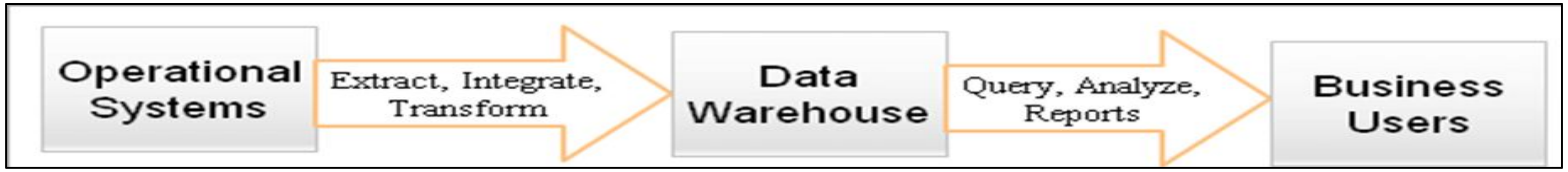
Database Design & Applications

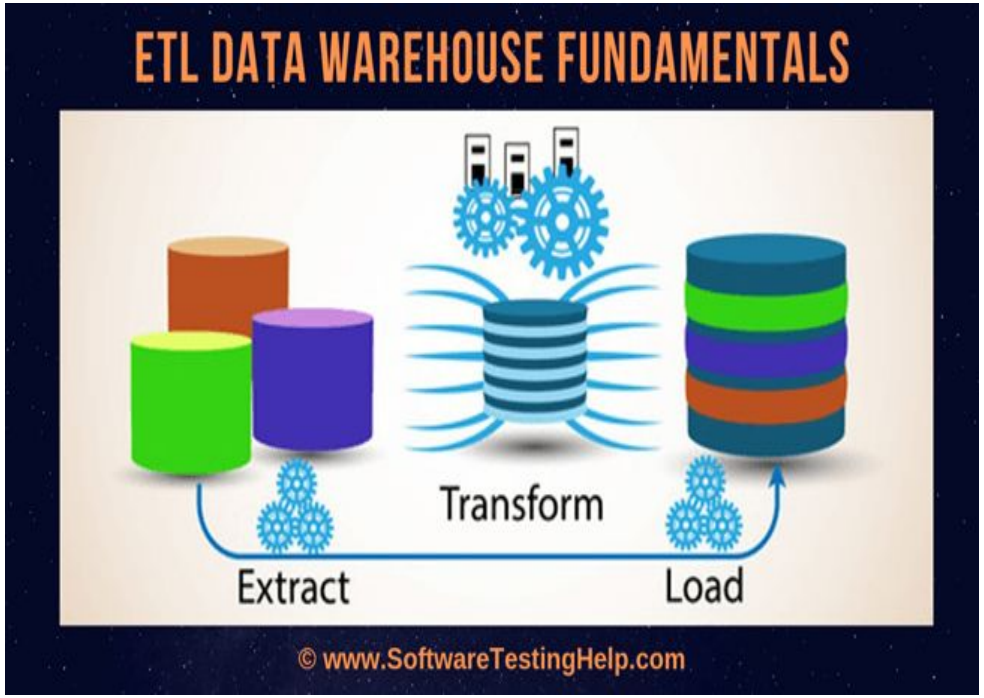
Data Warehousing and Data Mining

What Is Data Warehousing?

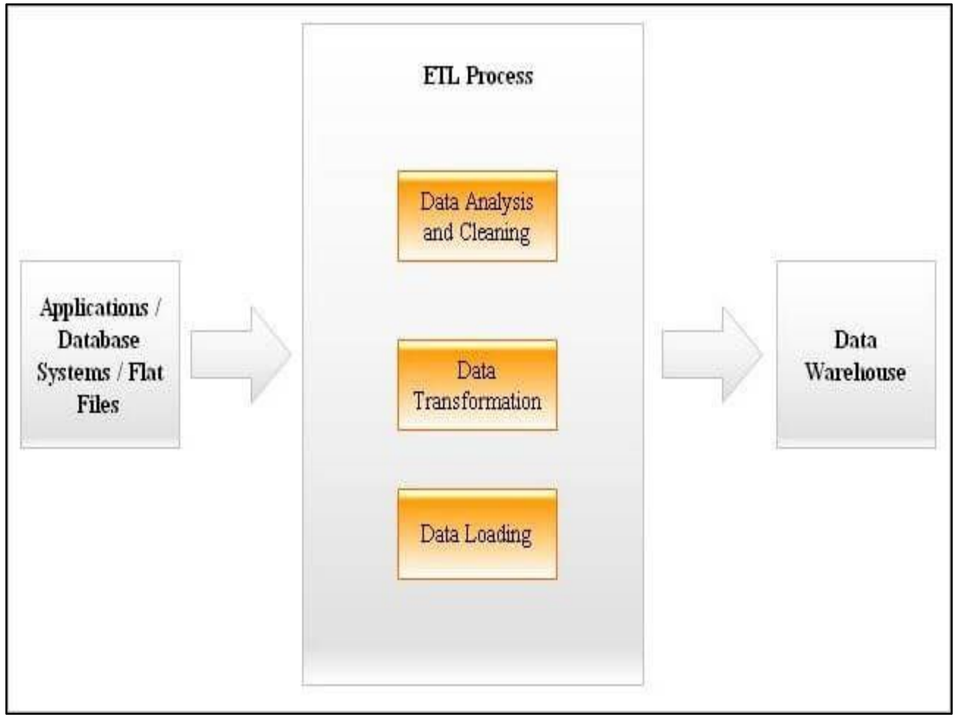
- A Data Warehouse (DW) is a repository of huge amount of organized data.
- This data is consolidated from one or more different data sources such as databases, applications, and flat files .
- DW is a relational database that is mainly designed for analytical reporting and on-time decision making in organizations.
- The data for this purpose is isolated and optimized from the source transaction data, which will not have any impact on the main business.
- If an organization introduces any business change, then DW is used to examine the effects of that change, and hence DW is also used to monitor the non-decision making process.
- The data warehouse is mostly a read-only system as operational data is very much separated from DW. This provides an environment to retrieve the highest amount of data with good query writing.
- Thus DW will act as the backend engine for Business Intelligence tools which shows the reports, dashboards for the business users. DW is extensively used in banking, financial, retail sectors, etc.

