

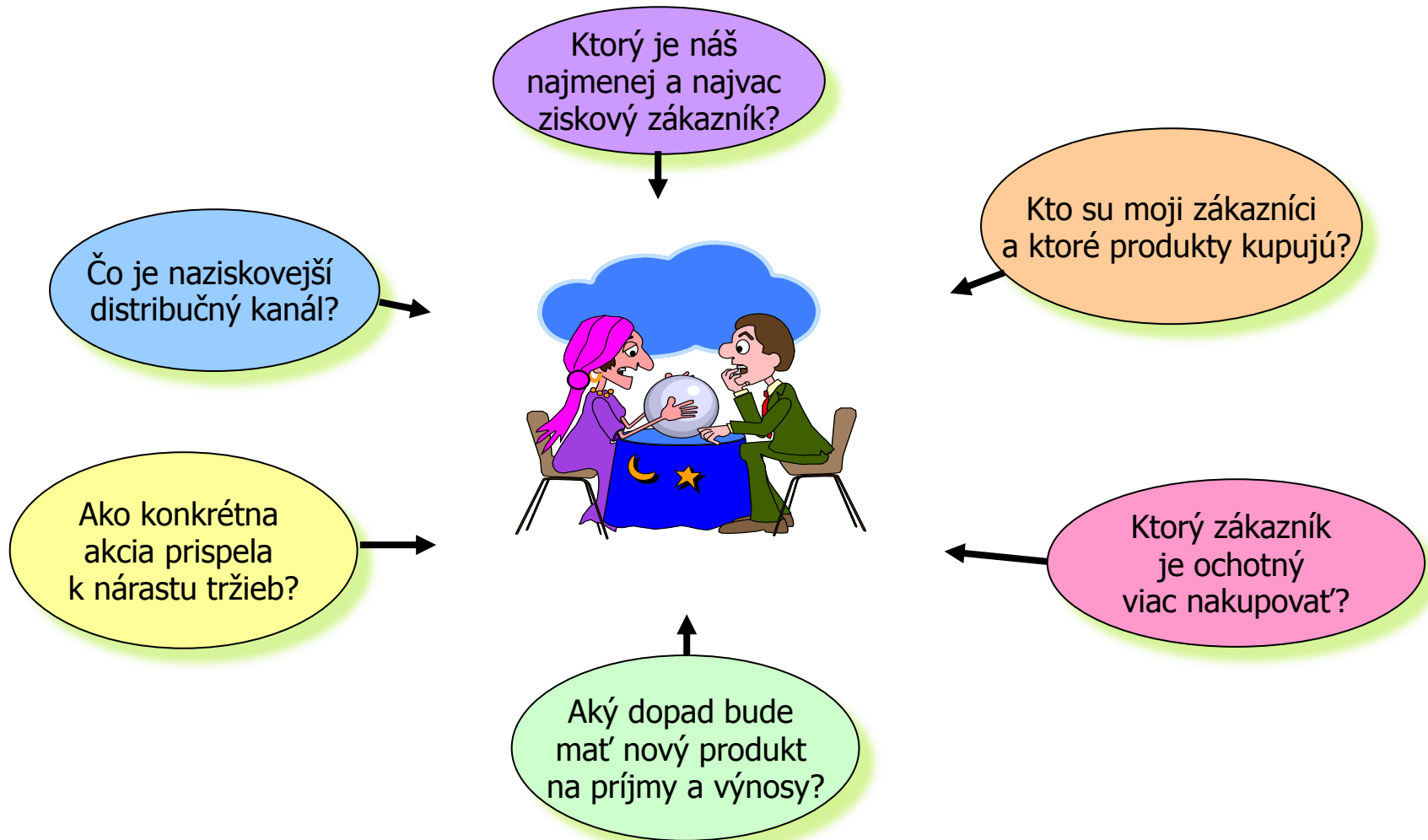


Dátové sklady

Ing. Vladimír Hanušniak



Výrobca potrebuje vedieť...



Čo je to dátový sklad

- Databáza na podporu rozhodovania, ktorá je udržiavaná mimo produkčnej databázy.
- Podpora **spracovania informácií** prostredníctvom konsolidovanej databázy s historickými dátami.
- “Dátový sklad je **subjektívne-orientovaná, integrovaná, časovo stála** kolekcia dát pre podporu manažmentu.” - W. H. Inmon
- Data warehousing – proces **vytvorenia** a používania dátového skladu

DWH – Subjektívne-orientovaný

- organizovaná okolo hlavných subjektov (zákazník, produkt, predaj)
- zameraná na modelovanie a analýzu dát pre riadiacich pracovníkov na báze transakčných dát
- poskytuje **jednoduchý** a **stručný** pohľad na čiastkové subjekty pomocou **vylúčenia dát**, ktoré sú **nepoužiteľné** pre rozhodovanie

DWH - Integrovaný

- Konštruovaný ako integrovaný priestor združujúci heterogénne zdroje
 - Relačné DB, textové súbory, on-line záznamy
- Sú implementované techniky integrácie a dátovej kvality
 - Zaistenie konzistencie v názvoch, atribútoch medzi rôznymi dátovými zdrojmi
 - Konvertovanie zdrojových dát podľa definícií

DWH – Časovo nezávislý

- Časový horizont pre dátový sklad je výrazne dlhší ako údaje uchovávané v operačných systémoch
 - Operačné databázy: aktuálna hodnota dát
 - Data warehouse data: poskytuje informácie z historickej perspektívy (napr., posledných 5-10 rokov)
- Každá kľúčová štruktúra v DW obsahuje časovú dimenziu, ale nie všetky dáta majú časovú dimenziu

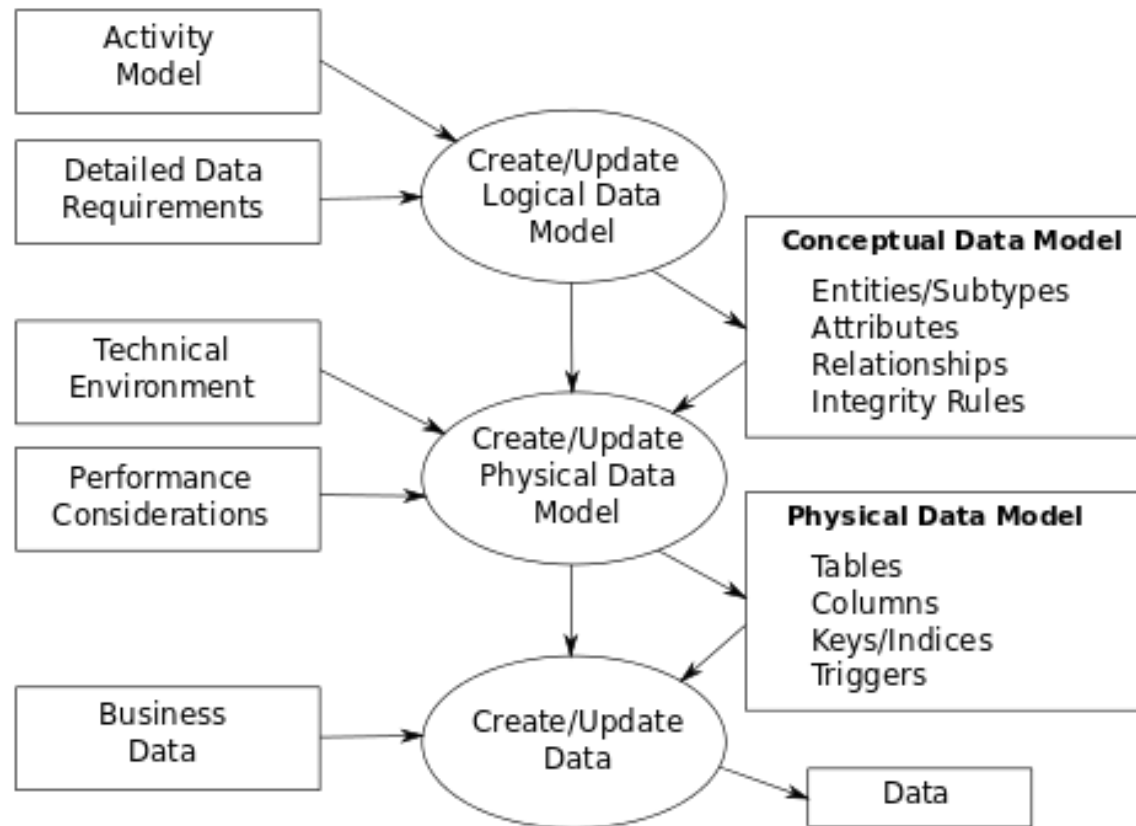
DWH – Stály

- Fyzicky oddelené uloženie dát transformovaných z operatívnej evidencie
- Operatívne uchovávanie dát v dátovom sklade nenastáva
 - DW nepotrebuje transakčné spracovanie a obnovu
 - Používa dve operácie:
 - Inicializačné nahratie dát
 - Prístup k dátam

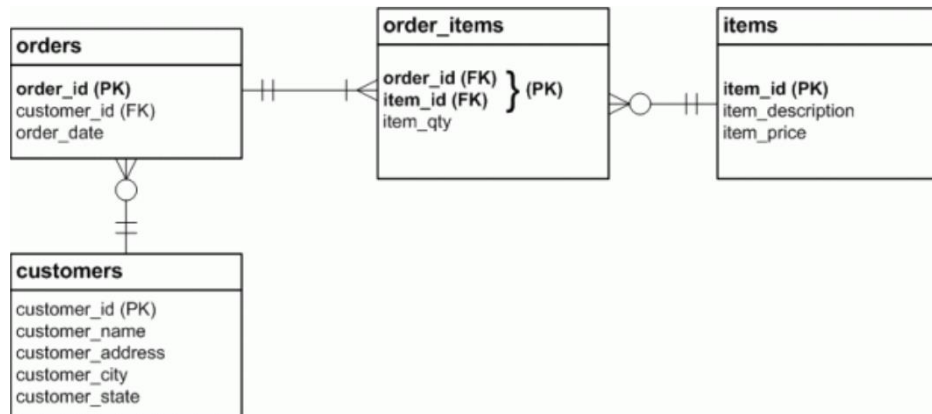
Dátové modelovanie

- Proces vytvorenia dátového modelu informačného systému použitím formálnych modelovacích techník
- Fázy návrhu DB:
 - Konceptuálny (koncept. model, Chenn)
 - Logický (normalizácia)
 - Fyzický (závislé na implementačnom prostr.)

