Datové sklady - struktura, důvody použití, rozdíl od běžné databáze, popis a rozdíl schémat Star/Snowflake, OLTP/OLAP význam

zdroje:
sešit Ka
https://en.wikipedia.org/wiki/Garbage_in,_garbage_out

Datový sklad

- Datový sklad je speciální typ databáze určený pro analytické zpracování dat
- Slouží k:
  - získávání business informací
  - tvorbě přehledů, statistik a reportů
- Obsahuje převážně historická data
- Data se v něm:
  - většinou jen čtou
  - minimálně upravují

Důvody použití

- oddělení provozních a analytických dotazů
- zrychlení analytických dotazů
- možnost práce s historickými daty
- jednodušší tvorba reportů
- sjednocení dat z více zdrojů

Rozdíl: běžná databáze × datový sklad

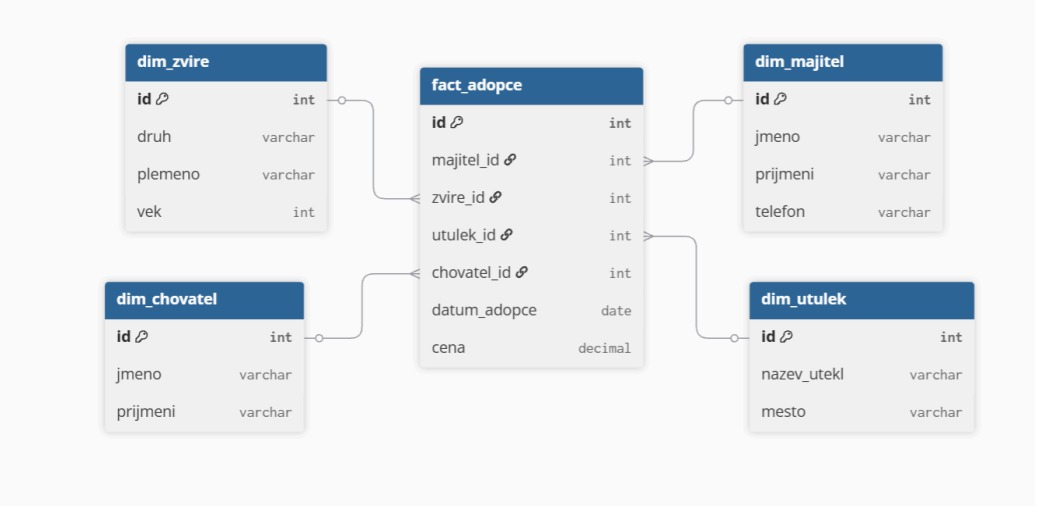
- Běžná databáze (OLTP)
  - určena pro provoz systému
  - plně normalizovaná (3. NF)
  - splňuje všechna integritní omezení
  - vhodná pro:
    - denní transakce (INSERT, UPDATE, DELETE)
  - nevhodná pro:
    - složité analytické dotazy
    - dlouhodobé uchovávání dat
- Datový sklad (OLAP)
  - určen pro analýzu dat
  - nenormalizovaná struktura - denormalizace
  - optimalizován pro:
    - složité SELECT dotazy
  - obsahuje:
    - agregovaná a historická data

Schémata datového skladu

Star

- hvězdná databáze
- skládá se z jedné fact table, která obsahuje “měnitelné” a “spočítatelné” hodnoty “tabulka faktů” a odkazy na dimenzionální tabulky a dalších dimenzionálních tabulek - dim tables (číselníky) obsahující informace: kdo, kde, kdy, co

