

Data Warehouse

A subject-Oriented, Integrated, Time-variant, and non-volatile collection of data is support of management's decision-making process.

Abbreviated *DW*, a collection of data designed to support management decision making. Data warehouses contain a wide variety of data that present a coherent picture of business conditions at a single point in time.

Development of a data warehouse includes development of systems to extract data from operating systems plus installation of a warehouse Data Base systems that provides managers flexible access to the data.

The term data warehousing generally refers to the combination of many different databases across an entire enterprise. Contrast with data mart.

Benefits of Date Warehouse

1. High return on investment
2. Competitive advantage
3. Increase productivity of corporate decision-makers

Comparison between On-Line Transaction Processing(OLTP) and Data Warehouse

OLTP Systems	Data Warehousing Systems
1. Holds current data	1. Holds historic data
2. Stores detailed data	2. Stores detailed, lightly, and highly summarized data
3. Data is dynamic	3. Data is largely static
4. Repetitive processing	4. Unstructured and heuristic processing
5. High level of transaction throughput	5. Medium to low level of transaction throughput
6. Predictable pattern of usage	6. Unpredictable pattern of usage
7. Transaction Driven	7. Analysis driven
8. Application Oriented	8. Subject oriented
9. Support day to day decisions	9. Support strategic decisions
10. Serves large number of clerical users	10. Serves relatively lower number of managerial users

Data warehouse vs. Relational database

- Relational database
 - Good for:
 - Transactional handling
 - Ensures consistency

- Bad for:
 - Reports
 - Joining different data sources
- Data Warehouse
 - Reports
 - Understanding
 - A common place for storing the organizational data
 - Consistent because no user enters data

Consider Registration Data Base with the following relations:

Student, Course, Instructor, Section, Grades, Averages

Some Data Base queries:

1. Get the grades for each student in semester 20051.
2. Get the courses offered by instructor “Kadhim”
3. Change the student information.

