**Features of Data Warehouse( DW)**

* **Subject oriented data**

**Rather than company operations, a data warehouse typically provides information on a specific topic (such as sales inventory or supply chain).**

* **Integrated data**

**Establishing a common unit of measurement for all related data in a data warehouse using data from different databases is the process of integrating data. You must store data within it in a simple and universally acceptable manner.**

**It must also be consistent in terms of nomenclature and layout. This type of application is useful for analyzing** [**big data**](https://intellipaat.com/blog/tutorial/big-data-and-hadoop-tutorial/introduction-to-big-data-2/)**.**

* **Time variant data**

**Time-Variant – In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems. The data resided in data warehouse is predictable with a specific interval of time and delivers information from the historical perspective. It comprises elements of time explicitly or implicitly. Another feature of time-variance is that once data is stored in the data warehouse then it cannot be modified, alter, or updated. Data is stored with a time dimension, allowing for analysis of data over time.**

* **Non volatile Data**

**The data warehouse is also non-volatile, which means that past data cannot be erased. The information is read-only and is only modified on a routine basis. It also helps with statistical data evaluation and comprehension of what and when events occurred. You don’t require any other complicated procedure.**

* **Data Granularity**

**Data granularity is a measure of the level of detail in a data structure. In time-series data, for example, the granularity of measurement might be based on intervals of years, months, weeks, days, or hours. For ordering transactions, granularity might be at the purchase order level, or line item level, or detailed configuration level for customised parts. The name field could represent the full name or have separate entries for first name, middle name, and last name.**

**Components of data warehouse architecture:**

* **Source data**

**Production Data:** This type of data comes from the different operating systems of the enterprise. Based on the data requirements in the data warehouse, we choose segments of the data from the various operational modes.

**Internal Data:** In each organisation, the client keeps their "private" spreadsheets, reports, customer profiles, and sometimes even department databases. This is the internal data, part of which could be useful in a data warehouse.

**Archived Data:** Operational systems are mainly intended to run the current business. In every operational system, we periodically take the old data and store it in achieved files.

**External Data:** Most executives depend on information from external sources for a large percentage of the information they use. They use statistics associating to their industry produced by the external department**.**

* **Data Staging**

**Data Extraction:** This method has to deal with numerous data sources. We have to employ the appropriate techniques for each data source.

**Data Transformation:** As we know, data for a data warehouse comes from many different sources. If data extraction for a data warehouse poses big challenges, data transformation presents even significant challenges. We perform several individual tasks as part of data transformation.

First, we clean the data extracted from each source. Cleaning may be the correction of misspellings or may deal with providing default values for missing data elements, or elimination of duplicates when we bring in the same data from various source systems.

Standardisation of data components forms a large part of data transformation. Data transformation contains many forms of combining pieces of data from different sources. We combine data from single source record or related data parts from many source records.

**Data Loading:** Two distinct categories of tasks form data loading functions. When we complete the structure and construction of the data warehouse and go live for the first time, we do the initial loading of the information into the data warehouse storage. The initial load moves high volumes of data using up a substantial amount of time.

* **Data Storage**:Data storage for the data warehousing is a split repository. The data repositories for the operational systems generally include only the current data. Also, these data repositories include the data structured in highly normalized for fast and efficient processing.