ETL (Extract, Transform, Load) Process In Data Warehouse?





ETL DATA WAREHOUSE FUNDAMENTALS
Transform
Extract
Load
© www.SoftwareTestingHelp.com



ETL Process
Data Analysis and Cleaning
Data Transformation
Data Loading
Applications / Database Systems / Flat Files
Data Warehouse

ETL Process In Data Warehouse?

- The data into the system is gathered from one or more operational systems, flat files, etc. The process which brings the data to DW is known as ETL Process. Extraction, Transformation, and Loading are the tasks of ETL.
- **Extraction:**
  - All the preferred data from various source systems such as databases, applications, and flat files is identified and extracted.
  - Data extraction can be completed by running jobs during non-business hours.
- **Transformation:**
  - Most of the extracted data can't be directly loaded into the target system. Based on the business rules, some transformations can be done before loading the data.
- **Loading:**
  - All the gathered information is loaded into the target Data Warehouse tables.

Introduction to Data Warehousing

- Data warehouse gathers all the operational data from several heterogeneous sources of "different formats" and through the process of extract, transform and load (ETL) it loads the data into DW in a "standardized dimensional format" across an organization.
- Data warehouse maintains both "current data and historical data" for analytical reporting and fact-based decision making.
- It helps organizations to take "smarter and quick decisions" on reducing costs and to increase the revenue, by comparing quarter and annual reports to improve their performance.

Types Of Data Warehouse Applications

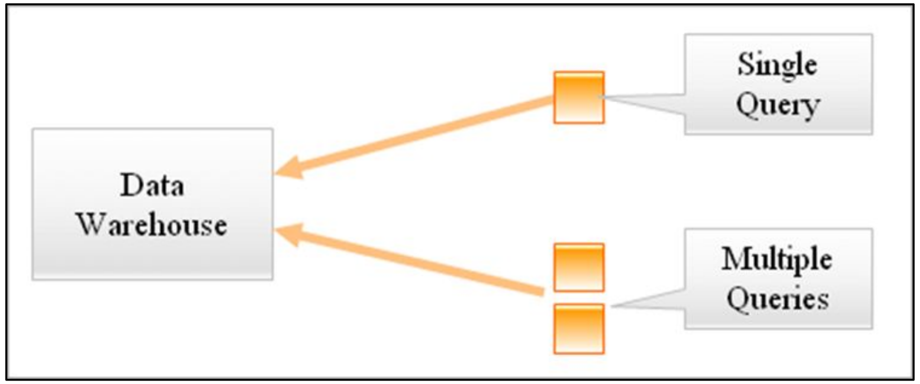
- Information processing
- Analytical processing
- Data mining which serves the purpose of BI

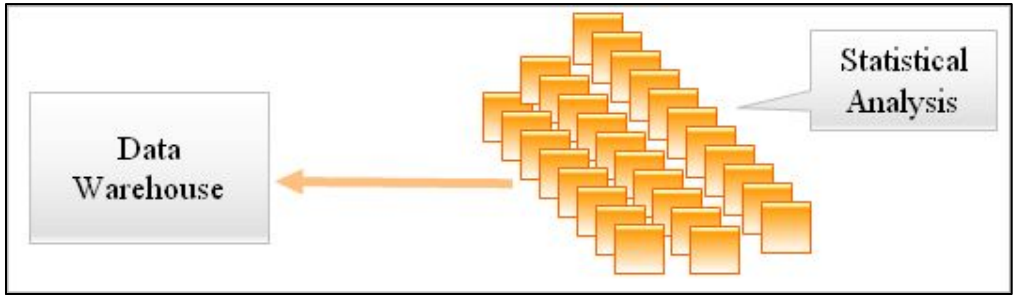
Information Processing

- This is a kind of application where the data warehouse allows direct one-one contact with the data stored in it.
- As the data can be processed by writing direct queries on the data (or) with a basic statistical analysis on the data and the end results will be reported to the business users in the form of reports, tables, charts or graphs.
- DW supports the following tools for Information Processing:
  - **Query Tool**
  - Reporting Tool
  - Statistics Tool

