CS685: Data Mining
Data Warehousing

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs685/

1st semester, 2021-22

Mon 1030-1200 (online)

Data Warehousing

- A data warehouse is a data storage system, usually separate from the original database

Data Warehousing

- A data warehouse is a data storage system, usually separate from the original database
- It has four important features

1. *Subject-oriented*: It is modeled around subjects, e.g., sales, customers, etc.
2. *Integrated*: It organizes information from multiple sources into a single storage
3. *Time-variant*: It stores information across different time points
4. *Non-volatile*: It stores data permanently and requires only two operations, construction and access

Data Warehousing

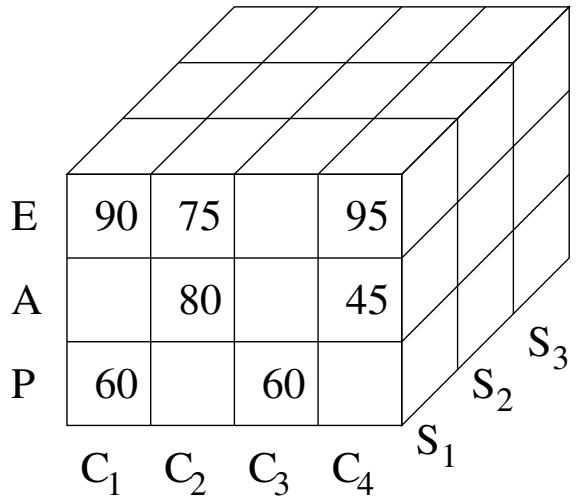
- A data warehouse is a data storage system, usually separate from the original database
- It has four important features

1. *Subject-oriented*: It is modeled around subjects, e.g., sales, customers, etc.
2. *Integrated*: It organizes information from multiple sources into a single storage
3. *Time-variant*: It stores information across different time points
4. *Non-volatile*: It stores data permanently and requires only two operations, construction and access

- A data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support model
- Data warehousing is the process of constructing and using data warehouses

Data Warehouse Model

- A data warehouse is modeled as a multidimensional data model or data cube
- *Dimensions* of a data cube are attributes important for that analysis
- Each dimension has a corresponding dimension table that stores metadata about the dimension
- Numeric values about the subject of the data warehouse are *facts*
- The fact table stores information about them



Cuboids

- Any subset of a data cube is a cuboid
- It is essentially the result of "group by" operator

Cuboids

- Any subset of a data cube is a cuboid
- It is essentially the result of "group by" operator
- All cuboids together form a lattice of cuboids
- Base cuboid: no summarization, at level nD
- Apex cuboid: full summarization, at level 0D

Cube Operations

- `compute cube` operator computes aggregation over *all* subsets of dimensions specified
- For example, specifying the dimensions as item, time and loc, the cuboids computed are (item, time, loc), (item, time), (time, loc), (loc, item), (item), (time), (loc) and ()
- Total of 2^n cuboids
- () implies empty group by, i.e., dimensions are not grouped

Cube Operations

- `compute cube` operator computes aggregation over *all* subsets of dimensions specified
- For example, specifying the dimensions as item, time and loc, the cuboids computed are (item, time, loc), (item, time), (time, loc), (loc, item), (item), (time), (loc) and ()
- Total of 2^n cuboids
- () implies empty group by, i.e., dimensions are not grouped
- Cuboids can be pre-computed and materialized
- *No materialization*: No non-base cuboid is precomputed
- *Full materialization*: Full cube is precomputed
- *Partial materialization*: Some *subcubes* are precomputed based on usage and storage
- Iceberg cube: computes those subcubes whose size (number of tuples) is above a threshold

OLAP Operations

- OLAP stands for *online analytical processing*
- OLTP stands for *online transactional processing*

OLAP Operations

- OLAP stands for *online analytical processing*
- OLTP stands for *online transactional processing*
- Different operations
  - Roll up (drill up): Summarize by going up the level
  - Drill down (roll down): Go down the level
  - Slice: Project operation; on only one dimension
  - Dice: Select operation; on more than one dimensions
  - Pivot (rotate): Rotate for better or alternate visualization
  - Drill across: Summarize across different fact tables
  - Drill through: Access underlying relational data through base cuboids

OLAP Operations

- OLAP stands for *online analytical processing*
- OLTP stands for *online transactional processing*
- Different operations
  - Roll up (drill up): Summarize by going up the level
  - Drill down (roll down): Go down the level
  - Slice: Project operation; on only one dimension
  - Dice: Select operation; on more than one dimensions
  - Pivot (rotate): Rotate for better or alternate visualization
  - Drill across: Summarize across different fact tables
  - Drill through: Access underlying relational data through base cuboids
- How is OLAP related to data mining?

OLAP Operations

- OLAP stands for *online analytical processing*
- OLTP stands for *online transactional processing*
- Different operations
  - Roll up (drill up): Summarize by going up the level
  - Drill down (roll down): Go down the level
  - Slice: Project operation; on only one dimension
  - Dice: Select operation; on more than one dimensions
  - Pivot (rotate): Rotate for better or alternate visualization
  - Drill across: Summarize across different fact tables
  - Drill through: Access underlying relational data through base cuboids
- How is OLAP related to data mining?
- It essentially facilitates data analysis by efficiently providing summaries, projections, etc.

OLAP Implementation

- Different server models to implement OLAP operations
- Relational OLAP (ROLAP): Uses a relational database backend
- Multidimensional OLAP (MOLAP): Uses multidimensional arrays
- Hybrid OLAP (HOLAP): Hybrid system that tries to exploit scalability of ROLAP in lower levels and efficiency of MOLAP in higher levels

OLAP Implementation

- Different server models to implement OLAP operations
- Relational OLAP (ROLAP): Uses a relational database backend
- Multidimensional OLAP (MOLAP): Uses multidimensional arrays
- Hybrid OLAP (HOLAP): Hybrid system that tries to exploit scalability of ROLAP in lower levels and efficiency of MOLAP in higher levels
- For data mining, OLAM systems
- OLAM stands for *online analytical mining*
- Integrates data mining operations directly into OLAP systems