales,[1991],Products.ALL)
```

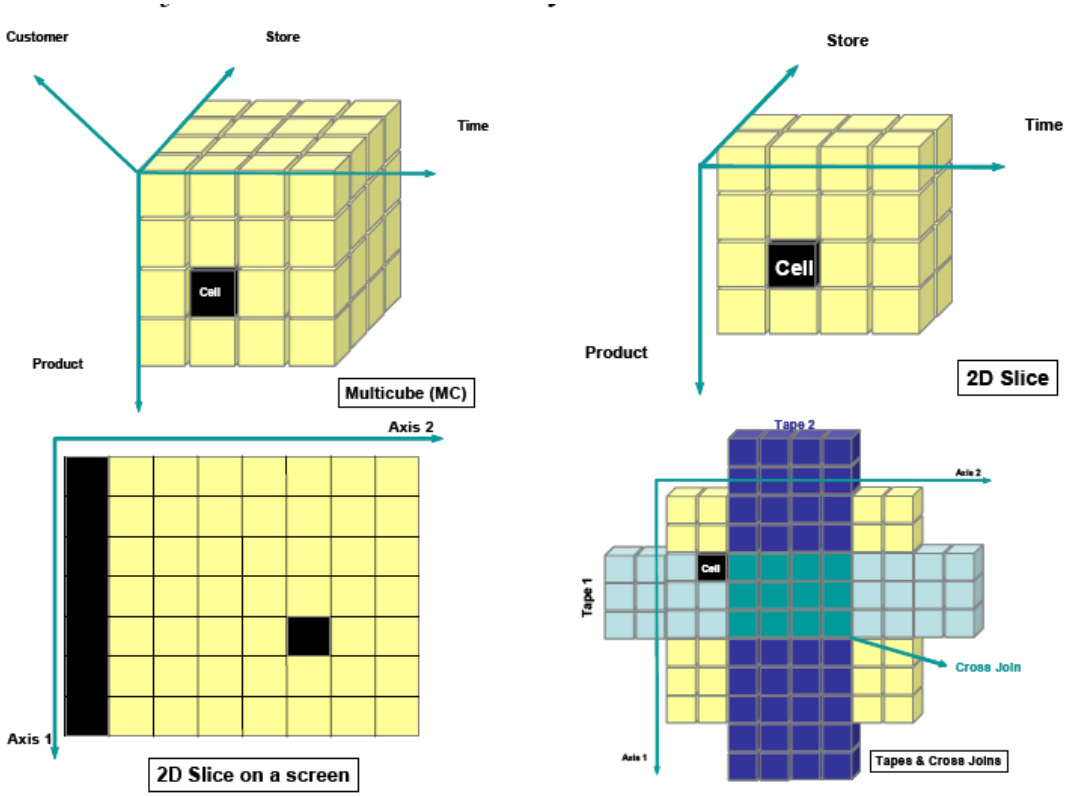
Tento dotaz v terminológií CPM reprezentuje 2D-Slice. Ten je definovaný štyrmi horizontálnymi páskami označenými ako *R1, R2, R3, R4* a šiestimi vertikálnymi páskami označenými od *C1* do *C2*. Význam horizontálnych pások prezentuje časovú dimenziu *(Time)* v štvrťrokoch, prípadne mesiacoch. Význam vertikálnych pások je komplexnejší, reprezentujú kombináciu dimenzií *Salesman*(Venk,Netz) s *Geography*, u niektorých s vyjadrením úrovní *Štátu(Country)*, *Oblasti(Region)* a *Mesta(City)*. Je nutné ale uviesť dve dimenzie, ktoré sú skryté a vo vyjadrení ignorované, tou je *Rok(Year=1991)* a dimenzia *Produktu(Prouct=all)*.