Dátové modelovanie



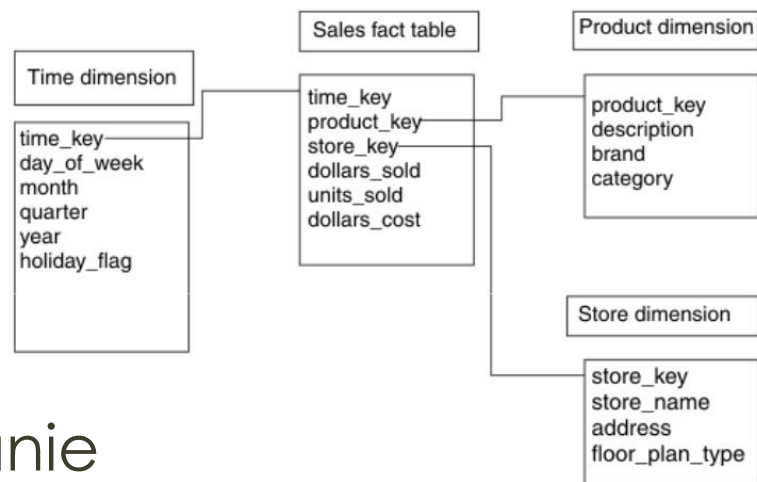
Relačný model v 3NF



- Jednoduché nahranie dát
 - Prenos dát zo zdrojových súborov a ich integrácia
- Zložité dotazy
 - Množstvo JOIN operácií
 - Ťažšie pochopiteľný bežnými používateľmi
- Model pre centrálnu úložisko dát podľa Inmona

Dimenzionálny dátový model

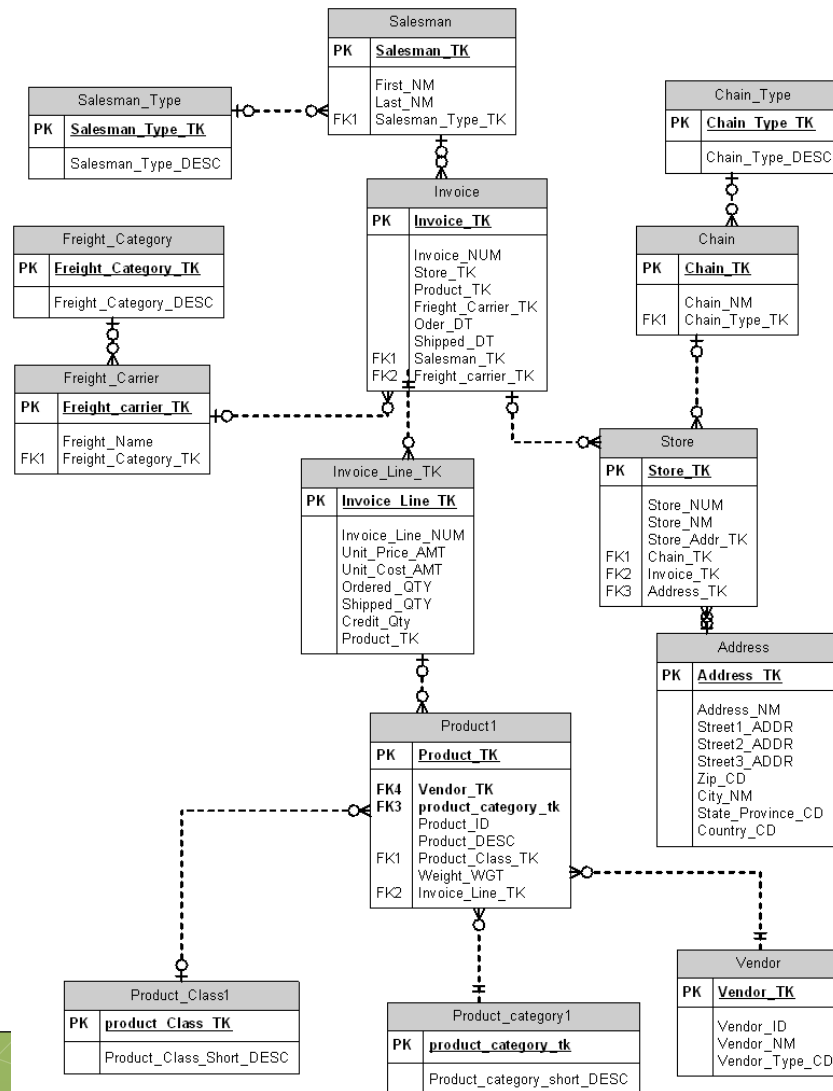
- ◉ Odporúčaný pre DWH
- ◉ Zložité ETL
 - ◉ Transformácie dát
 - ◉ Integrácia, ...
- ◉ Jednoduché reportovanie
 - ◉ Ľahšie pochopiteľný
 - ◉ Rýchlejšie analytické dotazy
- ◉ Model pre datamarty – podľa Inmona aj Kimballa



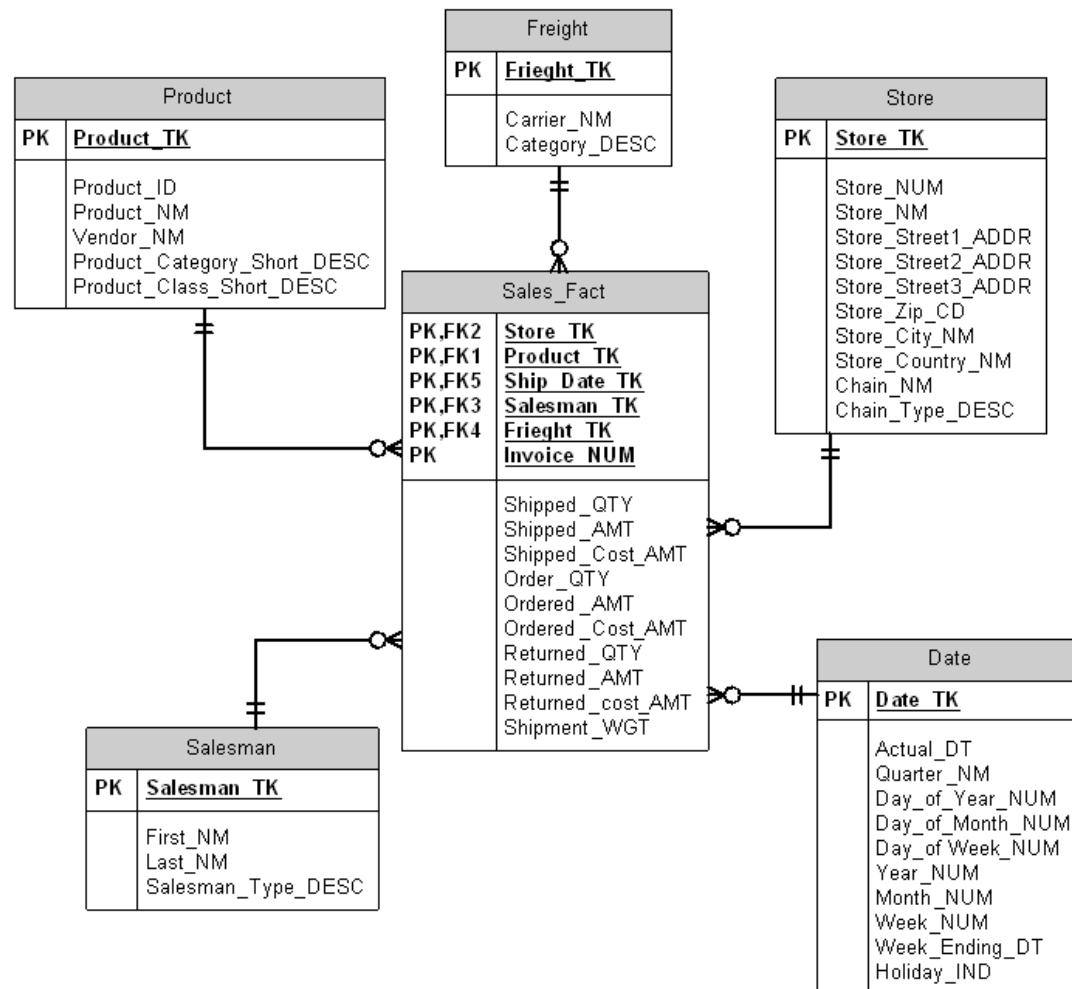
Relačný vs. dimenzionálny model

- Relačný dátový model v 3-NF
 - Odstránenie duplicitných dát – zmenšenie počtu záznamov
 - Zvýšenie počtu tabuliek
 - Prepojenie cez cudzie kľúče a tabuľky relácií
 - Efektívny insert/update, menej efektívne dotazovanie
- Dimenzionálny dátový model (nie je v 3NF)
 - Adaptácia relačného modelu
 - Faktové a dimenzionálne tabuľky
 - Denormalizovaný, duplicitné dáta
 - Menší počet tabuliek
 - Efektívne dotazovanie

Entitno-relačný dátový model



Dimenzionálny dátový model



DWH vs. DBMS

- OLTP (on-line transaction processing)
 - Hlavné úlohy tradičných DBMS
 - Denno-denné operácie: nákup, sklady, bankové transakcie, výroba, mzdy, účtovanie atď.
- OLAP (on-line analytical processing)
 - Hlavne využívané v dátových skladoch
 - Analýza dát a podpora rozhodovania
- Odelené funkcie (OLTP vs. OLAP):
 - Užívateľsky a systémovo orientovaný: zákazník vs. trh
 - obsah: aktuálne, detailné vs. historické, konsolidované
 - Design: ER + aplikačný vs. star + subjektívny
 - Pohľad: skutočnosť, lokálne vs. evolučne, integrované
 - Vzory prístupu: upravovanie vs. read-only, ale komplexné dotazy

OLTP a OLAP model

	OLTP	OLAP
Zdroj dát	Operatívne dáta: OLTP systémy sú vlastné zdroje dát.	Konsolidované dáta: dáta pochádzajú z viacerých OLTP databáz.
Účel	Kontrola základných činností.	Pomoc pri plánovaní, riešení problémov a podpora rozhodovania.
Cieľ	Vkladanie dát do systému, automatizácia operácií alebo procesov.	Získavanie informácií zo systému, poskytovanie optimálnych informácií pre rozhodovanie.
Čo reprezentujú dáta	Pohľady aktuálnych dát, obmedzené historické dáta.	Multidimenzionálne pohľady na dáta, historické dáta
Vkladanie a aktualizácia dát	Krátke a rýchle.	Užívateľ má možnosť iba čítať. Periodické obnovovanie, dopĺňanie dát.
Dotazy	Relatívne štandardizované a jednoduché. Návratová hodnota je niekoľko záznamov.	Komplexné dotazy umocňované agregáciami
Rýchlosť	Veľmi rýchla, zlomky sekúnd až sekundy.	Záleží na množstve požadovaných dát. Sekundy až hodiny. Možnosť optimalizácie vytvorením indexov.
Transakcie	Veľké množstvo transakcií.	Malý počet zložitých dotazov.

