MIS – Prednáška 2.

Ing. Ján Čabala

Obsah

- Dáta
  - Spôsob a možnosti ich získavania
  - Úprava a predspracovanie
  - ETL fáza tvorby dátového skladu (Extraction, Transformation, Loading)
- Relačná databáza
  - E-R model, fyzický model, vzťahy medzi nimi
  - Primárny a cudzí kľúč
  - Vzťahy medzi atribútmi
  - Normálne formy
- Problematika OLAP (multidimenzionálny model)
  - Fakty
  - Dimenzie
  - Hierarchie
  - Schémy
- Relačný vs. Multidimenzionálny model

Dáta

- V drvivej väčšine prípadov sú nehomogénne, pochádzajú z rôznych zdrojov
- Zdroje:
  - Databázové softvéry
  - Produkčné systémy
  - Externé súbory
- Pred zavedením dát do databázy, resp. do dátového skladu, je nevyhnutné dáta predspracovať (vyčistiť, upraviť)

ETL fáza

- Extraction, Transformation, Loading
- Pomerne časovo náročný proces (niekedy až polovica z celkovej implementácie)
- Extrakcia – výber dát prostredníctvom rôznych metód
- Transformácia – overenie,čistenie, integrovanie a časové označenie dát
- Loading – premiestnenie dát do databázy

1.Extrakcia

- Centralizácia dát z nehomogénnych zdrojov
  - Rôzne platformy
  - Rôzne softvéry
  - Archívne dáta
  - Externé súbory

2.Transformácia

- Zmena a úprava údajov do formy potrebnej pre ďalšiu prácu s dátami
- Problémy pri transformácii:
  - Chýbajúce hodnoty (ignorovať resp. doplniť z iných zdrojov)
  - Duplicitné záznamy (odstrániť)
  - Rôzne názvy rovnakých pojmov a objektov (zjednotiť)
  - Rôzne peňažné meny (previesť na rovnakú menu)
  - Chýbajúci dátum
  - Nejednoznačnosť údajov
  - Formáty čísel a textov

2.Transformácia
Nejednoznačnosť údajov

name	gender
Nowmer Sheri	Female
Whelply Derrick	male
Derry Jeanne	F
Spence Michael	man

- Po transformácii

name	gender
Nowmer Sheri	F
Whelply Derrick	M
Derry Jeanne	F
Spence Michael	M

2.Transformácia
Formáty čísel a textov

	Numerický formát	Textový formát
Rodné číslo	6254111547	625411/1547
PSČ	08701	087 01
	8701	08701

3.Loading + problémy ETL

- Loading - Prenos údajov a ich uloženie do databázových tabuliek

- Problémy ETL
  - Nutnosť overenia údajov
  - Nepresné údaje – nepresné výsledky analýz – nesprávne strategické rozhodnutia

Relačné databázy

- Databázy, v ktorých sa vykonáva veľké množstvo transakcii v reálnom čase (banky, supermarkety) zvyknú byť nazývané ako OLTP (On-Line Transaction Processing)
- Transakčné databázy sú uložené relačnou formou, teda v databázových tabuľkách, medzi ktorými sú isté logické vzťahy (relácie)

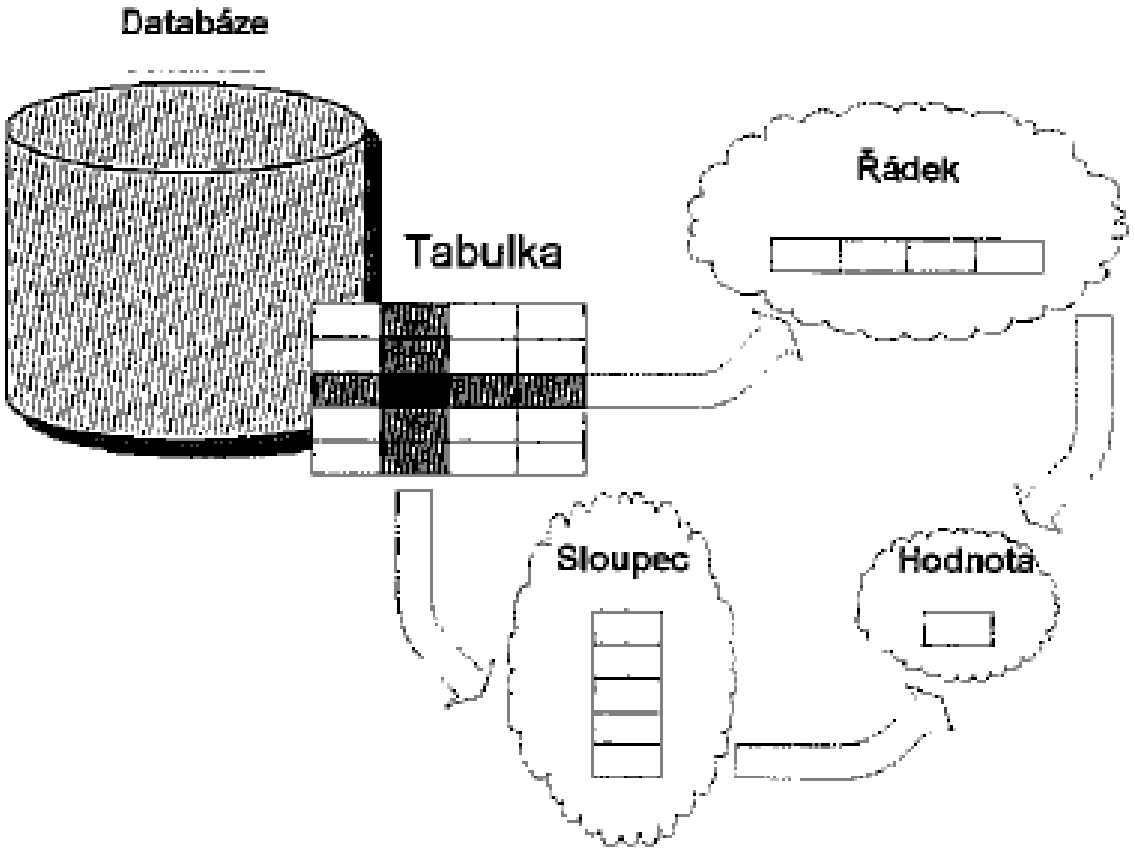
Transakčná databáza

Id	Produkt	Počet kusov	Dátum	Skupina	Názov predajne	Okres	Kraj
1	samorezna skrutka M12	408	1.7.2004	spotrebny material	Kovomat	Trencin	Tremcianske kraj
2	lanko	32.8	1.3.2004	spotrebny material	Urob si sam	Nové Zámky	Nitriansky kraj
3	brusny kotuc	154.4	1.1.2004	spotrebny material	Urob si sam	Detva	Banskobystricky kraj
4	rezny kotuc	190	1.7.2004	spotrebny material	Lapal	Myjava	Tremcianske kraj
5	zavlacka 6mm	138	1.12.2004	spotrebny material	Urob si sam	Topoľčany	Nitriansky kraj
6	poistny kruzok M10	110.556	1.4.2004	stavebny material	Urob si sam	Brezno	Banskobystricky kraj
7	poistny kruzok M12	89.04	1.7.2004	spotrebny material	Urob si sam	Martin	Zilinsky kraj
8	tesniaci kruzok M8	100.8	1.8.2004	spotrebny material	Hornbach	Roznava	Kosicky kraj
9	zavitova tyc M8	52.2	1.10.2004	spotrebny material	Mega zeleziarstvo	Nové Zámky	Nitriansky kraj
10	vrtak na	59.1	1.9.2004	spotrebny	Zeleziarst	Nové	Nitriansky

Databázová tabuľka

- V relačných databázach sú dáta uložené v 2-rozmerných tabuľkách, kde je hodnota obsahujúca potenciálne užitočnú informáciu uložená na priesečníku príslušného riadku a stĺpca
- Každý riadok by mal byť označený jednoznačným identifikátorom (ID)

Databázová tabuľka



Databázová tabuľka

DATUM_ID	DATUM	MESIAC	KVARTAL	ROK
1	1.1.2004	Január	Q1	2004
2	1.2.2004	Február	Q1	2004
3	1.3.2004	Marec	Q1	2004
4	1.4.2004	April	Q2	2004
5	1.5.2004	Maj	Q2	2004

Relačné databázy
Abstrakcia

- Hlavným účelom databázového systému je poskytnúť používateľom s abstraktný pohľad na systém.
- Systém skrýva určité detaily o tom, ako sú dáta uložené a vytvorené a spravované.
- Zložitosť by mala byť skrytá od užívateľov databázy.

