**Definition:** A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management’s decisions.

**A further definition:** The data in the data warehouse is separate, available, integrated, time-stamped, subject-oriented, nonvolatile, and accessible.

**CHARACTERISTICS OF DATA WAREHOUSE DATA**

**Subject Orientation:** In every company, data in a database or file for an operational system are organized according to the applications or business processes of the company. For an order processing system, the data may be organized by customer, product, order, invoice, and accounts. For a healthcare organization, the data may be organized according to the patients, appointments, claims, accounts, and so on.

In a data warehouse, data are organized by business subjects. Business subjects differ from company to company. These are the parameters critical for the enterprise to succeed. For a manufacturing company, the critical subjects are: inventory, sales, and shipments. For a retail store, the critical subjects are: inventory, sales, orders, and suppliers.
Integrated: This is the most important aspect of data in the data warehouse. Data in a data warehouse comes from many operational systems as well as from outside sources. Data structures in the database or files of these systems are designed by different architects and developed in different platforms. They are bound to be different. Naming conventions for files or tables and attributes, as well as data types will be different.

Before the data from various disparate sources can be usefully stored in a data warehouse, we have to remove the inconsistencies and standardize the various data elements. We have to go through the processes of data cleansing, transformation, consolidation, and integration.

Figure below shows an inconsistency of data in gender as stored in four operational systems by male/female, m/f, 1/0, and x/y. Whatever way we want to encode in the data warehouse, it has to be done consistently. The same consideration of consistency goes for all data in the warehouse.
**Nonvolatile:** Data in an operational system is meant to run day-to-day business of a company, and as such data is inserted, deleted, and updated as soon as there is a transaction. But once data is transferred into a data warehouse, they are not typically changed.

Data from the operational systems are moved into a data warehouse at specific intervals and en masse. Depending on the business requirements, these data are moved once a day, once a week, or once in two weeks. In fact, in a typical data warehouse, data movements of different data sets may take place at different frequencies. The units of sales may be transferred once a day and the customer data may be transferred once a month.

Once data are transferred to a data warehouse, it stays there – not deleted or generally updated, but accessed. In other words, data in a data warehouse is not as volatile as in operational systems. It is stored primarily for query and analysis.
**Time-Variant:** Data in an operational system stores data for current business operation, but data in a data warehouse is stored for historical purpose. If a user is looking for the buying pattern of a specific customer, the user needs data not only about the current purchase, but the past purchases as well. Thus data in a data warehouse contain both the current data as well as historical data.

The time horizon for the data warehouse is significantly longer than the operational systems. A 60-to-90-day time frame is normal for operational systems; a 5-to-10-year time horizon of data is normal for the data warehouse.

Data in an operational system is valid as of the moment of access. As such, they can change due to a transaction. Data in a data warehouse is nothing more than a sophisticated series of snapshots taken over past and current periods.

The key structure of operational data may or may not contain any time element such as year, month, week, or day. Every data structure in the data warehouse contains some element of time. Depending on the level of details, the sales quantity in a record may relate to a specific day, week, month, or quarter. The time-variant nature of data allows: analysis of the past, relate to current information, and predict the future.
**Data Granularity**: Granularity refers to the level of details or summarization held in the unit of data. Many times, granularity is considered to be a design issue rather than an inherent property of the data warehouse.

In an operational system, data is usually kept at the lowest level of detail. In a point-of-sale system for a grocery store, the units of sale of an item are captured and stored as individual transaction. Whenever the total sale of an item for a day or month is required, all transactions are added up. Summarized data are not kept in an operational system.

Data in a data warehouse are kept in summarized form in various levels. When a user queries a data warehouse for analysis, he or she usually starts with the highest level, then drills-down to lower levels. For example, a user may start with the units of sale of a product for an entire region, then look at the breakdown by states within a region. Next level might be the sales by individual stores.