OLTP a OLAP model

	OLTP	OLAP
Štruktúra	ER dátový model, vysoko normalizovaný s viacerými tabuľkami.	Menej de-normalizovaných tabuliek, Star alebo Snowflake dátový model.
Zaistenie integrity dát	Dátovým modelom.	Pri zavádzaní (transformácií) dát .
Indexovanie	Nadmerné indexovanie spomaľuje operácie modifikácie.	Môže byť použité neobmedzene.
Backup a Recovery	Pravidelné zálohovanie, strata dát znamená výraznú škodu.	Možnosť obnovenia údajov z OLTP dát v prípade straty, alebo poškodenia.
Použitie	Štruktúrované, opakované.	Ad hoc
Užívatelia	Úradník, IT profesionál.	Znalostný analytik
Orientácia	Customer - oriented	Market - oriented

OLTP vs. OLAP zhrnutie

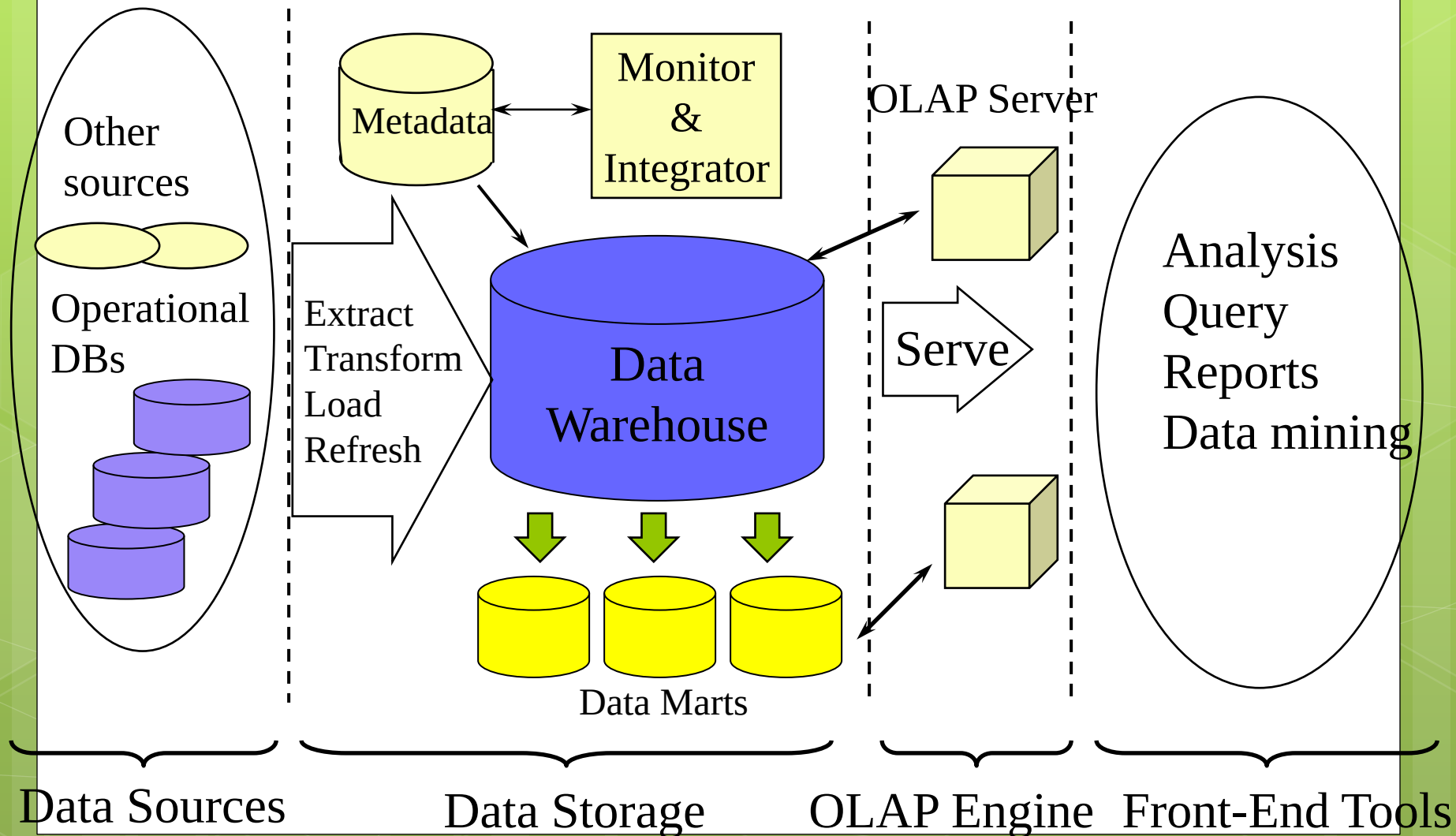
- OLTP sa používajú na beh firmy



- OLAP pomáha optimalizovať fungovanie



DWH architektúra



ETL (Extract, Transform, Load)

- Vykonáva sa automatizovane v pravidelných časových intervaloch
 - Denne (v noci)
 - Týždenne
- Extrakcia
 - Extrakcia dát z rozličných zdrojov a formátov
 - Validácia dát (správna forma/hodnota)

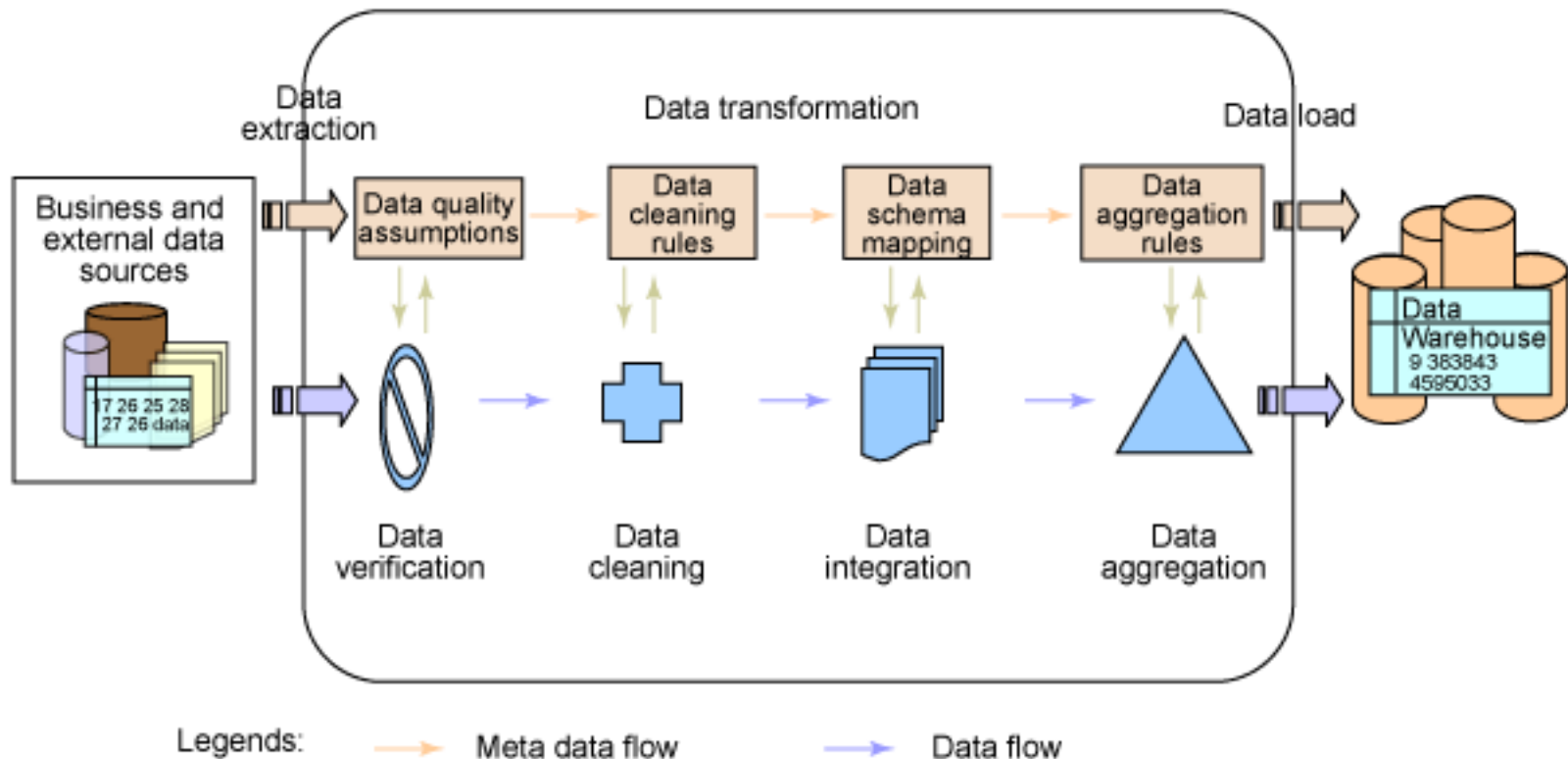
ETL (Extract, Transform, Load)

- Transformácia
 - Na dáta je aplikovaná séria funkcií a pravidiel pre prípravu ich nahratiu do DWH
 - Čistenie dát
 - Iba správne dáta môžu byť nahraté
 - Voľba iba určitých stĺpcov
 - Kódovanie („Male“ to „M“)
 - Odvodenie nových hodnôt (hodnota_predaja = cena * počet_predaných_produkto)
 - Agregácia (sumarizácia)
 - Rozdelenie stĺpcov (napr. na dátum a čas)