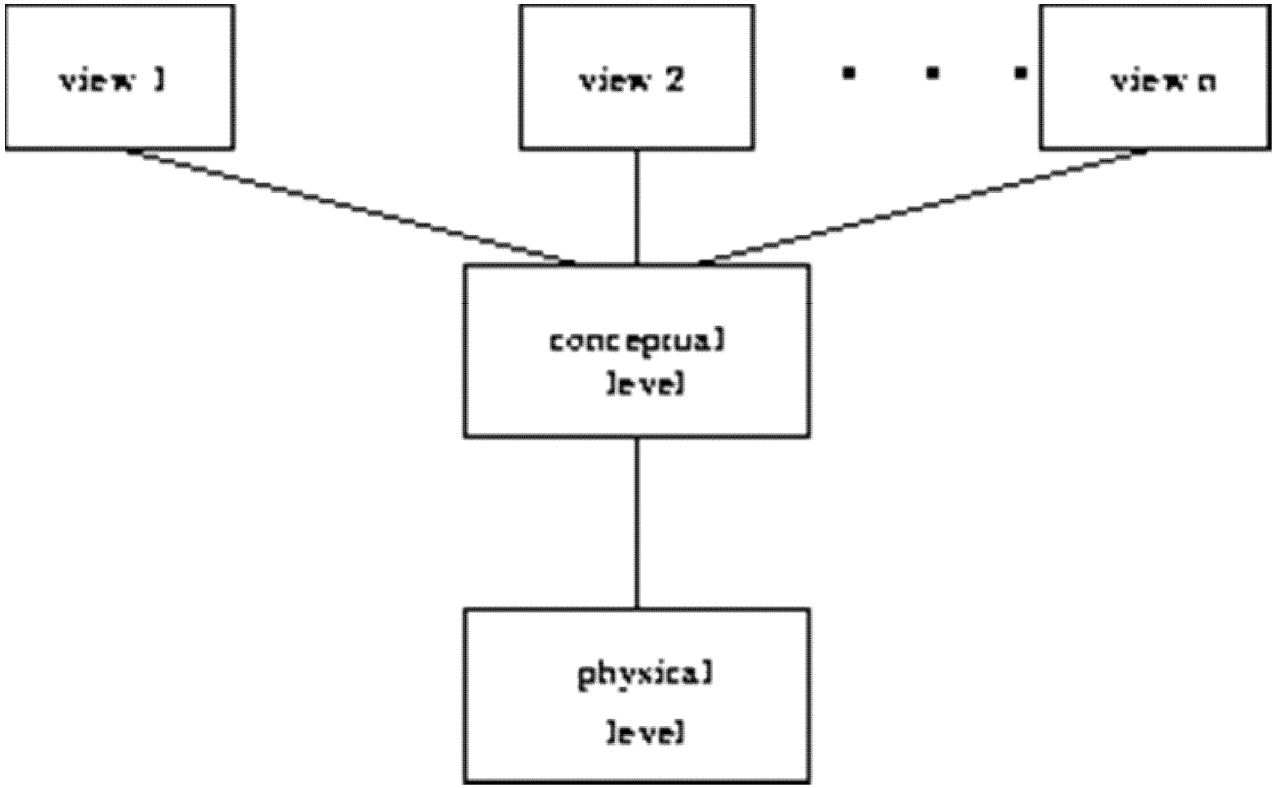
- Existuje niekoľko úrovní abstrakcie:
- 1. interná/fyzická úroveň:
  - Ako sú dáta uložené
  - Najnižšia úroveň abstrakcie.

Relačné databázy
Abstrakcia II.

2. Konceptuálna/logická úroveň: ďalšia vyššia úroveň abstrakcie.
   - Popisuje, aké sú dáta uložené.
   - Popisuje vzťahy medzi dátami.
   - Úroveň správcu databázy.

3. Externá/Pohľadová úroneň: najvyššia úroveň.
   - Popisuje časť databázy pre konkrétnu skupinu užívateľov.
   - môže byť mnoho rôznych pohľadov na databázu.

Relačné databázy
Abstrakcia III.



Relačné databázy
Entitno-relačný model

- Model konceptuálnej úrovne
- Logický model

  Modelovanie na logickej úrovni z ktorej je možné vytvoriť dátový model

  – Entita
  – Atribút
  – Relácia (1:1, 1:N,N:M)
  – Pohľad

Relačné databázy
Entitno-relačný model - pojmy

Entity

- bod záujmu pre návrhára
- zvyčajne podstatné mená a môžu sa vzťahovať na reálne objekty alebo pojmy
- Musia mať atribúty a jednoznačný identifikátor

Inštancia(Instance)

- je už popis skutočného objektu popísaného entitou. Inštancia sa stane vo fyzickom modeli riadkom tabuľky.

Atribút

- opisuje, kvantifikuje alebo určuje jednu vlasnosť entity. Entita sa zvyčajne skladá z niekoľkých atribútov. Vo fyzickom modeli se stane z atribútu stĺpec.
- povinnosť/voliteľnosť atribútov sa znázorňuje:
  - o nepovinný atribút, može obsahovať null hodnotu
  - * povinný atribút, hodnota musí byť rôzny od null

Jednoznačný identifikátor UID (Unique Identifier)

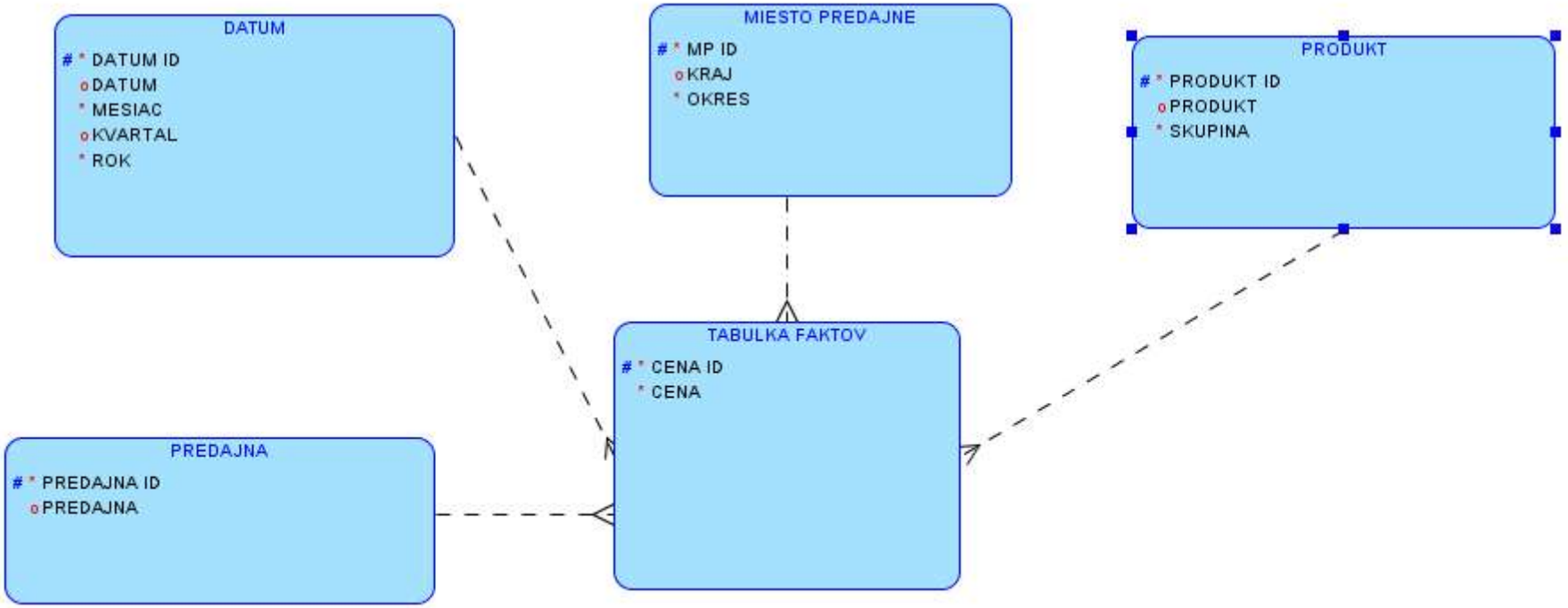
- je jeden alebo viacej atribútov, ktoré jednoznačne identifikujú inštanciu entity. Znazorňuje sa: #

Relačné databázy
Entitno-relačný model – pojmy II.

Relácie

- Relácie reprezentujú vzťahy medzi entitami
- V E-R modeli môžu byť reprezentované rozličné typy relácií dané ich ďalšími charakteristikami:
  - **Účasť(Participation) :**
    - povinná(mandatory)
    - alebo voliteľná(optional)
  - **Kardinalita(Cardinality):**
    - jedna - k - jednej(one - to - one) 1:1
      - Občan -> rodné číslo
    - jedna - k - mnoho(one - to - many) 1:N 
      - Klient -> konto, otec -> dieťa
    - mnoho - k - mnoho(many - to - many) M:N 
      - Učiteľ -> predmet
      - Väčšina databáz s nimi nevie priamo pracovať, využíva sa dekompozícia na viacero 1:N relácií

Relačné databázy
Entitno-relačný model



Relačné databázy
Fyzický model

- Návrh databázy na fyzickej úrovni (najnižsia úroveň abstrakcie)
- Na zakreslenie fyzického modelu nám slúžia Data Model Diagramy DMD
- DMD popísujú dáta, ktoré nás zaujímajú, v akom sú vzťahu a ako budú fyzicky uložené.
- Fyzický dátový model môže vznikať priamo jeho návrhom bez E-R modelu, alebo transformáciou E-R modelu do fyzického modelu pomocou nástrojov v závislosti od druhu databázy.

Relačné databázy
Fyzický model II.

- Hlavné kroky návrhu relačného modelu:

- Definícia tabuliek(table definition)
  - Tabuľky sú definované na základe E-R modelu použitím transformácie

- Definícia stĺpcov(column definition)
  - Stĺpce sú odvodené z definícii atribútov a relácií, ktoré zodpovedajú definíciam cudzích kľúcov

