**DATA ENGINEERING** -Data Engineering is all about building, scaling and organize the data for analytics.

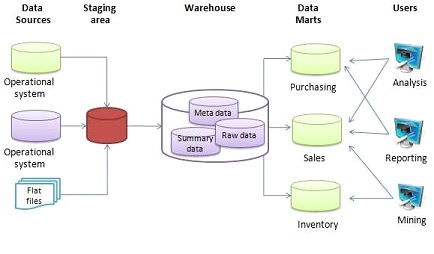
**DATA WAREHOUSING**

* For storing huge amount of data we can use data warehouse.
* It is centralized area where we can store the data from different location.
* We can use various tools and software to find meaningful data for using purpose.
* Some of the tools are:-1.Oracle data integration 2.Microsoft SSIS(Sequel Server Integration services**)**.

**Purpose of Data Warehouse**

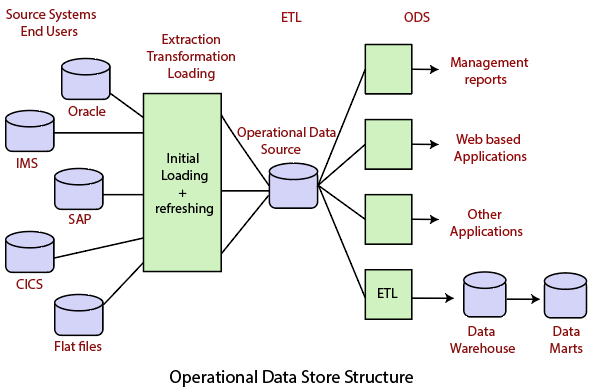
* The primary purpose of a data warehouse is to provide a central repository of information that can be quickly analyzed and queried to generate relevant insights. The specific types of insights generated from a data warehouse can vary.

**Data Warehouse Architecture**



**Operational Data Store**

An operational data store (ODS) is a central database that provides a snapshot 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.



**OLTP Vs Warehouse Applications**

|  |  |
| --- | --- |
| Data Warehousing DWH | Online transaction |
| It is technique that gathers or collect data from different sources into central repository. | It is technique that is used for detailed day to day transaction data which keep chaining on everyday. |
| It is designed for decision making process. | It is designed for business transaction process. |
| It stores large amount of data or historical data. | It holds current data. |
| It used for analyzing the business. | It used for running the business. |
| In Data warehousing, the size of database is around 100GB-2TB. | In Online transaction processing, the size of data base is around 10MB-100GB. |
| In Data warehousing, denormalized data is present. | In Online transaction processing, normalized data is present. |

**Data Marts**

* It is a data store which is designed for a particular department of an organization.
* Data mart is an subset of the data warehouse that is usually oriented to a specific purpose.

**Reasons for Creating Data Mart**

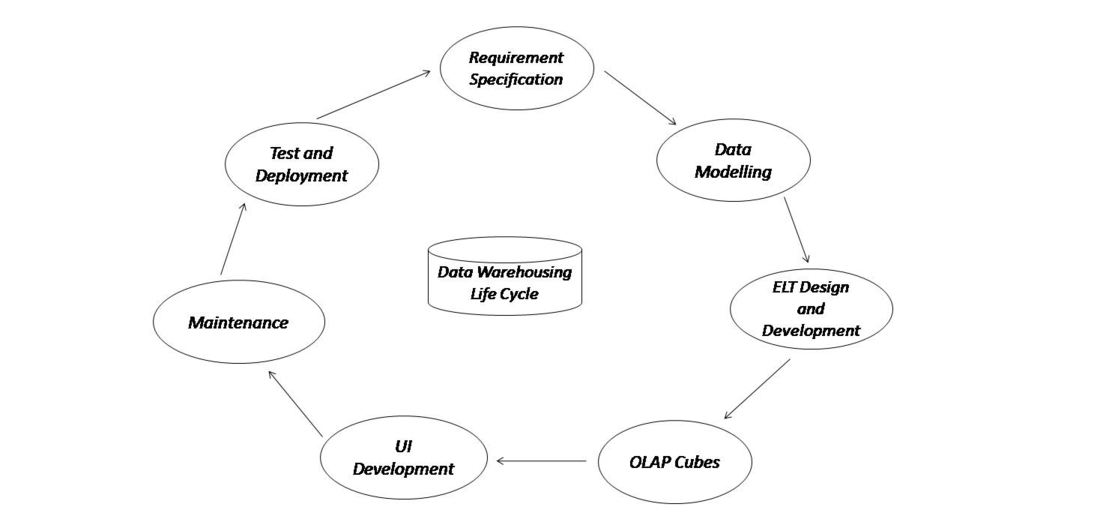
* Easy access of frequent data
* Improved end-user response time
* Easy creation of data mart
* Less cost