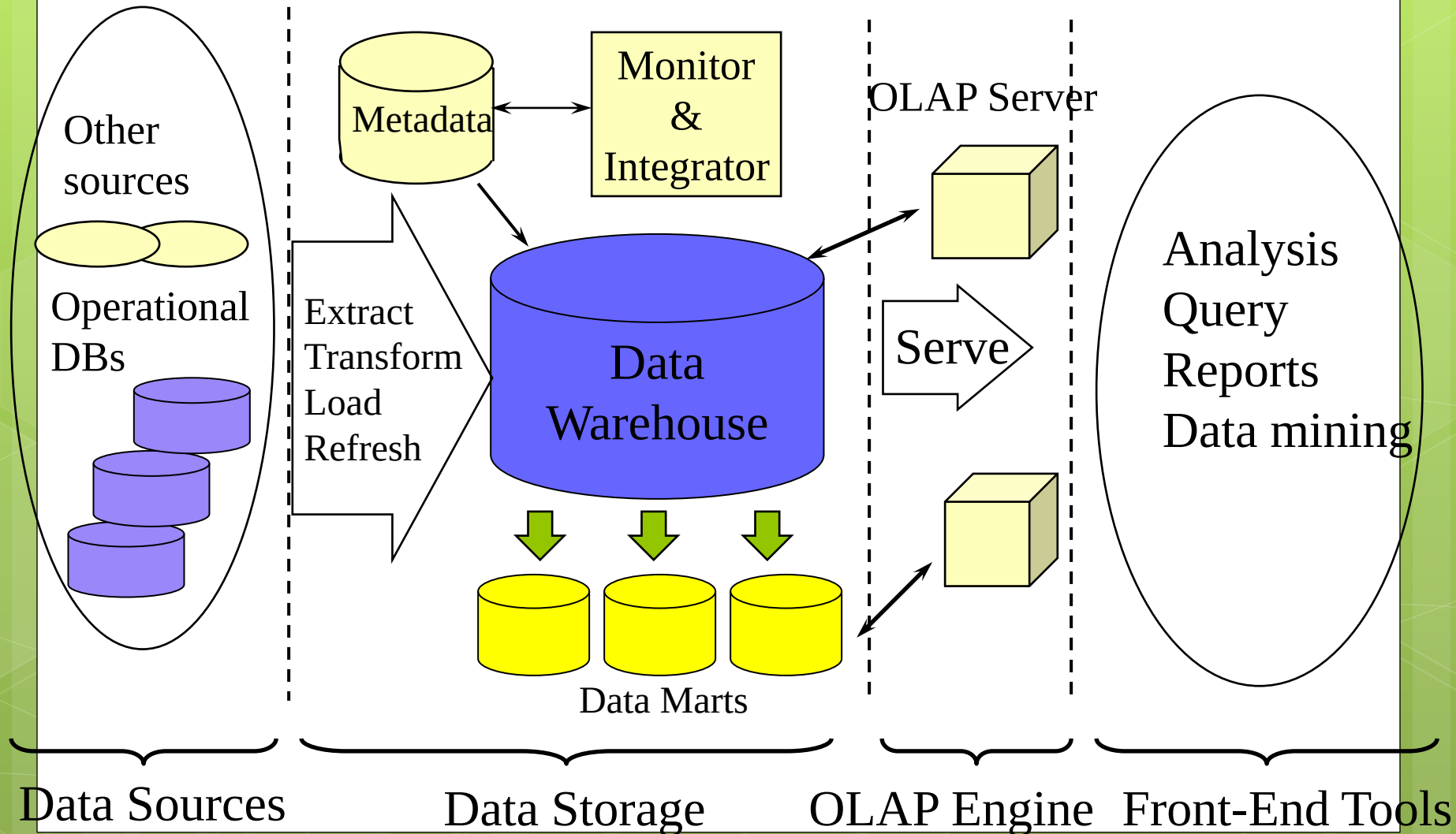
ETL (Extract, Transform, Load)

- Nahrание
 - Zabezpečuje nahrание extrahovaných a transformovaných dát do cieľového systému (DWH)

ETL (Extract, Transform, Load)



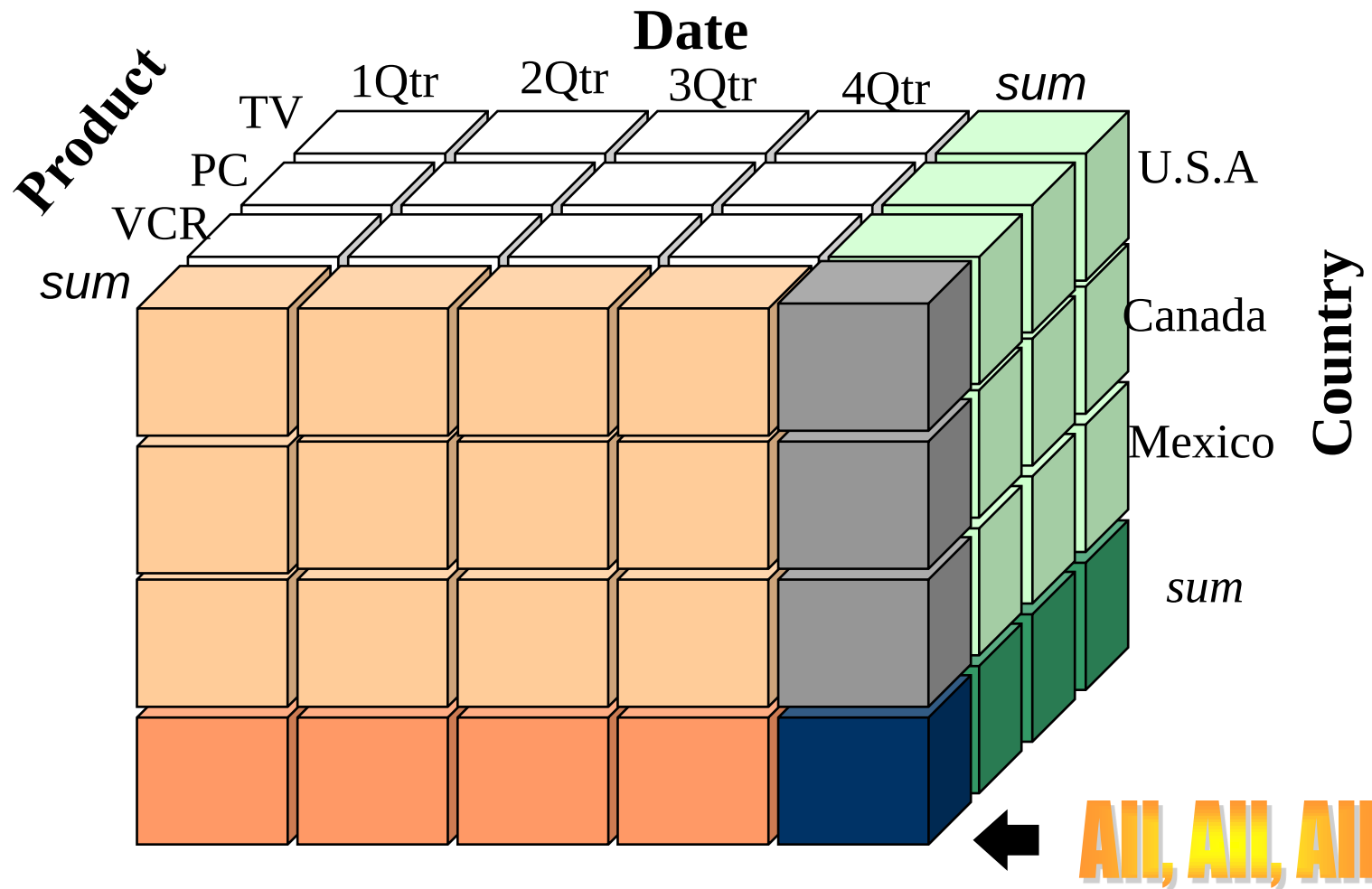
DWH architektúra



Multidimenziálna databáza

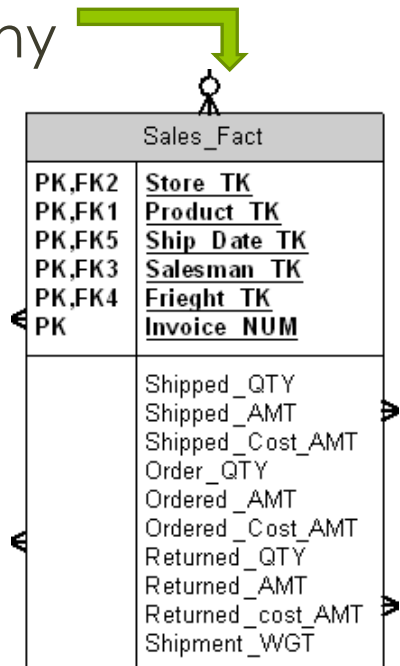
- DW je postavené na multidimenziom dátovom modeli, ktorý zobrazuje dáta z dátových kociek
- Dátová kocka, ako napr. predaj, dovoľuje modelovať informácie z rôznych dimenzií
- Dimenzionálne tabuľky položka (meno_položky, značka, typ), alebo čas(den, týždeň, mesiac, štvrťrok, rok)
- Tabuľka faktov obsahuje hodnoty (napr. predaná_cena) a kľúče vytvárajú reláciu s dimenziou

Multidimenzionálna kocka (OLAP kocka)



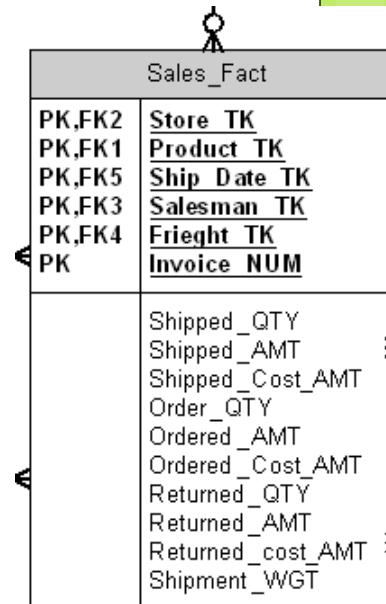
Tabuľky faktov

- Tabuľka faktov obsahuje dva druhy atribútov:
 - klúčové atribúty – sú to cudzie kľúče z príslušných tabuliek dimenzií. Primárny kľúč tabuľky faktov je zložený zo všetkých jej klúčových atribútov
 - neklúčové atribúty – sú to samotné fakty, ktoré sa sledujú pre každú kombináciu neklúčových atribútov



Tabuľky faktov

- Popisuje konkrétny subjekt podnikania
- Fakt, ktorý je uložený v tabuľke, sa sleduje pre jednotlivé kombinácie dimenzií
- Tabuľka faktov obsahuje veľké objemy údajov
- Maximálny objem tabuľky sa určuje ako karteziánsky súčin hodnôt primárnych kľúčov tabuliek dimenzií
- Zdrojom údajov sú bázické dáta z podnikových informačných zdrojov
- Údaje sa pri transformácii do tabuľky faktov sumarizujú a agregujú na požadovanej úrovni
- Každú tabuľku faktov charakterizuje granularita, ktorá vyjadruje stupeň agregácie faktov oproti zdrojovým údajom

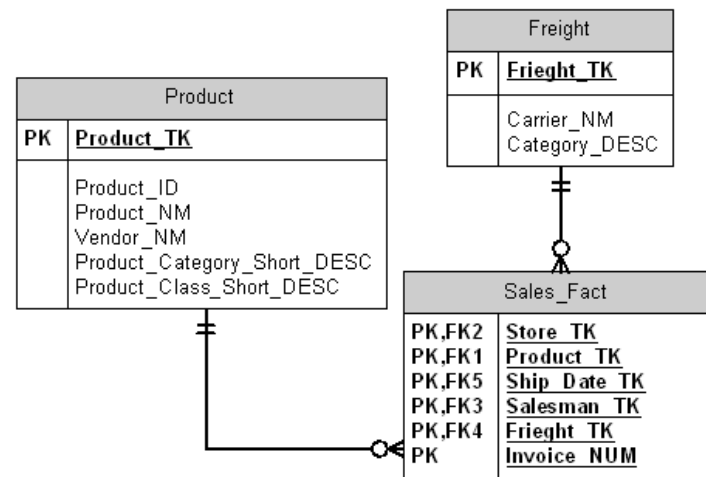


The diagram shows a fact table 'Sales_Fact' with dimensions and measures. Dimensions are listed on the left with their primary key (PK) and foreign key (FK) relationships, and measures are listed on the right.