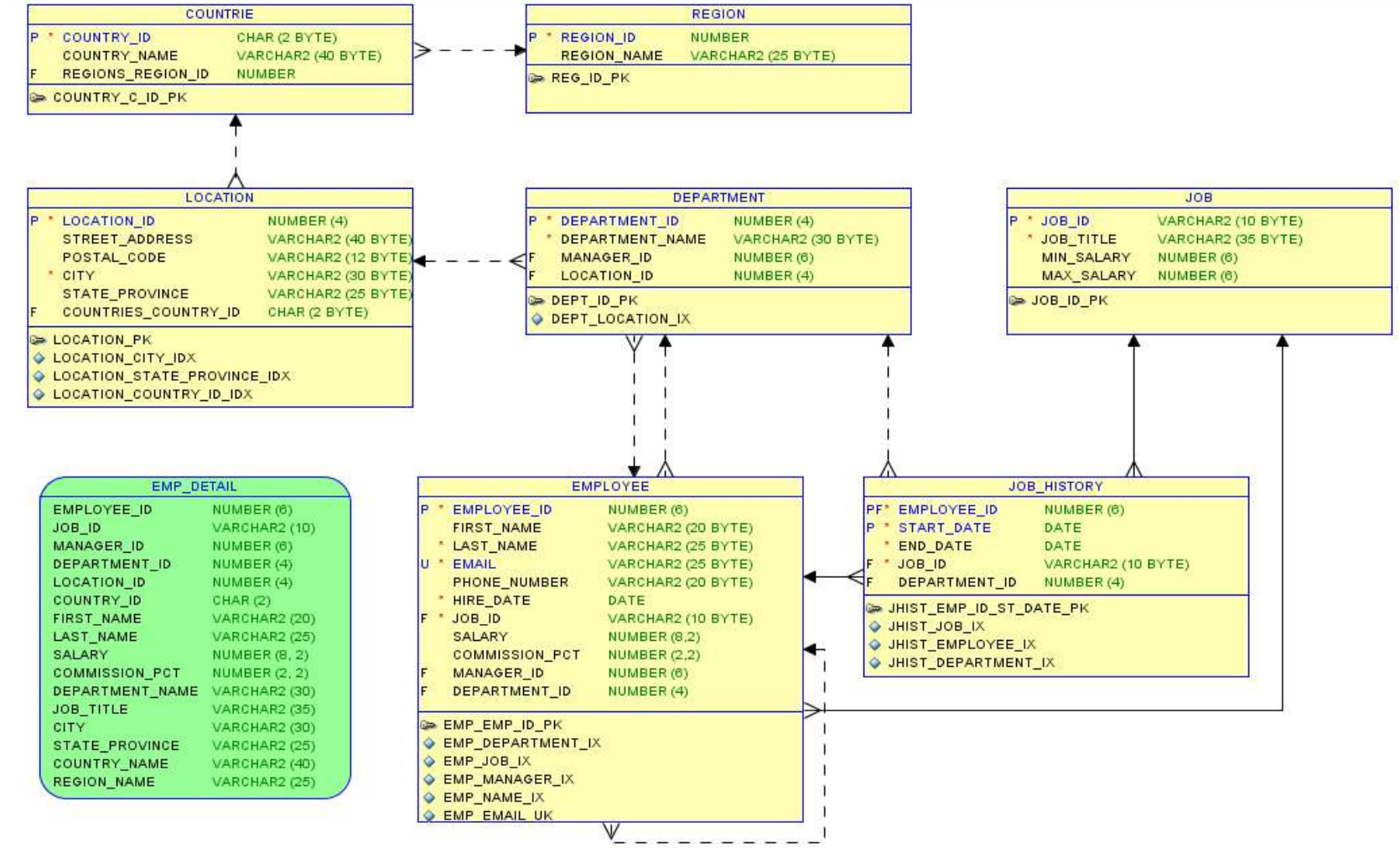
Relačné databázy
Fyzický model III.

- Definícia integrity
  - V čase databázového návrhu vieme zadefinovať väčšinu logiky uloženia dát, t.j integritu dát v databáze vo forme tabuliek a ich vzťahov, ale taktiež aj vo forme integritných obmedzení.
- Integritné obmedzenia v DB:
  - **Údajové typy**
    - Poskytujú základnú integritu dát. Špecfikáciou udajového typu definujeme typ, veľkosť a štruktúru ukladanej hodnoty.
  - **Primárny kľúč(Primary Key) PK**
    - Je stĺpec alebo kombinácia stĺpcov, ktorá jednoznačne identifikuje všetky riadky v tabuľke.
    - Stĺpce primárneho kľúča misia byť povinné,t.j. nesmú nadobúdať NULL hodnotu
    - Každá tabuľka môže mať len jeden PK

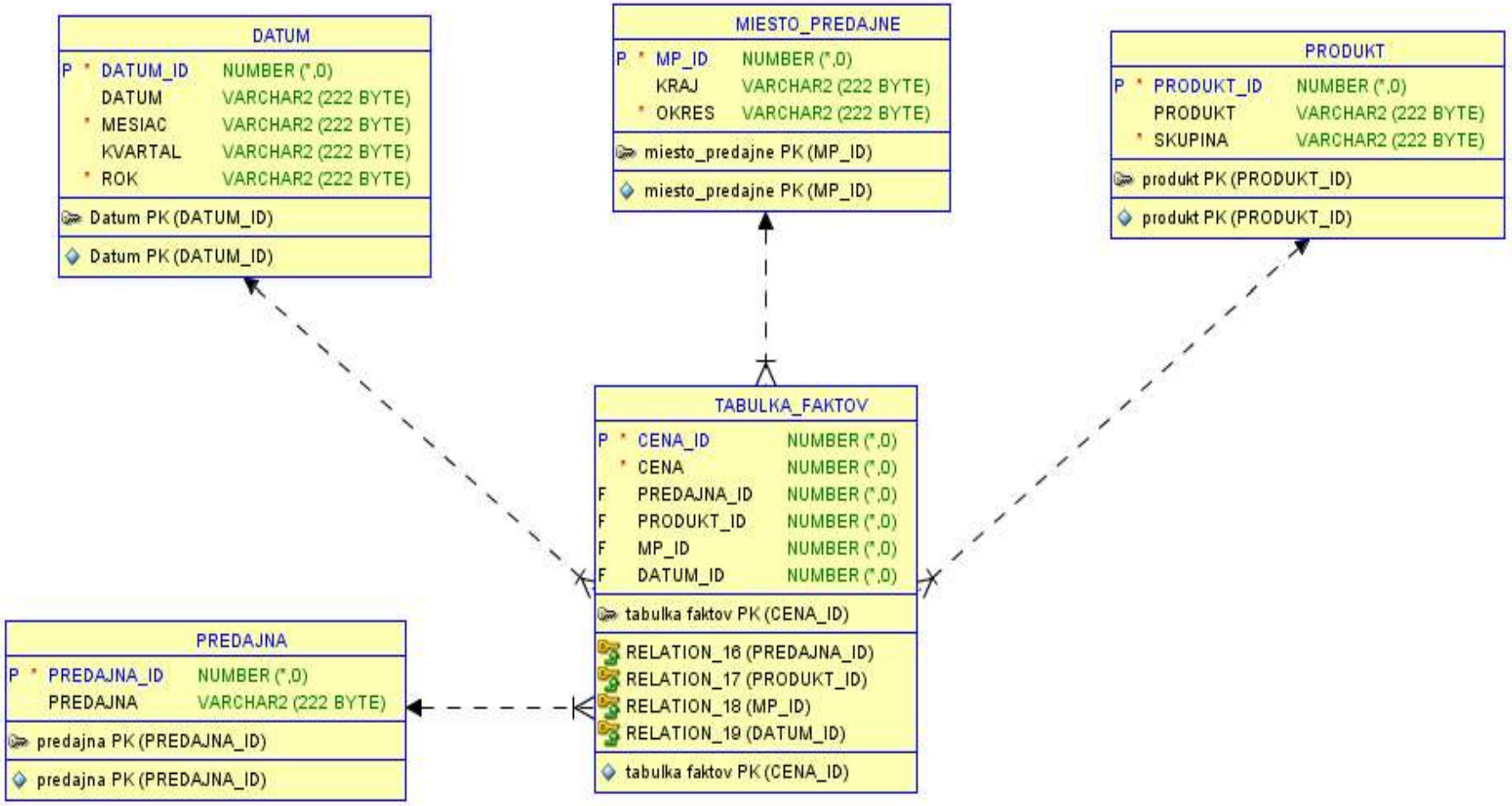
Relačné databázy
Fyzický model IV.

- **Unikátny kľúč(Unique Key) UK**
  - Je stĺpec alebo kombinácia stĺpcov, ktorá jednoznačne identifikuje všetky riadky v tabuľke.
  - Stĺpce unikátneho kľúča nemusia byť povinné,t.j. môžu nadobúdať aj NULL hodnotu
  - Každá tabuľka može mať viecej UK
- **Cudzí kľúč(Foreign Key) UK**
  - Je stĺpec alebo kombinácia stĺpcov, ktorá zodpovedá primárnemu kľúču v referencovanej tabuľke. Hodnoty cudzieho kľúča rovnať hodnotám primárneho kľúča v referencovanej tabuľke, alebo sú NULL.

Relačné databázy
Fyzciký model V.



Relačné databázy
Fyzický model VI.



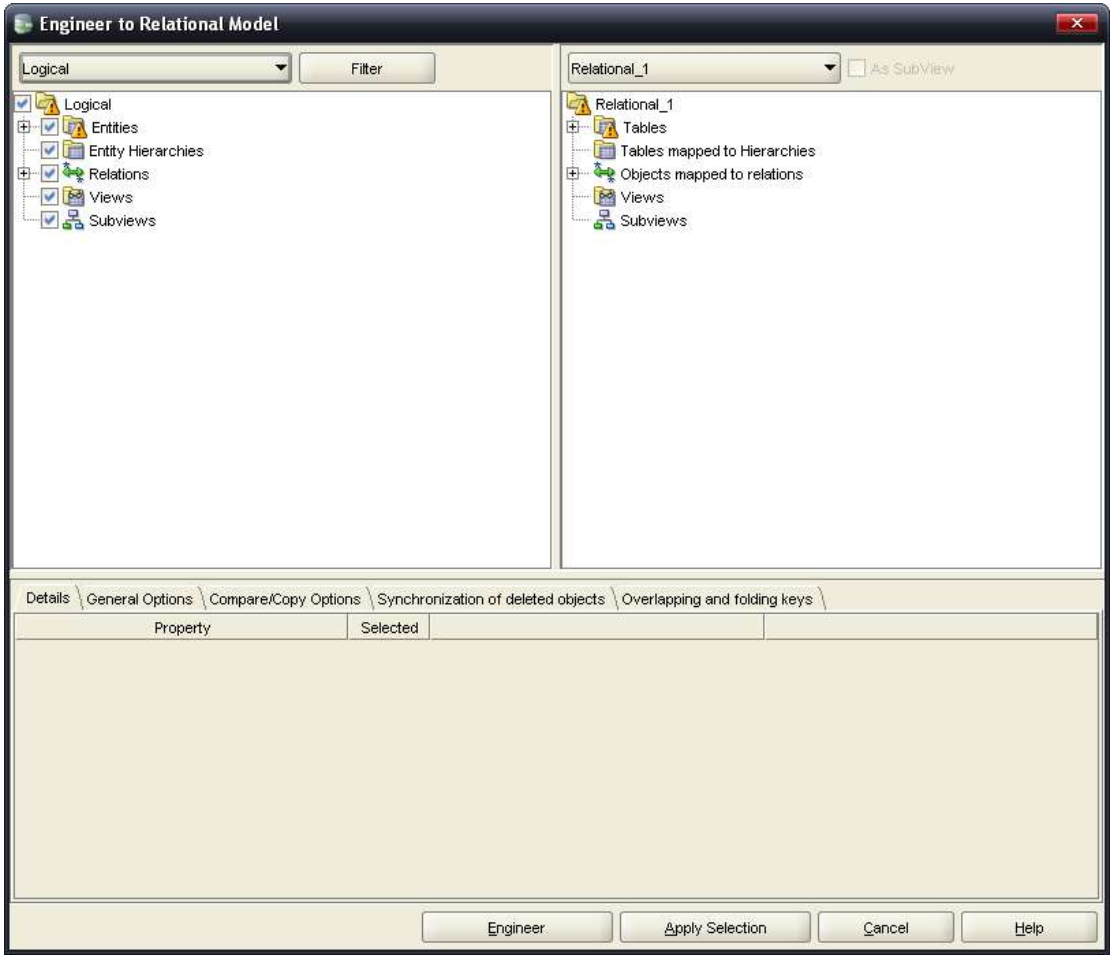
Relačné databázy
E-R (logický) model vs. Fyzický model