Data granularity in a data warehouse refers to the level of detail. The more detail there is, the higher the level of granularity. The less detail there is, the lower the granularity. Of course, one has to decide how much detail of data to store in a data warehouse to compromise between the granularity and the expected system performance for query analysis.
COMPONENTS OF A DATA WAREHOUSE: ARCHITECTURE

Data warehouse is not a single software or physical component. It is a collection of activities and components that are built and used on a continuing basis. Thus from the viewpoint of development and use, we can consider data warehouse as consisting of four basic components: Source data, data staging area, data presentation area, and data access tools.
Source Data: This is the area from where data warehouse gets operational data. It is not really a part of the data warehouse; however, it is important to have the operational data available for fast query processing. Source data can be divided into four categories: production data, internal data, archived data, and external data.

Production Data: This category of data comes from various operational systems of the enterprise. Based on the requirements of the data warehouse, we choose segments of data from different operational systems. When dealing with this data, we come across many variations of data formats.

Data may also reside on different software (operating system + database) and hardware platforms. The greatest challenge of a data warehouse analyst is to standardize these data in a common format before transferring them in the data warehouse.

Internal Data: These are the data that are kept by individuals in spreadsheets, databases, text files, and other formats in the departmental level. Internal data adds additional complexity such as how to collect data from spreadsheet and compare with database and text data, and so on, and then convert to a common format.

Archived Data: In every operational system, the old data is periodically transferred for archival such that the system runs efficiently. These data typically reside in flat files in tape cartridges or microfilms. These kind of data are useful discerning patterns and analyzing trends. But again, these data need to be retrieved and converted to a common format.

External Data: Most executives depend on data from external sources such as market share data of competitors, production schedule of the competitors, and statistics relating to the industry. These data will have different format, and thus routine conversion mechanisms must be created to standardize these data into a common format.
**Data Staging Area:** The data staging area is both a storage area and a set of processes commonly referred to as **Extract-Transformation-Load or ETL.** It is somewhat analogous to a kitchen where raw materials are combined to form a fine meal.

The extracted data from source data areas need to be changed, converted, and made ready in a format suitable for querying and analysis.

Three major functions need to be performed to get the data ready: *data extraction, data transformation, and data loading.*

**Data Extraction:** As data comes from various sources (files, databases, etc.), formats, and platforms, appropriate techniques must be used to extract these data and placed in a separate physical environment. The result might be a group of flat files. Any mechanism that is placed for data extraction, should be a continuous effort for future data source.

**Data Transformation:** Data transformation requires combining many forms of data. It requires several steps. First, the extracted data are cleaned, which might require elimination of duplicates, standardization of data such as ‘TX’ and ‘Texas’ should be in a common format ‘TX’, or providing a default value for missing data elements.

Standardization of data elements forms a large part of data transformation. Data transformation also requires purging unwanted data. Sorting and merging takes place on a large scale in a data staging area. When data transformation ends, we have a collection of integrated data that is cleaned, standardized, and summarized.

**Data Loading:** After data is transformed into a common format, they are ready to be loaded to the data warehouse. Two tasks are commonly involved. Initially, a large volume of data is loaded when the design of data warehouse is completed. As the data warehouse starts functioning, there is an ongoing loading of data as they become available.
Data Storage or Presentation Area: The data storage or presentation area is where data is organized, stored, and made available for direct querying by users, report writers, and other analytical applications. The storage area might be a single database or it might consist of multiple data marts. There can be various designs of the data storage: centralized database, independent data marts, federated, hub-and-spoke, and Bus.

Typically, data in a data warehouse are organized in a dimensional schema, which makes it simple for query processing a large volume of data as compared to the relational schema used for operational database that deals with fast transactions. Data in a data warehouse are “read-only” repositories of data.

Information Delivery/Data Access Tools: Data access tools are used to query data warehouse and provide necessary information. Various users may access through various mechanisms. Provision of ad hoc reports, complex queries, multidimensional analysis, and statistical analysis, and data mining, are part of this area.
Other Components

In addition to the four components mentioned above, there are other components of importance for a data warehouse such as: metadata, and management and control.

**Metadata:** Metadata is similar to the data dictionary in a database system. The data dictionary contains data about the data in the data warehouse, but not the data itself. Data in a data warehouse come from varied sources in varied timeframes. Everything that is done, from extracting to loading of data, are documented in the metadata, so that the processes can be repeated.

