Dated: 16/3/23

Name: Muqaddas Fatima

**DET- Day 3**

**Task 4**

**ETL:**

ETL stands for Extract, Transform, and Load, and it refers to the process of collecting data from various sources, transforming it into a format that is suitable for analysis, and loading it into a target system.

* **Extract:**

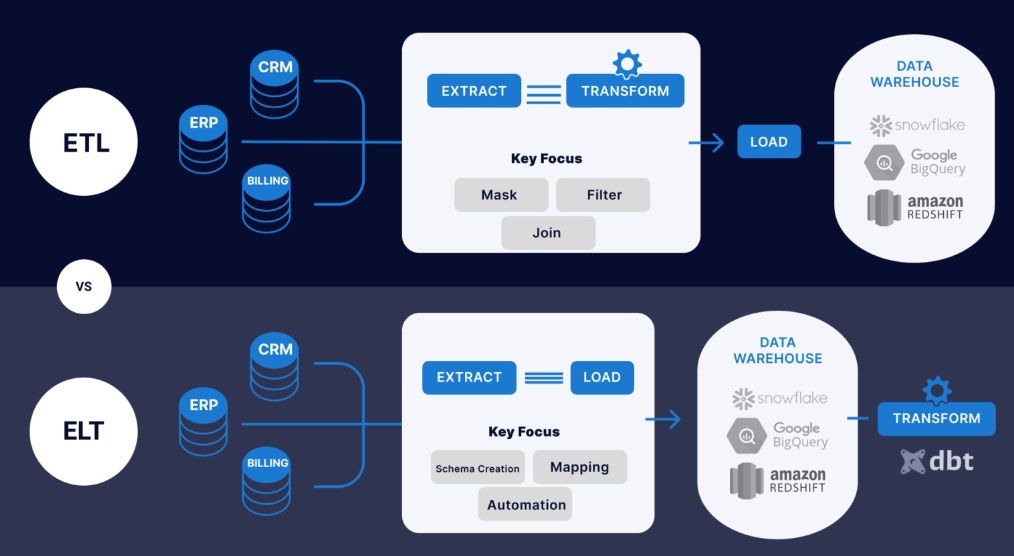
The first step of the ETL process is to extract data from various sources such as databases, flat files, APIs, or web services. This data could be structured or unstructured, depending on the source. The goal of this step is to retrieve the data that is required for analysis.

**Transform:**

The next step is to transform the extracted data into a format that is suitable for analysis. The transformation process involves cleaning the data, reformatting it, filtering it, or joining it with other datasets. This step is crucial because the quality of the data will determine the accuracy of the analysis.

**Load:**

The final step of the ETL process is to load the transformed data into a target system such as a data warehouse, a database, or a cloud-based storage solution. The loaded data is then ready for analysis by business intelligence tools or data analysts.



ELT:

ELT was developed as an alternative to ETL because of the growing popularity of cloud-based data warehouses, which provide powerful data processing capabilities and scalability.

In the traditional ETL approach, data is transformed into a format that is suitable for analysis before it is loaded into a target system, such as a data warehouse. However, with the rise of cloud-based data warehouses, it became possible to load large volumes of raw data into a target system and perform the transformation step within the system.

The ELT process involves three stages:

**Extraction:** Raw data is extracted from various sources, such as applications, SaaS, or databases.

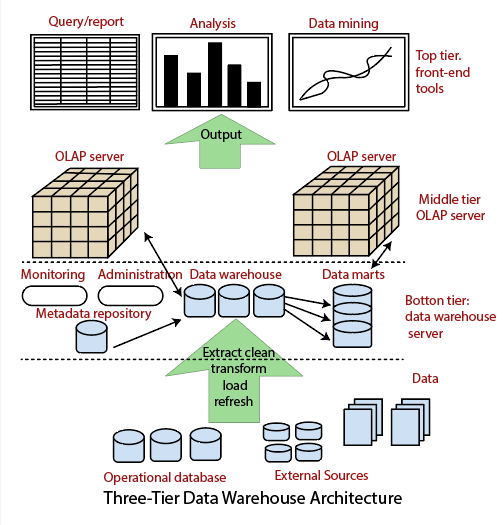
**Loading:** Data is delivered directly to the target system – typically with schema and data type migration factored into the process.

**Transformation:** The target platform can then transform data for reporting purposes. Some companies rely on tools like dbt for transformations on the target.

One of the main advantages of ELT over ETL is scalability. Cloud-based data warehouses have the ability to scale processing power up or down as needed, allowing organizations to handle large volumes of data and complex transformations. ELT also eliminates the need for intermediate storage and staging areas, which can reduce costs and improve performance.

**Three tier Architecrure in Data Engineering**

Data Warehouses usually have a three-level (tier) architecture that includes:



1. **Bottom Tier (Data Warehouse Server)**

A bottom-tier that consists of the Data Warehouse server, which is almost always an RDBMS. It may include several specialized data marts and a metadata repository.

1. **Middle Tier (OLAP Server)**

A middle-tier which consists of an OLAP server for fast querying of the data warehouse.

The OLAP server is implemented using either

- A Relational OLAP (ROLAP) model, i.e., an extended relational DBMS that maps functions on multidimensional data to standard relational operations.

- A Multidimensional OLAP (MOLAP) model, i.e., a particular purpose server that directly implements multidimensional information and operations.

1. **Top Tier (Front end Tools)**

A top-tier that contains front-end tools for displaying results provided by OLAP, as well as additional tools for data mining of the OLAP-generated data.

**ETL Tools**

ETL tools are a set of software tools that are used to extract, transform, and load data from one or more sources into a target system or database.

1. **Informatica PowerCenter** is one of the best ETL tools on the market. It has a wide range of connectors for cloud data warehouses and lakes, including AWS, Azure, Google Cloud, and SalesForce. Its low- and no-code tools are designed to save time and simplify workflows.
2. **Dataflow** is the serverless ETL service offered by Google Cloud. It allows for both stream and batch data processing and does not require companies to own a server or cluster. Instead, users only pay for the resources consumed, which scale automatically based on requirements and workload.
3. **Oracle Data Integrator** is an ETL tool that helps users build, deploy, and manage complex data warehouses. It comes with out-of-the-box connectors for many databases, including Hadoop, EREPs, CRMs, XML, JSON, LDAP, JDBC, and ODBC.

Ref:

https://www.youtube.com/watch?v=CPFd0Q0xecg