E-R model	Fyzický model
Entita	Tabuľka
Atribút	Stĺpec
Účasť	NULL resp. NOT NULL CONSTRAINT
Relácia	cudzí kľúč ...FK
M:N relácia	väzobná, resp. prepájacia tabuľka

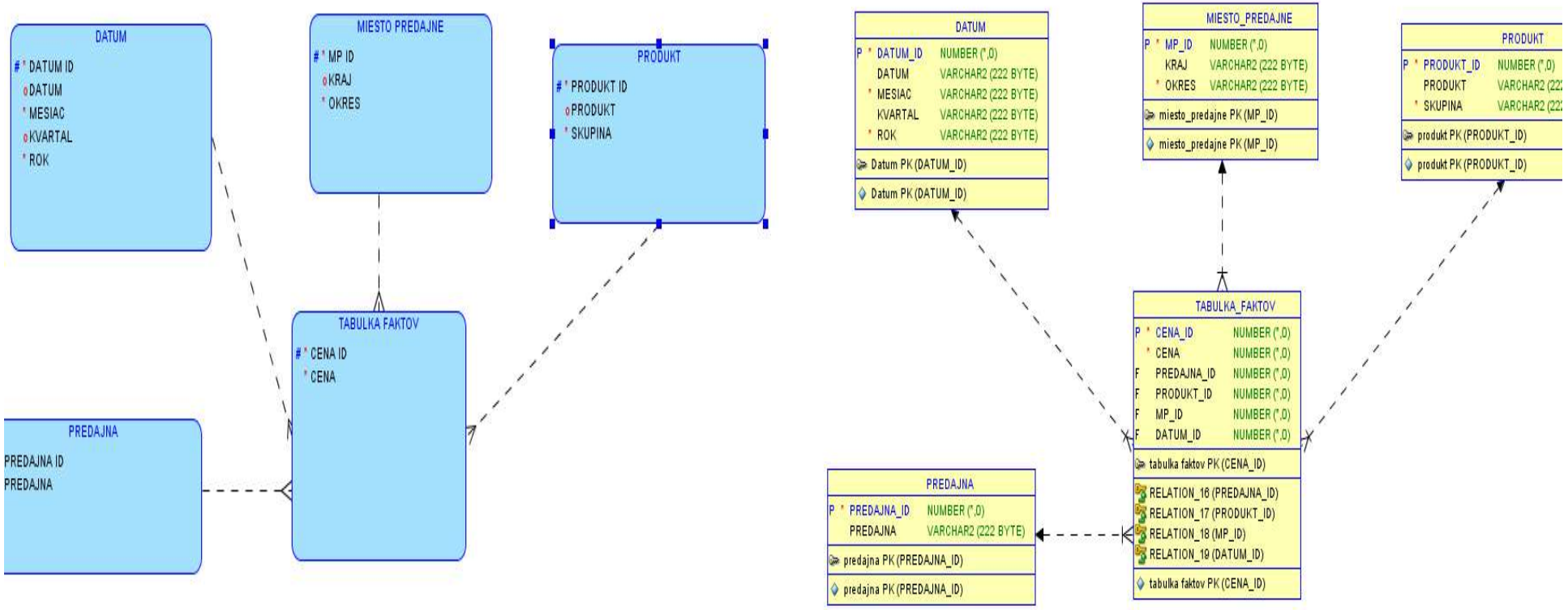
Jednoznačný identifikátor

E-R model	Fyzický model
Označený ako primárny PUID	primárny kľúč ..._PK
Ostatné UID	unikátny kľúč ..._UK

Relačné databázy
E-R (logický) model vs. Fyzický model - transformácia



Relačné databázy
E-R (logický) model vs. Fyzický model



Relačné databázy
Normalizácia

- Odstraňuje redundanciu, zjednodušuje prácu s databázou

- 3 normálne formy

Relačné databázy
Normalizácia – 1. NF

- Prvá normálová forma (1NF)
  - Trieda spĺňa prvú normálovú formu, pokiaľ sú všetky jej atribúty atomické, čiže ďalej nedeliteľné.
  - Jeden stĺpec triedy nesmie obsahovať viac druhov údajov, ale musí obsahovať skalárnu hodnotu.
  - Pokiaľ trieda nespĺňa podmienky prvej normálovej formy, je potrebné ju rozložiť.

Meno	Telefón
Ján Novák	09123465789,1234567890 ,2345678901
Linda Dvorská	3456789012,4567890123, 4578962130
Libor Kudláček	5678901234

Meno	Priezvisko	Telefón1	Telefón2	Telefón3
Ján	Novák	0912345678	1234567890	234567890
Linda	Dvorská	3456789012	4567890123	
Libor	Kudláček	5678901234		

Relačné databázy
Normalizácia – 2. NF

Druhá normálová forma (2NF)

- Trieda je v druhej normálovej forme, ak spĺňa podmienky prvej normálovej formy a každý jej atribút, ktorý nepatrí do žiadneho kľúča úplne závisí od každého kľúča.

- Týka sa len tabuliek, ktoré majú viac primárnych kľúčov, pri tabuľkách s jedným PK je splnená automaticky

- Dosahuje sa dekompozíciou

Relačné databázy
Normalizácia – 2. NF

- Pred

Produkt	Dodávateľ	E-mail	Cena
Cement	Cementáreň Bystré	cement@bystre.sk	270
Cement	Stavebniny Prešov	stavba@presov.sk	220
Piesok	Pieskovňa Svidník	piesok@svidnik.sk	45
Piesok	Stavebniny Prešov	stavba@presov.sk	35

- Po

Produkt	Kód dodávateľa	Cena
Cement	002	270
Cement	007	220
Piesok	004	45
Piesok	007	35

Kód	Dodávateľ	E-mail
002	Cementáreň Bystré	cement@bystre.sk
004	Pieskovňa Svidník	piesok@svidnik.sk
007	Stavebniny Prešov	stavba@presov.sk

Relačné databázy
Normalizácia – 3. NF

Tretia normálová forma (3NF)

- Trieda je v tretej normálovej forme ak spĺňa podmienky druhej normálovej formy a zároveň v nej neexistujú tranzitívne závislosti nekľúčových atribútov.

Priezvisko	Rodné číslo	PSČ	Mesto
Novák	120315/7895	05612	Praha
Dvorská	125504/1254	04811	Brno
Kudláček	740731/8520	02174	Plzeň

- Existuje závislosť medzi PSČ a mesto
- Riešením by bolo PSČ odstrániť a vyhľadávať ho z osobitnej tabuľky (otázkou je, či je to výhodné)

Multidimenzionálna databáza

- Organizuje údaje do multidimenzionálnej štruktúry
- Podklad pre získavanie agregovaných a sumarizovaných údajov -> informácií
- Výhody:
  - Rýchly komplexný prístup k veľkému množstvu údajov
  - Prístup k multidimenzionálnym aj relačným dátovým štruktúram
  - Možnosť komplexných analýz
  - Silné schopnosti pre modelovanie a prognózy
- OLAP (On-Line Analytical Processing) – štruktúry údajov a analytické služby slúžiace pre analýzu veľkého množstva údajov

