**Data Engineering**

**By: Shaharyar Sabir**

**Week#01**

**Task # 4:**

1. **What is ETL? in detail.**

ETL stands for extract, transform and load. THE data engineers extract data from different sources, they transform that data and load that data into the systems that can be accessible to end users to use to solve business problems.

**Extract:**

First step is to extract data from different sources such as business systems, APIs, sensor data, marketing tools, and transaction databases, and others.

There are different ways to extract data:

**Partial Extraction:**

Simple way is that when new record is maintained in system the source system will notify you. Not all systems notify you, so they indicate the changed record and provide extract of that.

**Full Extraction:**

Some system cannot identify data changes or give notifications the only option is to reload all data. It keeps the copy of last extract to identify new record and it is recommended to use this extraction on small tables as data transfer is high volume.

**Transform:**

In the staging area, the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. This phase can involve the following tasks:

Filtering, cleansing, de-duplicating, validating, and authenticating the data.

Performing calculations, translations, or summarizations based on the raw data. This can include changing row and column headers for consistency, converting currencies or other units of measurement, editing text strings, and more.

Conducting audits to ensure data quality and compliance

Removing, encrypting, or protecting data governed by industry or governmental regulators

Formatting the data into tables or joined tables to match the schema of the target data warehouse.

**Load:**

In this last step, the transformed data is moved from the staging area into a target data warehouse. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse. For most organizations that use ETL, the process is automated, well-defined, continuous and batch-driven. Typically, ETL takes place during off-hours when traffic on the source systems and the data warehouse is at its lowest.

1. **What is ELT? in detail.**

Extract, Load & Transform. It is best when dealing with huge amount of structured and unstructured data.

**Extract:**

During data extraction it make copy of data from source location to staging area. The data set consists of many data sets that comes from structured and unstructured source.

**Load:**

In this the transformed data is load from staging area into a data storage area. It takes place during business hours when traffic on source system and data ware house is on peak and consumers are waiting to use data for analysis and otherwise.

**Transform:**

It includes filtering, cleaning, de duplication and authenticating the data. Performs calculations, analysis and summery.

1. **3 Tier Architecture in DE**

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

**Bottom Tier (Data Warehouse Server)**

* It is the rational database system. It uses back-end tools to store data into the bottom tier.

**Middle Tier (OLAP Server)**

* In the middle tier we have OLAP server that can be implemented by two ways:
* Relational OLAP that map operations on Multi-dimensional OLAP.
* Multi-dimensional OLAP: that directly implement its data and operations.

**Top Tier (Front end Tools)**

* This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

1. **ETL Tools (any 3)**

Image Enhancers.

Image Generators.

Music Generators.

Video Enhancers.

Video Generators.

Voice Generators.

**Task # 5:**

**- What is Historical Load**

It is one-time initial load of data that the source already have before the creation of pipelines.

- **What is Full Load**

Loading of all the data from source to destiny. Before loading data from source the targeted tables are cut short.

- **What is Incremental Load**

Selective transfer of data from one system to another system. Search for modifies or newly created data for comparison to transfer data.