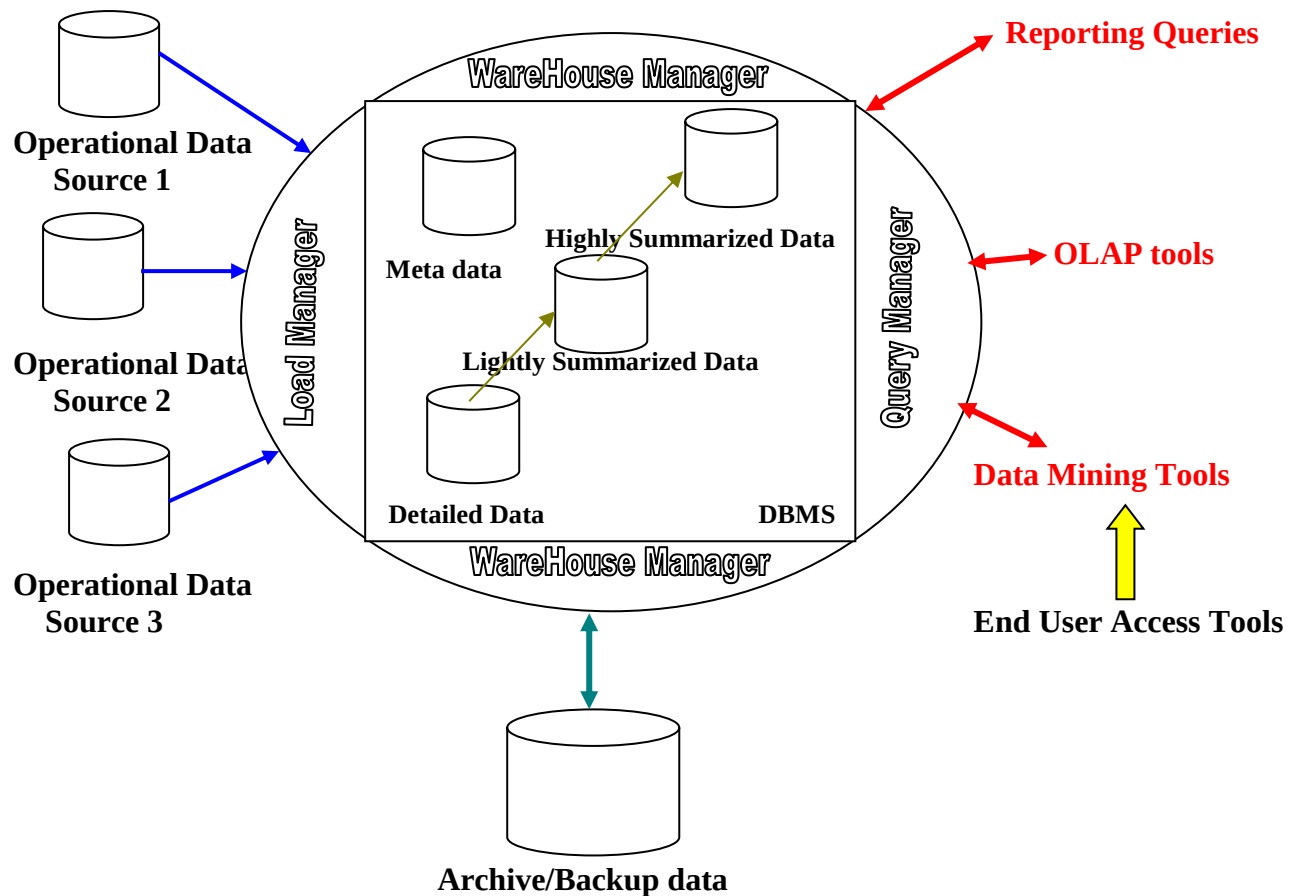
Some data warehouse queries

1. Obtain the students having maximum average in each semester for each department since 1995.
2. Obtain the courses with enrolment greater than the average enrolment criteria since 2001.
3. Suggest a time table for the next semester

Problems of Data Warehousing

1. Data ownership
2. High maintenance
3. Long duration projects
4. Complexity of integration
5. Underestimation of resources for data loading
6. Hidden problems with source systems
7. Required data not captured
8. Increased end user demands
9. Data homogenization
10. High demand of resources

Data Warehouse Architecture



Typical Architecture of Data Warehouse

Operational Data:

Daily transaction data on different systems (Oracle, SQL server, Informix, ...) and external data such as Internet, ... and from different sites and sources.

Load Manager:

The load manager (also called the front-end component) performs all the operations associated with the extraction and loading of data into the warehouse.

Warehouse Manager:

The warehouse manager performs all the operations associated with the management of the data in the warehouse.

1. Analysis of data to ensure consistency
2. Transformation and merging of source data from temporary storage into warehouse tables.
3. Creation of indexes and views on base tables.
4. Generation of aggregation (if necessary).
5. Backup and archiving data.

Query Manager:

The query +Manager (also called the back-end component) performs all the operations associated with the management of user queries.

Detailed Data:

This area of the Warehouse stores all the detailed data in the data base schema.

Lightly and Highly Summarized Data

This area of the warehouse stores all the predefined lightly and highly summarized (Aggregate) data generated by the warehouse manager. The purpose of the summarized data is to speed up the performance of the queries. The summary data is updated continuously as new data is loaded into the warehouse.

