TASK -1

1. WHAT IS ETL?

ETL (Extract Transform and Load) is a integration process that combines data from multiple data sources into single consistent data store that is loaded into data warehouse or other targeted system.

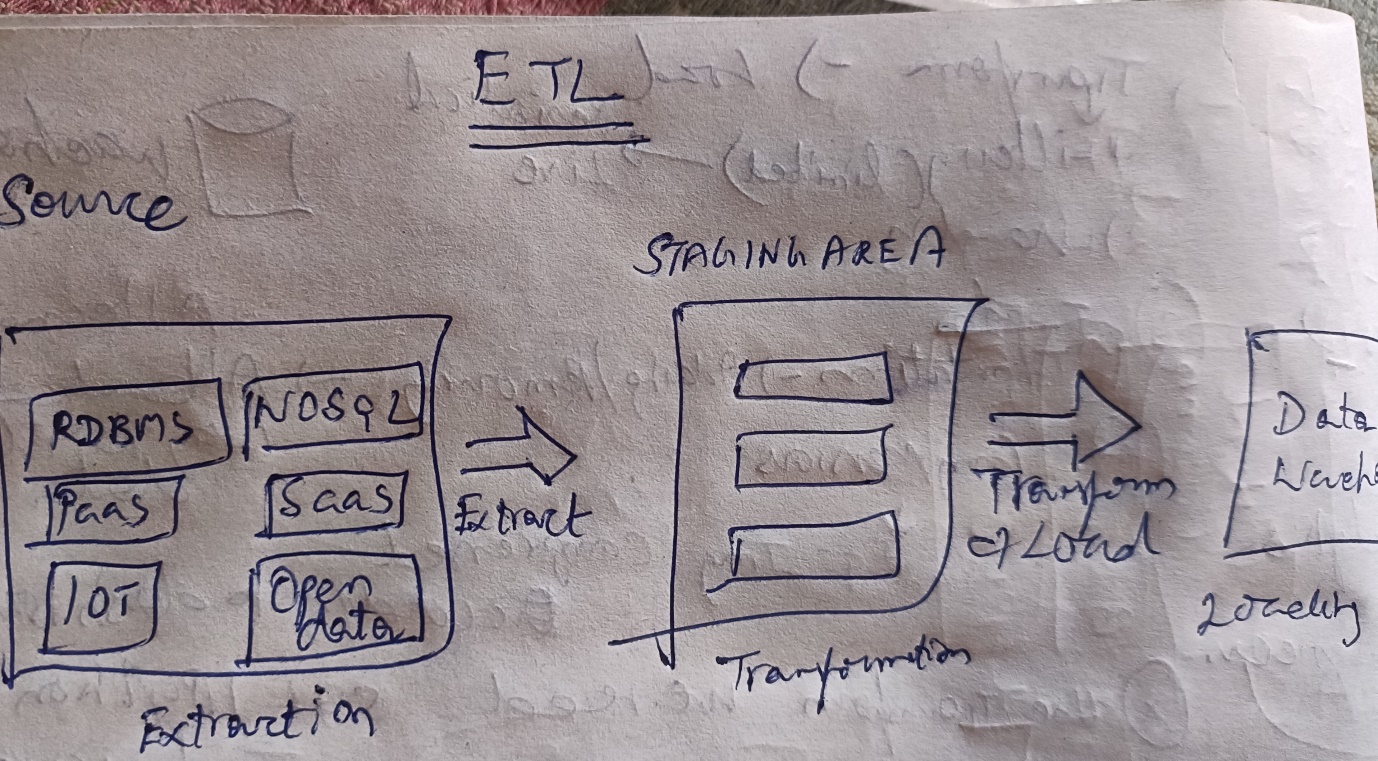
* **Extract**- Reading data from source system and storing it in a staging area
* **Transform-** Cleaning and validating extracted data into a format that is suitable for loading into the warehouse
* **Load**-Creating physical data structures and loading data into the warehouse.

1. WHAT IS ELT?

ELT (Extract Load Transform) is the process of extracting data from one or multiple sources and loading raw data directly it into a target data warehouse. You can transform the data whenever you need it.

1. List out various tools used in ETL?

* Informatica PowerCenter
* Business Objects Data Integrator
* Microsoft SQL Server Integration Services
* Oracle Data Integrator
* AWS GLUE
* Apache Air flow
* IBM Info Sphere DataStage

1. **Mechanism of ETL**
2. What is a Data Lake?

