profinit.
NEW FRONTIER GROUP

IT řešení, které přináší ovoce

NF SK | NEW FRONTIER SLOVAKIA

Dátové sklady
Modelovanie

Jana Dvořáková

1.10.2010

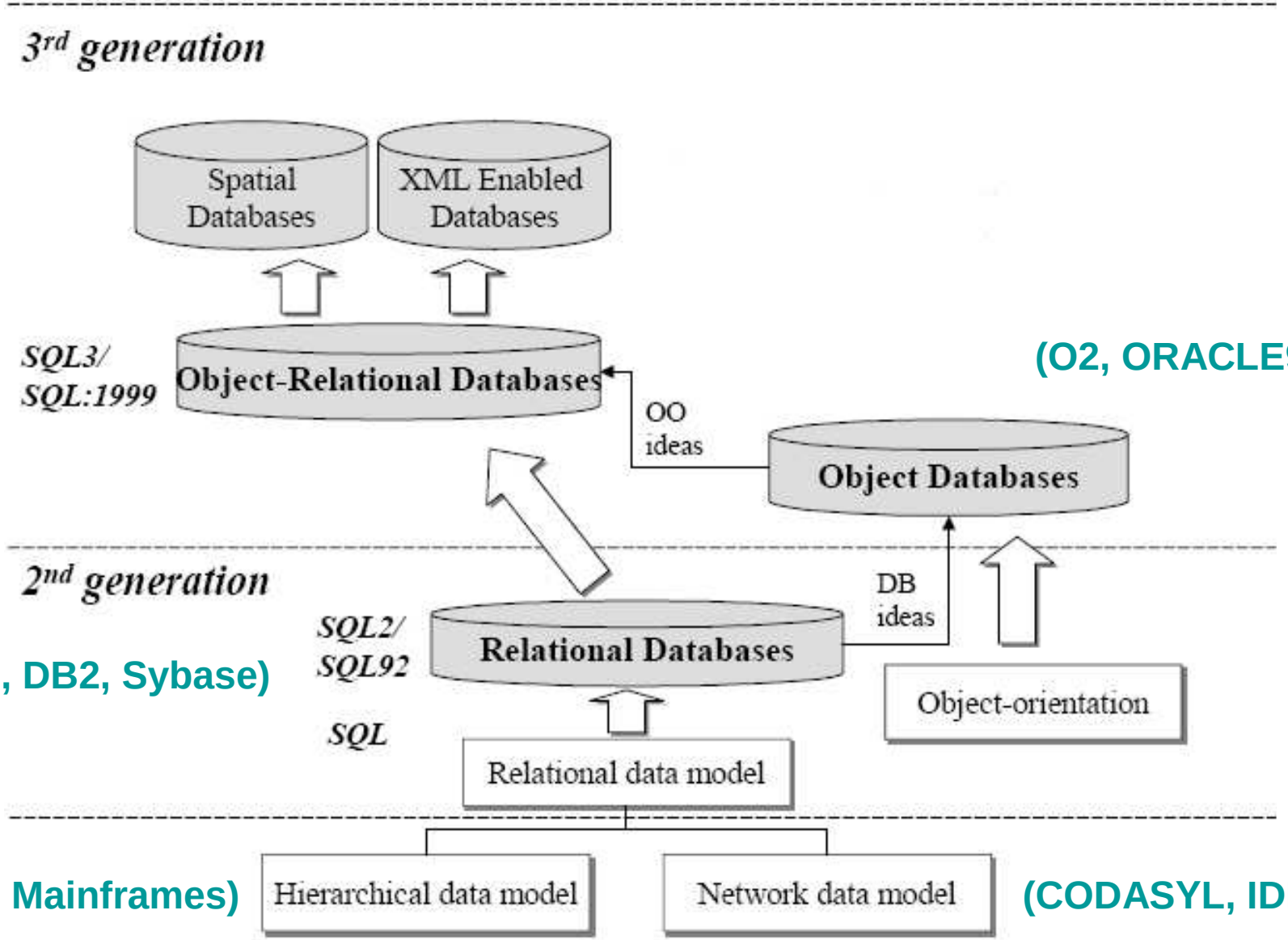
Pokročilé databázové technológie, FIIT STU

Obsah

Dátový model

- Grafická reprezentácia internej organizácia dát v DBMS
  - Entity
  - Vzájomné vzťahy

História DB modelov



(O2, ORACLE9i)