Example

Detailed data

Student Number	Course Number	Section Number	Semester	Grade

Lightly summarized data

Student Number	Semester	Average

Highly summarized data

Student Number	Overall Average

Archive/Backup Data

This area of the warehouse stores the detailed and summarized data for the purpose of archiving and backup

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 is used for a variety of purposes including:

1. Mapping data sources to a common view of the information within the warehouse.

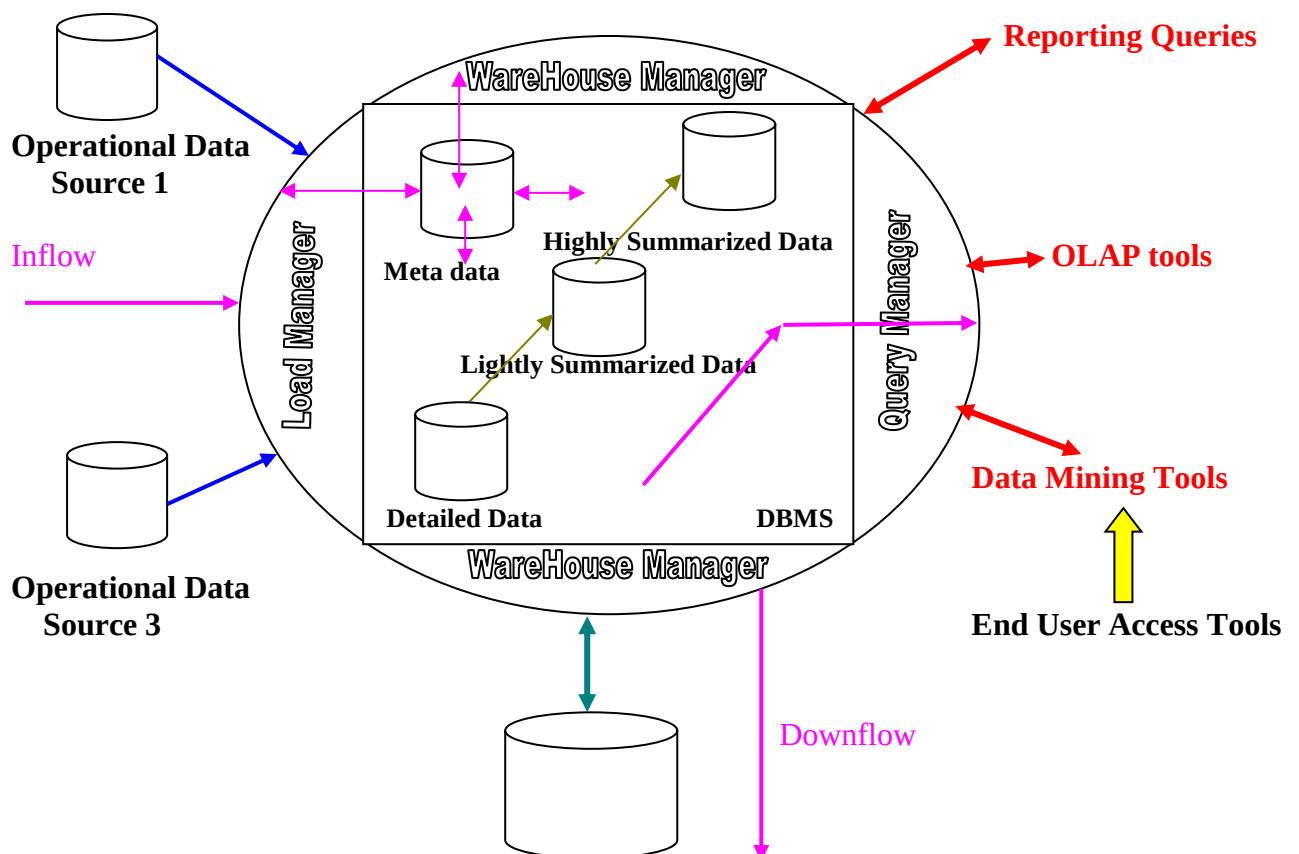
2. Automating the production of summary tables.
3. Directing a query to the most appropriate data source.

End User Access tools

The principle 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.

1. Reporting and query tools
2. Application development tools
3. Executive Information System (EIS) tools
4. On Line Analytical Processing (OLAP)
5. Data Mining Tools

Data Warehouse Information Flows



Archive/Backup data

Inflow: The process associated with the extraction, cleansing, and loading of the data from the source systems into the data warehouse.