It is a centralized repository designed to store, process and secure large amounts of structured semi structured and unstructured data.

It can any variety of it native format ignoring size limits. Eg: Saas platform

1. What is Datawarehouse?

It is a centralized storage system that allows for storing, analyzing and interpreting of data in order to facilitate better decision-making. It supports BI activities especially Analytics.

1. Difference between OLAP AND OLTP

|  |  |
| --- | --- |
| OLAP(ONLINE ANALYTICAL PROCESSING) | OLTP(ONLINE TRANSACTION PROCESSING) |
| It is a database query management system | It is online database modifying system |
| Consist of historical data from various database | Consist of operational current data |
| It makes use of a data warehouse | It makes use of DBMS |
| Used for data analytics , mining | Used for Business tasks |
| Data is used for decision making and problem solving | Data is used for day to day fundamental operations. |
| Processing time is lengthy because of complex queries | Processing time is faster compared to olap because of simple and straightforward queries. |
| Process focused on customer | Process focused on market |
| Improves the efficiency of business analysts | Enhances the user’s productivity |
| Used by CEO MD AND GM  Eg: Spotify – It analyse songs by users to come up with personalized playlist  Netflix-movie recommendation system | Used by managers and clerks  Eg: ATM Center  Online banking airline booking and adding items to the cart during purchase in online apps like Amazon, Myntra etc |

1. What are Various Analytical tools that can be used to connect with DataWarehouse?

* Amazon RedShift
* Big Query
* Tera Data
* Snow Flake
* Oracle
* Microsoft Azure
* SAS
* PostgreSQL
* Cloudera
* Amazon DynamoDB
* Microfocus Vertica

1. Explain Different stages in ETL

First stage of ETL is **extract**. It is used to extract data from various source system and storing it into the staging area first and not directly because extracted data in various formats can be corrupted so loading into data warehouse may damage.

Second stage is **Transformation**. It includes different processes /Tasks to convert into single standard format

Filtering, Cleaning, Joining, Splitting, Sorting

Third stage is **Loading**. It is the process where transformed data loaded and sometimes data is updated by loading into the data warehouse very frequently into the data ware house .but it purely depends on the requirements and period of loading varies from system to system.

ETL process is used in the pipelining Concept.

1. What are the Characteristics of a NoSQL

* Support for multiple data models
* They can process both unstructured and semi-structured data
* Easily scalable via peer-to peer architecture
* Distribution Capabilities
* They support flexible schema
* Zero downtime

1. What are the tools used to analyse Nosql?

* MongoDB
* Cassandra
* Elasticsearch
* HBase
* Orient DB
* Couch DB
* Neo4J
* Redis

1. Explain about 5 Use Case For NoSQL

**Personalization**:

It is fast becoming an opportunity and challenge. It is the right engagement to add visitor at the right time. But the ability to ingest, process and utilize the amount of data necessary to create personalized experience is a challenge for relational database. **Dominos** uses unified real time data to power personalized marketing campaigns leading to repeat customers.

**AR ANALYTICS:**

The ability to extract information from operational data in real time is critical for an agile enterprise. In particular the ability to increase operational efficiency, reduce costs and increase revenue by acting on current data immediately. **PayPal** monitors site traffic in real time with NoSQL.

**Ad-Targeting**:

Selecting and ad to display or an offer to present on a web page is a choice with direct revenue impact .To decide where to place such ads and what groups to target collects behavioral , demographic and psychographic characteristics of users at most 40 milliseconds to do so .A NoSQL database enables ad platforms to track users attributes and also access ads to place extremely quickly ,increasing the probability of a click. Companies like PayPal, Media mind and AOL.

**Mobile Application**:

App Developers ability to update and enhance mobile apps quickly and without service disruption is critical to user adoption and loyalty. Because No SQL database can store user information and application content in schema -less format developers can quickly modify apps without major database infrastructure changes. That means user experience no interruption to application uptime.

**Digital Communication:**

Today’s world relies on digital communication in the form of mobile text message online chat video calls etc. However enabling million of users to communicate in real time requires both performance and availability and challenge for relational database. Company like **Viber** implemented Couch Base In a multi layered AWS architecture to process up to 15 billion events per day.