Data warehouse

What is data warehouse

- Definition of DW
  - A storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources
    - data are combined in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs.
    - data are periodically updated and stored for read-only applications
  - A data warehouse is a set of facts perceived by a number of dimensions
- Characteristic of DW
  - Subject oriented
  - Nonvolatile — not usually subject tot changes
  - integrated — data is consistent and integrated from multiple courses
  - Time variant — historical data is recorded for analytical applications

Why we need data warehouse

- Traditional database issues
  - Traditional database applications consist of both updates and queries, some queries are large scale aggregation reports which can take long time to generate on-the-fly
  - Database updates and queries must lock data resources, large scale aggregation reports lock many resources for a long time
  - Organizations are analyzing current and historical data to identify useful patterns and support business strategies
- Benefits of data warehouse
  - Emphasis is on complex,interactive, exploratory analysis of very large datasets created by integrating data from across all parts of an enterprise
- Difference between DW and DB
  - Integrated data spanning long time periods, often augmented with summary information
  - very large volume
  - Interactive response times expected for complex queries
  - Ad-hoc updated uncommon(write-once and Read forever)

DW design

- Star schema — advantage — faster query processing speed
- Snowflake schema
  - fact constellation — A set of fact tables that share some dimension tables
  - advantage
    - less data integrity problem
    - less space consumption

DW implementation

- Data cube
  - OLAP
    - Pivot — Rotate data cube to show a different orientation of axes
    - Roll up — Move up concept hierarchy, grouping into larger units along a dimension with generalization
    - Drill-down — Disaggregate to a finer-grained view to show more details
    - Slice and dice — Perform projection operations on the dimensions
  - Cuboid — A cuboid is a denotation of one of the 2^d summarized views, which can be used for materialized view
- Materialized view
  - Advantages
    - OLAP queries are typically aggregate queries
    - Pre-aggregation is essential for interactive response time
      - Pre-calculate expensive joins
      - Speed up online OLAP queries
  - Disadvantages
    - It increases storage cost
    - The content of the materialized views must be maintained when the underlying details tables are modified
    - Trade off between query performance and accessibility to up-to-date data