An overview of Data Warehousing and OLAP Technology

Presenter: Pooyan Fazli
Discussion by Nguyet
Department of Computer Science
University of British Columbia

What is decision support?

• Decision support systems are a class of computerized information systems that support decision making activities.
• Decision support systems usually require consolidating data from many heterogeneous sources: these might include external sources. Such as stock market feeds.

What is a Data Warehouse?

Defined in many different ways:

• In simplest terms Data Warehouse can be defined as collection of Data marts
• A data warehouse is a “subject-oriented, integrated, time-variant, and nonvolatile” collection of data in support of management’s decision-making process.”—W. H. Inmon
• A data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker to make better decisions

Data Warehouse vs. Operational DBMS

• OLTP (on-line transaction processing)
  – Major task of traditional relational DBMS
  – Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
• OLAP (on-line analytical processing)
  – Major task of data warehouse system
  – Data analysis and decision making

Difference between OLAP and OLTP

<table>
<thead>
<tr>
<th></th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>Clerk, IT professional</td>
<td>Knowledge worker</td>
</tr>
<tr>
<td>Function</td>
<td>Day to day operations</td>
<td>Decision support</td>
</tr>
<tr>
<td>DB Design</td>
<td>Application-oriented</td>
<td>Subject-oriented</td>
</tr>
<tr>
<td>Data</td>
<td>Current, up-to-date,</td>
<td>Historical, summarized,</td>
</tr>
<tr>
<td></td>
<td>detailed</td>
<td>multidimensional...</td>
</tr>
<tr>
<td>Usage</td>
<td>Repetitive</td>
<td>Ad-hoc</td>
</tr>
<tr>
<td>Access</td>
<td>Read/write</td>
<td></td>
</tr>
<tr>
<td>Unit of work</td>
<td>Simple, single transaction</td>
<td>Complex query</td>
</tr>
<tr>
<td># rec accessed</td>
<td>100 MB-GB</td>
<td>100 GB-TR</td>
</tr>
<tr>
<td># users</td>
<td>Thousands</td>
<td>Hundreds</td>
</tr>
<tr>
<td>Metric</td>
<td>Transaction throughput</td>
<td>Query throughput</td>
</tr>
</tbody>
</table>

Why Separate Data Warehouse?

• High performance for both systems
  – DBMS—tuned for OLTP: access methods, indexing, concurrency control, recovery
  – Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
• Different functions and different data
  – OLTP: decision support requires historical data which operational DBs do not typically maintain
  – OLAP: data consolidation requires consolidation (aggregation, summarization) of data from heterogeneous sources
  – Data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
A Data Warehousing Architecture

Data Sources → Tier 1: Data Warehouse Server → Tier 2: OLAP Server → Tier 3: Clients

Data Extraction
- get data from multiple, heterogeneous, and external sources

Data Cleaning
- detect errors in the data and rectify them when possible

Data Transformation
- convert data from legacy or host format to warehouse format

Load
- sort, summarize, consolidate, compute views, check integrity, and build indices and partitions

Refresh
- propagate the updates from the data sources to the warehouse

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: taking a projection of the data on a subset of dimensions for selected values of the other dimension
- Pivot (rotate): reorient the cube, visualization, 3D to series of 2D planes

From Tables to Data Cubes
- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- In a multidimensional data model, there is a set of numeric measures that are the objects of analysis.
- Each of the numeric measures depends on a set of dimensions, which provide the context for the measure.

From Tables to Data Cubes
- The dimensions together are assumed to uniquely determine the measure.

- Each dimension is described by a set of attributes.

- The attributes of a dimension may be related via a hierarchy of relationships.

A Sample Data Cube
**Database Design Methodology**

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

**Snowflake Schema**
- A refinement of star schema where hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake

**Database Designs**

**Star Schema**
- Time
  - T_key
  - T_day
  - T_month
  - T_quarter
  - T_year
- Branch
  - B_key
  - B_name
  - B_type
- Location
  - L_key
  - L_name
  - L_type
- Sales Fact Table
  - S_key
  - S_type
- Item
  - I_key
  - I_name
  - I_brand
  - I_type
  - I_supplier_type
- Time
  - T_day
  - T_day_week
  - T_month
  - T_quarter
  - T_year
- Measures
  - Location_key
  - City
  - Province
  - Country
  - Location
  - Unit_sold
  - Dollars_sold
  - Avg_sales

**Snowflake Schema**
- Time
  - T_key
  - T_day
  - T_month
  - T_quarter
  - T_year
- Branch
  - B_key
  - B_name
  - B_type
- Location
  - L_key
  - L_name
  - L_type
- Sales Fact Table
  - S_key
  - S_type
- Item
  - I_key
  - I_name
  - I_brand
  - I_type
  - I_supplier_type
- Time
  - T_day
  - T_day_week
  - T_month
  - T_quarter
  - T_year
- Measures
  - Location_key
  - City
  - Province
  - Country
  - Location
  - Unit_sold
  - Dollars_sold
  - Avg_sales

**Challenges in exploiting materialized views**
- identify the views to materialize
- exploit the materialized views to answer queries,
- efficiently update the materialized views during load and refresh.

**Metadata Requirements**
- Administrative metadata
  - Source database and their contents
  - Source database and their contents
  - Back-end and front-end tools
  - Definitions of the warehouse schema
  - Pre-defined queries and reports
  - Data mart locations and contents
  - Data refresh and purging policies
  - User profiles and user access control policies
Metadata Requirements

- **Business metadata**
  - Business terms and definitions
  - Ownership of data
  - Charging policies

- **Operational metadata**
  - Data lineage: history of migrated data and sequence of transformations applied
  - Currency of data: active, archived, purged
  - Monitoring information: warehouse usage statistics, error reports, audit trails