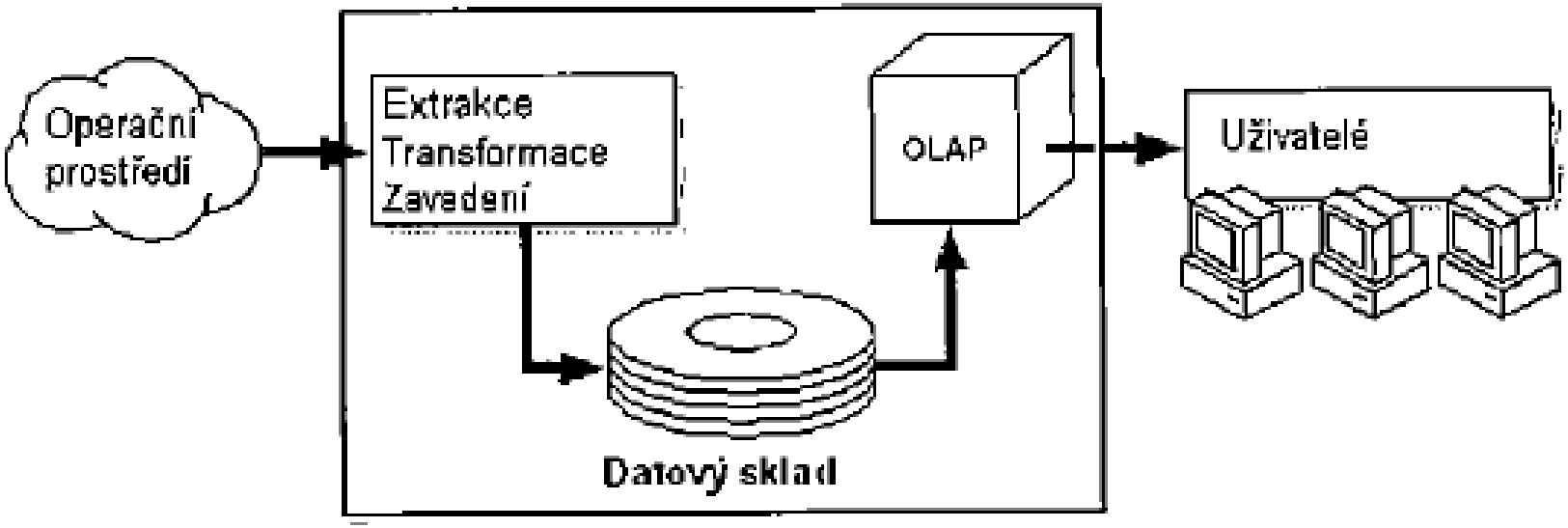
Multidimenzionálny databázový model

- Väčšina údajov uložených v relačnej databáze v dvojrozmerných tabuľkách
- Výsledkom agregácie a analýzy býva multidimenzionálna dátová štruktúra – kocka
- Výsledkom analýz sú väčšinou reporty, ktoré slúžia ako podklady pre manažérov, ktorí na základe nich prijímajú rozhodnutia

Dátový sklad

- B.Inmon: Dátový sklad je úložisko subjektovo orientovaných (orientované na predmet záujmu, nie na aplikáciu, v ktorej boli vytvorené), integrovaných(jednotná terminológia), časovo premenlivých (viac časových úsekov) a nemenných(dáta sa nemenia a neodstraňujú) historických dát použitých pre podporu rozhodovania
- Atomické a sumárne dáta

Dátový sklad



Dátový trh

- Podmnožina dátového skladu, určená pre menšie organizačné jednotky firmy (účtovníctvo, marketing)
- Vytvárajú sa pre jednoduchšiu orientáciu používateľov, aj kvôli bezpečnosti citlivých údajov

Budovanie dátového skladu

- ***Metóda „veľkého tresku“***
  - Vytvorenie dátového skladu implementáciou jedného projektu
  - Fázy:
    - Analýza požiadaviek zo strany podniku
    - Vytvorenie dátového skladu
    - Vytvorenie prístupu buď priamo, alebo cez dátové trhy
  - Nie je to veľmi dobrý prístup, pretože proces vytvárania dátového skladu je dynamický, môžu sa zmeniť požiadavky zákazníka, ako aj využiteľné technológie
  - Je potrebné dlhé čakanie na návratnosť investície

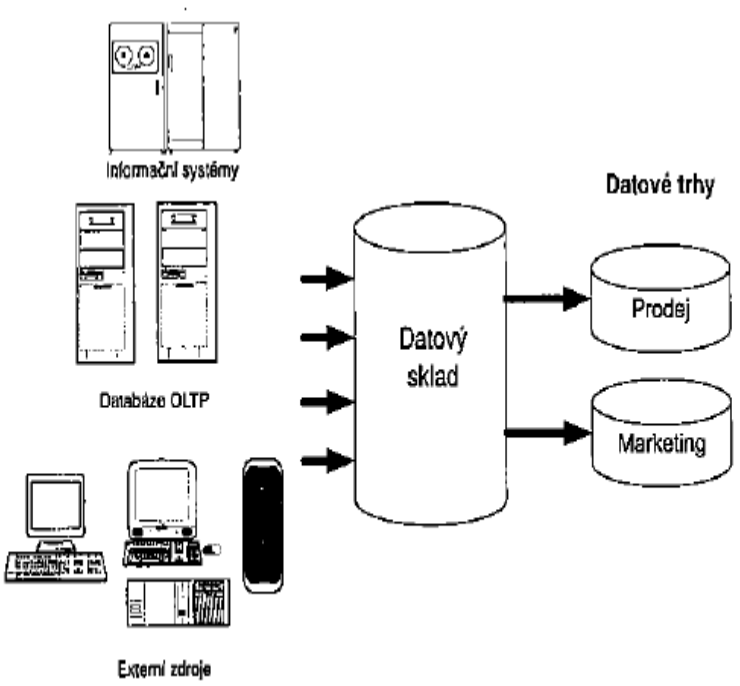
Budovanie dátového skladu II.

- ***Prírastková metóda***
  - Vytvorenie dátového skladu po etapách
  - Vytvorenie niekoľkých dátových trhov, postupné dopĺňanie projektu
  - Iteratívny proces
  - Výhody:
    - Kontinuita budovaného projektu
    - Implementácia škálovateľnej, rozšíriteľnej architektúry
    - Rýchlejšia návratnosť investícií
  - Fázy budovania:
    - stratégia
    - Definícia
    - analýza
    - návrh
    - Zostavenie
    - produkcia

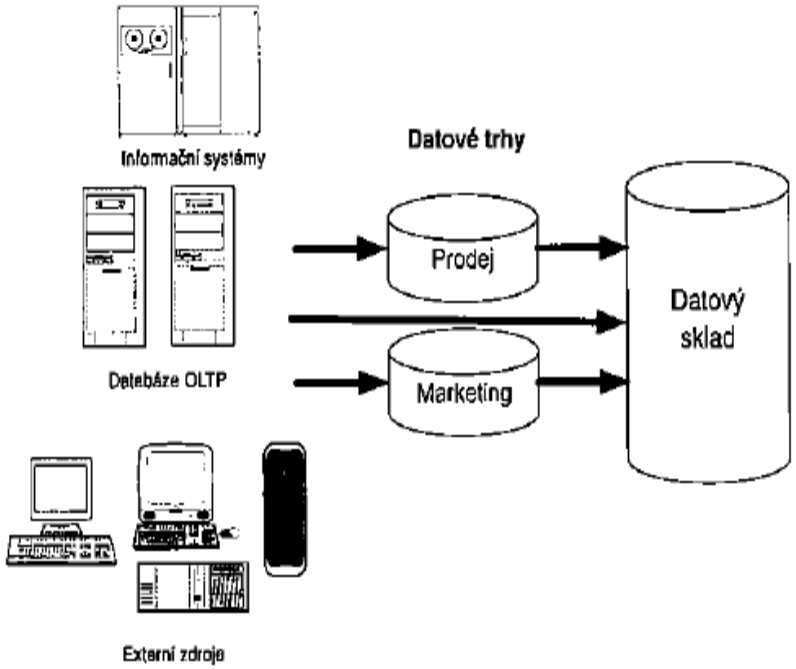
Budovanie dátového skladu III.

– Metódy budovania prírastkovou metódou:

Zhora nadol



Zdola nahor



OLAP

- On-Line Analytical Processing
- Voľne definované princípy tvoriace rámec pre podporu rozhodovania

12 pravidiel OLAP

- Multidimenzionálny konceptuálny pohľad
- Transparentnosť
- Dostupnosť
- Konzistentné vykazovanie
- Architektúra klient - server
- Generická dimenzionalita
- Dynamické ošetrenie riedkych matíc
- Podpora pre viac užívateľov
- Neobmedzené krížové dimenzionálne operácie
- Intuitívna manipulácia s údajmi
- Flexibilné vykazovanie
- Neobmedzené dimenzie a úrovne agregácie

Analýza OLAP – Pojmy
Fakt

- Fakt – numerická merná jednotka
- Tabuľka faktov
  - najväčšia tabuľka v databáze, obsahuje veľký objem dát
  - Sledovaný fakt a odkazy (cudzie kľúče) na tabuľky jednotlivých dimenzií

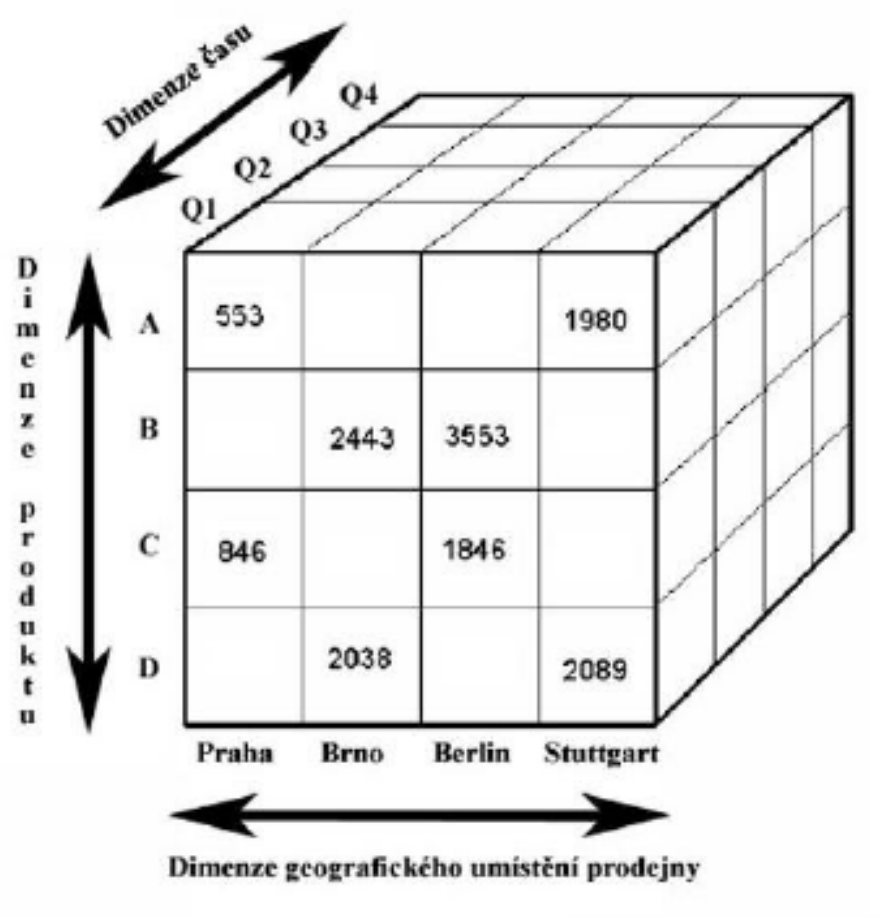
Analýza OLAP – Pojmy
Tabuľka Faktov

POCET_ID	POCET	FK_PREDAJNA_ID	FK_DATUM_ID	FK_PRODUKT_ID	FK_MIESTO_PREDAJNE_ID
1	6.8	84	1	15	2
2	32.8	84	1	16	2
3	154.4	84	1	17	2
4	1.9	84	1	18	2
5	4.6	84	1	22	2
6	66.6	84	1	24	2
7	74.2	84	1	25	2
8	6.3	84	1	26	2
9	8.7	84	1	1	2
10	19.7	84	1	29	2
11	29	84	1	3	2
12	29.8	84	1	30	2
13	2	84	1	4	2
14	2.5	84	1	5	2
15	2.4	84	1	6	2
16	1.4	84	1	7	2

