**1. Basics of DWH**

**1.1 Need of Data Warehouse in modern era.**

Every Organization is in need of maintaining the data, this means maintaining data in organization can be different type such as Books, tapes, excels. As time goes on changing, the importance of the data and its maintenance and handling the data is also changing during modern era. Let’s take an example of handling data when there are more than 1000 customers or employees in any organization and their data needs to be maintained and handle in future. All this has not been possible in practice until data warehousing came into reality.

In challenging times good decision-making becomes critical. The best decisions can be made when all the relevant data available is taken into consideration. The best possible source for that data is a well- designed data warehouse.

Some Organizations are really small which cannot afford a data warehouse. They can actually use data marts which interpret as data warehouse where their transactional data is very low. Data warehouse has become more efficient when handling data for larger organizations.

**1.2 What is Data warehouse?**

Data warehousing combines data from multiple, variable sources into one database which can be easily manipulated. This can be used for Analysis, Transformation and Reporting. Usually Larger companies use this data warehousing for analyzing trends over a period of time for viewing transactional data and plan accordingly.

**Data warehouse is defined as subject oriented, Integrated, Time Variant and Non-Volatile collection of data for the management for decision making processing.**

**Subject Oriented:** This is used to analyze particular subject area.

**Integrated:** This shows that integrates data from different sources.

**Time variant:** Historical data is usually maintained in a Data warehouse, i.e. retrieval can be for any period.

**Non-Volatile:** Once the data is placed in the data warehouse, it cannot be altered, which means we will never be able to alter the historical data.

**1.3 Data Ware Housing Concepts**

**OLTP (Online Transaction Processing System)**

OLTP is nothing but a database which actually stores the daily transactions which is called as the current data. Usually OLTP is used for more of the online applications where the application needs to update very frequently in order to maintain consistency in the data.

OLTP deals with the large number of data.

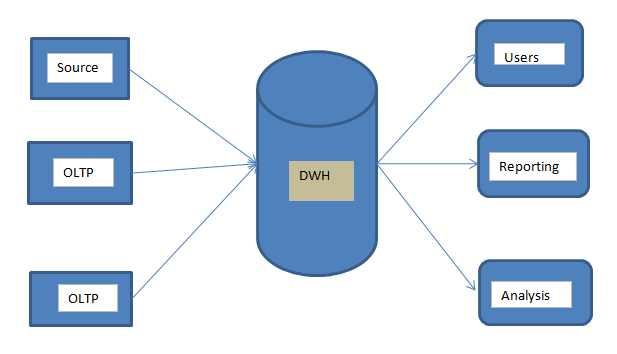
**OLAP (Online Analytical Processing System**

OLAP deals with analyzing the data for Decision making and planning. This actually deals with the aggregations and the data in OLAP will be from different data sources Compared to OLTP, OLAP deals with relatively small amount of data.

**1.4 Difference between OLTP v/s OLAP**

|  |  |
| --- | --- |
| **OLTP** | **OLAP** |
| It is an online transactional system and manages database modification. | It is an online data retrieving and data analysis system. |
| Insert, Update and Delete information from the database. | Extract data for analyzing that helps in decision making. |
| OLTP and its transactions are the original source of data. | Different OLTPs database becomes the source of data for OLAP. |
| OLTP has short transactions. | OLAP has long transactions. |
| The processing time of a transaction is comparatively less in OLTP. | The processing time of a transaction is comparatively more in OLAP. |
| Simpler queries. | Complex queries. |
| Tables in OLTP database are normalized. | Tables in OLAP database are de-normalized. |
| OLTP database must maintain data integrity constraint. | OLAP database does not get frequently modified. Hence, data integrity is not affected. |

**1.5 Data Warehouse Architecture**

****

**1.6 Data Marts and its Types**

Data Mart is nothing but the smallest version of the DWH. Data Mart deals with single subject area.

Data Mart focuses on one area hence they draw data from limited data source.

**Example:**

In Single Enterprise level which has multiple departments i.e. HR, Finance and Supply chain management those are nothing but Data Marts.

**Types of Data Mart:**

* Dependent DM
* Independent DM
* Hybrid DM

**Dependent DM:**

* Dependent data marts are created by drawing data directly from operational and DWH.



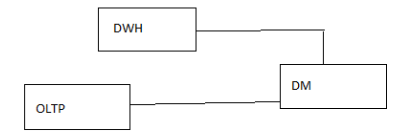
**Independent DM:**

* Independent data mart is created without the use of a central data warehouse.
* It has been created directly from source system and suitable for small organization.



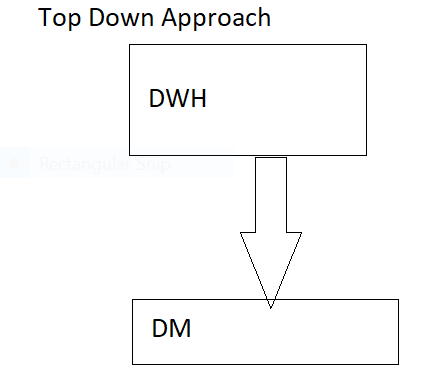
**Hybrid DM:**

* In this type of data marts, data can feed from data warehouses or operational systems (OLTP).

