The Evolution of Data Warehousing

- Since 1970s, organizations gained competitive advantage through systems that automate business processes to offer more efficient and cost-effective services to the customer.

- This resulted in accumulation of growing amounts of data in operational databases.

- Organizations now focus on ways to use operational data to support decision-making, as a means of gaining competitive advantage.

- However, operational systems were never designed to support such business activities.

- Businesses typically have numerous operational systems with overlapping and sometimes contradictory definitions.

- Organizations need to turn their archives of data into a source of knowledge, so that a single integrated / consolidated view of the organization’s data is presented to the user.

- A data warehouse was deemed the solution to meet the requirements of a system capable of supporting decision-making, receiving data from multiple operational data sources.
Data Warehousing Concepts

- A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management’s decision-making process (Inmon, 1993).

Subject-oriented Data

- Warehouse is organized around major subjects of the enterprise (e.g. customers, products, sales) rather than major application areas (e.g. customer invoicing, stock control, product sales).
- This is reflected in the need to store decision-support data rather than application-oriented data.

Integrated Data

- The data warehouse integrates corporate application-oriented data from different source systems, which often includes data that is inconsistent.
- The integrated data source must be made consistent to present a unified view of the data to the users.

Time-Variant Data

- Data in the warehouse is only accurate and valid at some point in time or over some time interval.
- Time-variance is also shown in the extended time that data is held, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots.
**Non-Volatile Data**

- Data in the warehouse is not updated in real-time but is refreshed from operational systems on a regular basis.
- New data is always added as a supplement to the database, rather than a replacement.

**Data Webhouse**

- Web is an immense source of behavioral data as individuals interact through their Web browsers with remote Web sites. Data generated by this behavior is called **clickstream**.
- A data webhouse is a distributed data warehouse with no central data repository that is implemented over the Web to harness clickstream data.

**Benefits of Data Warehousing**

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers

**Comparison of OLTP Systems and Data Warehousing**

<table>
<thead>
<tr>
<th>OLTP systems</th>
<th>Data warehousing systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holds current data</td>
<td>Holds historical data</td>
</tr>
<tr>
<td>Stores detailed data</td>
<td>Stores detailed, lightly, and highly summarized data</td>
</tr>
<tr>
<td>Data is dynamic</td>
<td>Data is largely static</td>
</tr>
<tr>
<td>Repetitive processing</td>
<td><em>Ad hoc</em>, unstructured, and heuristic processing</td>
</tr>
<tr>
<td>High level of transaction throughput</td>
<td>Medium to low level of transaction throughput</td>
</tr>
<tr>
<td>Predictable pattern of usage</td>
<td>Unpredictable pattern of usage</td>
</tr>
<tr>
<td>Transaction-driven</td>
<td>Analysis driven</td>
</tr>
<tr>
<td>Application-oriented</td>
<td>Subject-oriented</td>
</tr>
<tr>
<td>Supports day-to-day decisions</td>
<td>Supports strategic decisions</td>
</tr>
<tr>
<td>Serves large number of clerical/operational users</td>
<td>Serves relatively low number of managerial users</td>
</tr>
</tbody>
</table>
Data Warehouse Queries

- Types of queries that a data warehouse is expected to answer ranges from the relatively simple to the highly complex and is dependent on the type of end-user access tools used.

- End-user access tools include:
  - Reporting, query, and application development tools
  - Executive information systems (EIS)
  - OLAP tools
  - Data mining tools

Examples of Typical Data Warehouse Queries

- What was total revenue for Scotland in third quarter of 2001?
- What was total revenue for property sales for each type of property in Great Britain in 2000?
- What are the three most popular areas in each city for the renting of property in 2001 and how does this compare with the figures for the previous two years?
- What is monthly revenue for property sales at each branch office, compared with rolling 12-monthly prior figures?
- What would be effect on property sales in the different regions of Britain if legal costs went up by 3.5% and Government taxes went down by 1.5% for properties over £100,000?

Problems of Data Warehousing

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization

Problems of Data Warehousing

- High demand for resources
- Data ownership
- High maintenance
- Long duration projects
- Complexity of integration
**Operational Data Sources**

- Mainframe first generation hierarchical and network databases.
- Departmental proprietary file systems (e.g., VSAM, RMS) and relational DBMSs (e.g., Informix, Oracle).
- Private workstations and servers.
- External systems such as the Internet, commercially available databases, or databases associated with an organization’s suppliers or customers.