**Data warehouses versus Data Marts**

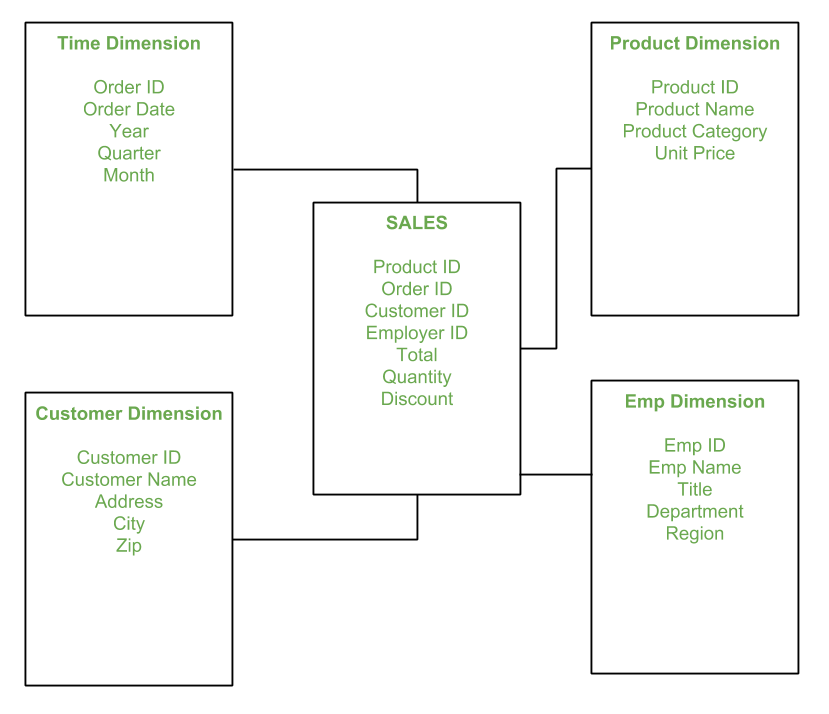
| **Parameter** | **Data Warehouse** | **Data Mart** |
| --- | --- | --- |
| **Definition** | **A Data Warehouse is a large repository of data collected from different organizations or departments within a corporation.** | **A data mart is an only subtype of a Data Warehouse. It is designed to meet the need of a certain user group.** |
| **Usage** | **It helps to take a strategic decision.** | **It helps to take tactical decisions for the business.** |
| **Objective** | **The main objective of Data Warehouse is to provide an integrated environment and coherent picture of the business at a point in time.** | **A data mart mostly used in a business division at the department level.** |
| **Designing** | **The designing process of Data Warehouse is quite difficult.** | **The designing process of Data Mart is easy.** |
|  | **May or may not use in a dimensional model. However, it can feed dimensional models.** | **It is built focused on a dimensional model using a start schema.** |
| **Data Handling** | **Data warehousing includes large area of the corporation which is why it takes a long time to process it.** | **Data marts are easy to use, design and implement as it can only handle small amounts of data.** |
| **Focus** | **Data warehousing is broadly focused all the departments. It is possible that it can even represent the entire company.** | **Data Mart is subject-oriented, and it is used at a department level.** |
| **Data type** | **The data stored inside the Data Warehouse are always detailed when compared with data mart.** | **Data Marts are built for particular user groups. Therefore, data short and limited.** |
| **Subject-area** | **The main objective of Data Warehouse is to provide an integrated environment and coherent picture of the business at a point in time.** | **Mostly hold only one subject area- for example, Sales figure.** |
| **Data storing** | **Designed to store enterprise-wide decision data, not just marketing data.** | **Dimensional modeling and star schema design employed for optimizing the performance of access layer.** |
| **Data type** | **Time variance and non-volatile design are strictly enforced.** | **Mostly includes consolidation data structures to meet subject area’s query and reporting needs.** |
| **Data value** | **Read-Only from the end-users standpoint.** | **Transaction data regardless of grain fed directly from the Data Warehouse.** |
| **Scope** | **Data warehousing is more helpful as it can bring information from any department.** | **Data mart contains data, of a specific department of a company. There are maybe separate data marts for sales, finance, marketing, etc. Has limited usage** |
| **Source** | **In Data Warehouse Data comes from many sources.** | **In Data Mart data comes from very few sources.** |
| **Size** | **The size of the Data Warehouse may range from 100 GB to 1 TB+.** | **The Size of Data Mart is less than 100 GB.** |
| **Implementation time** | **The implementation process of Data Warehouse can be extended from months to years.** | **The implementation process of Data Mart is restricted to few months.** |

| **S.N o** | **ER Modeling** | **Dimensional Modeling** |
| --- | --- | --- |
| **1** | **It is transaction-oriented.** | **It is subject-oriented.** |
| **2** | **Entities and Relationships.** | **Fact Tables and Dimension Tables.** |
| **3** | **Few levels of granularity.** | **Multiple levels of granularity.** |
| **4** | **Real-time information.** | **Historical information.** |
| **5** | **It eliminates redundancy.** | **It plans for redundancy.** |
| **6** | **High transaction volumes using few records at a time.** | **Low transaction volumes using many records at a time.** |
| **7** | **Highly Volatile data.** | **Non-volatile data.** |
| **8** | **Physical and Logical Model.** | **Physical Model.** |
| **9** | [**Normalization**](https://www.geeksforgeeks.org/introduction-of-database-normalization/) **is suggested.** | [**De-Normalization**](https://www.geeksforgeeks.org/denormalization-in-databases/) **is suggested.** |
| **10** | [**OLTP**](https://www.geeksforgeeks.org/on-line-transaction-processing-oltp-system-in-dbms/) **Application.** | [**OLAP**](https://www.geeksforgeeks.org/olap-operations-in-dbms/) **Application.** |
| **Ex** | **The application is used for buying products from e-commerce websites like Amazon.** | **Application to analyze buying patterns of the customer of the various cities over the past 10 years.** |

