CSC4008 Data Mining

117010279 Ziren WANG

March 2020

2 Data Warehousing & OLAP

2.1 Overview

A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process. It is regarded as a decision support database that is maintained separately from the organization's operational database.

1. **Subject-oriented**
   Organized around major subjects, provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

2. **Integrated**
   Constructed by integrating multiple, heterogeneous data sources. In this process, data cleaning and data integration techniques are applied. When data is moved to the warehouse, it is converted.

3. **Time-variant**
   The time horizon for the data warehouse is significantly longer than that of operational systems. Every key structure in the data warehouse Contains an element of time, explicitly or implicitly. But the key of operational data may or may not contain "time element".

4. **Nonvolatile**
   A physically separate store of data transformed from the operational environment. So, Operational update of data does not occur in the data warehouse environment.

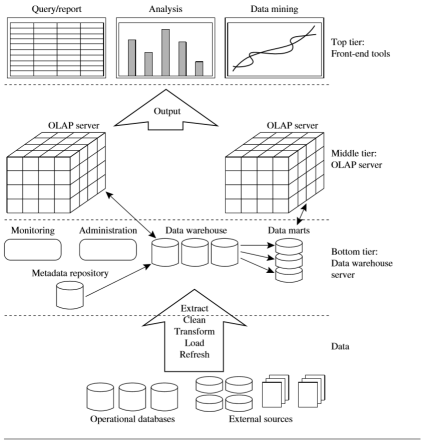

Figure 1: data warehousing

There are 3 data warehouse models:

- Enterprise Warehouse: Collects all of the information about subjects spanning the entire organization
- Data Mart: A subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart

- Virtual Warehouse: A set of views over operational databases. Only some of the possible summary views may be materialized

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Online Transactional Processing (OLTP); Online Analytical Processing (OLAP)

Figure 2: OLTP v.s. OLAP

2.2 Database Systems

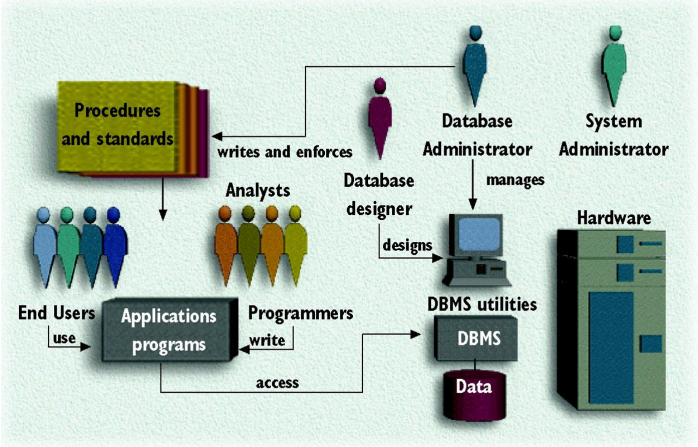

Figure 3: Database System Environment

2.2.1 Structured Query Language (SQL)

SQL coverage fits into three categories:
(1) Data definition.
(2) Data management.
(3) Data query.
There are 3 keywords in a SQL statement: SELECT, FROM, WHERE, relate to: attributes, tables, filters respectively.

2.3 Data Cube and OLAP

A data warehouse is based on a multidimensional data model which views data in the form of a data cube.
Base cuboid: an n-D base cube in data warehousing literature.

2.3.1 Conceptual Modeling of Data Warehouses

- **Star schema**: A fact table in the middle connected to a set of dimension tables.

- **Snowflake schema**: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables.
- **Fact constellations**: Multiple fact tables share dimension tables, viewed as a collection of stars.

2.3.2 Data Cube Measures: Three Categories

- Distributive: if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning.
  E.g. count(), sum(), min(), max()
- Algebraic: if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate function.
  E.g. avg(), min_N(), standard$_d$*eviation*()
- Holistic: if there is no constant bound on the storage size needed to describe a sub-aggregate.
  E.g. median(), mode(), rank()

2.3.3 Typical OLAP Operations

- **Roll up (drill-up)**: summarize data by climbing up hierarchy or by dimension reduction.
- **Drill down (roll down)**: reverse of roll-up, from higher level summary to lower level summary or detailed data, or introducing new dimensions.
- **Slice and dice**: project and select.
- **Pivot (rotate)**: reorient the cube, visualization, 3D to series of 2D planes.
- **Drill across**: involving (across) more than one fact table.
- **Drill through**: through the bottom level of the cube to its back - end relational tables (using SQL).

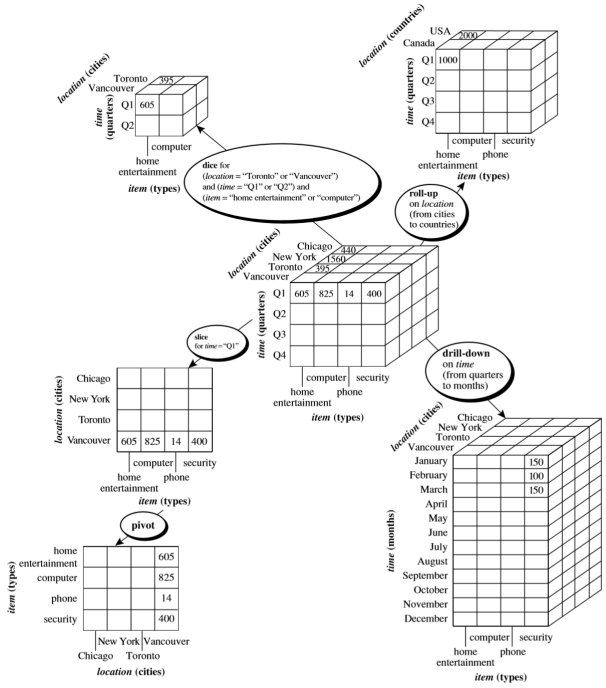

Figure 4: Examples of typical OLAP operations

2.4 Data Warehouse Design and Usage

2.4.1 Four views regarding the design of a data warehouse

- Top-down view: allows selection of the relevant information necessary for the data warehouse.
- Data source view: exposes the information being captured, stored, and managed by operational systems.
- Data warehouse view: consists of fact tables and dimension tables.
- Business query view: sees the perspectives of data in the warehouse from the view of end-user.

2.4.2 Data Warehouse Usage

- Information processing:supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs;
- Analytical processing: multidimensional analysis of data warehouse data. Or supports basic OLAP operations, slice-dice, drilling, pivoting;
- Data mining: knowledge discovery from hidden patterns. Supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

2.5 Data Warehouse Implementation

2.6 Data Generalization by Attribute-Oriented Induction

Data generalization: summarizes data by replacing relative low-level values with high-level concepts. Or, by reducing the number of dimensions to summarize data in concept space involving fewer dimensions.
Concept Description: generate descriptions for data characterization and comparison.

2.6.1 Basic Principles of Attribute-Oriented Induction

- Data focusing: task-relevant data, including dimensions, and the result is the initial relation;
- Attribute-removal: remove attribute A if there is a large set of distinct values for A but
  (1): There is no generalization operator on A , or
  (2): A 's higher level concepts are expressed in terms of other attributes;
- Attribute-generalization: If there is a large set of distinct values for A , and there exists a set of generalization operators on A , then select an operator and generalize A;
- Attribute-threshold control: typical 2-8, specified/default;
- Generalized relation threshold control: control the final relation/rule size.