Sales_Fact	
PK,FK2	<u>Store</u> TK
PK,FK1	<u>Product</u> TK
PK,FK5	<u>Ship Date</u> TK
PK,FK3	<u>Salesman</u> TK
PK,FK4	<u>Frieght</u> TK
PK	<u>Invoice</u> NUM
	Shipped_QTY
	Shipped_AMT
	Shipped_Cost_AMT
	Order_QTY
	Ordered_AMT
	Ordered_Cost_AMT
	Returned_QTY
	Returned_AMT
	Returned_cost_AMT
	Shipment_WGT

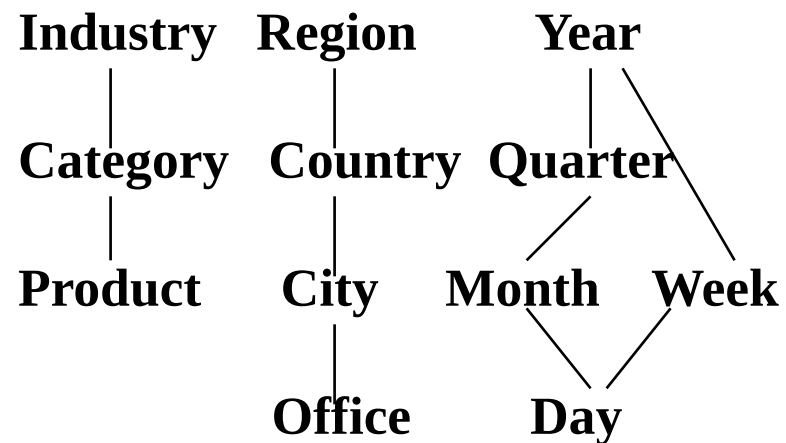
Tabuľky dimenzií

- Ohraničujú výber konkrétneho faktu z tabuľky faktov.
- Obsahujú atribúty, ktoré popisujú vybranú dimenziu a charakterizujú význam jednotlivých faktov v tabuľke faktov.
- Príkladmi dimenzií sú:
 - produkt
 - čas
 - geografická lokalita

