DATA ENGINEERING

Name: Ramireddy Preethi

Batch: Python Batch 2

DAY1:

Concepts:

- An introduction to Data Warehousing
- Purpose of Data Warehouse
- Data Warehouse Architecture
- Operational Data Store
- OLTP vs Warehouse Applications
- Data Marts
- Data Marts vs Data Warehouses
- Data Warehouse Life cycle

Introduction to Data Warehousing and Purpose of Data Warehouse

What is Data Warehouse?

Data Warehouse is single or group of databases stores historical data in structured manner which is large in size when compared to normal database.

Why Data Warehouse?

1) Integrates cleaned data into centralized location.

2) Contains historical data unlike operational database which contains recent transactions.

3) Fast and Efficient to retrieve data because it contains data mart which categorizes the data.

4) Scalability (even data grows it handle data efficiently).

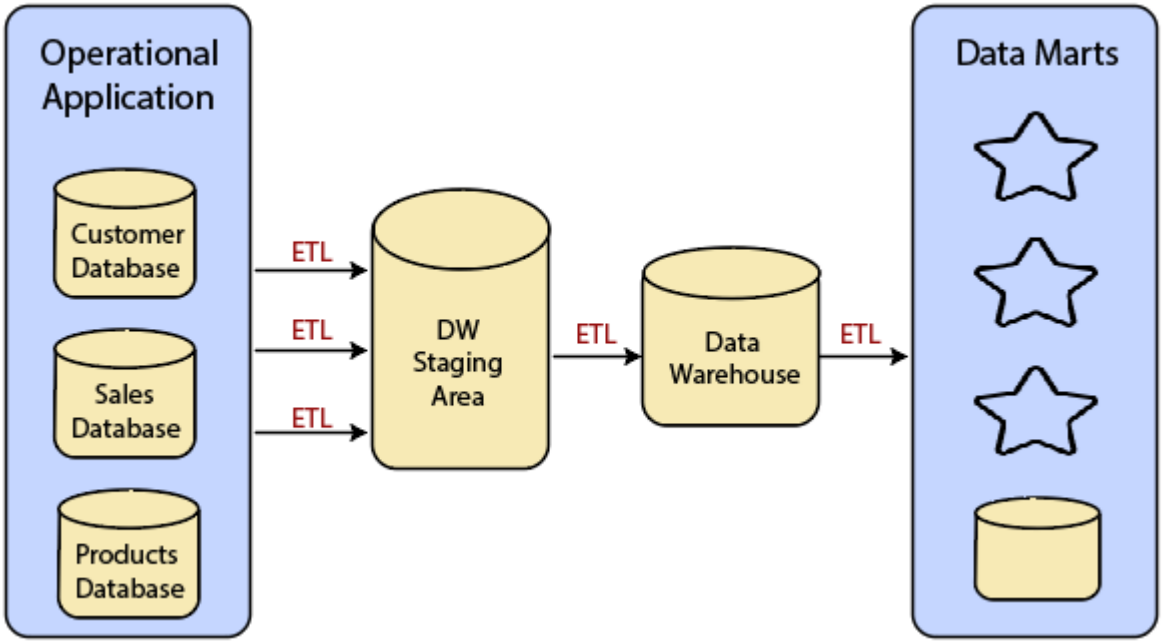
5) Analysis can done easily.

6) Helps in decision making.

What is Data Warehousing?

Process of setting up, organizing, and managing the database so it is ready for analysis and decision making.

Data Warehouse Architecture



Improving quality of data and performance with staging area and data mart.

Staging Area: It is also a database where we store data that need to be cleaned and all data cleaning is done here. In order to extract data from sources we use SSIS for extracting and then it is sent to staging area and then again cleaned data is extracted from the staging area and then transform and load to the data warehouse.

Data Mart: Subset of Data Warehouse, which categorize the data in order to retrieve the data in fast manner.

Operational Data Store (ODS)

1) ODS is a type of database used for operational reporting and supports current or near real time reporting.

2) ODS support OLTP systems, for managing and processing current operations.

3) ODS integrates data from various systems and update frequently to reflect current operations.

OLTP (Online transaction processing):

1) It manages day to day transaction data (fresh data)

2) It is small in size and more users use this.

3) In this read (20%) and write (80%)

4) Data Modification can be done frequently.

5) It follows entity relational diagram (ERD) and contains Master and Transactional tables.

6) Operations on OLTP is performed by Data Engineer

MASTER: It is a plain text and cannot perform any calculations or comparisons and it contains unique data.

TRANSACTIONAL: It can perform any calculations and comparisons

OLAP (Online Analytical Processing):

1) It manages historical data.

2) It is large in size and less users will use this.

3) In this read (80%) and write (20%)

4) Data Modification can be done rarely.

5) It contains Dimensional and Facts tables.

6) Operations on OLAP is performed by Data Analyst and Data Scientist.

DIMENSIONAL: This table is same as master table but its naming convention is like “Dimtablename”

FACTS: This table is same like transactional table but its naming convention is like “Facttablename”

OLTP vs Warehouse Applications

OLTP	Warehouse Applications
Manages day to day transaction data(fresh data)	Manages historical data.
More write operations are done	More read operations are done
More Users	Less Users
Data Modifications are done frequently	Data Modifications are done Infrequently
Fast and simple queries	Complex queries
MYSQL,SQL SERVER,ORACLE ETC	Snowflake ,Redshift etc

Data Marts vs Data Warehouses

Data Marts	Data Warehouses
Subsets of data Warehouse	Single or group of databases
It contains detailed and relavent data based on specific needs	It contains integrated , historical and current data from various sources in catogerized form.
Small in size	Large in size
Purpose is reporting and analyzing	Purpose is reporting and analyzing
Less complex	More Complex
Scope is specific to particular department	Scope is organization wide

Data Warehouse Life cycle

1) Requirements Gathering and Analysis: All requirements are gathered from stakeholders and analyze their needs.

2) Data Modelling: Visualize and design databases in suitable to meet specific requirements and to store in data warehouse.

3 data models (schemas):

1) Star Schema: Dimensional tables directly connected to single Fact table

2) Snowflake Schema: Dimensional tables indirectly connected to single Fact table

3) Galaxy Schema: Dimensional tables connected to multiple Fact tables

3) ELT Design and Development: In this phase data is extracted from sources and then loaded into datalake which stores all the data and then transform data into required format i.e this step contains data cleaning, aggregating and whatever required.

4) OLAP: Designed for fast and efficient analysis

5) UI Development: Developing user interface in order to users interact with data in data warehouse

6) Maintanence: Regular maintanence helps for data accuracy, integrity and performance which is easy for analysis and accessing data in data warehouse

7) Testing: Checks whether data warehouse correctly implements or not. And also checks errors or any performance issues.