(ORACLE, DB2, Sybase)

(IMS IBM Mainframes)

(CODASYL, IDMS)

Dátové modely v DWH

NEW FRONTIER SLOVAKIA

Vznik modelu DWH:

- Požiadavky používateľov (požadované analýzy, reporty)
- Informácia o zdrojových systémoch, ich dátach a štruktúre

Používané spôsoby modelovania:

- Dimenzionálny dátový model
- Relačný 3-NF dátový model

.. alebo niečo medzi tým

3-NF vs. dimenzionálny model

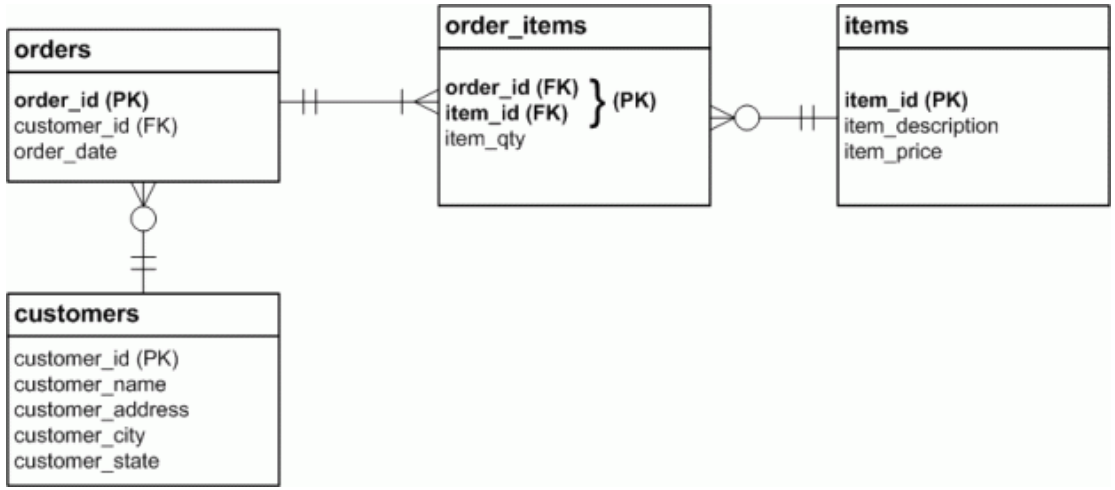
NEW FRONTIER SLOVAKIA

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

Dimenzionálny model typicky nie je v 3-NF (aj keď definícia to nevylučuje)

DWH - relačný dátový model v 3-NF

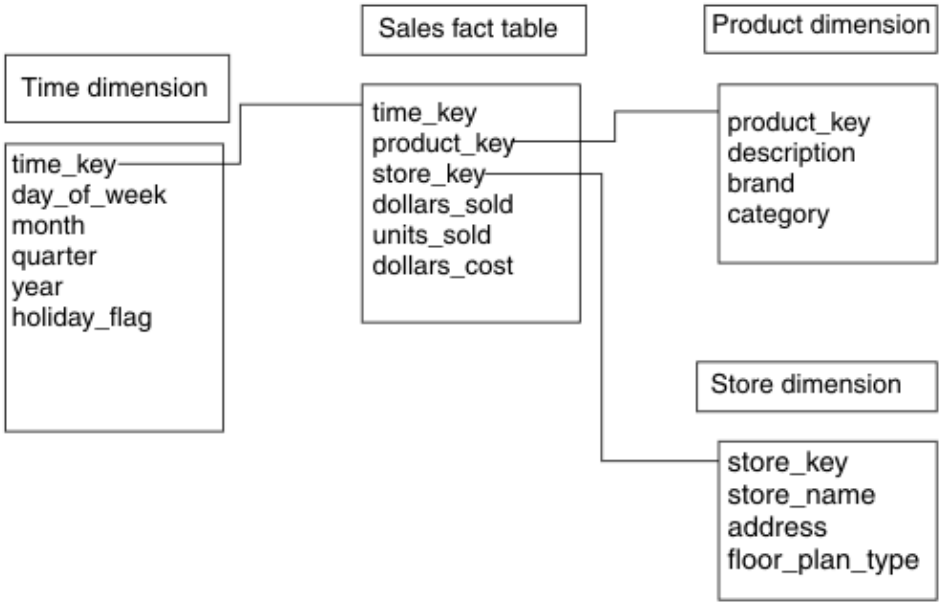
- 3-NF



- Jednoduché ETL
  - Prenos dát zo zdrojových systémov a ich integrácia
- Zložité reportovanie
  - Veľké množstvo JOIN´operácií
  - Ťažsie pochopiteľný bussiness používateľmi
- Vhodné pre niektoré typy databáz
  - TERADATA
- ´Model pre centrálne úložište dát – podľa B. Inmona