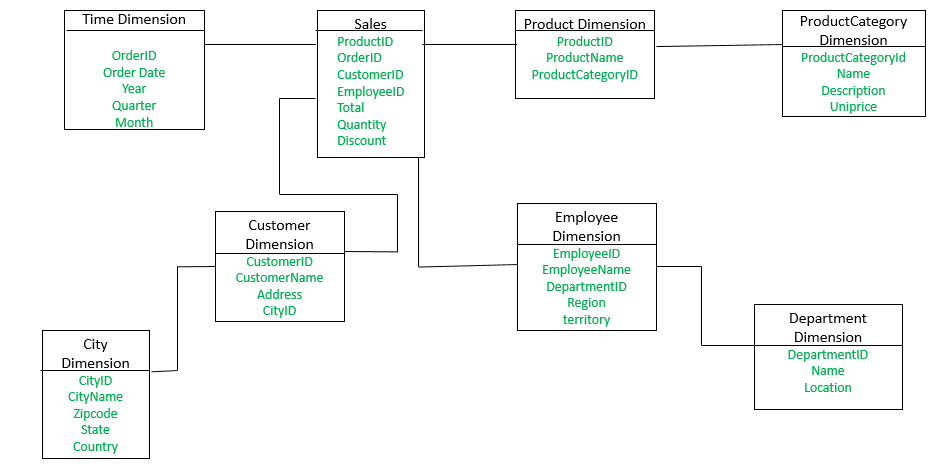
**Data WareHouse Schemas.**

* **Star Schema:** Star schema is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables. The star schema is a necessary cause of the snowflake schema. It is also efficient for handling basic queries.

It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star’s points. Below is an example to demonstrate the Star Schema:

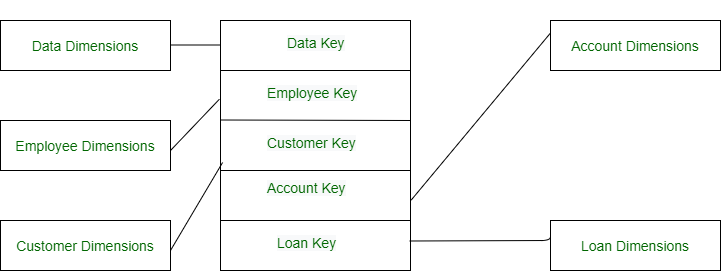


* **Snow Flake Schema: The snowflake schema is a variant of the star schema. Here, the centralized fact table is connected to multiple dimensions. In the snowflake schema, dimensions are present in a normalized form in multiple related tables. The snowflake structure materialized when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent tables. The snowflake effect affects only the dimension tables and does not affect the fact tables.**

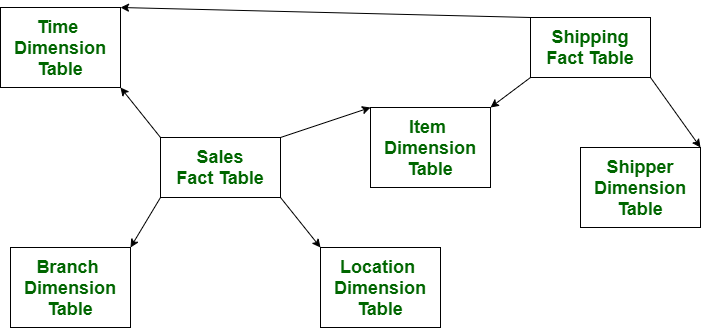
****

* **Factless fact table: Factless tables simply mean the key available in the fact that no remedies are available. Factless fact tables are only used to establish relationships between elements of different dimensions. And are also useful for describing events and coverage, meaning tables contain information that nothing has happened. It often represents many-to-many relationships.**