Information Processing

- **Query Tools:** The business (or) the analyst runs the queries using query tools to explore the data and generate the output in the form of reports or graphics as per the business requirement.
- **Reporting Tools:** If the business wants to see the results in any defined format and on a scheduled basis i.e. daily, weekly or monthly then reporting tools will be used. These kinds of reports can be saved and reviewed at any time.
- **Statistics Tools:** If the business wants to do an analysis on a broad view of data then statistics tools will be used to generate such results. Businesses can make conclusions and predictions by understanding these strategic results.





Analytical Processing

This is a kind of application where a data warehouse allows the analytical processing of data stored in it. The data can be analyzed by the following operations:

- Slice-and-Dice
- Drill Down
- Roll Up
- Pivoting.

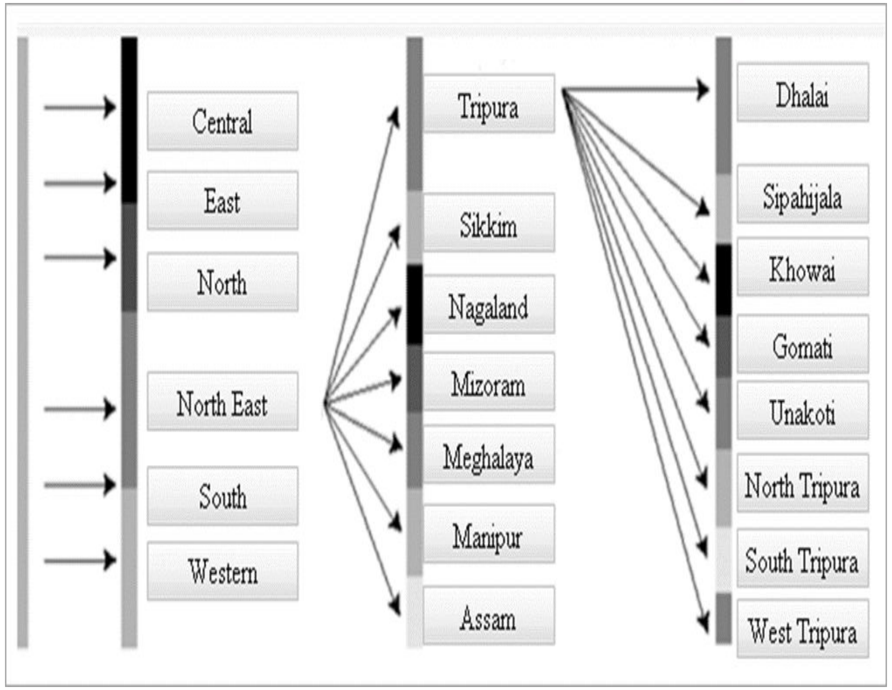
1. **Slice-and-Dice:**
    - To analyze the data accessed from many levels with a combination of different perspectives.
    - The slice-and-dice operation internally uses the drill-down mechanism.
    - Slicing works on dimensional data.
    - As a part of the business requirement, if we focus on a single area then slicing analyzes the dimensions of that particular area as per the requirements and gives the results.
    - Dicing works on analytic operations. Dicing zooms for a specific set of attributes over all the dimensions to provide diverse perspectives. The dimensions are considered from one or more consecutive slices.

Analytical Processing

2. **Drill Down:**

- If the business wants to go to a more detailed level of any summary number, then drill down is an operation for navigating down that summary to minor detailed levels.
- This gives a great idea of what is happening and where the business has to be focused more closely.
- Drill down tracks from the hierarchy level until the minor detail level for the root cause analysis.
- Example of sales drill down can happen from :

Country-level -> Region level -> State-level -> District level -> Store level.



Analytical Processing

3. **Roll up:**

- Roll up works opposite to the drill-down operation.
- If the business wants any summarized data, then roll up comes into the picture.
- It aggregates the detail level data by moving up in the dimensional hierarchy.
- Roll-ups are used to analyze the development and performance of a system.
- Example: In a sales roll up where the totals can be rolled up from

City level -> State-level -> Region level -> Country level.

4. **Pivot:**

- Pivoting analyzes dimension data by rotating the data on the cubes.
- For Example, the row dimension can be swapped into the column dimension and vice versa.

Data Mining

- This is a kind of application where the data warehouse allows knowledge discovery of the data and results will be represented with visualization tools.
- In the analytical and application types of applications, the information can be driven by the users.
- As the data goes vast in various businesses, it is difficult to query and drill down the data warehouse to get all possible insights into data.
- Then data mining accomplishes the discovery of knowledge.
- The data can be discovered by finding hidden patterns, associations, classifications, and predictions.
- Data mining goes in-depth with the data to predict the future. Based on the predictions, it also suggests the actions to take.