DWH - dimenzionálny dátový model

- Štandardne odporúčaný pre DWH



- Zložité ETL
  - Transformácie dát
  - Integrácia, ...
- Jednoduché reportovanie
  - Ľahšie pochopiteľný business používateľmi

- Vhodný (aj) pre relačné databázy
  - Oracle, MSSQL, ...

- Model pre datamarty – podľa B. Inmona aj R. Kimballa

Obsah

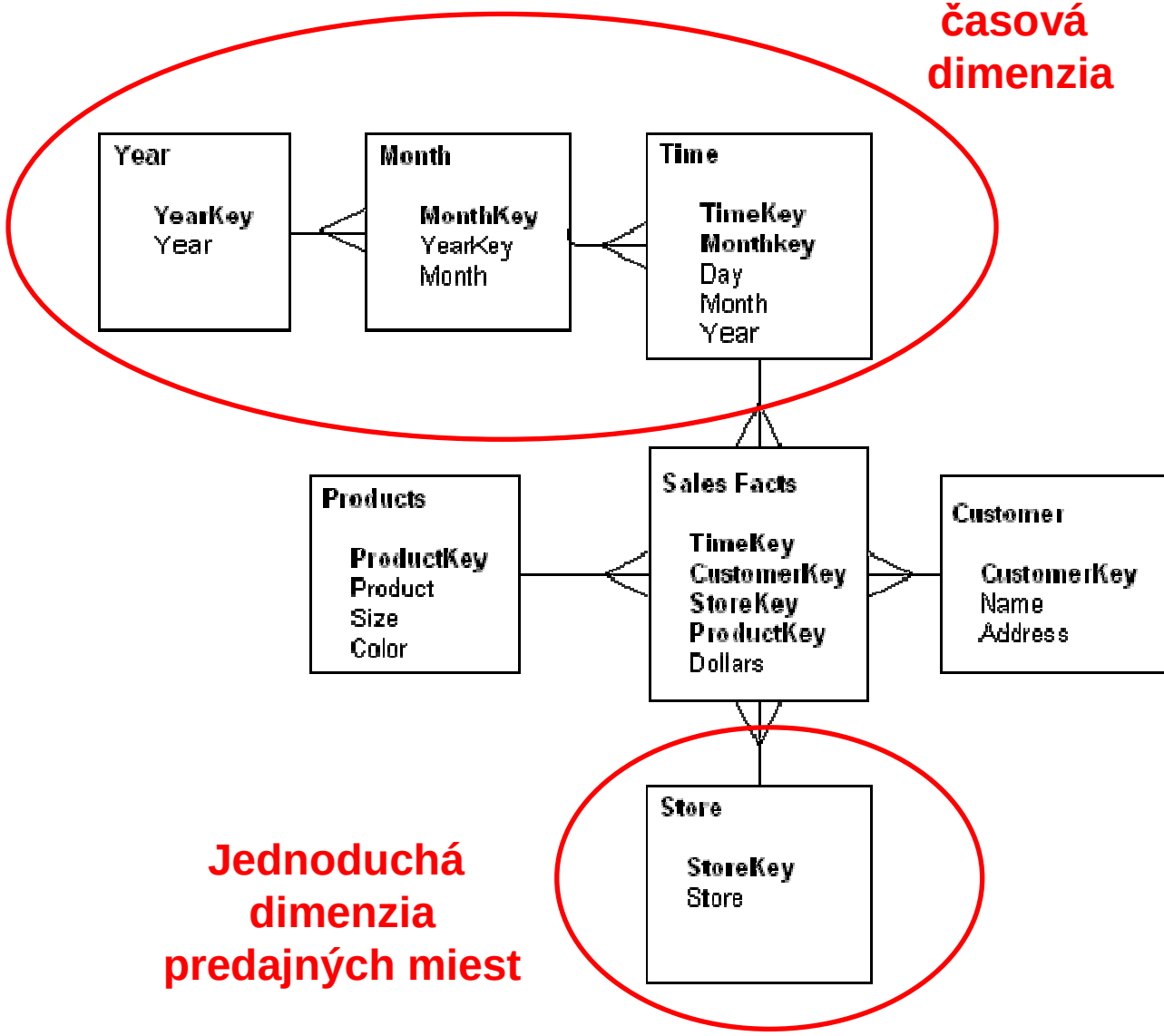
Typy tabuliek DWH

- Dimenzionálne tabuľky – dimenzie
- Faktové tabuľky – fakty
- Agregačné tabuľky

- Ostaté tabuľky – LNK, TMP, ...

