**Task #4**

**Extract, Transform, Load**

ETL is short for Extract, Transform, Load. This is a common process used in data integration and data warehousing. ETL involves taking data from different sources, changing it to a common format, and then putting it into a target data warehouse or data mart.

The first phase, "extract", is all about getting data from places like databases, files, APIs, or other systems.

The "transform" phase is when the data is cleaned, checked, and changed so that it matches the target data warehouse or data mart. This step can also make the data more useful.

The "load" phase is about putting the changed data into the target system, usually a data warehouse or data mart. Then the data can be used for business analysis and reporting.

**Extract, Load, Transform**

ELT means Extract, Load, Transform, which is a way to put data together that's similar to ETL but with a few changes.

First, in ELT, data is taken from different places and put into a storage area like a data lake or data warehouse. It's not changed much at this stage so it can be loaded quickly.

Next, the "transform" part happens where the data is changed to make it more useful. This can involve things like filtering, combining, and cleaning the data. This part happens at the end stage after loading of data.

The best thing about ELT is that it lets companies use really powerful storage areas like cloud-based data lakes and data warehouses to process lots of data quickly. It also makes it easier to explore and analyze data using different tools and techniques.

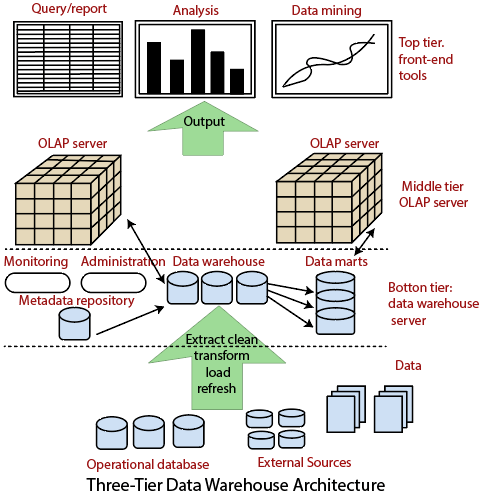
**3-Tier Architecture of Data Engineering (Data Warehouse)**

A data warehouse can be divided into three tiers,

The bottom tier contains the data warehouse server, where the data is stored. This tier may also include data marts and a metadata repository.

The middle tier consists of an OLAP that makes it easier to query the data warehouse.

The top tier contains front-end tools that let people interact with the data and create reports, like Power BI etc.



**ETL Tools**

ETL Tools basically help in performing ETL (ingesting data from different sources, performing any kind of transformations as per the need and then loading it into the target/consumption layer)

Examples include,

1. Apache Ni-fi: Provided by Apache.
2. AWS Glue: An ETL tool managed by AWS Services.
3. Snowpipe: An offering of Snowflake for continuous data ingestion and loading.

Data Mesh is a new way of organizing data within a company. Instead of having one big team manage all the data, each department has its own team responsible for their own data. These teams treat data like a product, and are responsible for making sure the data is high quality and easy to use.

For example, if a company sells products online, they might have a team responsible for managing the product data, another team responsible for managing customer data, and so on. Each team would make sure their data is easy to access and use for the other teams. This makes it easier for everyone in the company to use the data they need to make good decisions.

Data Mesh provides each department more control over their own data and this further helps in maintaining high quality of data within the organization. This high quality of data then can be used effectively for further analytical or Machine Learning purposes.

**Task #5**

**Historical Load**

It involves loading historical data into the target system. It is helpful to ensure that the target system has a complete record of historical data. Historical data can be loaded in batches or all at once using Full Load.

**Full Load**

It involves loading all the data from the source system into the target system at once. Full Load is generally used for initial data loading or when a major change occurs in the source system, such as a change in data format or schema.

It is different from Historical Load in a sense that Full Load means loading the complete data into target system, but Historical Load doesn’t involve loading complete data. Instead, it involves loading all the data up to a certain point in time

**Incremental Load**

It involves loading only the changes that have occurred in the source system since the last data load. Incremental Load is used for regular updates and ensures that the target system is always up-to-date with the latest data.

Incremental Load is faster than both the other approaches because it only loads the latest/changed data.