**Data warehouse**

**What is a data warehouse?**

* it is a single, complete and consistent store of data that is collected from different variety of Sources and made available to the end users in such a way they can understand and use in different business context
* it is designed for query and analysis rather than transaction process.
* it is like a central repository stores historical data derived from transactional data .

**Need for Data Warehouse**

* Business users for analysis
* To store historical data
* Strategic decisions
* Data consistency
* Response time

**Characteristics of data warehouse**

* A data warehouse is **subject-oriented**, as it organizes data around specific subjects or areas of interest for analysis and reporting. Eg: sales, products, customers
* A data warehouse is **integrated**, meaning it combines data from multiple sources into a unified and consistent format for analysis and reporting purposes.
* A data warehouse is **time variant**, meaning it stores historical data and allows for analysis of data changes over time.
* A data warehouse is **non-volatile**, meaning that once data is stored in the warehouse, it is not modified or updated, preserving its integrity for historical analysis.

|  |  |
| --- | --- |
| OLTP | OLAP |
| * Handles more no users and more no small transactions * Runs simple queries * Used for transaction process * Used to run a business * Deals with present data * Less resposnsive time * Database size is small 50 to 100 GB * Read and update access | * Handles more data and less no customers * Runs complex queries * Used for analysis * Used to optimize a business * Deals with kistorical data * More responsive time than oltp as it has to analyse * Large data base size in tera bites * Only read access |

**Components of data warehouse :**

**Source component:**

Source component is noting but the files where the data is extracted

The sources like data from operational database, client private data, external data, cloud applications, spread sheets ,flat files ,disks etc..

**Data staging component**

After the data has coming from different sources the data need to be cleaned and then loaded into the ware house

It involves   *Data extraction*

*Data transformation*

*Data loading*

It does not maintain any history it is just temporary area to etl operations

**Data storage component**

It is like a central repository where the transformed data is loaded. It is the key component that holds historical data and integrates the data

It enable the users to access and analyze the data through different data optimization tools

**Data marts**

Data mart are small and specialized subset of a data storage component where it focuses on a specific subject and the subject related data is stored in that data mart

E.g.: an inventory company data warehouse has diff data marts like sales, customers, products, Inventory and etc.…

|  |  |
| --- | --- |
| Datawarehouse | Data mart |
| * A Data Warehouse is a vast repository of information collected from various organizations or departments within a corporation. * Holds multiple subject are * Holds details information | * A data mart is an only subtype of a Data Warehouses. It is architecture to meet the requirement of a specific user group. * Holds one subject like sales ,customer * Holds summarized information |

**Meta data component**

It contains data about data

It is like a data dictionary or data catalog in dbms, it contains information about like data structure, records, address, and info about indexes ect…

**Olap component**

**Data mining**

It analyze the data that has been compiled by computer or downloaded into computer, it analyze the data and extract the useful information for it

It also extract the hidden patterns in the database

**What is etl?**

**Extraction**

That data is extracted from different sources like os, cloud , on premises server

The extraction should be done several times periodically to supply the changes and newly added data

**Data cleansing**

It is a technique in etl process to improve the data quality and to ensure that the data is accurate it has process like removing duplicate values, handling the missed data , standardizing data formats ,checking with the schema of the data etc….

**Transformation**

Transformation is the core of the reconciliation phase. It converts records from its operational source format into a particular data warehouse format.

**Loading**

The Load is the process of writing the data into the target database. During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible.

Loading can be carried in two ways:

1. refresh the data in target database

2.update the existing data in target

**Dimensional modeling ?**

It is a data structure technique for optimizing the data storage in data warehouse for fast retrieval of data.

A dimensional model in data warehouse is designed to read, summarize, analyse numeric information like values, balances, counts, weights, etc. in a data warehouse.

*The dimensional modelling arranges data in such a way that it is easier to retrieve data and generate report.*

**Elements of dimensional data modelling**

**Facts**: facts are the quantitative and measurable data that represent the business performances or activities being analysed

For a sales business process, a measurement would be quarterly sales.

**Dimensions**:

Dimensions provide details about the facts,the give the info of who, where, what of a fact

For a sales business process, a fact would be quarterly sales and dimensions would be Who – Customer Names

Where – Location

What – Product Name

Dimensions table is like a window to view the facts

**Attributes**

The Attributes are the various characteristics of the dimension in dimensional data modelling.

In the Location dimension, the attributes can be

State

Country

Zip code etc.

Attributes are used to search, filter, or classify facts. Dimension Tables contain Attributes.

**Steps of dimensional modeling**

**Identify the business process:**

Understanding of the business process and requirement to determine what need to be captured and analysed to design data in warehouse

**Identify the grain:**

it is the process for identifying the level of information for any table that need to be stored in data ware house

identification and design of facts and dimension tables

**Build schema:**

a schema refers to the logical structure or blueprint that defines how data is organized, stored, and represented within the data warehouse.

It provides a framework for understanding the relationships between different data elements.

**Star schema**

* Each dimensions in star schema represented with only one schema
* There is a fact table at the centre, it contains keys to each dimensions
* The fact table and dimension table contains attributes
* Fact table linked to dimension table through forienkeys

**Snow flake schema**

The snowflake schema is an extension of the star schema. In a snowflake schema, each dimension are normalized and connected to more dimension tables.

When to use snowflake

* + very large dimensions
  + some attributes not common to all the records

SCDs are techniques used to handle changes to dimension attributes over time in a data warehouse.

Type 1 SCD :

The dimension attributes are simply overwritten with the new values when changes occur.

This method is useful when historical data is not essential, and the focus is on the most recent information.

Before change

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| surid | empid | name | city | dept |
| 1 | 1234 | ashoak | hyderabad | qa |

After change

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| surid | empid | name | city | dept |
| 1 | 1234 | ashoak | chennai | qa |

Type 2 SCD :

creates new records for each change in the dimension attribute, preserving historical information.

It introduces new rows with different surrogate keys to track changes over time.

Before change

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| surid | empid | name | city | dept |
| 1 | 1234 | ashoak | hyderabad | qa |

After change

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| surid | empid | name | city | dept |
| 1 | 1234 | ashoak | chennai | qa |
| 2 | 1234 | ashoak | hyderabad | qa |

Type 3 SCD:

keeps track of a limited number of changes for each dimension attribute, usually storing both the old and new values in separate columns within the same record.

This is used only when recent changes are needed and not much historical data is stored

It will a add a new column instead of adding row

**Surrogate key**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| surrid | empids | name | dept | salary |
| 1 | 1234 | ravi | dev | 33000 |
| 2 | 1234 | ravi | dev | 45000 |

It is used to replace primary key

It is used in slowly changing dimension

To maintain historical data the primary key should be duplicate so it will make the primary key as dummy and surrogate key as primary key

In scd the primary key is not unique and surrogate key is unique