Palestinian Central Bureau of Statistics

Databases, Data Warehouses and Statistical Dissemination Systems

Haitham Zeidan
Dissemination and Documentation Department
The Definitions

- **Database**: Organized collection of data. The data are typically organized to model relevant aspects of reality in a way that supports processes requiring this information.

- **Data Warehouse**: A database used for reporting and data analysis. Integrating data from one or more disparate sources creates a central repository of data, a data warehouse (DW). Data warehouses store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons.

- **Statistical Dissemination System / Statistical Data Warehouse**: **???**
Data warehouse basic terminology

- **Aggregate data**: obtained by applying aggregations (count, sum, avg, etc.) over elementary data (aka raw data or microdata)

- **Fact tables (D1, D2, …, Dn; M)**
  - dimension codes (used to group data and/or to consider only specific subsets of data)
  - measure(s) (possibly to be aggregated and deriving from microdata quantitative variables)
Dimensions and dimension levels: dimensions are often articulated in different dimension levels, e.g. a territorial dimension may comprise the levels: national, regional, municipality.

Data cube: the association between dimension code combination and measure is represented by a n-dimensional hypercube.
A relational perspective for data cubes

- **Star schema:** the codes in the fact table/data cube are decoded by (possibly denormalized) single dimensional tables.
DATA WAREHOUSES

- **Data Warehouses:** Data spread in several databases — physically located at numerous sites.
- Data warehouse — repository of multiple DBs in single schema; resides at single site.
- Data warehousing processes:
  - ✓ Data Cleaning
  - ✓ Data Integration
  - ✓ Data Transformation
  - ✓ Data Loading
  - ✓ Periodic data refreshing
Data warehouse diagram
Data warehousing processes

- **Data cleaning:** Data Cleaning includes filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

- **Data integration:** Data Integration includes integration of multiple databases, data cubes, or files.

- **Data transformation:** Convert data from legacy or host format to warehouse format.

- **Load:** Sort, summarize, consolidate; compute views; check integrity. Build indices and partitions.

- **Refresh:** Propagates the update from data sources to the warehouse.
Data warehousing processes

Data Cleaning

Data Integration

Data Transformation
Components of a data warehouse

- **Sources** — Data source interaction
- **Data Transformation**
- **Data warehouse** (data storage)
- **Reporting** (Data presentation)
- **Metadata**
Data Warehouse Advantages

- Complete control over the four main areas of data management systems:
  - Sources — Data source interaction
    - Clean data
    - Query processing: multiple options
    - Indexes: multiple types
    - Security: data and access
Data Warehousing Disadvantages

- Adding new data sources takes time and associated high cost.
- Data owners lose control over their data, raising ownership, security and privacy issues.
- Long initial implementation time and associated high cost.
- Difficult to accommodate changes in data types and ranges, data source schema, indexes and queries.
Characteristics of Data Warehousing

- Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example: "sales" can be a particular subject.

- Integrated: A data warehouse integrates data from multiple data sources. For example: Source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

- Time Variant: Historical data is kept in a data warehouse. For example: One can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse.

- Non volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

- It must be optimized for access to very large amount of data.

- It is based on client server architecture.

- It is capable of handling dynamic matrices.

- It maintains transparency.

- It is consistent and flexible.
Three kinds of data warehouse applications

- **Information processing**: Supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs.

- **Analytical processing**:
  - Multidimensional analysis of data warehouse data
  - 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.

- Differences among the three tasks
In the next few years, data warehousing is expected to make big strides in software, especially for optimizing queries:

- indexing very large tables
- enhancing SQL
- improving data compression
- expanding dimensional modeling
- Real-Time Data Warehousing
- Multiple Data Types
- Adding Unstructured Data
DATA WAREHOUSE USAG (2)

- Searching Unstructured Data
- Spatial Data
- Data Visualization
- Major Visualization Trends
- Visualization Types
- Advanced Visualization Techniques Chart Manipulation.
- Drill Down.
- Advanced Interaction
Thank you for your kind Attention