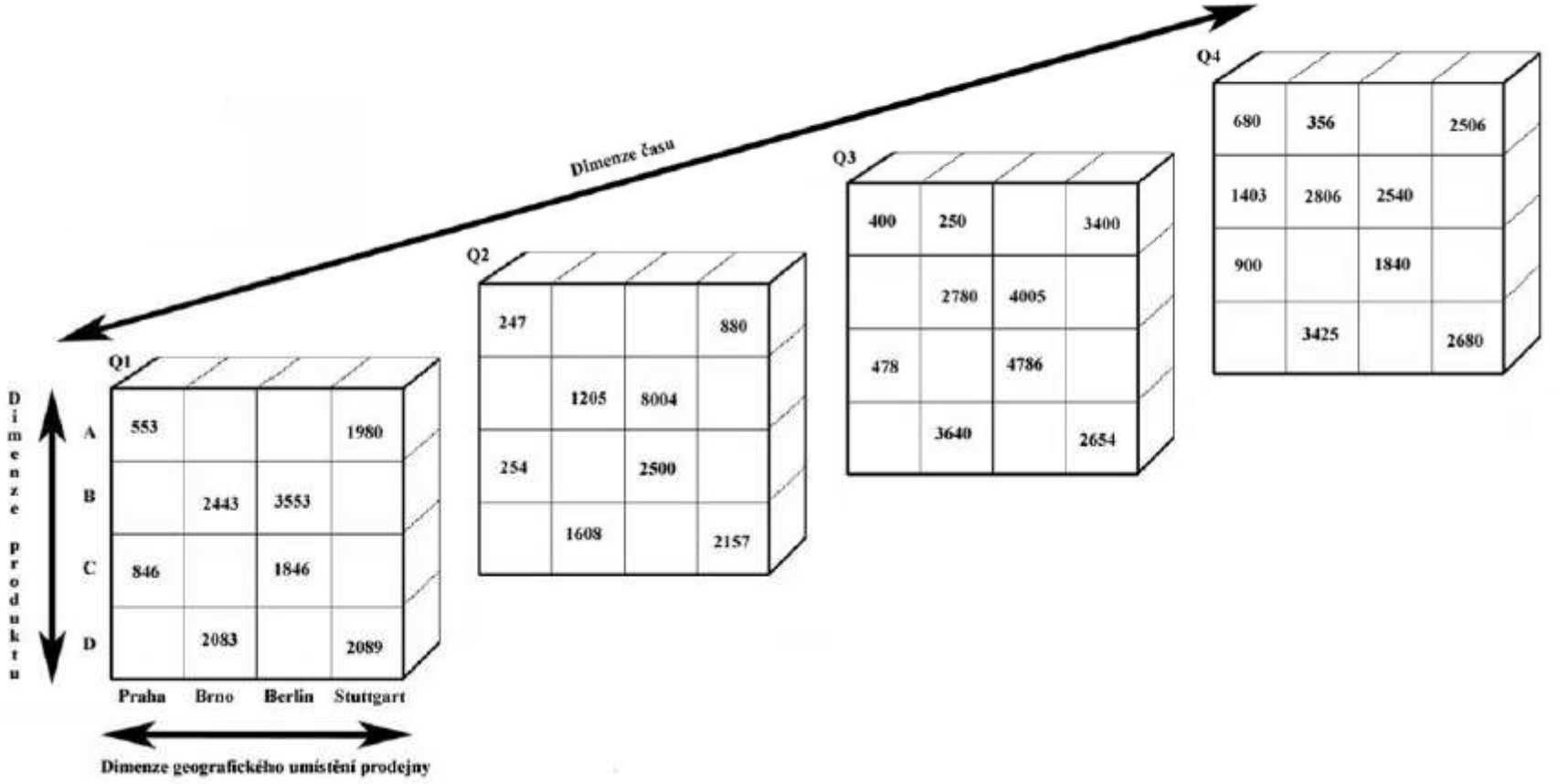
Analýza OLAP – Pojmy II.
Dimenzia

- Dimenzia – obsahuje logicky alebo hierarchicky usporiadané údaje
- Tabuľky dimenzií
  - Menšie než tabuľky faktov
  - Dáta sa v nich často nemenia (výnimkou napr. dáta o zákazníkoch)
  - Často sa používajú časové, produktové a geografické dimenzie

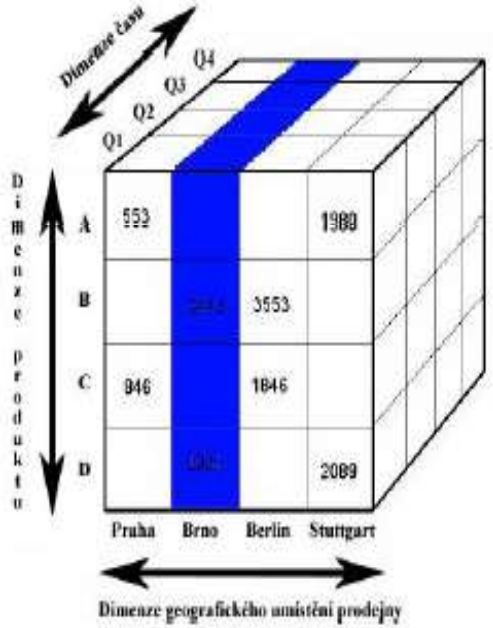
Analýza OLAP – pojmy III.
Dimenzia



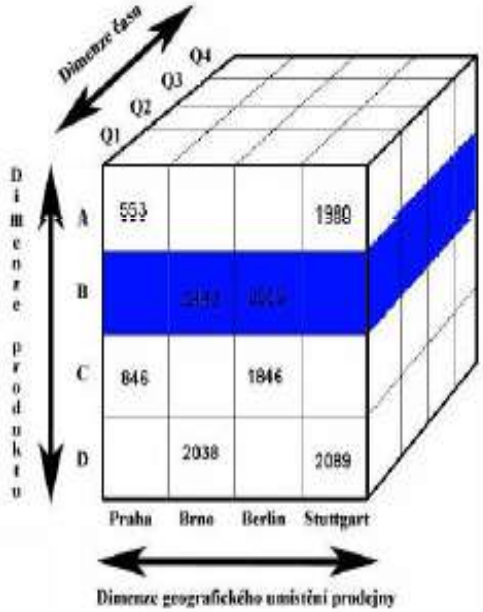
Analýza OLAP – pojmy IV.
Dimenzia



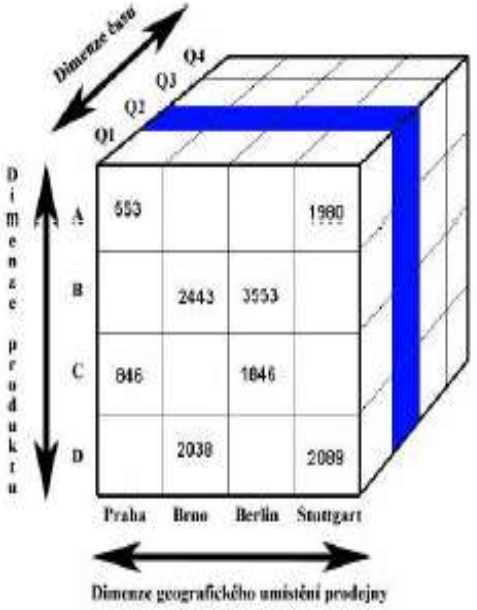
Analýza OLAP – pojmy V.
Dimenzia



A - Analýza údajů podle geografických kritérií



B - Analýza údajů podle produktu



C - Analýza údajů podle časových kritérií

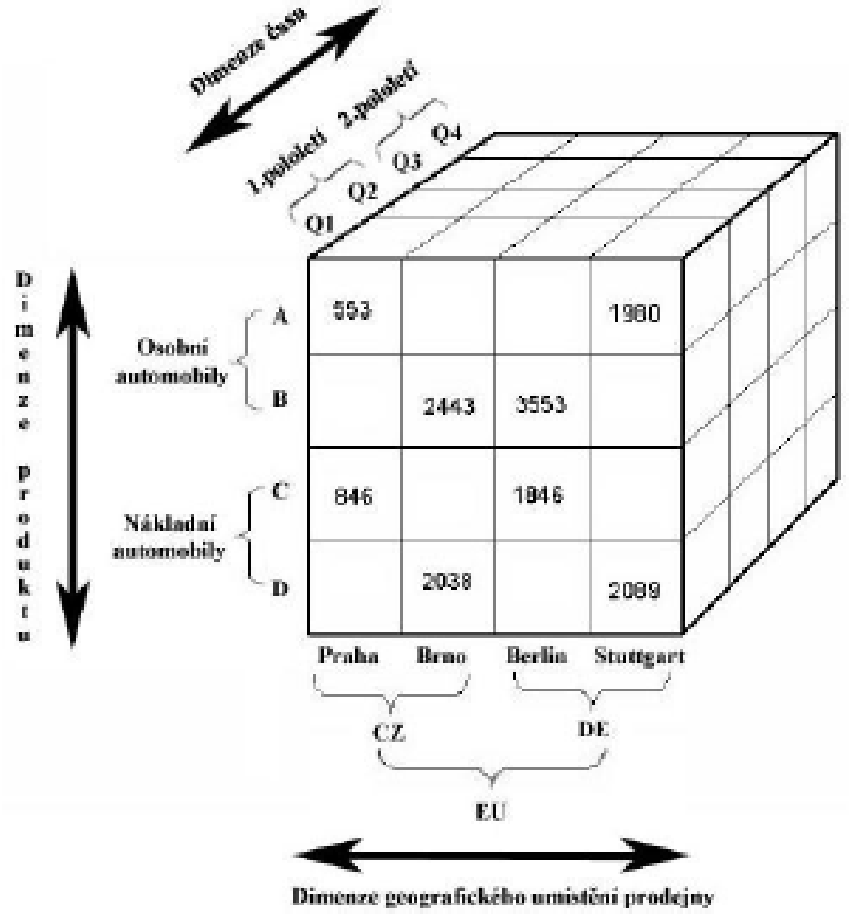
Analýza OLAP – pojmy VI.
Hierarchie

- Hierarchie – dimenzie majú obvykle stromovú (hierarchickú štruktúru)
- Napr.:

Dimenzia/ úrovne	REGIÓN	PRODUKT	ČAS
NAJVYŠŠIA	Kontinent	Druh	Rok
....	Štát	Kategória	Kvartál
....	Kraj	Subkategória	Mesiac
NAJNIŽŠIA	Mesto	Názov	Deň

Analýza OLAP – pojmy VII.
Hierarchie II.

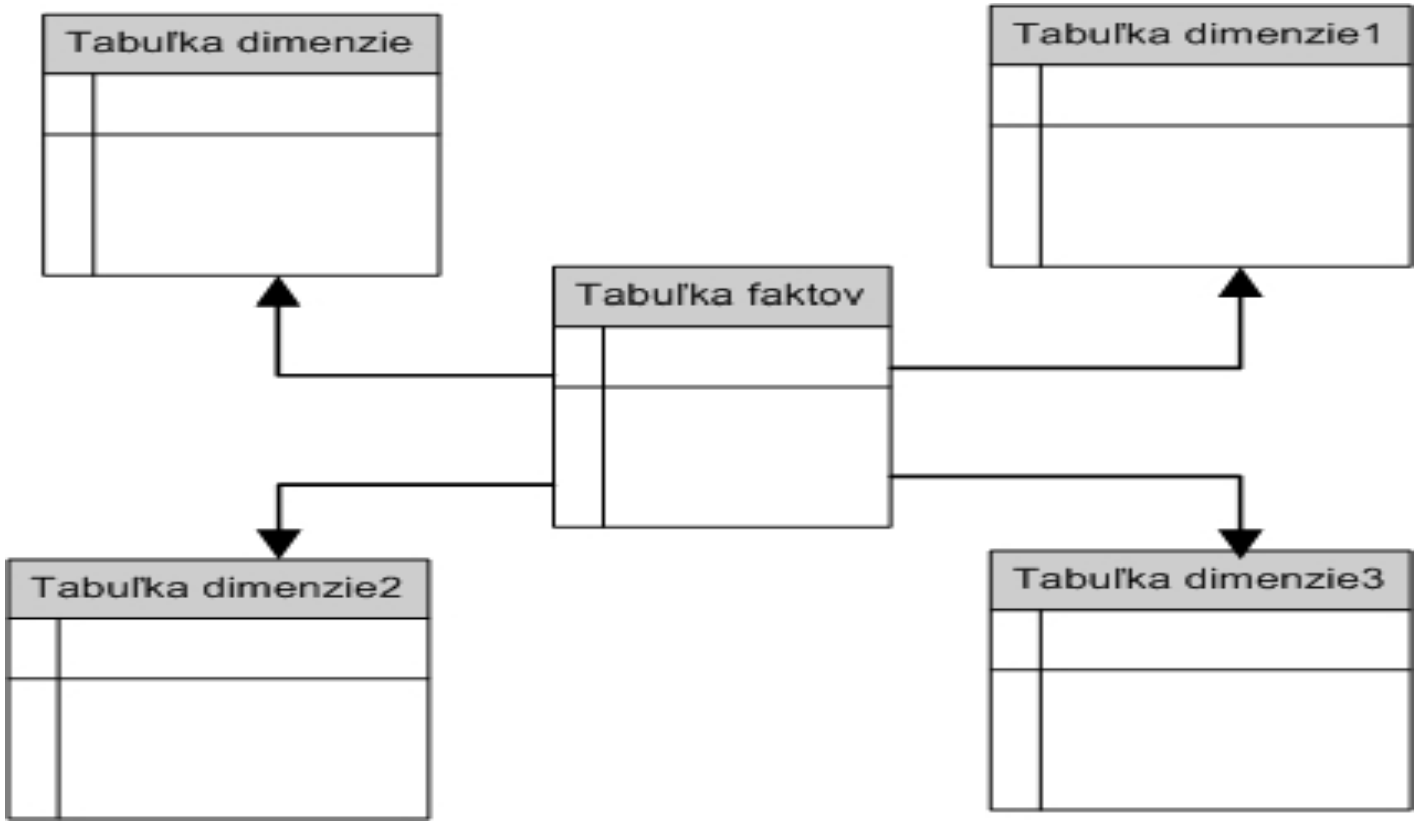
- ***Roll-up*** – vystupovanie na vyššie úrovne danej dimenzie
- ***Drill-down*** – zostupovanie na nižšie úrovne
- ***Slice*** – projekcia cez jednu dimenziu
- ***Dice*** – projekcia cez viac dimenzií



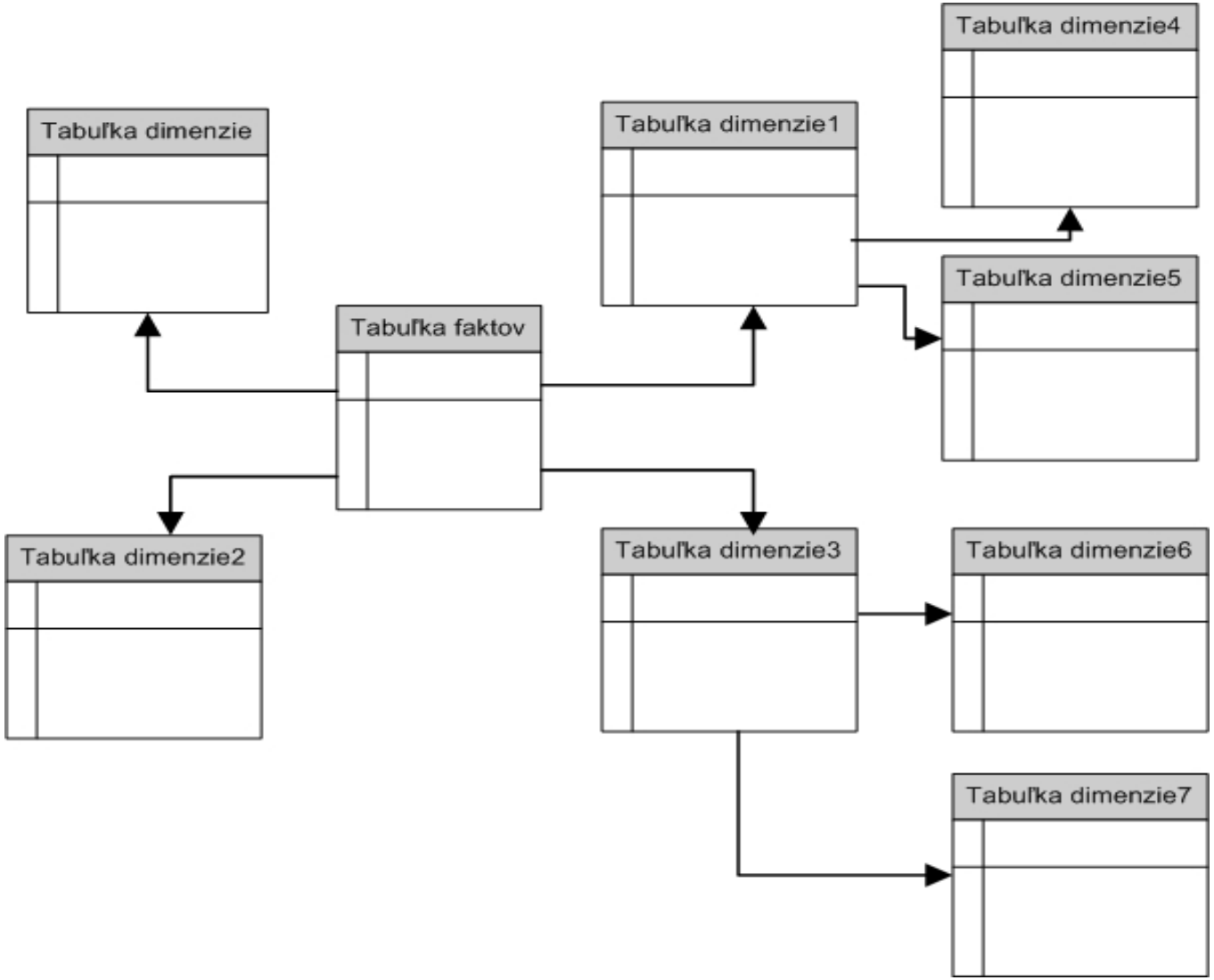
Analýza OLAP – pojmy VIII.
Schémy

- Schéma – dimenzionálny model, ktorý má topologické usporiadanie
- Typy schém:
- Hviezdicová (star) schéma
- Schéma snehovej vločky (snowflake)
- Súhvezdie faktov