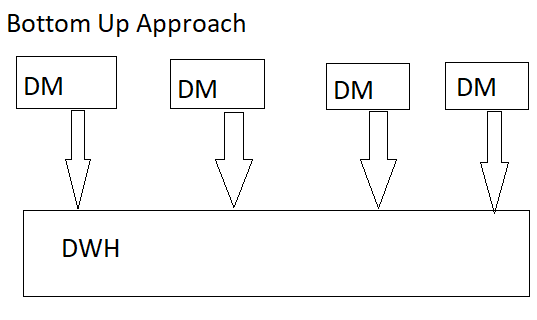


**Models:**

* **Top Down:** In top down approach first build DWH and then DM.



* **Bottom Up:** In Bottom up approach first build DM and then DWH.



**Difference between DWH and DM:**

|  |  |
| --- | --- |
| Data Warehouse | Data Mart |
| Data warehouse is Multiple subject area. | Data mart is single subject area. |
| In data warehouse, Multiple data source | Data mart, limited data source. |
| Data warehouse is top-down model. | Data Mart is a bottom-up model. |
| To build a warehouse is difficult. | To build a data mart is easy. |
| In data warehouse, Fact constellation schema is used. | In Data Mart, Star schema and snowflake schema are used. |
| Data Warehouse is flexible. | Data Mart is not flexible. |
| Data Ware house has long life. | Data Mart has short life than warehouse. |
| In DWH, Data are contained in detail form. | In DM, Data are contained in summarized form. |
| Data Warehouse is vast in size. | Data Mart is smaller than warehouse |

**1.7 ODS (Operational Data Store) for Data Warehousing**

An operational data store (ODS) is a central database that provides a understanding of the latest data from multiple transactional systems for operational reporting. It enables organizations to combine data in its original format from various sources into a single destination to make it available for business reporting.

**Difference between ODS and DWH:**

|  |  |
| --- | --- |
| **Operational Data Stores(ODS)** | **Data Warehouse(DWH)** |
| ODS means for operational reporting and supports current or near real-time reporting requirements. | A data warehouse is intended for historical and trend analysis, usually reporting on a large volume of data. |
| An ODS consist of only a short window of data. | A data warehouse includes the entire history of data. |
| It is typically detailed data only. | It contains summarized and detailed data. |
| It is used for detailed decision making and operational reporting. | It is used for long term decision making and management reporting. |
| It is used at the operational level. | It is used at the managerial level. |
| It serves as conduct for data between operational and analytics system. | It serves as a repository for cleansed and consolidated data sets. |
| It is updated often as the transactions system generates new data. | It is usually updated in batch processing mode on a set schedule. |

**1.8 Difference between Database and Data Warehousing**

|  |  |
| --- | --- |
| **Database** | **Data Warehouse** |
| Database has the current data which has a chance of Updating day by day. | Data warehouse Stores the Historical data where the accuracy is maintained over period of time. |
| Contains the Day to day operations data. | Contains the Long term Operations data. |
| Database professionals, agents access this particular data. | Managers, Analysts access the data warehouse data |
| Data in the Database will always be in the form of Input. | Data in the Data warehouse will always be the output data where it is used for Analyzing trends. |
| Database is used for Transactions. | Data warehouse is used for Analytics. |
| Database has both Read/Write access. | Data warehouse has only the Read access. |
| Database contains only few number of records compared to a Data warehouse. | Data warehouse contains millions of records as the DWH gets refreshed with historic data. |
| Database is always Normalized. | Data warehouse is Demoralized. |
| The data view in the database will always be Relational. | The data view in the Data warehouse id always Multidimensional |
| Database contains the detailed data. | Data warehouse contains the consolidated data. |

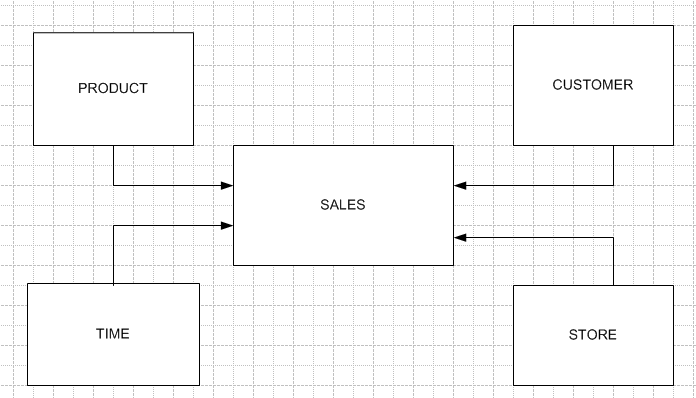
**2. Data Warehousing Concepts**

**2.1 Data Models:**

Data model tells how the logical structure of a database is modeled. Data Models are fundamental entities to introduce ideas in DBMS. Data models define how data is connected to each other and how it will be processed and stored inside the system.