Data Mining

- **Activities of Data Mining:**
  - **Patterns**: Data mining discovers patterns that occur in the database. Users can provide the business inputs on which some knowledge of the patterns is expected for decision making.
  - **Associations/Relationships**: Data mining discovers relationships between the objects with the frequency of their association rules. This relationship may be between two or more objects (or) it may discover the rules within the properties of the same object.
  - **Classification**: Data mining organizes data in a set of predefined classes. So if any object is picked up from the data, classification associates the respective class label to that object.
  - **Prediction**: Data mining compares a set of existing values to find the best possible future values/trends in business.

Characteristics Of A Data Warehouse

- A data warehouse is built based on the following characteristics of data:
- **Subject Oriented:**
  - data warehouse we can analyze data with respect to a specific subject area rather than the application of wise data.
  - This provides results that are more defined for easy decision making.
  - Example : In an education system, the subject areas could be students, subjects, marks, teachers, etc.
- **Integrated:**
  - The data in the data warehouse is integrated from distinct sources such as other relational databases, flat files, etc.
  - The data warehouse brings all this data in a consistent format across the whole system.

Characteristics Of A Data Warehouse

- **Non-volatile:**
  - Once the data is loaded into the data warehouse, it can’t be changed.
  - The frequent changes in the operational database can be loaded into a data warehouse on a scheduled basis, during this process, new data gets added, however, the earlier data is not erased and it remains as historical data.
- **Time-Variant:**
  - All the historical data along with the recent data in the Data warehouse play a crucial role to retrieve data of any duration of time.
  - If the business wants any reports, graphs, etc then for comparing it with the previous years and to analyze the trends, all the old data that are 6 months old, 1-year-old or even older data, etc. are required.

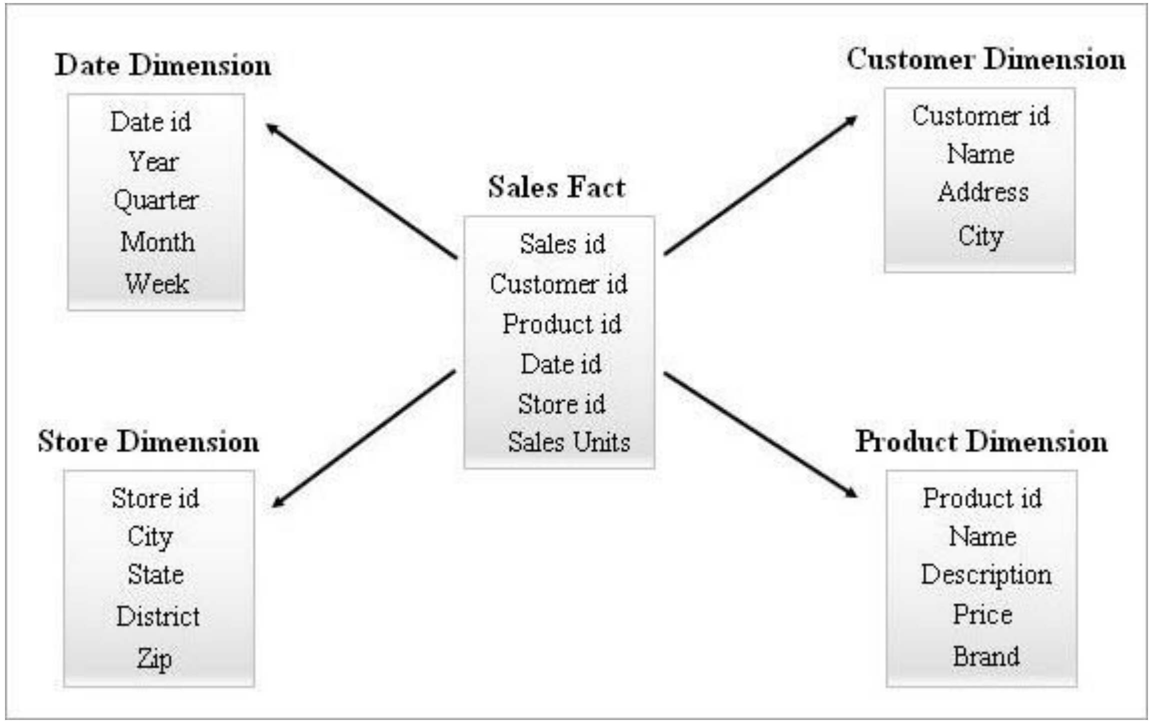
Data Warehouse Schema

- In a data warehouse, a schema is used to define the way to organize the system with all the database entities (fact tables, dimension tables) and their logical association.
- **Here are the different types of Schemas in DW:**
  - Star Schema
  - SnowFlake Schema
  - Galaxy Schema
  - Star Cluster Schema