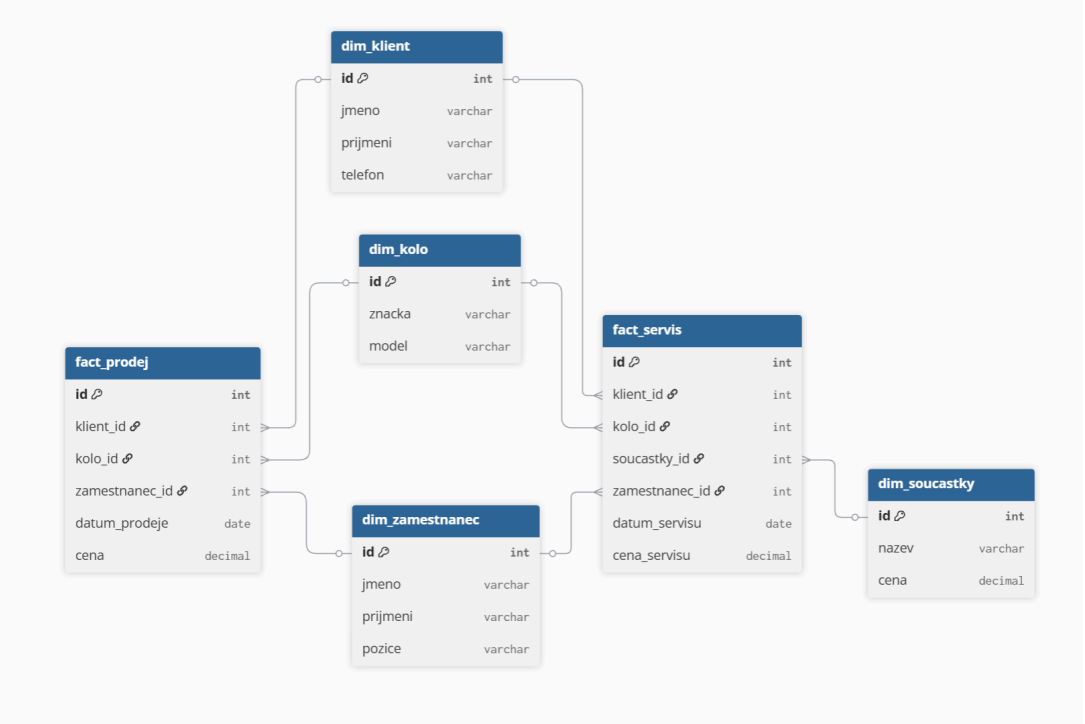
Příklad útulek



Galaxy

- obsahuje více fact tabulek

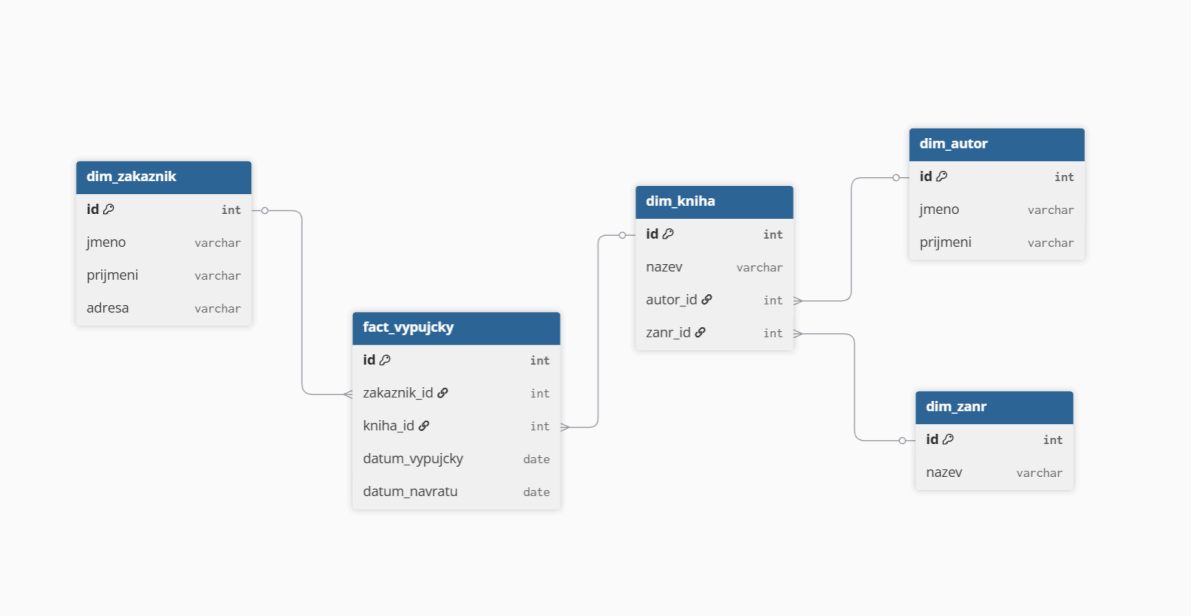
Příklad cykloservis



Snowflake

- potřeba, aby dim tabulky byli částečně normalizované

Příklad knihovna



Data Pipeline

- Data pipeline je proces, který zajišťuje tok dat od jejich zdrojů přes jejich zpracování a transformaci až do cílového systému určeného pro analýzu.
- GIGO (Garbage In, Garbage Out)
    - Princip, který říká, že pokud jsou vstupní data nekvalitní, budou nekvalitní i výsledky analýzy.
    - Data pipeline (zejména ETL proces) slouží k omezení dopadů GIGO pomocí čištění, kontroly a úpravy dat.

0. Zdrojová data

- uživatelské uložení dat (např.: excel, soubory)
- data jsou redundantní, obsahují chyby
- špatně se z nich získávají analytické informace

- Transformace ELT/ETL

- E - Extract
- L - Load
- T - Transform
- ELT - větší objem dat, dočasné tabulky
- ETL - menší objem dat, úprava napřed, přímo do nových tabulek

1. Vytvoření OLTP databáze

- OLTP (Online Transaction Processing)
    - způsob práce s daty
        - ➢ provádění denních transakcí
        - ➢ ukládání nových záznamů, úpravy existujících
        - ➢ rychlé dotazy pro provoz systému

OLTP databáze

- plně normalizovaná databáze (3. NF) , která splňuje všechna integritní omezení
- vhodné pro denní transakce
- není vhodné pro dlouhodobé uchovávání dat a získávání analytických informací (složité selecty)

2. Business informace

- definování analytických požadavků
- otázky typu
  - kdo?
  - kdy?
  - kde?
  - proč?

3. OLAP systém

- OLAP (Online Analytical Processing)
  - způsob práce s daty
    - ➢ získávat business informace
    - ➢ analyzovat historická a agregovaná data
    - ➢ tvořit přehledy, statistiky a reporty
- proces získávání business informací (analytické požadavky), pro které není původní OLTP databáze vhodná a proto se používá OLAP databáze

OLAP databáze

- nenormalizovaná struktura
- optimalizována pro SELECT / analytické dotazy
- obsahuje agregovaná a historická data
- používá schémata: Star, Snowflake, Galaxy