**Data Marts vs warehouse**

**Warehouse**  **Data Marts**

* It is a centralized system. It is a decentralized system
* It is top-down model It is bottom-up model
* To build a warehouse is difficult To build a mart is easy
* It is flexible It is not flexible
* Long life Short life

**Data warehouse development life cycle**

* **Requirement Specification:**It is the first step in the development of the Data Warehouse and is done by business analysts. In this step, Business Analysts prepare business requirement specification documents. More than 50% of requirements are collected from the client side and it takes 3-4 months to collect all the requirements. After the requirements are gathered, the data modeler starts recognizing the dimensions, facts & combinations based on the requirements. We can say that this is an overall blueprint of the data warehouse. But, this phase is more about determining business needs and placing them in the data warehouse.
* **Data Modelling:** This is the second step in the development of the Data Warehouse. Data Modelling is the process of visualizing data distribution and designing databases by fulfilling the requirements to transform the data into a format that can be stored in the data warehouse. For example, whenever we start building a house, we put all the things in the correct position as specified in the blueprint. That’s what data modeling is for data warehouses. Data modelling helps to organize data, creates connections between data sets, and it’s useful for establishing data compliance and its security that line up with data warehousing goals. There are three data models for data warehouses:
  + Star Schema
  + Snowflake Schema
  + Galaxy Schema.
* **ELT Design and Development:** This is the third step in the development of the Data Warehouse. ETL or Extract, Transfer, Load tool may extract data from various source systems and store it in a data lake. An ETL process can extract the data from the lake, after that transform it and load it into a data warehouse for reporting. For optimal speeds, good visualization, and the ability to build easy, replicable, and consistent data pipelines between all of the existing architecture and the new data warehouse, we need ELT tools.
* **OLAP Cubes:** This is the fourth step in the development of the Data Warehouse. An OLAP cube, also known as a multidimensional cube or hypercube, is a data structure that allows fast analysis of data according to the multiple dimensions that define a business problem. A data warehouse would extract information from multiple data sources and formats like text files, excel sheets, multimedia files, etc. The extracted data is cleaned and transformed and is loaded into an OLAP server (or OLAP cube) where information is pre-processed in advance for further analysis.
* **UI Development:**This is the fifth step in the development of the Data Warehouse. So far, the processes discussed have taken place at the backend. There is a need for a user interface for how the user and a computer system interact, in particular the use of input devices and software, to immediately access the data warehouse for analysis and generating reports. The main aim of a UI is to enable a user to effectively manage a device or machine they’re interacting with. There are plenty of tools in the market that helps with UI development. BI tools like Tableau or PowerBI for those using BigQuery are great choices.
* **Maintenance:**This is the sixth step in the development of the Data Warehouse. In this phase, we can update or make changes to the schema and data warehouse’s application domain or requirements. Data warehouse maintenance systems must provide means to keep track of schema modifications as well, for instance, modifications. At the schema level, we can perform operations for the Insertion, and change dimensions and categories. Changes are, for example, adding or deleting user-defined attributes.
* **Test and Deployment:** This is often the ultimate step in the Data Warehouse development cycle. Businesses and organizations test data warehouses to ensure whether the required business problems are implemented successfully or not. At the time of data warehouse deployment, most of its functions are implemented. The data warehouses can be deployed at their own data center or on the cloud.