Star Schema

- **Characteristics :**
  - This is the simplest and most effective schema in a data warehouse.
  - fact table in the center surrounded by multiple dimension tables resembles a star in the Star Schema model.
  - The fact table maintains one-to-many relations with all the dimension tables. Every row in a fact table is associated with its dimension table rows with a foreign key reference.
  - All the Business Intelligence (BI) tools greatly support the Star schema model.
  - While designing star schemas the dimension tables are purposefully de-normalized.
- **Benefits Of Star Schema:**
  - Queries use very simple joins while retrieving the data and thereby query performance is increased.
  - It is simple to retrieve data for reporting, at any point of time for any period.
- **Disadvantages Of Star Schema:**
  - If there are many changes in the requirements, the existing star schema is not recommended to modify and reuse in the long run.
  - Data redundancy is more as tables are not hierarchically divided.

Star Schema



Date Dimension

Date id
Year
Quarter
Month
Week

Customer Dimension

Customer id
Name
Address
City

Sales Fact

Sales id
Customer id
Product id
Date id
Store id
Sales Units

Store Dimension

Store id
City
State
District
Zip

Product Dimension

Product id
Name
Description
Price
Brand

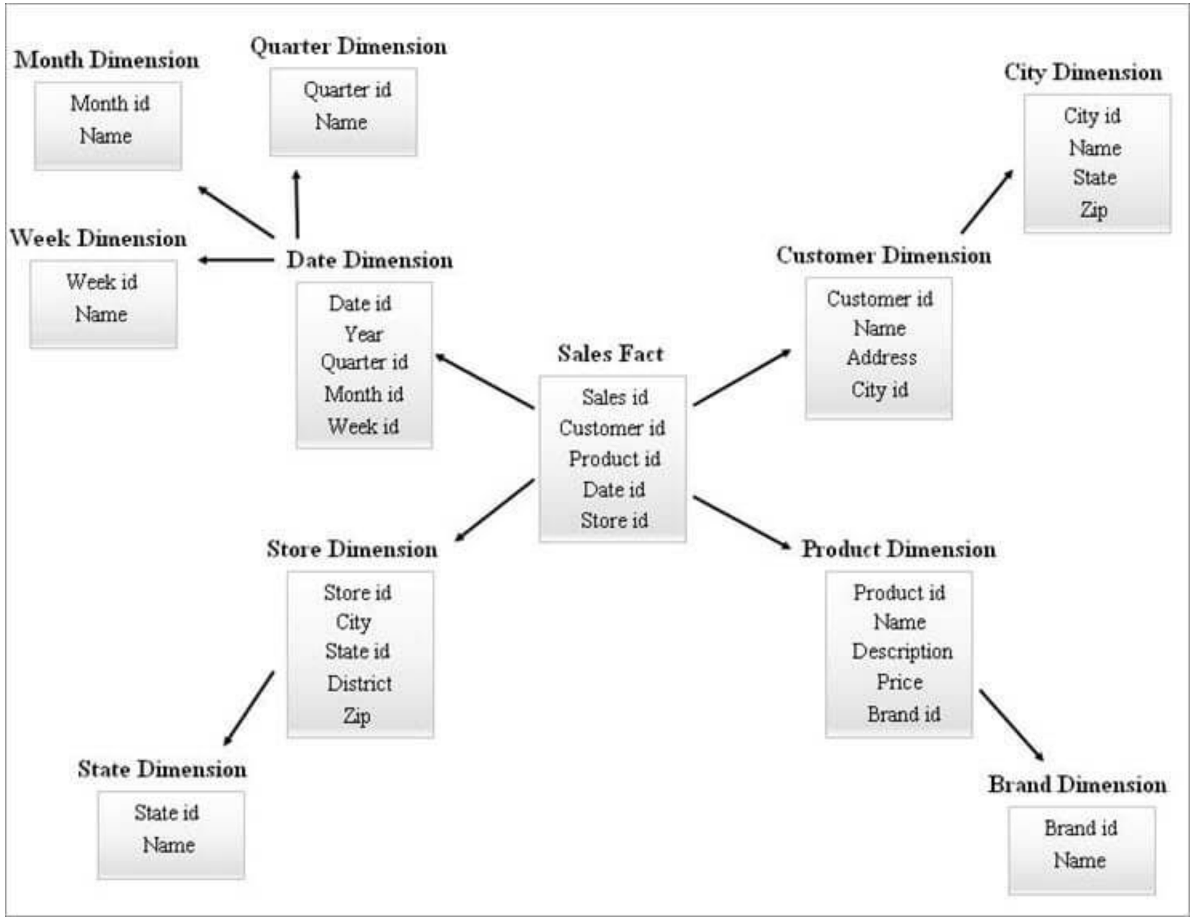
SnowFlake Schema

- **Characteristics :**
  - Star schema acts as an input to design a SnowFlake schema. Snow flaking is a process that completely normalizes all the dimension tables from a star schema.
  - The arrangement of a fact table in the center surrounded by multiple hierarchies of dimension tables looks like a SnowFlake in the SnowFlake schema model.
  - Every fact table row is associated with its dimension table rows with a foreign key reference.
  - While designing SnowFlake schemas the dimension tables are purposefully normalized. Foreign keys will be added to each level of the dimension tables to link to its parent attribute.
  - The complexity of the SnowFlake schema is directly proportional to the hierarchy levels of the dimension tables.