DIM tabuľky - dimenzie

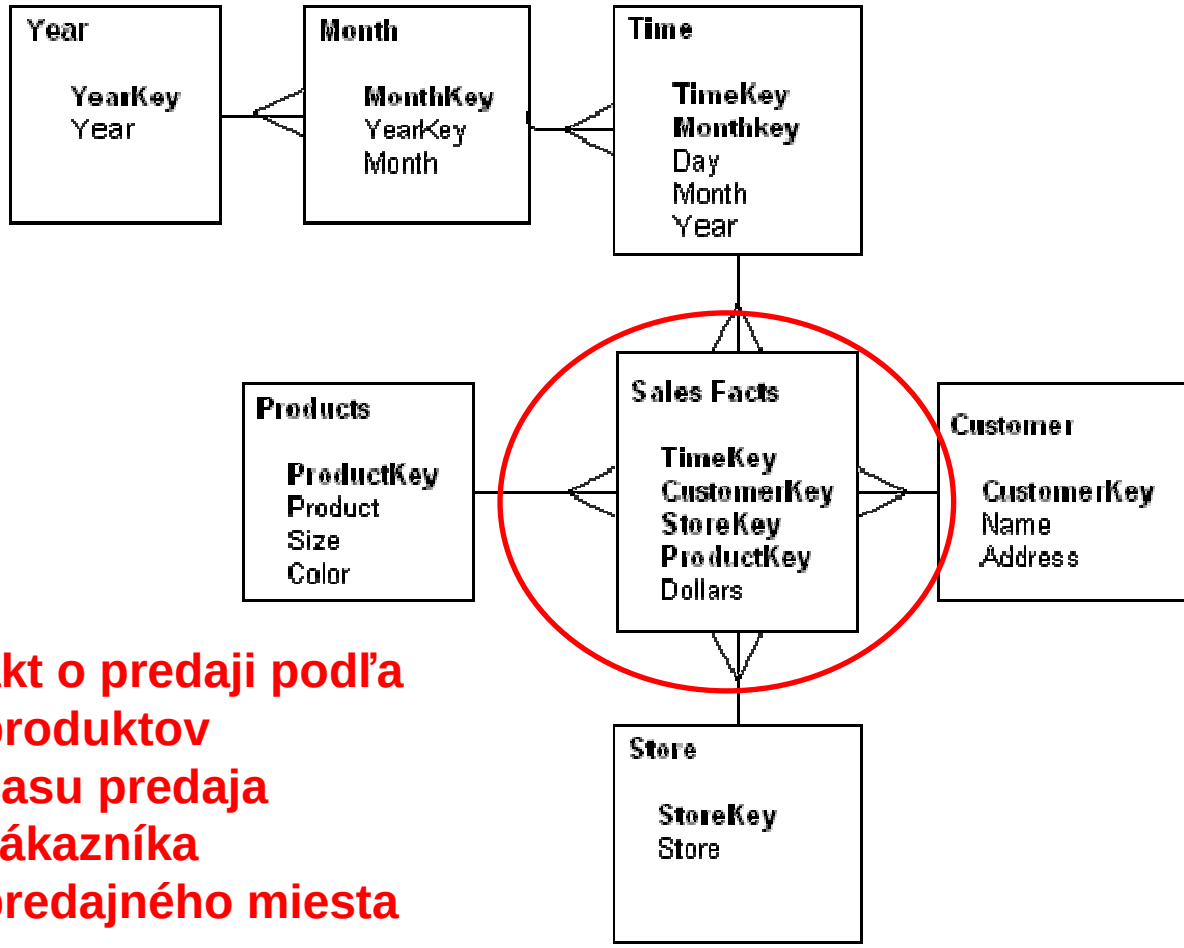
- Obsahujú **atribúty**
  - Napr. zákazník, produkt, dátum
- Majú prirodzený primárny kľúč
- Poskytujú opisné informácie k faktom
  - „Prístupové cesty“ k faktom
- Môžu obsahovať hierarchie
- Conformed dimensions
  - Zdieľané viacerými FCT tabuľkami



