Data Warehousing

(From: Database Systems: The whole book (Chapter 10.6 and 10.7))

On-Line Analytic Processing (OLAP)

- Examination of data for patterns or trends
- Queries („OLAP queries“ or „decision-support queries“)
  - Are **highly complex queries** that use one or more aggregations
  - Typically examine **very large amounts** of data (even if results are small)
    —> in contrast to OLTP, which might only retrieve a single tuple of data
- **Data Warehouse**:
  - Common that OLAP applications take place in a separate copy of the master database -
    Why:
    - Can integrate and organize data from separate databases
    - Queries take too much time be executed in a transaction-processing system with high
      throughput requirements
      —> Typical scenario: Warehouse updated overnight, frozen during the day to answer
        queries
  - Example: Sales data
    - Accumulated data from every store in warehouse
    - Aggregate sales and identify future problems and opportunities for the company
- **Fact table**:
  - Central relation or collection of data (e.g. Sales)
  - Can be thought of as multidimensional space —> **raw-data cube** (e.g. with dimensions
    car, dealer, date)
  - (Formal) data cube vs raw-data cube:
    - Formal data cube includes aggregations of data in all subsets of dimensions + data itself
    - Points in formal cube may represent already aggregated data (e.g. car sales aggregated
      by model)
- ROLAP (R for Relational):
  - Data stored in relations with **star schema** structure —> Star schema:
    - Fact table links to several other relations called **dimension tables** (dimension
      attributes reference dimension tables as a foreign key)
    - Fact table has several attributes that represent dimensions, and one or more
      **dependent** attributes that represent properties of interest for the point as a whole
      (dependent attribute will typically be requested in OLAP query)
    - Example: dimensions: date of sale, store VS dependent attributes: sales price, tax
- MOLAP (M for Multidimensional):
  - Data cube as specialized structure is used to hold the data
  - Non-relational operators may be implemented in systems
- Slicing and dicing:
  - Partitioning (group by) divides cube into smaller cubes —> **dicing**
  - WHERE clause can focus on partition —> **slicing**
- Example: slice by date, dice in two other dimensions (car and dealer)
- Drill-down and roll-up:
  - **Drill-down:**
    - Partitioning more finely
    - Focusing on specific values in certain dimensions
  - **Roll-up:**
    - Partitioning more coarsely
    - e.g. group by year instead of months to eliminate the effect of randomness in the data

**Data Cubes (formal data cube)**

- Precomputes **all possible aggregates** in a formal way
- Amount of extra storage needed is often tolerable (if number of values reasonable large, the border does not add much to the volume of the cube)
- It is common that the raw data is already aggregated before it is entered in the cube
- Augmented table that adds an additional value denoted *, in each dimension (~adding border to cube in each direction)
- BUT: With cube operator it is not possible to aggregate at intermediate levels of granularity based on values in dimension table (e.g. date attribute, with related dimension table with columns year, month, day —> cannot aggregate by month or year)
- SQL:
  - GROUP BY … WITH CUBE —> also get tuples that represent aggregations (with null where we used *)
  - WITH ROLLUP: yields only the aggregated tuples if they aggregate over a tail of the sequence of grouping attributes (Not yielding tuple like (Gobi, red, NULL, 2001-05-21, NULL, 4500) which would be the count of all red Gobi cars sold on 2001-05-21)