SnowFlake Schema

- **Benefits:**
  - Data redundancy is completely removed by creating new dimension tables.
  - When compared with star schema, less storage space is used by the Snow Flaking dimension tables.
  - It is easy to update (or) maintain the Snow Flaking tables.
- **Disadvantages:**
  - Due to normalized dimension tables, the ETL system has to load the number of tables.
  - You may need complex joins to perform a query due to the number of tables added. Hence query performance will be degraded.

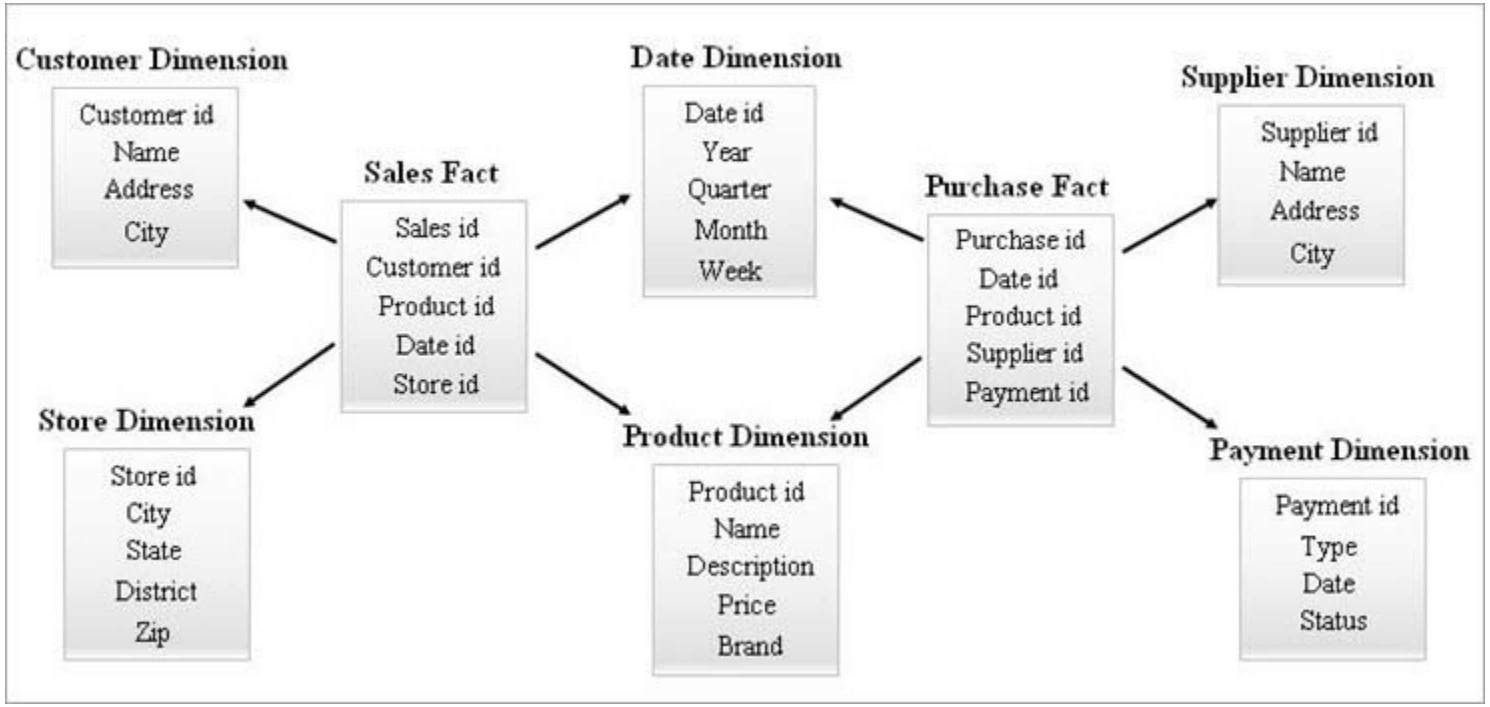
SnowFlake Schema



Galaxy Schema

- A galaxy schema is also known as Fact Constellation Schema.
- In this schema, multiple fact tables share the same dimension tables.
- The arrangement of fact tables and dimension tables looks like a collection of stars in the Galaxy schema model.
- The shared dimensions in this model are known as Conformed dimensions.
- This type of schema is used for sophisticated requirements and for aggregated fact tables that are more complex to be supported by the Star schema (or) SnowFlake schema.
- This schema is difficult to maintain due to its complexity.