**Operational Data Store (ODS)**

- Repository of current and integrated operational data used for analysis.
- Often structured and supplied with data in the same way as the data warehouse.
- May act simply as a staging area for data to be moved into the warehouse.
- Often created when legacy operational systems are found to be incapable of achieving reporting requirements.
- Provides users with the ease of use of a relational database while remaining distant from the decision support functions of the data warehouse.

**Load Manager**

- Performs all the operations associated with the extraction and loading of data into the warehouse.
- Size and complexity will vary between data warehouses and may be constructed using a combination of vendor data loading tools and custom-built programs.
Warehouse Manager

- Performs all the operations associated with the management of the data in the warehouse.
- Constructed using vendor data management tools and custom-built programs.

- Operations performed include
  - Analysis of data to ensure consistency.
  - Transformation and merging of source data from temporary storage into data warehouse tables.
  - Creation of indexes and views on base tables.
  - Generation of denormalizations (if necessary).
  - Generation of aggregations (if necessary).
  - Backing-up and archiving data.

Warehouse Manager

- In some cases, also generates query profiles to determine which indexes and aggregations are appropriate.

- Query profile can be generated for each user, group of users, or data warehouse and is based on information that describes characteristics of the queries such as frequency, target table(s), and size of results set.

Query Manager

- Performs all the operations associated with the management of user queries.

- Typically constructed using vendor end-user data access tools, data warehouse monitoring tools, database facilities, and custom-built programs.

- Complexity determined by the facilities provided by the end-user access tools and the database.
Query Manager

- The operations performed by this component include directing queries to the appropriate tables and scheduling the execution of queries.

- In some cases, the query manager also generates query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Detailed Data

- Stores all the detailed data in the database schema.

- In most cases, the detailed data is not stored online but aggregated to the next level of detail.

- On a regular basis, detailed data is added to the warehouse to supplement the aggregated data.

Lightly and Highly Summarized Data

- Stores all the pre-defined lightly and highly aggregated data generated by the warehouse manager.

- Transient as it will be subject to change on an on-going basis in order to respond to changing query profiles.

Lightly and Highly Summarized Data

- The purpose of summary information is to speed up the performance of queries.

- Removes the requirement to continually perform summary operations (such as sort or group by) in answering user queries.

- The summary data is updated continuously as new data is loaded into the warehouse.
**Archive / Backup Data**

- Stores detailed and summarized data for the purposes of archiving and backup.
- May be necessary to backup online summary data if this data is kept beyond the retention period for detailed data.
- The data is transferred to storage archives such as magnetic tape or optical disk.

**Meta-data**

- This area of the warehouse stores all the meta-data (data about data) definitions used by all the processes in the warehouse.

**Meta-data**

- Used for a variety of purposes
  - Extraction and loading processes – meta-data is used to map data sources to a common view of information within the warehouse.
  - Warehouse management process – meta-data is used to automate the production of summary tables.
  - Query management process – meta-data is used to direct a query to the most appropriate data source.

- The structure of meta-data will differ between each process, because the purpose is different.
- This means that multiple copies of meta-data describing the same data item are held within the data warehouse.
- Most vendor tools for copy management and end-user data access use their own versions of meta-data.
**Meta-data**

- Copy management tools use meta-data to understand the mapping rules to apply in order to convert the source data into a common form.

- End-user access tools use meta-data to understand how to build a query.

- The management of meta-data within the data warehouse is a very complex task that should not be underestimated.

**End-User Access Tools**

- The principal purpose of data warehousing is to provide information to business users for strategic decision-making.

- These users interact with the warehouse using end-user access tools.

- The data warehouse must efficiently support *ad hoc* and routine analysis.

**End-User Access Tools**

- High performance is achieved by pre-planning the requirements for joins, summations, and periodic reports by end-users (where possible).

- There are five main groups of access tools:
  - Data reporting and query tools
  - Application development tools
  - Executive information system (EIS) tools
  - Online analytical processing (OLAP) tools
  - Data mining tools

**Data Warehouse Information Flows**
Data Warehouse Information Flows

- **Inflow** - Processes associated with the extraction, cleansing, and loading of the data from the source systems into the data warehouse.