FCT tabuľky - fakty

- Obsahujú **merania, metriky** alebo **fakty k business procesom**
  - Napr.obrat obchodu, množstvo tovaru na sklade
- Primárny kľuč - dimenzie
- Na základe odkazov do dimenzíí poskytujú súbor hodnôt

- Conformed FACTs – Zdieľané údaje medzi viacerými business oblasťami

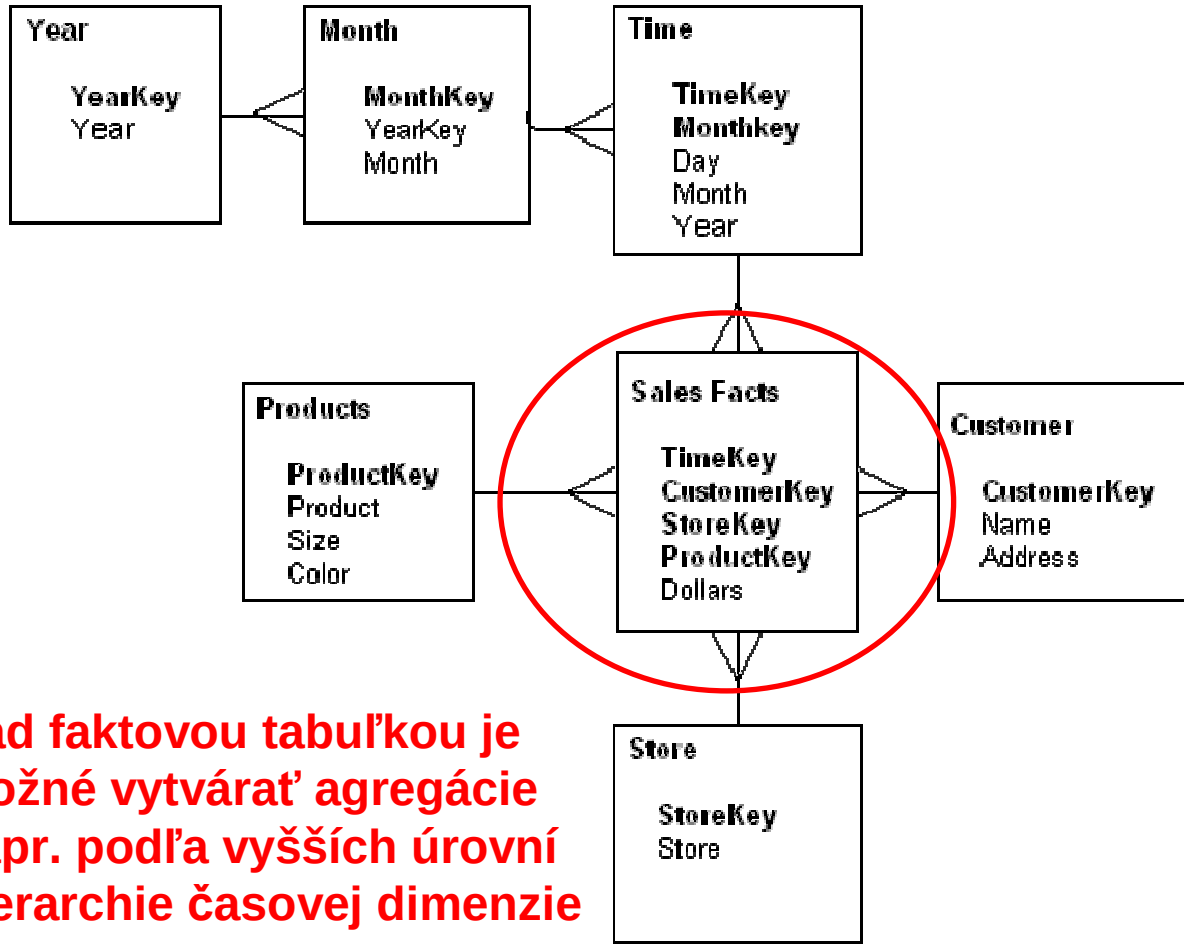


Fakt o predaji podľa
- **produktov**
- **času predaja**
- **zákazníka**
- **predajného miesta**