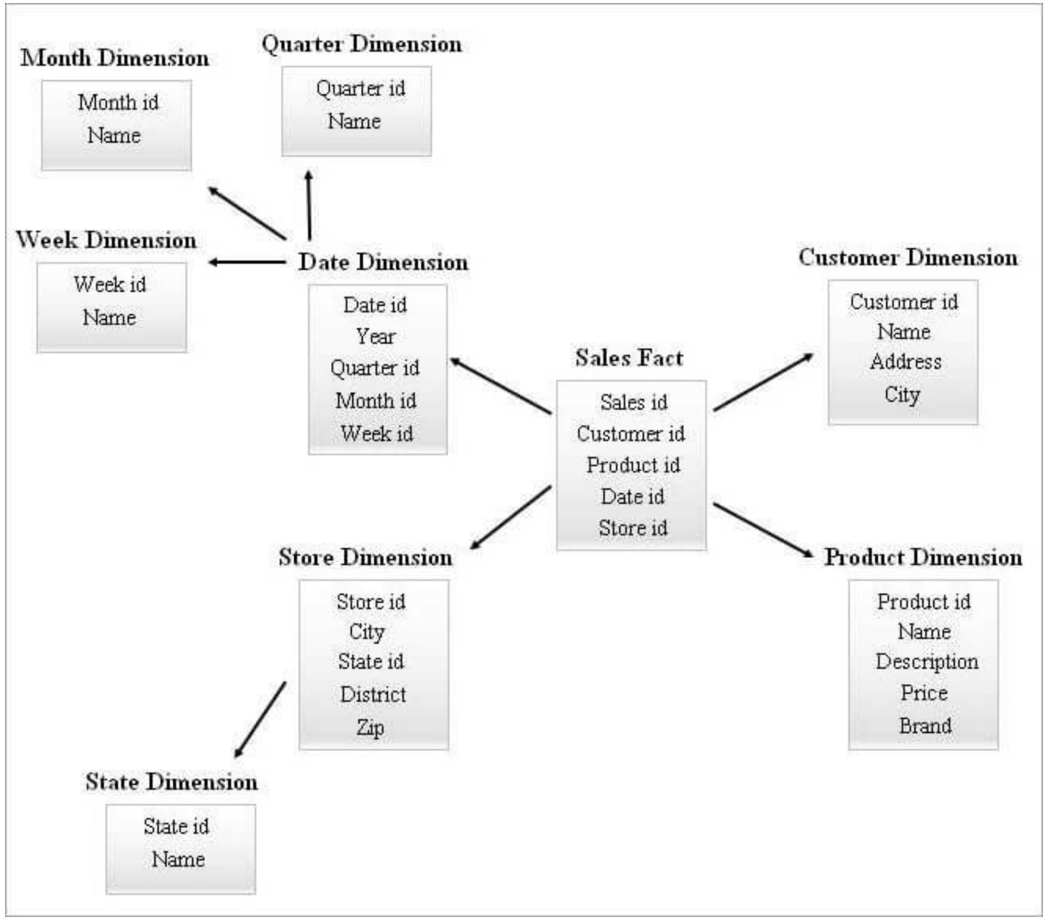
Galaxy Schema



Star Cluster Schema

- A SnowFlake schema with many dimension tables may need more complex joins while querying.
- A star schema with fewer dimension tables may have more redundancy.
- Star cluster schema combines the features of the above snowflake and star schema.
- Star schema is the base to design a star cluster schema and few essential dimension tables from the star schema are snowflaked and this, in turn, forms a more stable schema structure.

Star Cluster Schema



Month Dimension
Month id
Name

Quarter Dimension
Quarter id
Name

Week Dimension
Week id
Name

Date Dimension
Date id
Year
Quarter id
Month id
Week id

Customer Dimension
Customer id
Name
Address
City

Sales Fact
Sales id
Customer id
Product id
Date id
Store id

Store Dimension
Store id
City
State id
District
Zip

Product Dimension
Product id
Name
Description
Price
Brand

State Dimension
State id
Name

Benefits Of A Data Warehouse

- **Enhanced Business Intelligence:**
  - In the earlier days when Data Warehousing and Business Intelligence were not in, the business users and analysts used to take the decisions with a limited amount of data and with their own gut feeling.
  - DW & BI have brought a change by giving insights with real facts and with the real organization data which is gathered over a period of time.
  - Business users can directly query any of the business processes data such as marketing, finance, sales, etc., based on their needs for strategic decision making and smart business decisions.
- **Increased System and Query Performance:**
  - Data warehousing gathers bulky information from heterogeneous systems and places it under one system so that a single query engine can be used for fast data retrieval.
- **Timely Access to Data:**
  - Business users will get benefited by spending less time on data retrieval.
  - The business tools helps in retrieval of date with minimal technical knowledge and generate the reports.
  - This makes business users spend sufficient time on data analysis rather than data gathering.