Year = 1991 Product = ALL			Venk USA USA_N		USA_S	Japan	Netz USA USA_N		USA_S	Japan
			Seattle	Boston			Seattle	Boston		
		Size(city)								
R1	Qtr1	Jan								
		Feb	**C1**		**C2**	**C3**		**C4**	**C5**	**C6**
		Mar								
R2	Qtr2									
R3	Qtr3									
R4	Qtr4	Oct								
		Nov								
		Dec								

Obrázok 13 Príklad modelu kocky



Obrázok 14 2D-Slide uvedeného príkladu



Obrázok 15 Komplexná vizualizácia multikociek, bodov, ôs, 2D-plátov a krížení v 2D a 3D

4.3.4 Vizualizácia CPM

Uvedieme ako príklad vizualizácie CPM spoluprácu s *Table Leens(TL)* modelom.[13] TL predstavuje model tradičného zobrazenia cez niekoľko tabuliek, a spoluprácu s prostriedkami interakcie počítača a užívateľa. Tento model sa používa na vizualizáciu tabuľkových, variantných a multidimenzionálnych dát a je vhodný aj na účely OLAP. Je založený na princípe „*focus plus context*“, ktorý umožňuje manipuláciu a prezentáciu 2D tabuliek, ich zobrazenie detailov a agregácií bez straty globálneho pohľadu na dáta.

4.3.4.1 Table Lens Model

Pre lepšie pochopenie je nutné ozrejmenie TL modelu, ktorý sa skladá s niekoľkých základných konštrukcií:

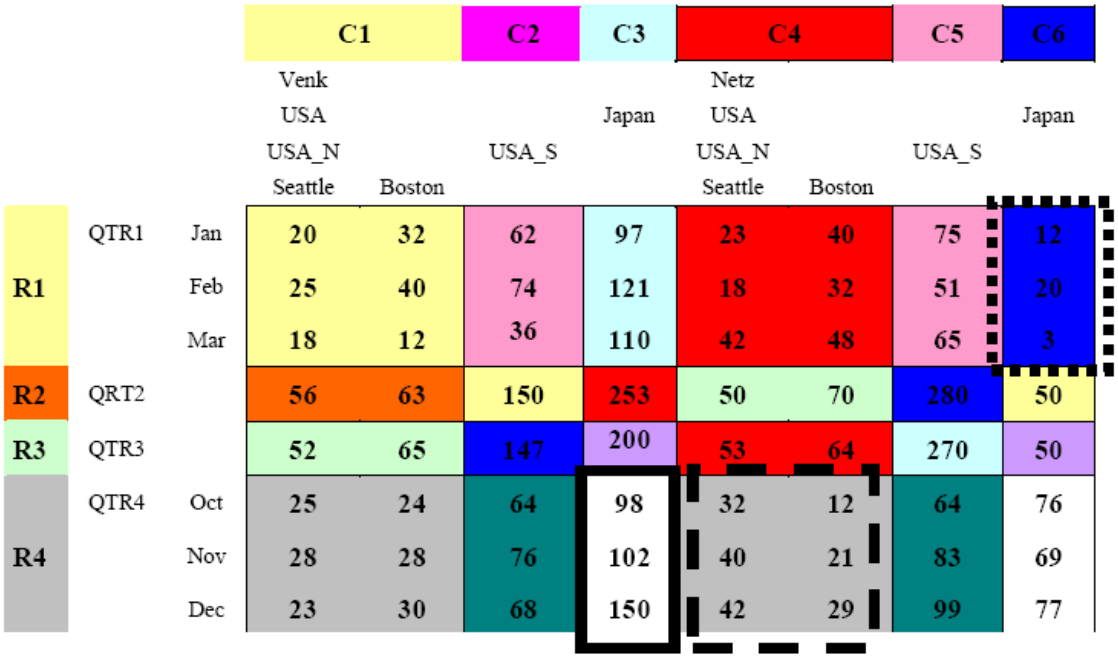
- **Osa***(Axe)*: TL model uvažuje o dvoch osách. Pre jednoduchosť budeme hovoriť o stĺpcoch a riadkoch.
- **2D priestor***(2D space)*: 2D priestor je zostrojený z kartezkého súčinu oboch osí. Je to konečná matica buniek.

- **Funkcia stupňu záujmu**(*Degree of Interest Function-DOI*): DOI je funkcia, ktorá mapuje každý bod osi, na hodnotu ktorá predstavuje úroveň záujmu pre daný bod. Pre každú osu iná DOI je definovaná.