AGG tabuľky - agregácie

- Obsahujú **agregované/ sumárne** dáta
- Špecifický typ FCT tabuliek
- Dôvod existencie – performance, drill down

- Analytici DWH často pristupujú
  - Najskôr k sumárnym dátam – rýchly prístup
  - Neskôr k detailným dátam – len pre určitú časť dát



Nad faktovou tabuľkou je možné vytvárať agregácie napr. podľa vyšších úrovní hierarchie časovej dimenzie

Granularita dát

Granularita = úroveň detailu dát (faktov aj dimenzií)

Nízka granularita

(-) obrovské množstvo dát

(+) väčšia množina realizovateľných dotazov - 15-30% dotazov je postavených na detailných dátach

Vysoká granularita

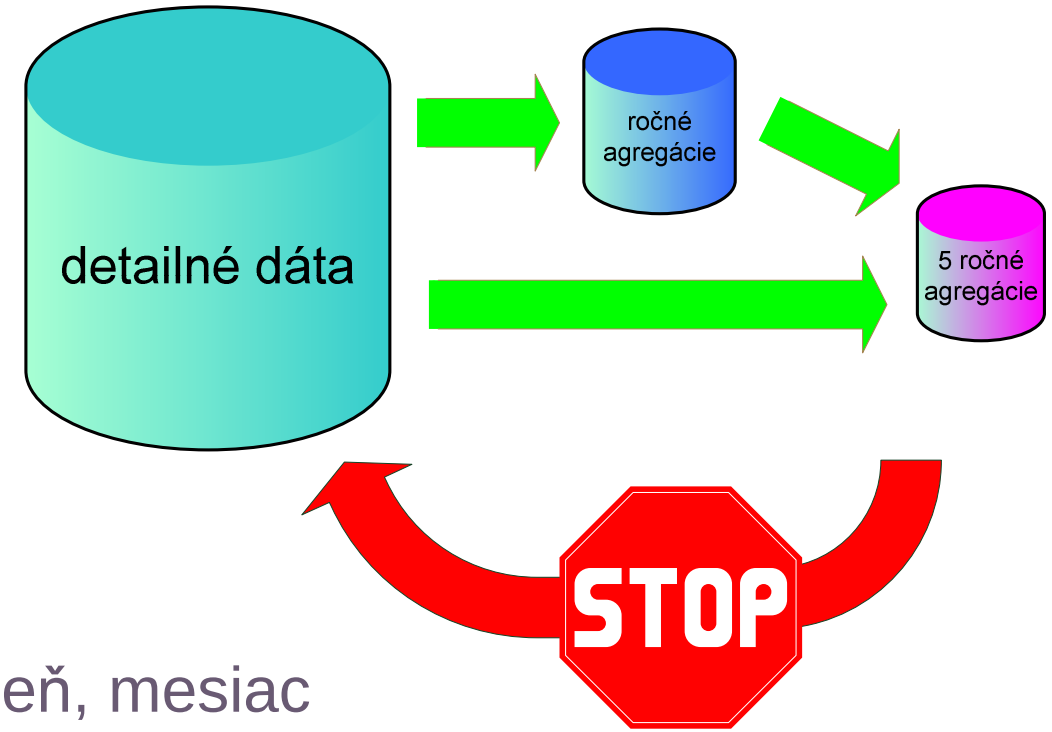
(+) menšie množstvo dát

(-) niektoré informácie sú už nedostupné

(-) niektoré reporty nie je možné realizovať

Granularita dát cont.

- Granularitu je možné len zvyšovať
  - Získavať z viac detailných menej detailné dáta
  - Okrem generovania dát



- Úroveň granularity
  - Detailné dáta – posledný deň, mesiac
  - Mierne sumarizované – posledný rok
  - Sumarizované – posledných 5 rokov
  - Archivované - ostatné