Advantages of Data Warehouse

- **Enhanced Data Quality and Consistency:**
  - Data warehousing transforms data with dissimilar source system formats into a single format
  - Different business units of same organization will stand by with consistent results/reports.
  - Thus this good quality and consistent data help to run a successful business.
- **Historical Intelligence:**
  - Data warehouse maintains all the historical data that are not maintained by any transactional systems.
  - This large amount of data is used to analyze data for specific time duration and to report it, and to analyze the trends to predict the future.
- **High Return on Investment:**
  - In the real data world, many studies have proved that implementing the data warehouse and Business Intelligence systems generated high revenues and saved the cost.

Disadvantages of Data Warehousing

- Creating a Data Warehouse is definitely a time-consuming and complex process.
- The maintenance cost is heavy as the system needs continuous upgrades. It might also increase if it is not properly utilized.
- Proper training should be given to the developers, testers, and users to understand the DW system and to implement it technically.
- There may be sensitive data that can’t be loaded into DW for decision making.
- Restructuring of any business processes (or) source systems has a major effect on DW.

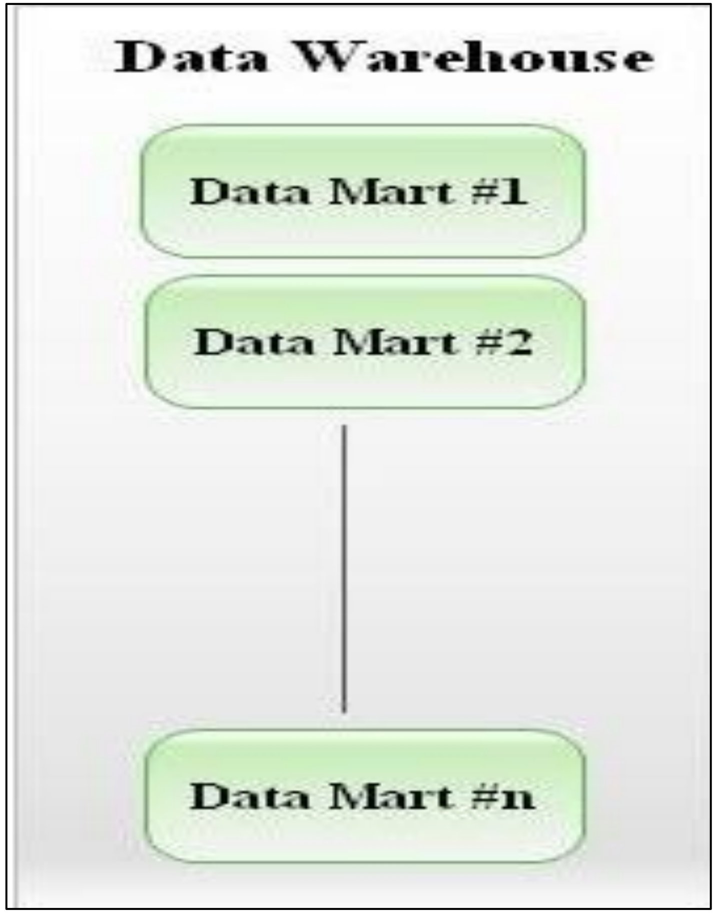
Data Mart

1. A data mart is a small portion of the data warehouse that is mainly related to a particular business domain as marketing (or) sales etc.
2. The data stored in the DW system is huge hence data marts are designed with a subset of data that belongs to individual departments. Thus a specific group of users can easily utilize this data for their analysis.
3. Unlike a data warehouse that has many combinations of users, each data mart will have a particular set of end-users.
4. Data marts are also accessible to business intelligence (BI) tools.
5. Data marts do not contain duplicated (or) unused data.
6. They do get updated at regular intervals.

Data Mart

7. They are subject-oriented and flexible databases. Each team has the right to develop and maintain its data marts without modifying data warehouse (or) other data mart's data.

8. A data mart is more suitable for small businesses as it costs very less than a data warehouse system. The time required to build a data mart is also lesser than the time required for building a data warehouse.



Need of Data Mart

- Based on the necessity, plan and design a data mart for your department by engaging the stakeholders because the operational cost of data mart may be high sometimes.
- **Consider the below reasons to build a data mart:**
  - If you want to partition the data with a set of user access control strategy.
  - If a particular department wants to see the query results much faster instead of scanning huge DW data.
  - If a department wants data to be built on other hardware (or) software platforms.
  - If a department wants data to be designed in a manner that is suitable for its tools.

Comparison Of Data Warehouse Vs Data Mart

S.No	Data Warehouse	Data Mart
1	Complex and costs more to implement.	Simple and cheaper to implement.
2	Works at the organization level for the entire business.	The scope is limited to a particular department.
3	Querying the DW is difficult for business users because of huge data dependencies.	Querying the data mart is easy for business users because of limited data.
4	Implementation time is more may be in months or years.	Implementation time is less may be in days, weeks or months.
5	Gathers data from various external source systems.	Gathers data from a few centralized DW (or) internal (or) external source systems.
6	Strategic decisions can be made.	Business decisions can be made.

THANK YOU!