Upflow: The processes associated with adding value to the data in the warehouse through summarizing, packaging, and distribution of the data.

Downflow: The processes associated with archiving and backing up/recovery of data in the warehouse.

Outflow: The processes associated with making the data available to the end users.

Metaflow: The processes associated with the management of the meta-data.

Data Mart:

It is a subset of data warehouse that supports the requirements of a particular department or business function.

1. A data mart focuses on only the requirements of users associated with one department or business function.
2. Data mart contains less information compared with data warehouse, data marts are more easily understood and navigated.
3. Data marts do not normally contain detailed operational data unlike data warehouses

Reasons for creating a Data Mart

1. To give users access to the data they need to analyze most often.
2. 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.
3. To improve end – users response time due to the reduction in the volume of the data to be accessed.
4. The cost of implementing data marts is normally less than that required to establish a data warehouse.
5. Operations such as cleansing, loading, transformation, and integration are easier.
6. The potential users of a data mart are more clearly defined.

Data Marts Issues

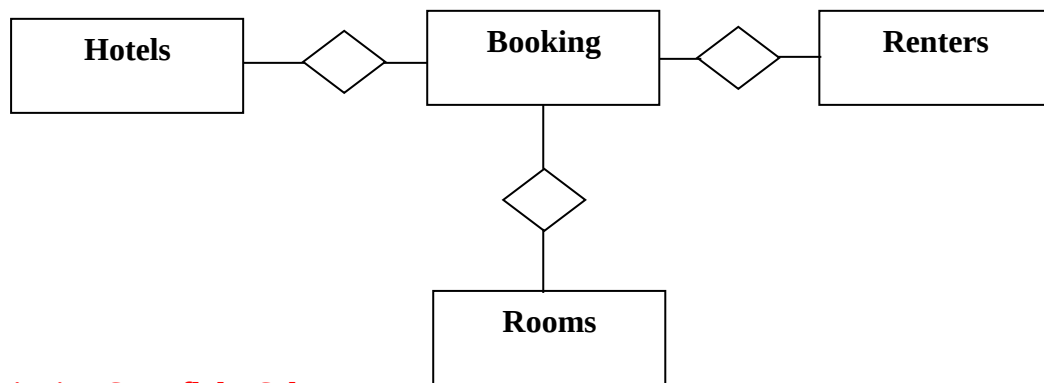
1. Data mart functionality
2. Data mart size
3. Data mart load performance
4. Users access to data in multiple data marts
5. Data mart Internet/Intranet access
6. Data mart administration
7. Data mart installation

Designing Data Warehouses

1. Designing Star Schemas

It is a logical structure that has a fact table in the center, surrounded by dimension tables (containing reference data).