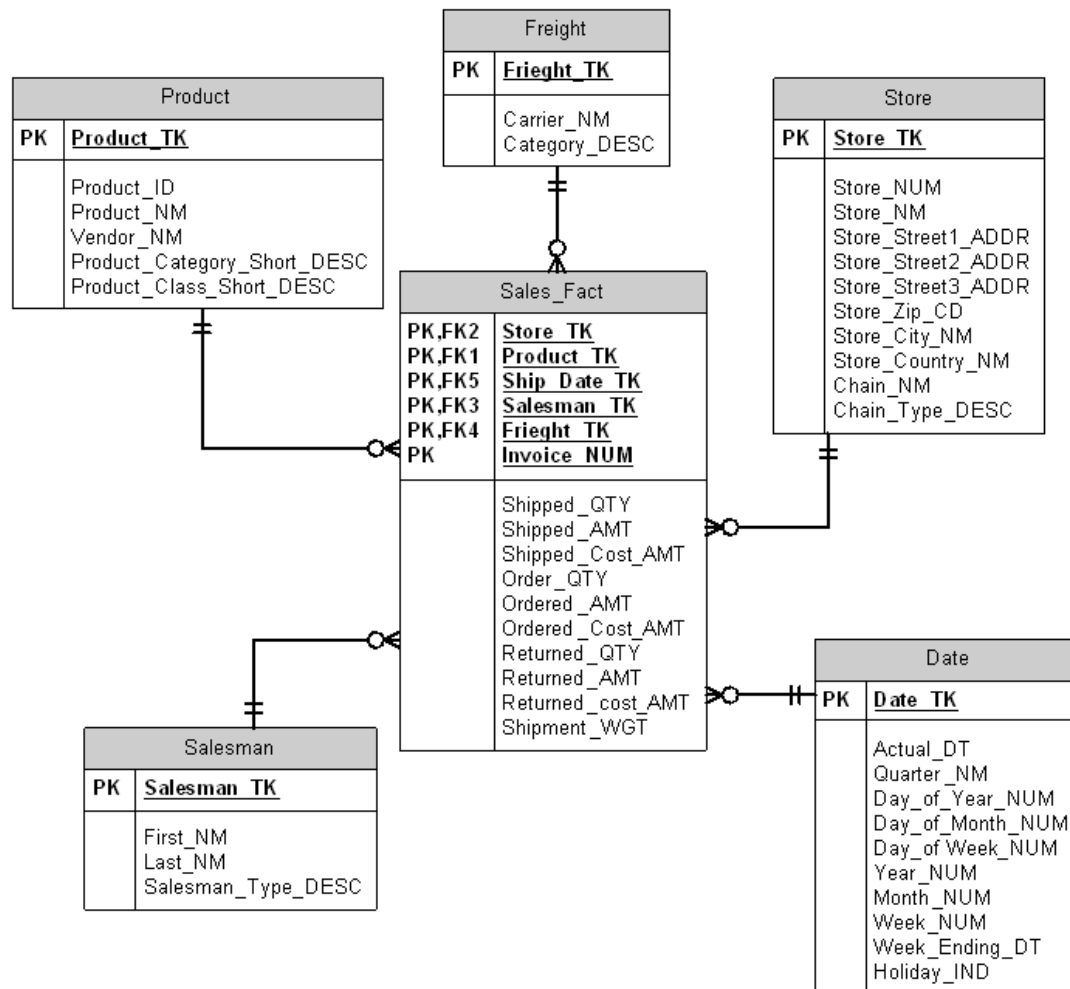


Tabul'ky dimenzií

- Granularita dimenzií



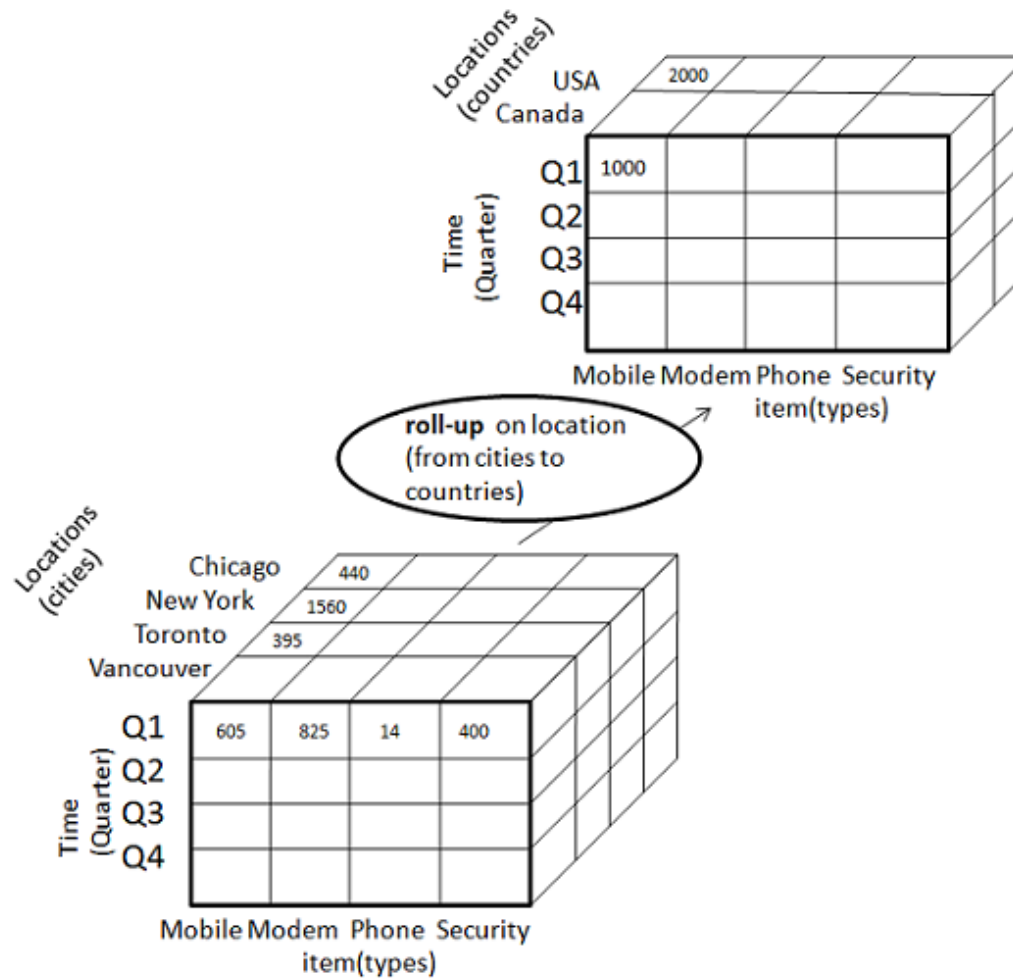
Dimenzionálny dátový model



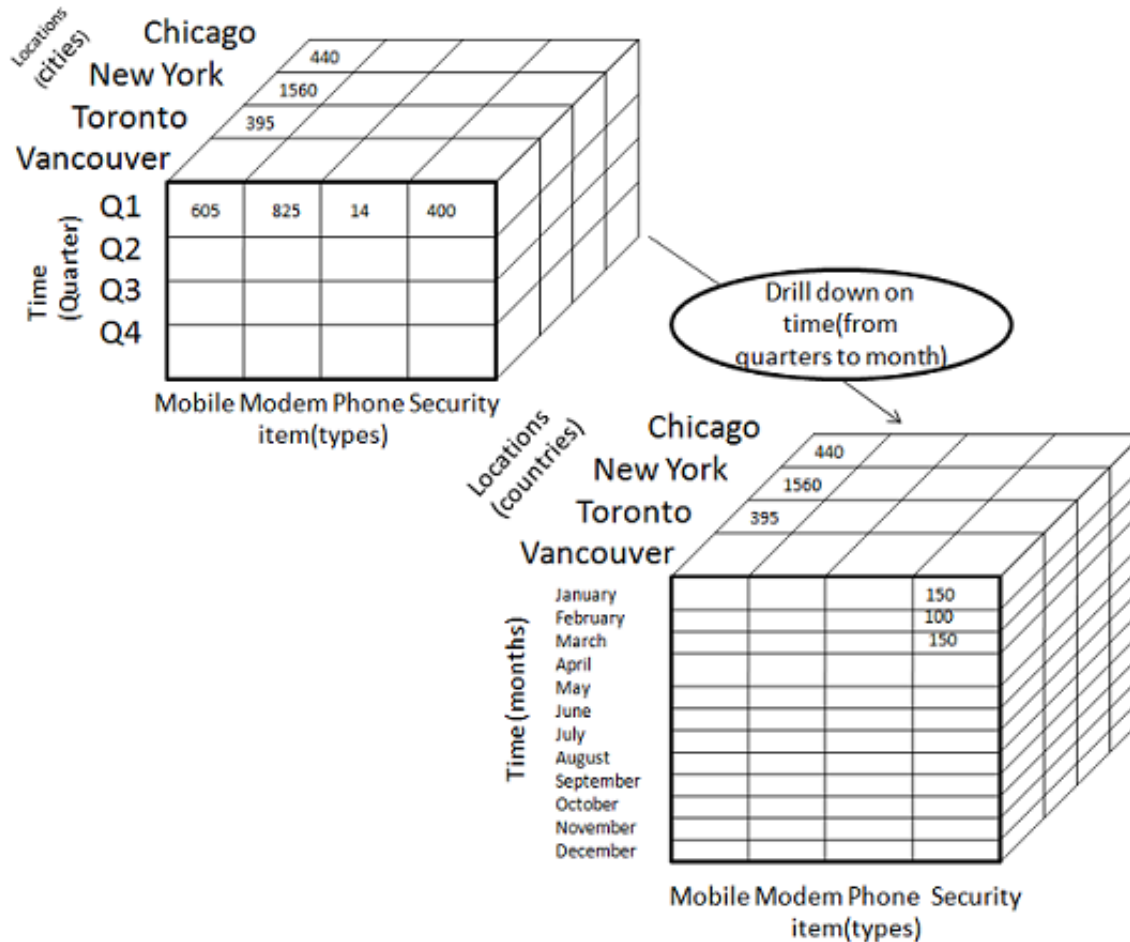
OLAP operácie

- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

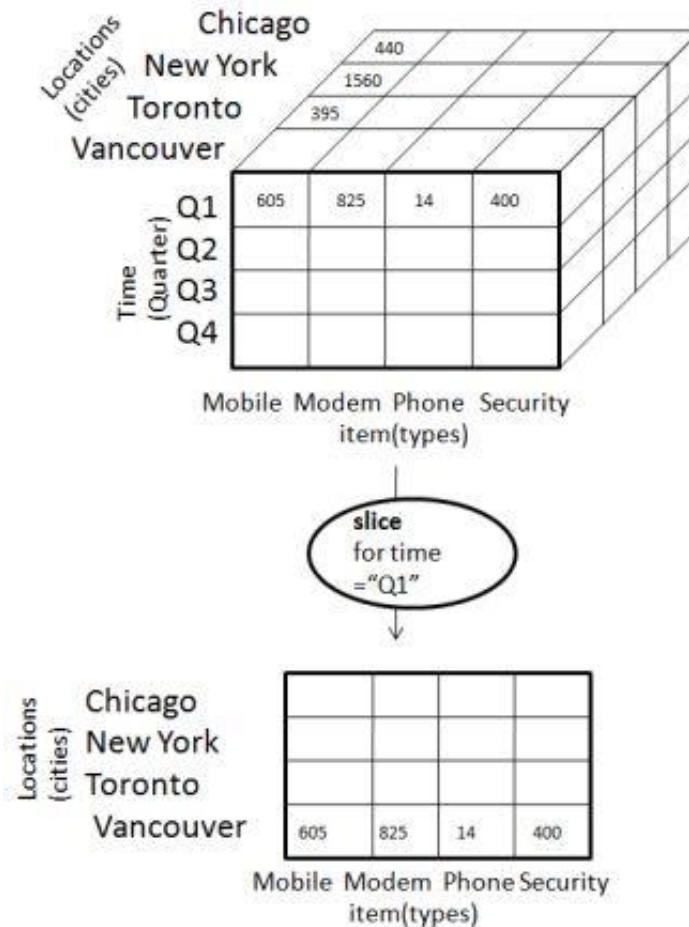
OLAP operácie Roll-up



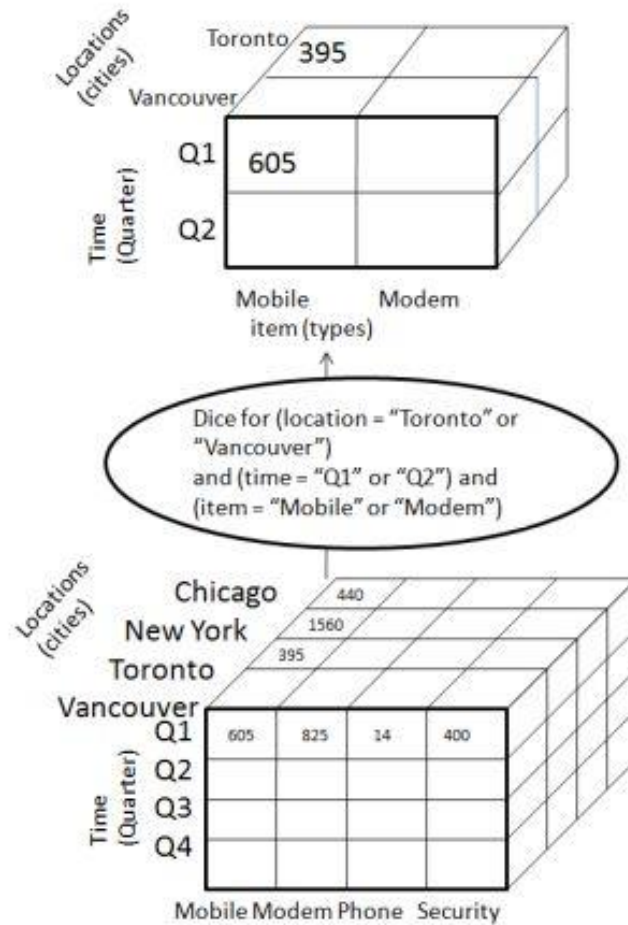
OLAP operácie – Drill-down



OLAP operácie - Slice



OLAP operácie - Dice



OLAP operácie – Pivot (rotation)

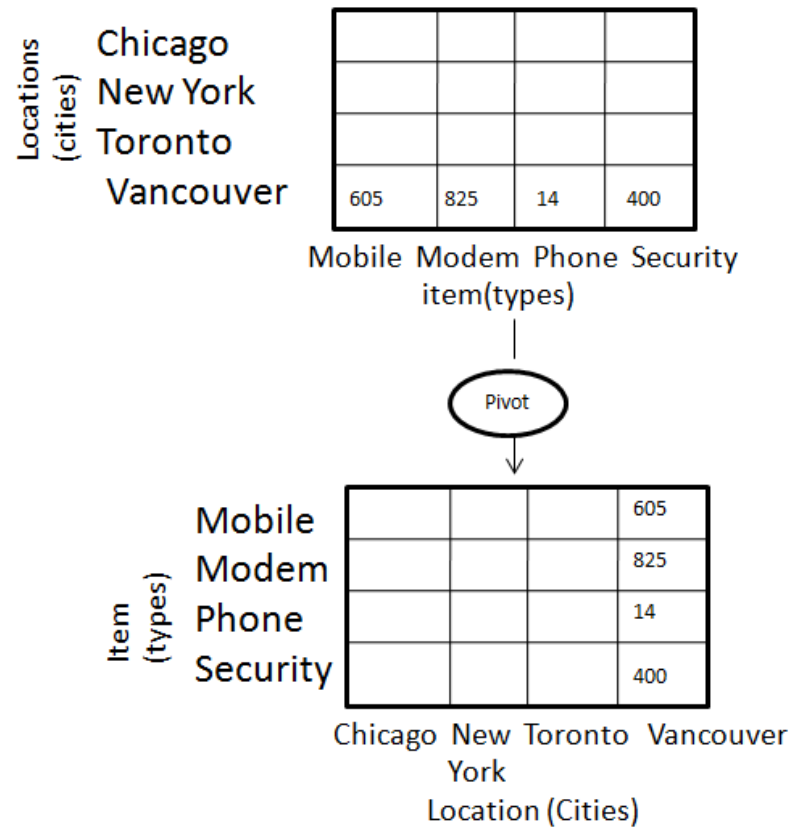
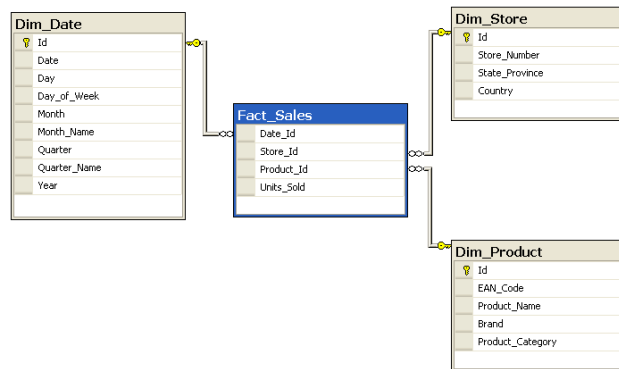
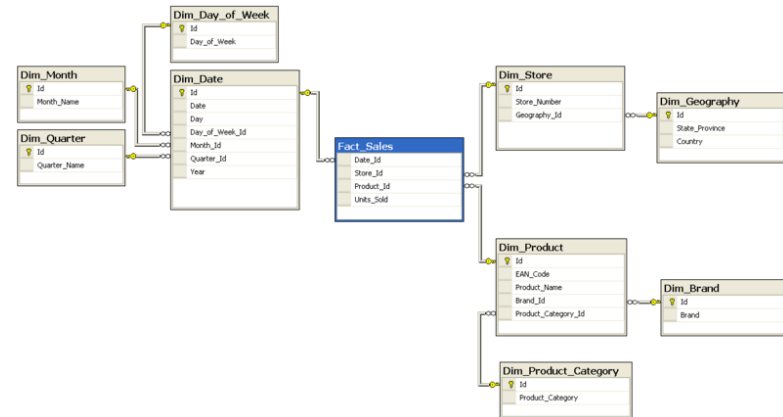