**2.1.1 Conceptual Data Model**

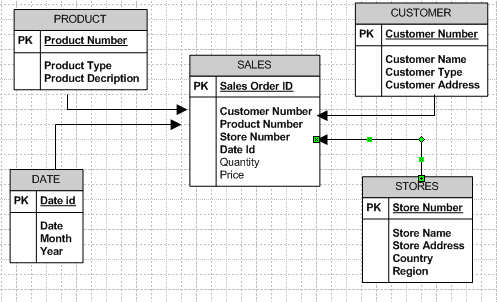
This usually pictures the highest level of relationship between the entities.

* Displays the important entities and the relationships among them.
* No attribute is specified.
* No primary key is specified.

**2.2.2 Logical Data Model**

Logical Data Model defines the data as much as possible, to show how they can be physically implemented in the database.

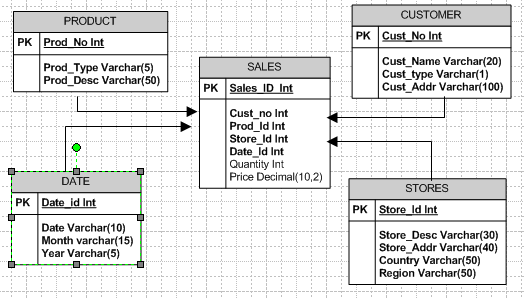
* Displays all the entities and the attributes and the relationships between them.
* Primary key for each entity is specified.
* Foreign keys for each entity if exists is specified.
* Normalization is performed.



**2.1.3 Physical Data Model**

Physical Model defines how the model physically exists in the system.

* Displays all the tables and columns.
* Displays foreign keys.
* Displays lookup tables.
* Change the relationships into foreign keys.
* Entity now becomes table.
* Attribute now becomes column Data types are also shown in this model.



* 1. **Data Schemas**

A database schema defines its entities and the relationship among them. Database schema is a descriptive detail of the database, which can be implemented by means of schema diagrams. All these activities are done by database designer to principle architect in order to give some ease of understanding of detailed database.

**Dimensions**

A dimension table contains dimensions of a fact. They are joined to fact table via a foreign key. Dimension tables are de-normalized tables. The Dimension Attributes are the various columns in a dimension table Dimensions offers **descriptive characteristics** of the facts with the help of their attributes. The dimension can also contain one or more hierarchical relationships.

Table which describes dimension involved are called as dimension table.

**Types of Dimensions**

1. SCD

2. Confirmed dimension

3. Degenerated dimension

4. Junk dimension

5. Roll playing dimension

6. Static dimension

7. Shrunken dimension

**1. SCD (Slowly changing dimension):**

Dimension that change slowly over a period of time rather than changing regularly is group in SCD.

**For ex: Location, Name, Address, City etc.**

Slowly Changing Dimensions are often categorized into three types namely

* SCD Type1 (SCD 1)
* SCD Type2 (SCD 2)
* SCD Type3 (SCD 3)

**Ex:** Consider a customer with name SOHAN who is living in Sydney from 2018.This can be depicted in a table as below:

|  |  |  |  |
| --- | --- | --- | --- |
| CUSTOMER NUMBER | CUTOMER NAME | YEAR | location |
| 1 | SOHAN | 2018 | SYDNEY |

**SCD Type 1(SCD 1)**

Replaces the old entry with the new value.

The customer SOHAN has moved from Sydney to Perth in the year 2021

In this Type 1 SCD, the table will be changed as below:

|  |  |  |  |
| --- | --- | --- | --- |
| CUSTOMER NUMBER | CUSTOMER NAME | YEAR | LOCATION |
| 1 | SOHAN | 2021 | PERTH |

In this case, the previous entry which is treated as history is lost.

**SCD Type 2 (SCD 2)**

The New record is inserted into the same table, In this Type 2 SCD, the table will be changed as below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| CUSTOMER NUMBER | CUSTOMER NAME | YEAR | LOCATION | ETL\_flag | START\_DATE | END\_DATE |
| 1 | SOHAN | 2020 | Sydney | N | 1/2/2020 | 1/2/2021 |
| 1 | SOHAN | 2021 | Perth | Y | 1/2/2021 | NULL |

To identify the current record and history record the columns like **ETL\_Flag or START\_DATE and END\_DATE** is used.

In this case, each record will be treated differently, which makes the table to grow fast and complicates the ETL process. This is mainly for tracking the historical changes.

**SCD Type 3 (SCD 3)**

New fields/Columns are added to the table to main the history.

In this Type 3 SCD, the table will be changed as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CUSTOMER NUMBER | CUSTOMER NAME | YEAR | LOCATION | OLD YEAR | OLD LOCATION |
| 1 | SOHAN | 2021 | Perth | 2020 | Sydney |

The previous record itself is modified such that neither the history is lost and even the new record is also displayed. But this can accommodate only one change. This Type3 SCD is rarely used, when the changes are prone to change only once.