Denormalizácia

= redundancia dát

Cieľ

- Redukcia JOIN operácií → zvýšenie výkonnosti

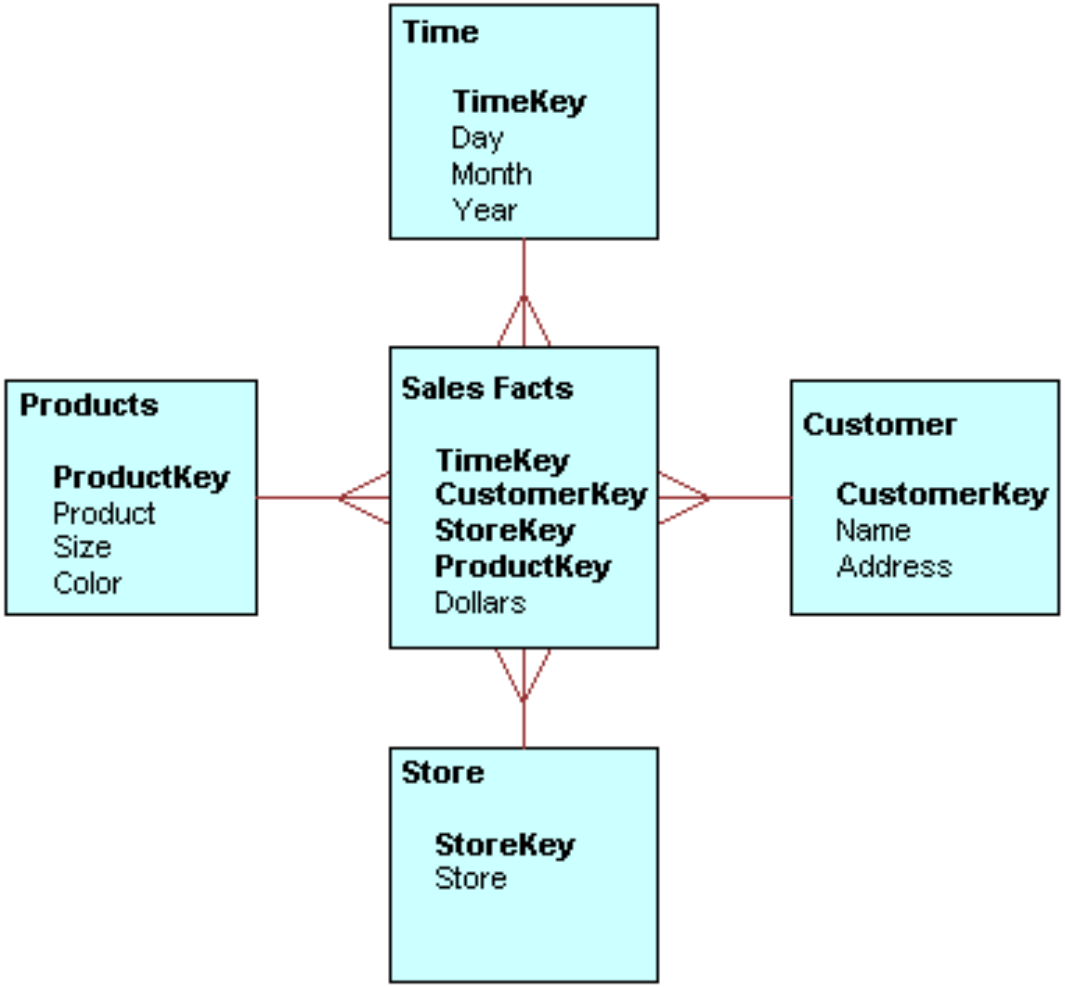
Spôsoby

- Indikátory z dimenzií uložené aj pri faktoch
  - Nie sú nutné JOIN operácie faktov s dimenziami
- Dopredu vypočítané hodnoty
  - Rýchlejšie dotazy
- Predpočítané výstupné reporty
  - Rýchlejšie dostupné reporty

Ostatné tabuľky

- HST, ACT
  - Historizácia záznamov (bližšie pri ETL)
- Role-playing dimension
  - Používanie dimenzie vo viacerých rôznych významoch
  - Napr. dimenzia Date (dátum predaja, dátum objednávky, ..)
- Factless fact table
  - Len prepojenie dimenzií
  - Napr. evidencia udalosti, ktorá nastala ale nie je k nej merateľný údaj