Schéma datového skladu

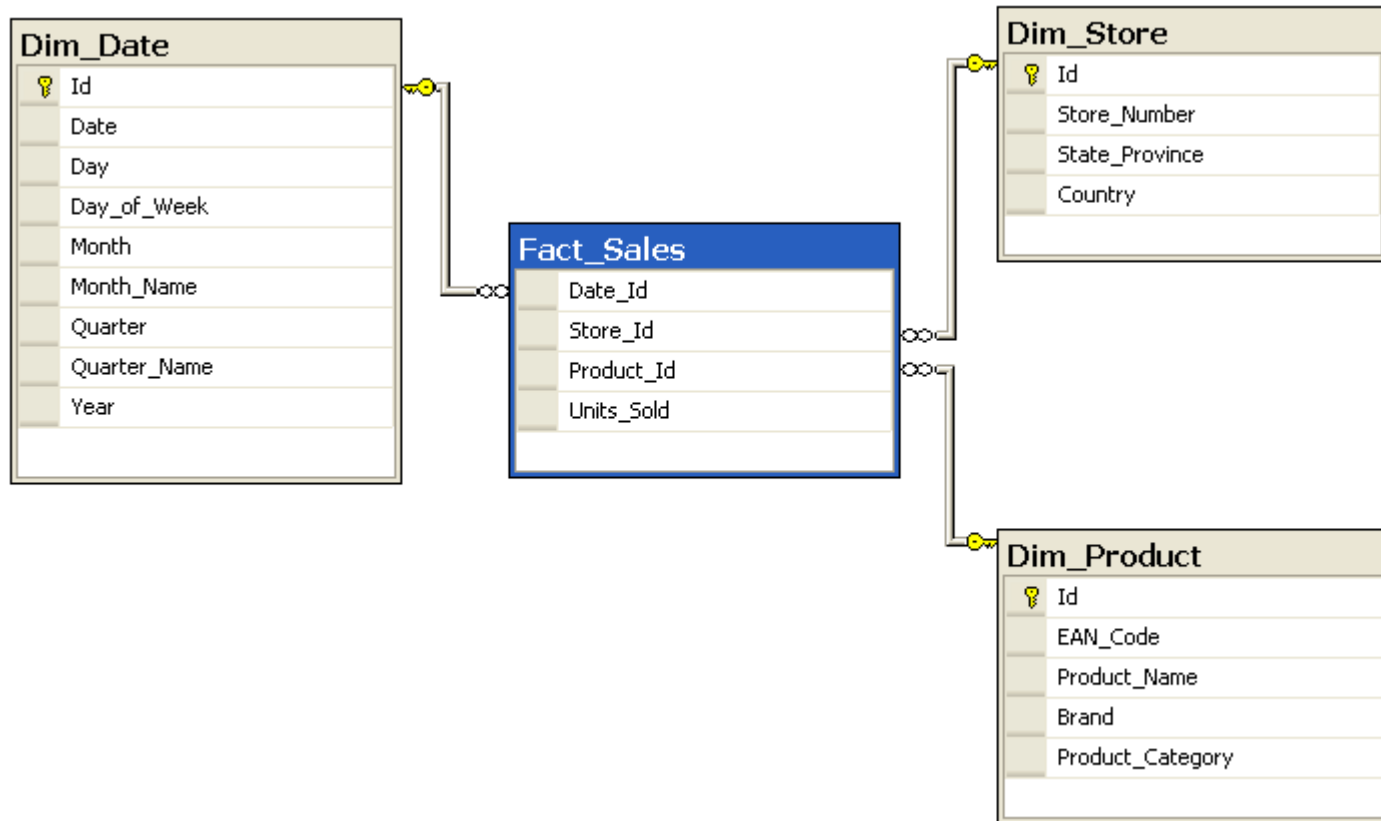
STAR



SNOWFLAKE



Star - schéma datového skladu



Star - schéma datového skladu

SELECT

P.Brand,
S.Country **AS** Countries,
SUM(F.Units_Sold)

FROM Fact_Sales F

INNER JOIN Dim_Date D **ON** F.Date_Id = D.Id

INNER JOIN Dim_Store S **ON** F.Store_Id = S.Id

INNER JOIN Dim_Product P **ON** F.Product_Id = P.Id

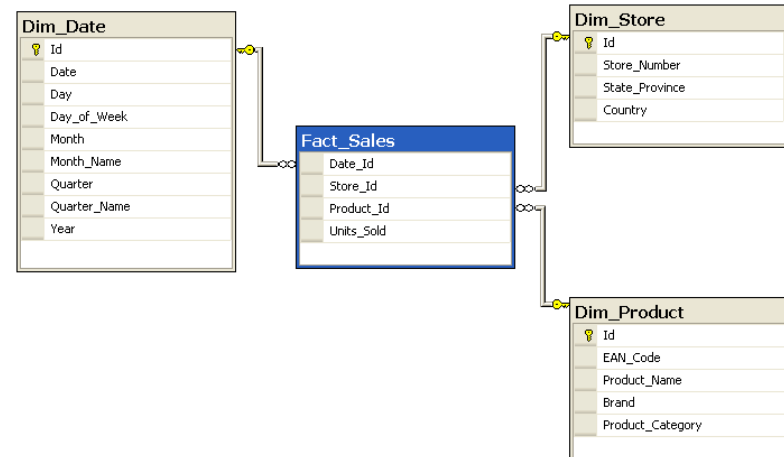
WHERE

D.YEAR = 1997

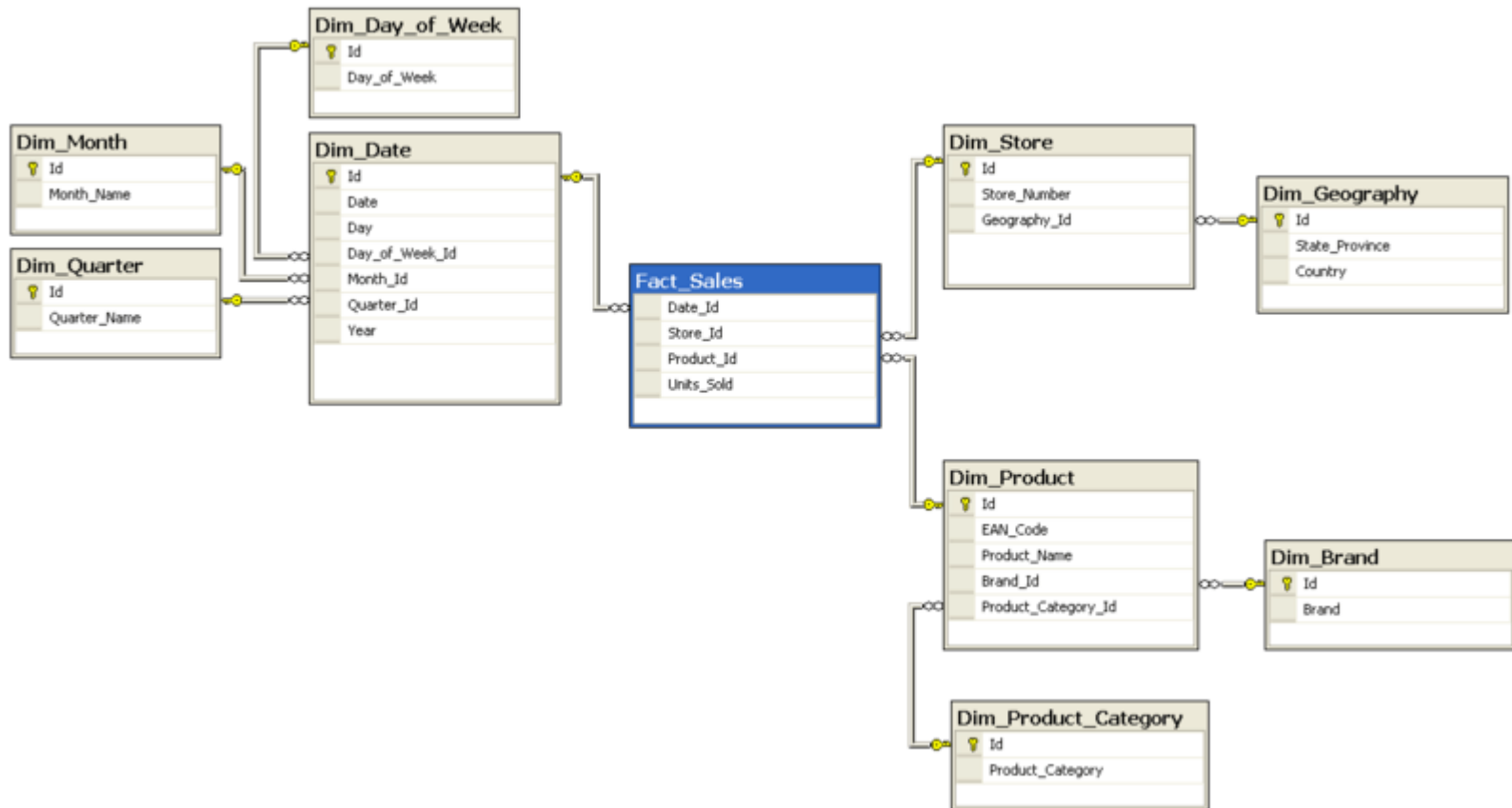
AND P.Product_Category = 'tv'