- **Funkcia prenosu***(Transfer function)*: funkcia prenosu mapuje každú bunku na svoju fyzickú pozíciu, indikuje úroveň zobrazených detailov pre každú bunku. Prakticky funkcia prenosu je ekvivalentom funkcie stupňu záujmu v pixelovom zobrazení.

Jednou zo základných myšlienok TL je, že nie všetky bunky sú si významovo rovné. V skutočnosti niektoré bunky zahrnujú konkrétnu oblasť 2D-plátu a zaberajú v zobrazení väčšie miesto ako iné.

Samotné mapovane CPM na TL nie je zložité. Osy CPM sa namapujú na osy TL a 2D-plát CPM je implementovaný ako 2D-plát TL. CPM je generický dosť na to, aby poskytol potrebné údaje pre DOI funkcie. Užívateľ často volí tzv.: „Okno záujmu *(Window of Interest)*“, ktoré je určitá časť 2D-plátu, ktorá je špecifikovaná a zvýraznené svojou veľkosťou alebo formátovaním.

Ostáva otázkou ako automaticky poskytovať užívateľovi okno záujmu pre lepšie skúmanie OLAP reportu. Je vhodné priaktívne určiť takéto okno a zobraziť ho užívateľovi. Dokážeme zostaviť algoritmus, podľa volieb užívateľa, ktorý môže určiť podmienky zastavenie, obsahu chýb a iných parametrov. Taktiež samotnou voľbou užívateľa zostáva, či sa má zobraziť okno záujmu s hodnotami, alebo len s agregovanými výpočtami, ako napr.: maximum, minimum, priemer, suma.



			C1		C2	C3	C4		C5	C6
			Venk				Netz			
			USA			Japan	USA			Japan
			USA_N		USA_S		USA_N		USA_S	
			Seattle	Boston			Seattle	Boston		
R1	QTR1	Jan	20	32	62	97	23	40	75	12
		Feb	25	40	74	121	18	32	51	20
		Mar	18	12	36	110	42	48	65	3
R2	QRT2		56	63	150	253	50	70	280	50
R3	QTR3		52	65	147	200	53	64	270	50
R4	QTR4	Oct	25	24	64	98	32	12	64	76
		Nov	28	28	76	102	40	21	83	69
		Dec	23	30	68	150	42	29	99	77

Obrázok 16 Zobrazenie príkladu s hodnotami, kde rozdielne farby zobrazujú rozdielne kríženia, a zvýraznené čiary určujú hodnoty najbližšie k najvyšším, najnižším a priemerným hodnotám

Metóda CPM je analogicky spojená so samotnou dátovou reprezentáciou OLAP. Zložitou zostáva je vizualizácia, kde *Table Lens* poskytuje vhodnú alternatívu. Postupný rozvoj by mohla zaznamenať so zavadzaním lepších vizualizačných techník programových ale aj hardwarových, ktoré dokážu zobrazovať prehľadne veľké ‚množstvo dát, vďaka svojej veľkosti, rozlíšeniu, či výkonu. Podstatná ostáva tiež jednoduchosť manipulácie s dátami, na ktorej je nutné ešte pracovať, aby boli tieto nástroje prakticky nasadené a poskytovali užívateľom jednoduchú a prehľadné rozhranie.

5. Záver

Postupné komerčné využitie aplikácií vizualizácie, spolu s výskumom, podporujú vývoj v niektorých hlavných trendoch. Jedná sa o dostupnosť väčšieho množstva typov vizualizácie. Taktiež dochádza k zmene zo statickej vizualizácie k dynamickým reprezentáciám informácií. Vzrastá možnosť vizualizovať väčšie množstvo a viac komplexných dát. A v neposlednom rade je nutne spomenúť vytváranie štandardov v oblastí vizualizácie.