Star schéma



Zjednodušene

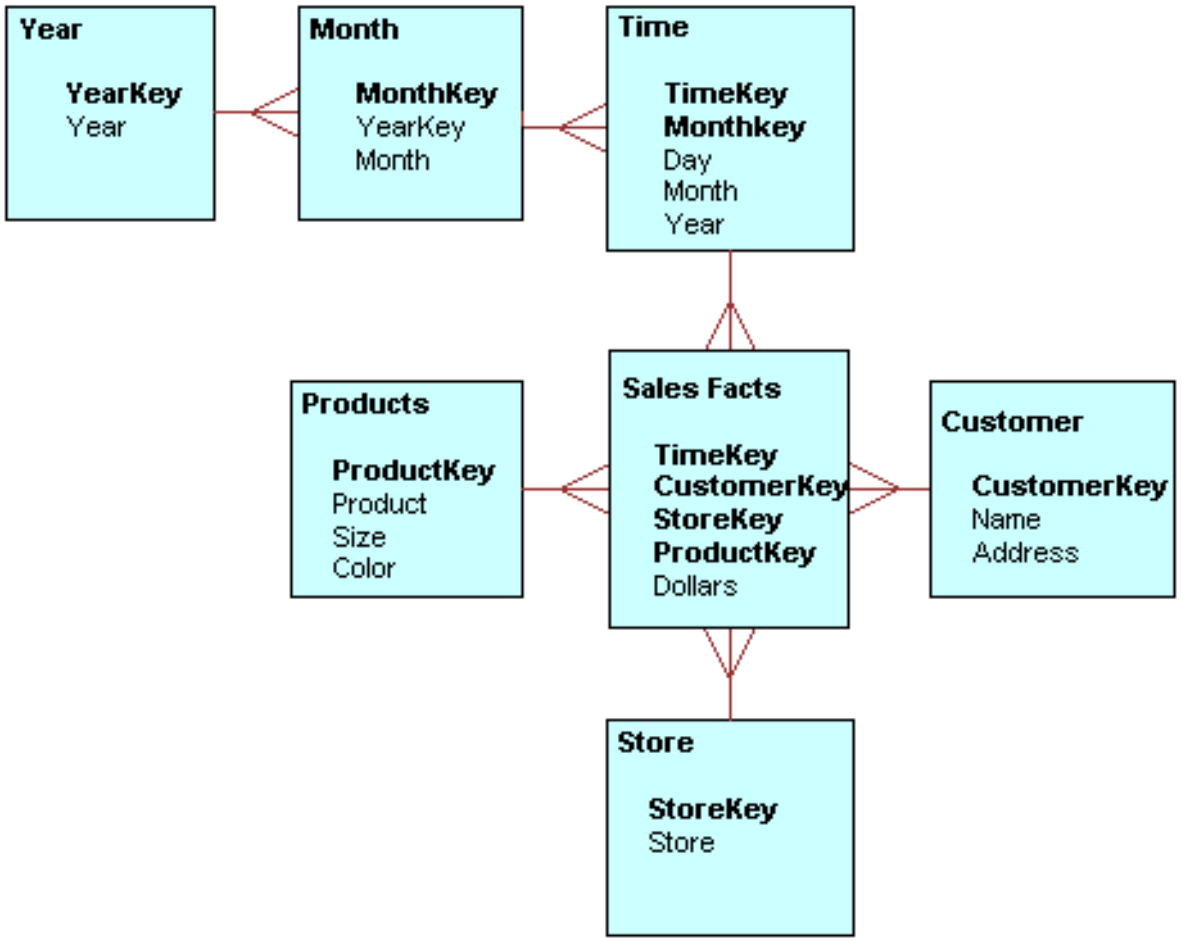
- Fakty v 3-NF
- Niektoré dimenzie v 2-NF

Otázka: Ktorá z uvedených dimenzií nie je v 3-NF?

Constellation schéma

= star schéma s viacerými FCT tabuľkami (zdieľajúcimi dimenzie)

Snowflake schéma



Zjednodušene

- Fakty v 3-NF
- Dimenzie v 3-NF

Snowstorm schéma

= snowflake schéma s viacerými FCT tabuľkami (zdieľajúcimi dimenzie)

Obsah

NEW FRONTIER SLOVAKIA

Dátové modely – história a použitie v DWH

Dimenzionálny model – typy tabuliek

Ukážka modelu dátového skladu

Nástroje

Zhrnutie a diskusia

Obsah

NEW FRONTIER SLOVAKIA

Dátové modely – história a použitie v DWH

Dimenzionálny model – typy tabuliek

Ukážka modelu dátového skladu

Nástroje

Zhrnutie a diskusia

Nástroje

- Sybase PowerDesigner
- Rational Rose (IBM)
- Enterprise Architect
- Microsoft Visio

- Eclipse UML2 Tools

Diskusia

profinit.
PROFESSIONALS IN IT