**The only thing they have is an abbreviated key. They still represent a focal phenomenon that is identified by the combination referenced in the dimension tables.**

****

* **Fact Constellation schema: This schema is a type of multidimensional model. In this, dimension tables are shared by many fact tables. The fact constellation schema consists of more than one star schema at a time. Unlike the star schema, it is not easy to operate, as it has more joins between the tables. Unlike the Star schema, fact constellation schema uses heavily complex queries to access data from the database.**

****

**OLAP vs OLTP**

**Online Analytical Processing (OLAP): Online Analytical Processing consists of a type of software tools that are used for data analysis for business decisions. OLAP provides an environment to get insights from the database retrieved from multiple database systems at one time. Examples – Any type of Data warehouse system is an OLAP system. The uses of OLAP are as follows:**

**Spotify analyzed songs by users to come up with a personalized homepage of their songs and playlist.**

**Netflix movie recommendation system.**

**Online transaction processing (OLTP): Online transaction processing provides transaction-oriented applications in a 3-tier architecture. OLTP administers the day-to-day transactions of an organization.**

**Examples: Uses of OLTP are as follows:**

**ATM center is an OLTP application.**

**OLTP handles the ACID properties during data transactions via the application.**

**It’s also used for Online banking, Online airline ticket booking, sending a text message, add a book to the shopping cart.**

| **No.** | **Category** | **OLAP (Online analytical processing)** | **OLTP (Online transaction processing)** |
| --- | --- | --- | --- |
| **1.** | **Definition** | **It is well-known as an online database query management system.** | **It is well-known as an online database modifying system.** |
| **2.** | **Data source** | **Consists of historical data from various Databases.** | **Consists of only of operational current data.** |
| **3.** | **Method used** | **It makes use of a data warehouse.** | **It makes use of a standard database management system (DBMS).** |
| **4.** | **Application** | **It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc.** | **It is application-oriented. Used for business tasks.** |
| **5.** | **Normalized** | **In an OLAP database, tables are not normalized.** | **In an OLTP database, tables are normalized (3NF).** |
| **6.** | **Usage of data** | **The data is used in planning, problem-solving, and decision-making.** | **The data is used to perform day-to-day fundamental operations.** |
| **7.** | **Task** | **It provides a multi-dimensional view of different business tasks.** | **It reveals a snapshot of present business tasks.** |
| **8.** | **Purpose** | **It serves the purpose to extract information for analysis and decision-making.** | **It serves the purpose to Insert, Update, and Delete information from the database.** |
| **9.** | **Volume of data** | **A large amount of data is stored typically in TB, PB** | **The size of the data is relatively small as the historical data is archived. For ex MB, GB** |
| **10.** | **Queries** | **Relatively slow as the amount of data involved is large. Queries may take hours.** | **Very Fast as the queries operate on 5% of the data.** |
| **11.** | **Update** | **The OLAP database is not often updated. As a result, data integrity is unaffected.** | **The data integrity constraint must be maintained in an OLTP database.** |
| **12.** | **Backup and Recovery** | **It only need backup from time to time as compared to OLTP.** | **Backup and recovery process is maintained rigorously** |
| **13.** | **Processing time** | **The processing of complex queries can take a lengthy time.** | **It is comparatively fast in processing because of simple and straightforward queries.** |
| **14.** | **Types of users** | **This data is generally managed by CEO, MD, GM.** | **This data is managed by clerks, managers.** |
| **15.** | **Operations** | **Only read and rarely write operation.** | **Both read and write operations.** |
| **16.** | **Updates** | **With lengthy, scheduled batch operations, data is refreshed on a regular basis.** | **The user initiates data updates, which are brief and quick.** |
| **17.** | **Nature of audience** | **Process that is focused on the customer.** | **Process that is focused on the market.** |
| **18.** | **Database Design** | **Design with a focus on the subject.** | **Design that is focused on the application.** |
| **19.** | **Productivity** | **Improves the efficiency of business analysts.** | **Enhances the user’s productivity.** |

**Steps in ETL**

**ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse. The process of ETL can be broken down into the following three stages:**

* **Extract: The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area.**
* **Transform: In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.**
* **Load: After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.**