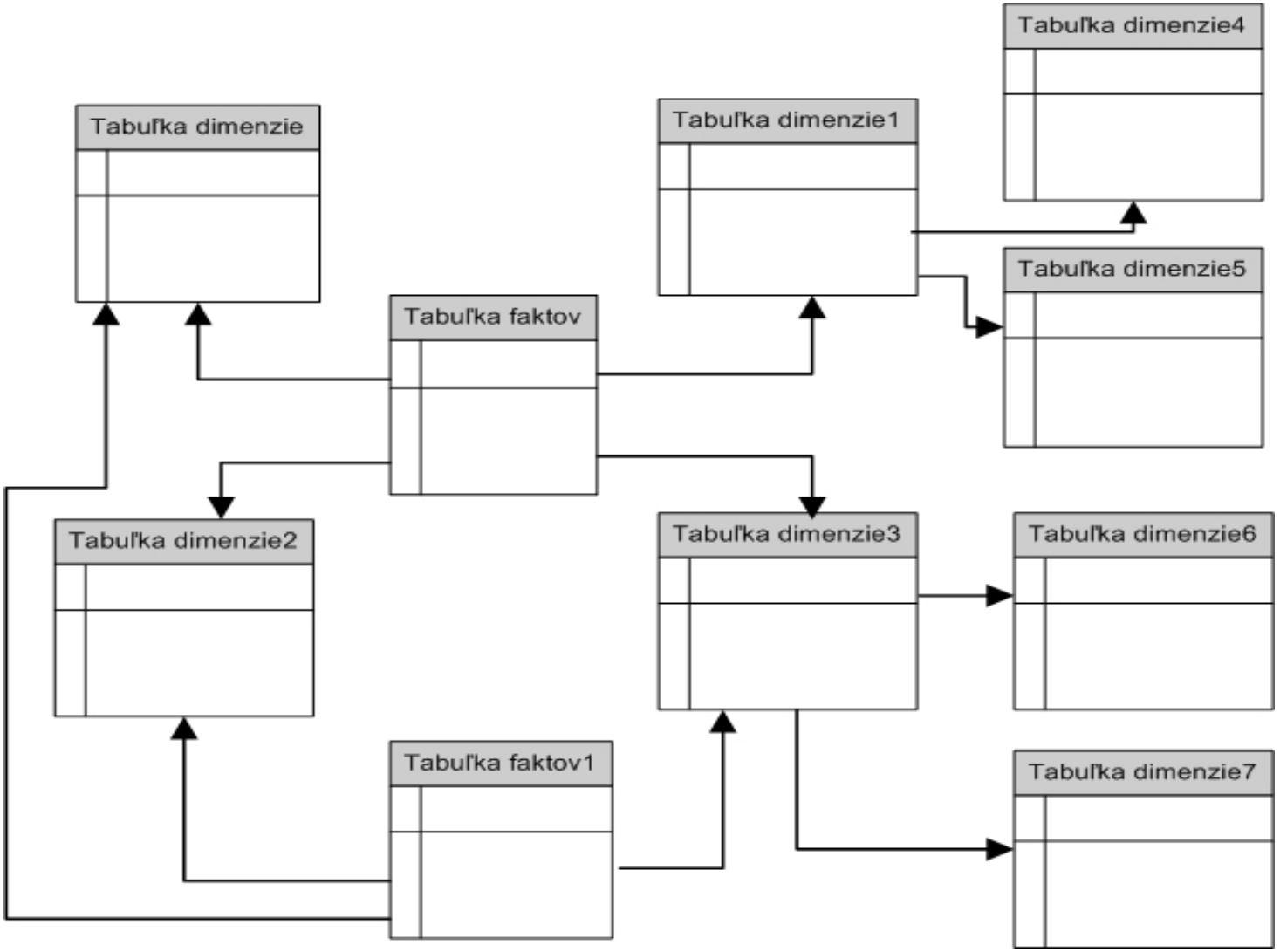
Analýza OLAP – pojmy IX.
Star schéma



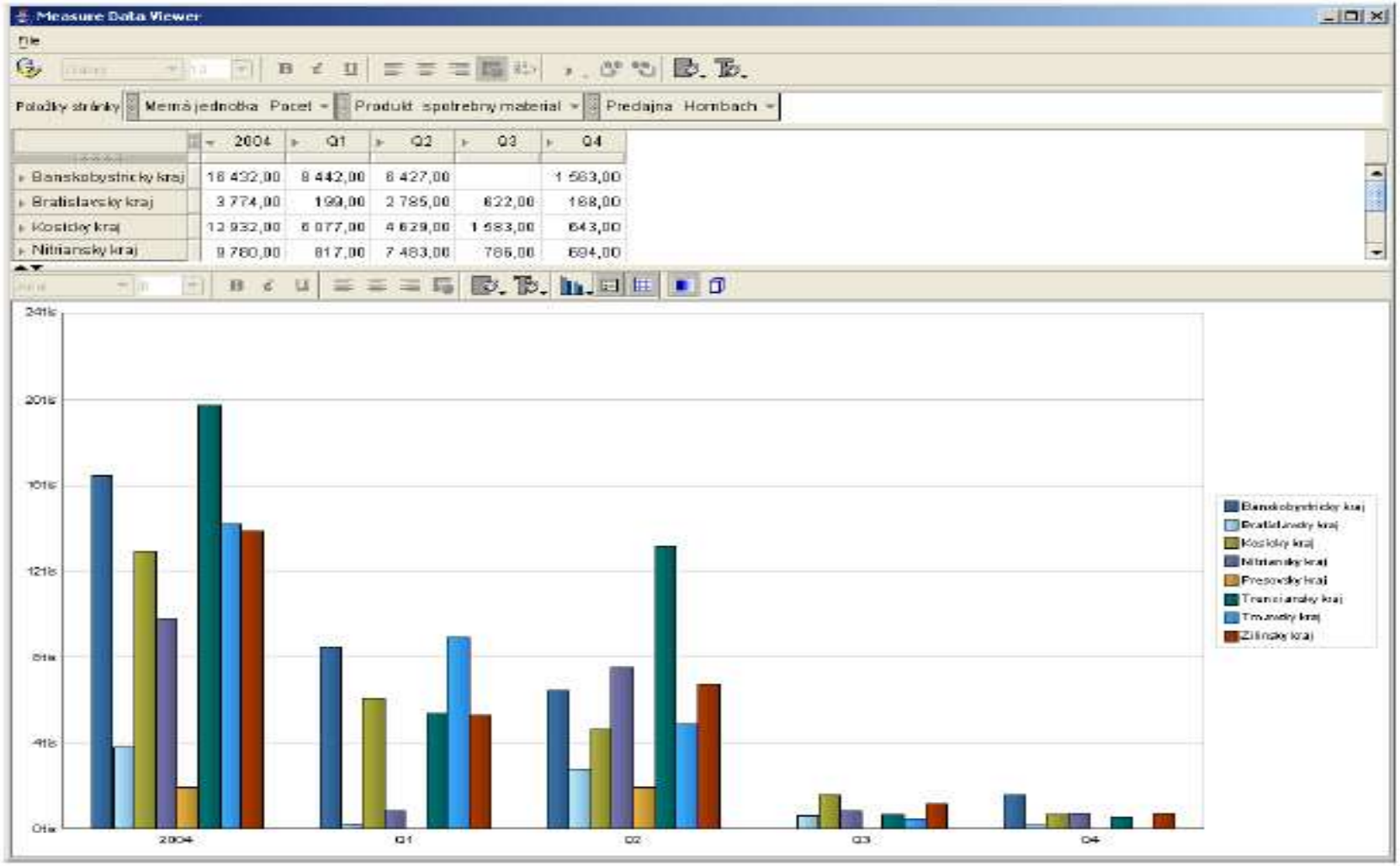
Analýza OLAP – pojmy X.
Snowflake schéma



Analýza OLAP – pojmy XI.
Súhvezdie faktov



Analýza OLAP
Ukážka



Ročný a kvartálny pohľad spotrebného materiálu v predajni Hornbach.

Relačný vs. Multidimenzionálny model

Relačný model
+ potenciál odborníkov vo firmách, rutinná práca s modelom
+využitie v transakčných databázach, aj dátových skladoch

Relačný model
- Absencia konkrétnych analytických nástrojov
- Potenciálne kapacitné obmedzenia
- Obmedzenia dostupnosti dát v rozumnou čase

Multidimenzionálny model
+ rýchly a komplexný prístup k veľkému množstvu údajov
+ prístup k multidimenzionálnym aj relačným dátovým štruktúram
+ možnosť komplexných analýz
+ schopnosti pre modelovanie a prognózy

Multidimenzionálny model
- Problémy pri zmenách dimenzií, menšia flexibilita
- Vyššie kapacitné nároky

Zdroje

1. OCELÍKOVÁ, Eva - LIGUŠOVÁ, Jana - TAKÁČ, Ladislav: Databázové systémy a jazyk SQL / - 1. vyd - Košice : FEI TU - 2013. - 165 s.. - ISBN 978-80-553-1266-8.
2. NOVOTNÝ, Ota; POUR, Jan; SLÁNSKÝ, David. Business Intelliegence : Jak využít bohatství ve vašich datech. Praha : Grada Publishing, a.s., 2005. 256 s. ISBN 80-247-1094-3.
3. LACKO, Luboslav. Databáze: datové sklady, OLAP a dolování dat s příklady v Microsoft SQL Serveru a Oracle . Brno : Computer Press, a.s., 2003. 2003 s. ISBN 80-7226-969-0.
4. Computer Science 831: Knowledge Discovery in Databases[online]. 2007 [cit. 2011-07-25]. Introduction to Data Cubes. Dostupné z WWW: .
5. VLÁŠKOVÁ, Markéta. Návrh hybridního úložiště dat [online]. Plzeň, 2006. 65 s. Diplomová práce. Západočeská univerzita v Plzni. Dostupné z WWW: .
6. http://czm.fel.cvut.cz/vyuka/A4M33CPM/Download/DatoveKostky.pdf

ĎAKUJEM ZA POZORNOSŤ