GROUP BY

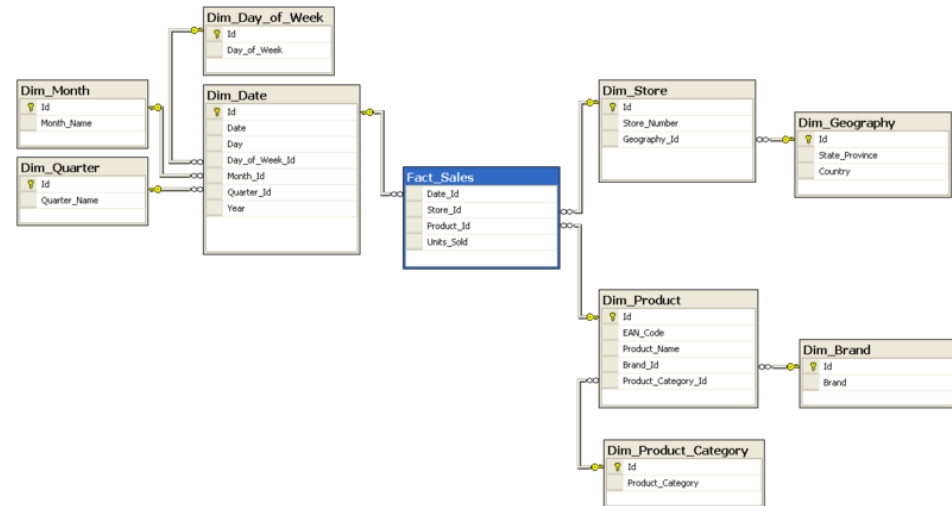
P.Brand,
S.Country



Snowflake - schéma datového skladu



Snowflake - schéma datového skladu



```

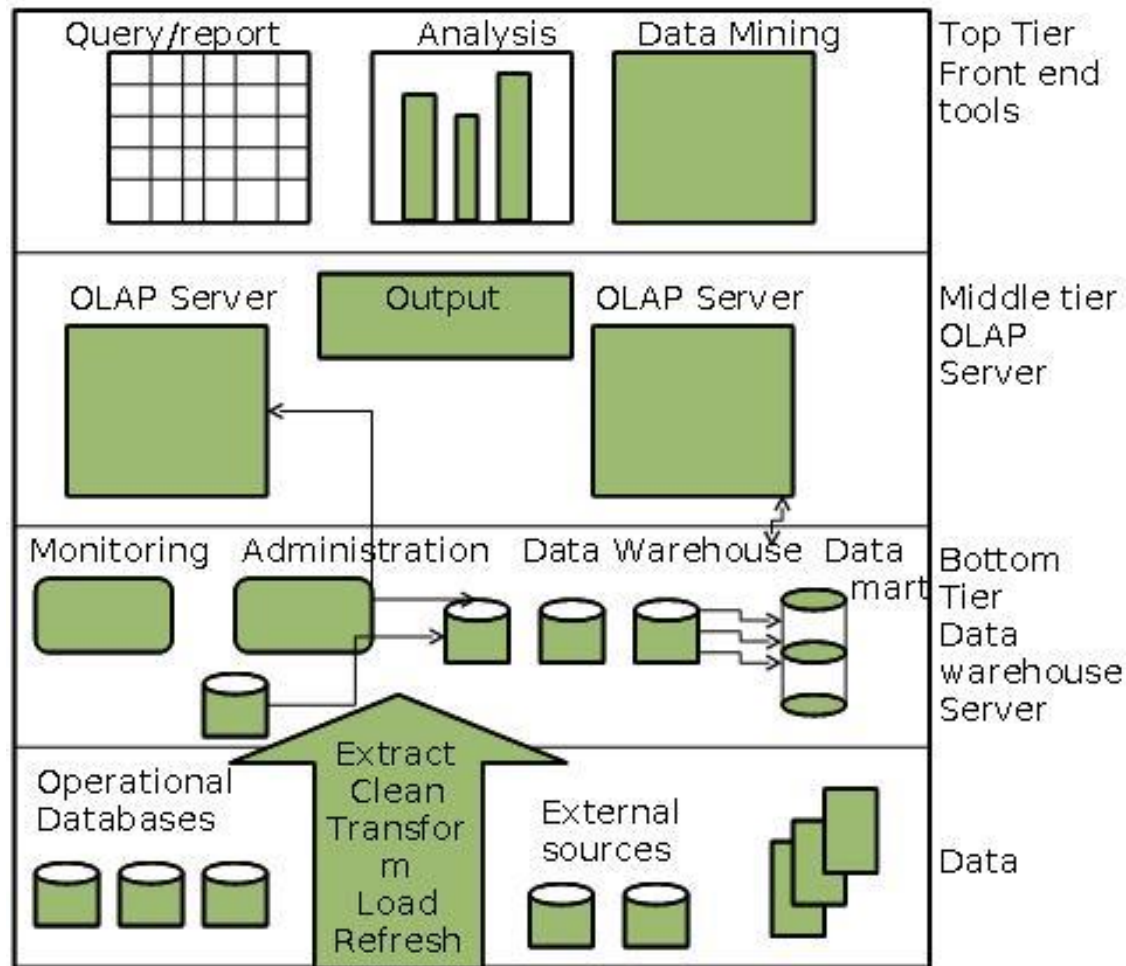
SELECT
    B.Brand,
    G.Country,
    SUM(F.Units_Sold)
FROM Fact_Sales F
INNER JOIN Dim_Date D
INNER JOIN Dim_Store S
INNER JOIN Dim_Geography G
INNER JOIN Dim_Product P
INNER JOIN Dim_Brand B
INNER JOIN Dim_Product_Category C
WHERE
    D.YEAR = 1997 AND
    C.Product_Category = 'tv'
GROUP BY
    B.Brand,
    G.Country
  
```

ON F.Store_Id = S.Id
 ON S.Geography_Id = G.Id
 ON F.Product_Id = P.Id
 ON P.Brand_Id = B.Id
 ON P.Product_Category_Id = C.Id

STAR vs. SNOWFLAKE

	SNOWFLAKE	STAR
Jednoduchosť údržby/zmeny	Neobsahuje redundanciu takže je ľahšie modifikovateľná	Obsahuje redundantné dáta
Jednoduchosť používania	Komplexnejšie dotazy a tým aj menej zrozumiteľné	Menšia komplexnosť dotazov a jednoduchšia na pochopenie
Rýchlosť vykonania dotazov	Väľa cudzích kľúčov, čo spôsobuje dlhší čas vykonania dotazov	Menší počet cudzích kľúčov a tým aj rýchlejšie vykonanie dotazov
DWH typ	Komplexné vzťahy (many:many)	Jednoduché vzťahy (1:1 alebo 1:many)
Join	Veľký počet	Malý počet
Počet tabuliek dimenzií	Môže obsahovať viac ako jednu tabuľku pre každú dimenziu	Obsahuje iba jednu tabuľku pre dimenziu
Kedy použiť	Ak je DT veľká, šetrí miesto	Ak tabuľka dimenzií obsahuje menší počet riadkov
Normalizácia schémy	DT je normalizovaná, FT denormalizovaná	Obe DT aj FT sú denormalizované

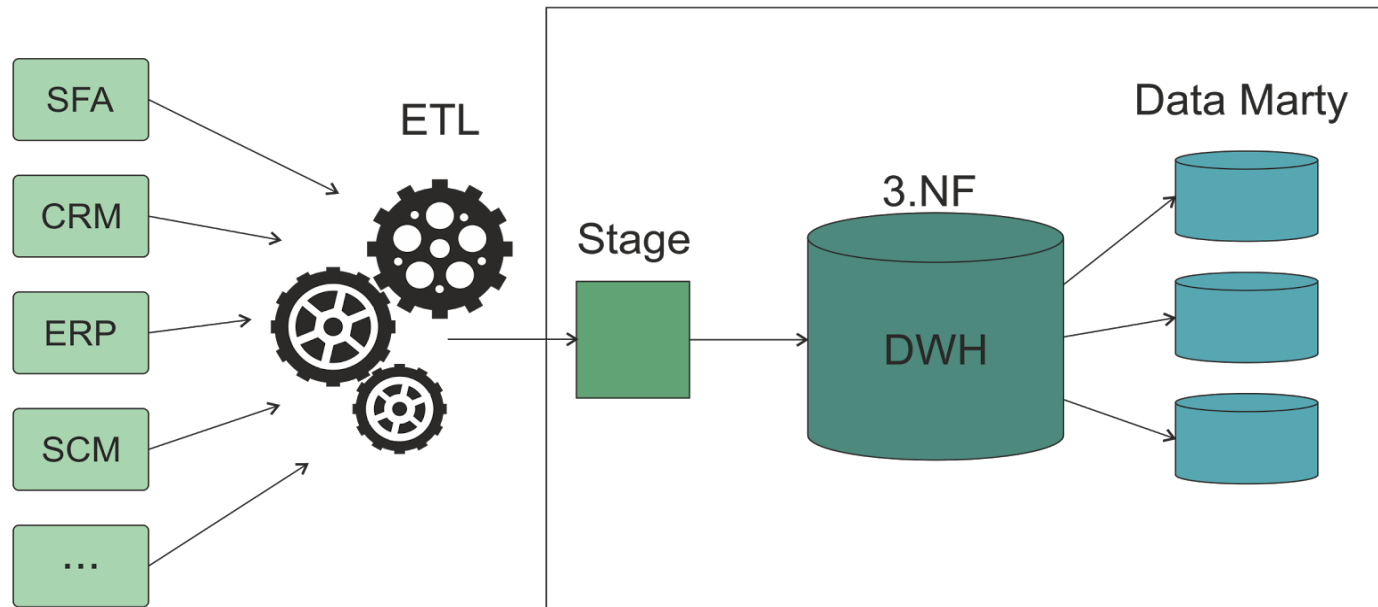
Architektúra DWH (3-vrstvová)



Architektúra DWH podľa Inmona

○ Pohľad zhora-nadol

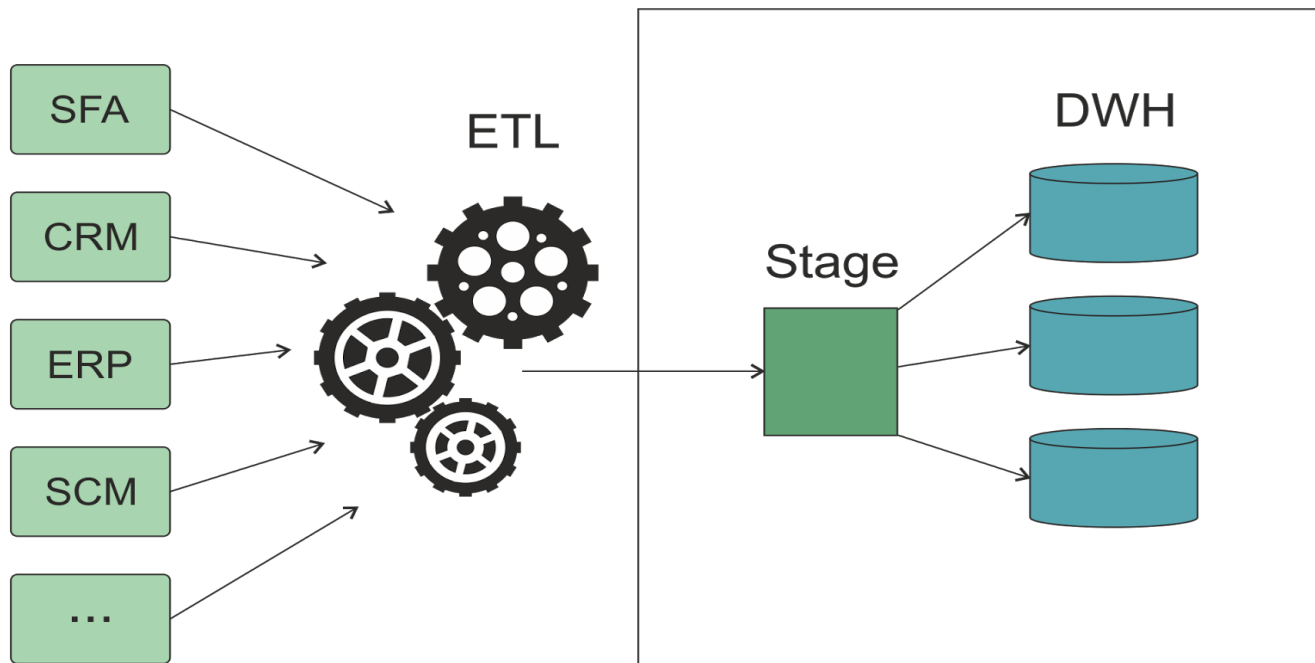
Zdrojové systémy



Architektúra DWH podľa Kimballa

- Pohľad zdola-nahor

Zdrojové systémy



Využitie DWH

- Spracovanie informácií
 - Podporuje dotazovanie a základné štatistické analýzy, reporting a grafy a tabulky
- Spracovanie analýz
 - Multidimenzionálna analýza v DW
 - Podporuje základné OLAP operácie, slice-dice, drillovanie, pivoting
- Data mining
 - Objavovanie znalostí z ukrytých vzorov správania
 - Podporuje asociácie, vytváranie analytických modelov, vykonáva klasifikáciu informácií vizualizáciu, ...

Záver

- Prečo data warehousing?
- Dátové modelovanie (E-R vs. dimenzionálne)
- **Multidimenzionálny model** dátového skladu
 - Star schema, snowflake schema, fakty
 - Dátová kocka - dimenzie & hodnoty
- **ETL** proces (Extract, Transform, Load)
- **OLAP** operácie: drilling, rolling, slicing, dicing and pivoting

Ďakujem za pozornosť

Použité zdroje

- <http://www.diffen.com/difference/Image:Snowflake-schema-4.png>
- <http://www.diffen.com/difference/Snowflake Schema vs Star Schema>
- http://www.tutorialspoint.com/dwh/dwh_olap.htm4
- http://www.maind.sk/dimenzionalne_modelovanie.php
- http://new.dcs.fmph.uniba.sk/files/mit/20111004_DWH_Modeling_JDvorakova.pdf