- **Upflow** - Processes associated with adding value to the data in the warehouse through summarizing, packaging, and distribution of the data.

- **Downflow** - Processes associated with archiving and backing-up/recovery of data in the warehouse.

- **Outflow** - Processes associated with making the data available to the end-users.

- **Metaflow** - Processes associated with the management of the meta-data.

Data Warehousing Tools and Technologies

- Building a data warehouse is a complex task because there is no vendor that provides an ‘end-to-end’ set of tools.

- Necessitates that a data warehouse is built using multiple products from different vendors.

- Ensuring that these products work well together and are fully integrated is a major challenge.

Extraction, Cleansing, and Transformation Tools

- Tasks of capturing data from source systems, cleansing and transforming it, and loading results into target system can be carried out either by separate products, or by a single integrated solution.

- Integrated solutions include:
  - Code Generators
  - Database Data Replication Tools
  - Dynamic Transformation Engines
Data Warehouse DBMS Requirements

- Load performance
- Load processing
- Data quality management
- Query performance
- Terabyte scalability
- Mass user scalability
- Networked data warehouse
- Warehouse administration
- Integrated dimensional analysis
- Advanced query functionality

Data Warehouse Parallel Database Technologies

- Aims to solve decision-support problems using multiple nodes working on the same problem.

- Performs many database operations simultaneously, splitting individual tasks into smaller parts so that tasks can be spread across multiple processors.

- Parallel DBMSs must be capable of running parallel queries, parallel data loading, table scanning, and data archiving and back up.

Data Warehouse Parallel Database Technologies

- Two main parallel hardware architectures include:
  - Symmetric Multi-processing (SMP)
  - Massively Parallel Processing (MPP)

- SMP - A set of tightly coupled processors that share memory and disk storage.

- MPP - A set of loosely coupled processors, each of which has its own memory and disk storage.

Data Warehouse Meta-data

- Meta-data is used for a variety of purposes and management of meta-data is a critical issue in achieving a fully integrated data warehouse.

- Problem is that meta-data has several functions in the data warehouse:
  - Data transformation and loading.
  - Data warehouse management.
  - Query generation.
Data Warehouse Meta-data

- Tools generate and use their own meta-data.
- Challenge is to synchronize meta-data between different products from different vendors using different meta-data stores.
- Two industry organizations: Meta Data Coalition (MDC) and Object Management Group (OMG) have merged to propose a single standard for meta-data and modeling in data warehousing called the Common Warehouse Metamodel (CWM).

Administration and Management Tools

- Monitoring data loading from multiple sources.
- Data quality and integrity checks.
- Managing and updating meta-data.
- Monitoring database performance to ensure efficient query response times and resource utilization.
- Auditing data warehouse usage to provide user chargeback information.

Administration and Management Tools

- Replicating, subsetting, and distributing data.
- Maintaining efficient data storage management.
- Purging data.
- Archiving and backing-up data.
- Implementing recovery following failure.
- Security management.

Typical Data Warehouse and Data Mart Architecture
Data Mart

◆ A subset of a data warehouse that supports the requirements of a particular department or business function.

◆ Characteristics include:
  – Focuses on only the requirements of one department or business function.
  – Do not normally contain detailed operational data unlike data warehouses.
  – More easily understood and navigated.

Reasons for Creating a Data Mart

◆ To give users access to the data they need to analyze most often.

◆ To provide data in a form that matches the collective view of the data by a group of users in a department or business function area.

◆ To improve end-user response time due to the reduction in the volume of data to be accessed.

Reasons for Creating a Data Mart

◆ To provide appropriately structured data as dictated by the requirements of the end-user access tools.

◆ Building a data mart is simpler compared with establishing a corporate data warehouse.

◆ The cost of implementing data marts is normally less than that required to establish a data warehouse.

Reasons for Creating a Data Mart

◆ Potential users of a data mart are more clearly defined and can be more easily targeted to obtain support for a data mart project rather than a corporate data warehouse project.
Data Marts Issues

- Data mart functionality
- Data mart size
- Data mart load performance
- Users access to data in multiple data marts
- Data mart Internet / Intranet access
- Data mart administration
- Data mart installation