Tieto trendy sa navzájom nevylučujú. Súčasný komerčný vizualizačaný software používa iba niekoľko vizualizačných techník, z veľkého množstva vytvorených a dokumentovaných. Postupný vývoj vizualizácie spolu so zobrazovacími technikami a technikami interakcie s užívateľom, smeruje k interaktívnej vizualizácií, kde užívateľ bude môcť do zobrazovanie priamo a dynamicky zasahovať a dané dáta dynamicky meniť, čo umožní ich lepšie a názornejšie pochopenie. Dynamická vizualizácie by užívateľovi umožňovala operácie nad dátami, ako napr.: rotácie, zahorovanie do detailov, výber skúmaných dát a získavanie ich agregovaných hodnôt, a pod.

Ďalšiu oblasť predstavujú dáta ktoré nie sú reprezentované v jednoduchom formáte, ktorý by mohol byť zobrazený ako tabuľka. Aj keby bolo možné uvedené dáta previesť do podoby tabuľky, tak by stratili svoj podstatný sémantický význam, čím by neposkytli potrebné informácie. Do tejto oblasti je taktiež zameraný databázový výskum.

Predmetom skúmania sú aj vizualizácie veľkého objemu dát. Nie sú výnimkou terabytové databáze u ktorých zobrazenia dokážeme niektoré hodnoty ignorovať, agregovať a zanedbávať, ale súčasná pixlovo orientovaná vizualizácie dokáže zobraziť len istý počet rôznych dát. Je zrejmé, že napr. výsekový graf, len veľmi ťažko dokáže zobraziť niekoľko tisíc delení, a by strácal svoj význam a prehľadnosť.

Cieľom tejto práce bolo uviesť niekoľko štandardných, známych ale aj menej rozšírených postupov vizualizácie dát. Je zložité obsiahnuť skúmané metodológie ktorých je značné množstvo. Môžeme predpokladať, že vizualizácia sa bude ďalej rozvíjať a nachádzať široké uplatnenie. Častým problém však zostáva reálne nasadenie vzniknutých techník, pretože mnoho nápadov, zostáva len v teoretickej rovine, na papieri, a väčšina ľudí používa bežné grafy a tabuľky a o nových technikách nemajú informácie. Podľa môjho názoru je žiadané aby databázová komunita pracovala aj v tomto poli a ukazovala potenciálne možnosti vizualizácie dát a jej výhody bežným užívateľom.

6. Literatúra

[1] Soukup, J.,Davidson I., Visual Data Mining, Wiley Publishing, 2002

[2] Fayyad, U. M. - Piatetsky-Shapiro, G. a kol.: Advances in Knowledge Discovery and Data Mining. Cambridge MA, 1996

[3] OLAP Council. The APB-1 Benchmark. 1997. http://www.olapcouncil.org/research/bmarkly.htm

[4] S. Chaudhuri, U. Dayal: An overview of Data Warehousing and OLAP technology. ACM SIGMOD Record, 26(1), March 1997.

[5] Matiaško, K., Vnuk, L., Ševčíková, K., 2001. Dátové sklady ako informačný zdroj pre podporu rozhodovania. Fakulta riadenia a informatiky, Žilinská univerzita, Bulletin SIS. - č. 2, 2001.

[6] 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.

[7] Harris, R., Information graphics:A Comprehensice Illustrated Reference, Oxford University Press, 1999

[8] Keim, D.A., Visual Data Mining. Tutorials of the 23rd International Conference on Very Large Data Bases, Athens, Greece, 1997.

[9] http://msdn2.microsoft.com/en-us/library/ms717005.aspx, 14.12.2007

[10] Gebhardt, M., Jarke, M., Jacobs, S.: A Toolkit for Negotiation Support Interfaces to Multi-Dimensional Data. ACM SIGMOD 1997, pp. 348 – 356.

[11] Maniatis, A., Vassiliadis, P., Spiros Skiadopoulos, Vassiliou, Y.: CPM: A Cube Presentation Model for OLAP. DaWaK 2003, Prague, Czech Republic, September 3-5 2003.

[12] Gebhardt, M., Jarke, M., Jacobs, S., A Toolkit for Negotiation Support Interfaces to Multidimensional Data. In Proc. of the 1997 ACM SIGMOD Conf., Arizona, USA, 1997.

[13] Maniatis, A., Vassiliadis, P., Spiros Skiadopoulos, Vassiliou, Y.: Advanced Visualization for OLAP, National Technical University of Athens, 2003