Example



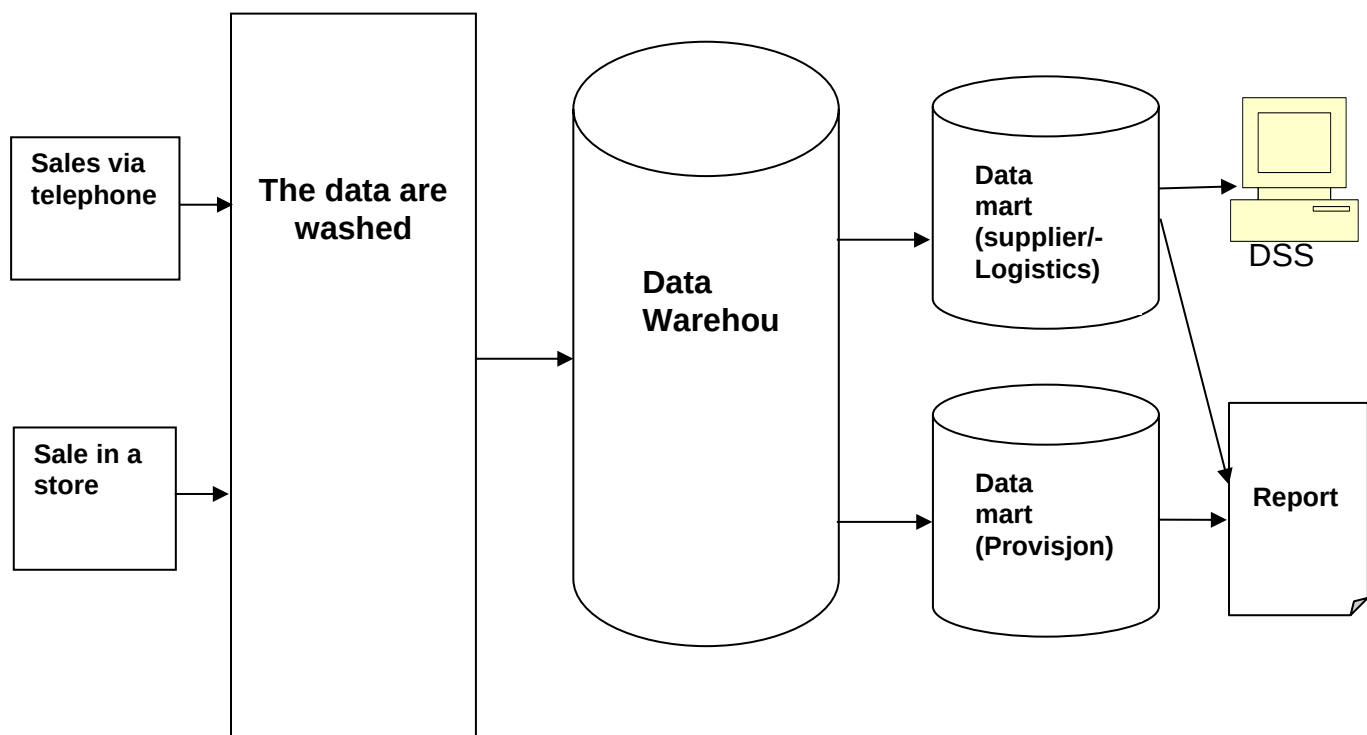
2. Designing Snowflake Schemas

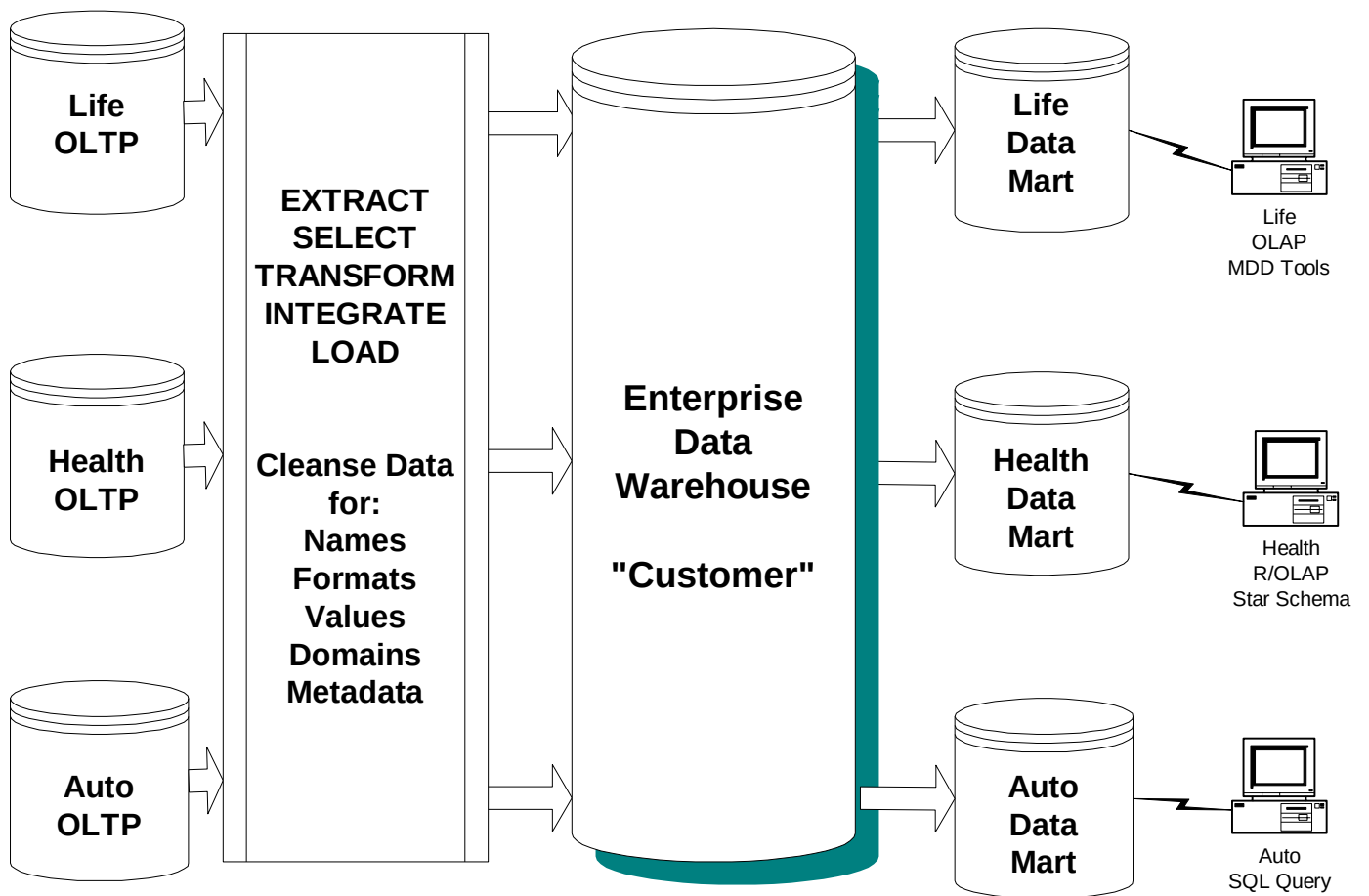
It is a variant of the star schema where each dimension can have its own dimensions.

3. Designing Starflake Schemas

Hybrid structure that contains a mixture of demoralized star and normalized snowflake schemas.

Examples: Data Warehouse and Data Mart





Online Transaction Processing (OLTP)

- The Online Transaction Processing is characterized by
 - Handling high transaction throughput with transaction typically making small changes to organization's operational data
 - Making changes to the data to handle day to day operations
- The OLTP systems can range from small database of few megabytes to large databases requiring terabytes or Pet bytes of storage

From OLTP To Data Warehousing for Decision Making

- Corporate decision-makers access to all organization's data wherever it is located
- To Provide comprehensive analysis of the organization, its business, its requirements and any trends requires:
 - Access not only to the current values in database but also historical data,
- To facilitate this type of analysis, the data warehouse has been created to:
 - hold data drawn from several data sources and
 - Maintained by different operating units, together with historical and summary transformation

The Data Warehouse (DW)

- It is based on the extension of the databases technology to provide the management of the data store or historical data
- To satisfy the decision-makers requirement of analysis tools, two powerful analysis tools associated to DW have emerged over the last few years:
 - Online analytical Processing (OLAP) and
 - Data mining tools

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.

The Evolution of Data Warehousing

- 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.

The Evolution of Data Warehousing

- 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

- The warehouse is organized around the major subjects of the enterprise (e.g. customers, products, and sales) rather than the major application areas (e.g. customer invoicing, stock control, and 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 the 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 Warehouse

A data warehouse is data management and data analysis technology

Benefits of Data Warehousing

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