Once data are in the staging area, metadata guides an analyst in the transformation and loading processes, including staging files and target table layouts, transformation and cleansing rules, conformed dimension and fact definitions, aggregation definitions, and extract-transformation-load (ETL) schedules and run-log results.

**Management and Control:** This component sits on top of all the other components. It is like the management console of a database. It coordinates the services and activities within the data warehouse. This component may control the data transformation and data transfer into the data warehouse. It works with the metadata to make sure incoming data conforms to standards. It also moderates the information delivery to the users.
Differences between the Operational Database Systems and Data Warehouses

The major task of operational database systems is to provide the capabilities of on-line transaction and query processing for day-to-day business operations such as purchasing, inventory, banking, payroll, registration, and accounting. These systems are called on-line transaction processing (OLTP) systems.

Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. These systems are called on-line analytical processing (OLAP) systems. The major differences between the two systems are:

- **Users and Systems Orientation:** An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is market-oriented and it is used for data analysis by knowledge workers, including managers, executives, and analysts.

- **Data Content:** An OLTP system manages current data, that are typically too detailed, which cannot be easily used for decision making. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at various levels of granularity.

- **Database Design:** An OLTP system usually adopts an entity-relationship (ER) model and an application-oriented database design. An OLAP system typically adopts either a *star* or *snowflake* model and a subject-oriented database design.

- **Data Types:** An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in other organizations. In contrast, an OLAP system deals mainly with historical data and that originating from many sources.

- **Access Patterns:** The access patterns of an OLTP system consists mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, access to OLAP systems are mostly read-only operations, although many could be complex queries.
Table 2.1  Comparison between OLTP and OLAP systems.

<table>
<thead>
<tr>
<th>Feature</th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
<td>operational processing</td>
<td>informational processing</td>
</tr>
<tr>
<td>Orientation</td>
<td>transaction</td>
<td>analysis</td>
</tr>
<tr>
<td>User</td>
<td>clerk, DBA, database professional</td>
<td>knowledge worker (e.g., manager, executive, analyst)</td>
</tr>
<tr>
<td>Function</td>
<td>day-to-day operations</td>
<td>long-term informational requirements, decision support</td>
</tr>
<tr>
<td>DB design</td>
<td>ER based, application-oriented</td>
<td>star/snowflake, subject-oriented</td>
</tr>
<tr>
<td>Data</td>
<td>current; guaranteed up-to-date</td>
<td>historical; accuracy maintained over time</td>
</tr>
<tr>
<td>Summarization</td>
<td>primitive, highly detailed</td>
<td>summarized, consolidated</td>
</tr>
<tr>
<td>View</td>
<td>detailed, flat relational</td>
<td>summarized, multidimensional</td>
</tr>
<tr>
<td>Unit of work</td>
<td>short, simple transaction</td>
<td>complex query</td>
</tr>
<tr>
<td>Access</td>
<td>read/write</td>
<td>mostly read</td>
</tr>
<tr>
<td>Focus</td>
<td>data in</td>
<td>information out</td>
</tr>
<tr>
<td>Operations</td>
<td>index/hash on primary key</td>
<td>lots of scans</td>
</tr>
<tr>
<td>Number of records</td>
<td>tens</td>
<td>millions</td>
</tr>
<tr>
<td>accessed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of users</td>
<td>thousands</td>
<td>hundreds</td>
</tr>
<tr>
<td>DB size</td>
<td>100 MB to GB</td>
<td>100 GB to TB</td>
</tr>
<tr>
<td>Priority</td>
<td>high performance, high availability</td>
<td>high flexibility, end-user autonomy</td>
</tr>
<tr>
<td>Metric</td>
<td>transaction throughput</td>
<td>query throughput, response time</td>
</tr>
</tbody>
</table>

**Assignment:**

Go to the web and search for words like *business intelligence (BI)*, *data warehouse*, *data mining*, *OLAP*, and etc. Find a suitable company that provides software or services on any of the areas mentioned. Write a report on your finding – include graphs and tables as necessary. Limit your report to within 5 pages – single or double-spaced. Submit your report at the beginning of next class period.