**Hands-on Practice**

I completed and submitted Homework.

**Self-Learning**

I completed Agile, Jira & Confluence.

SQL basics as from shared resources in the group – In Progress

**Certification:**

I started going through the pdf document of DP-900.

**EOD Updates:**

I have done homework as given and started reading the DP-900 question for certificate preparation, also revised the topics which are explained in yesterday's session and prepared notes as well. Working on to complete the Weekend Project which is assigned.

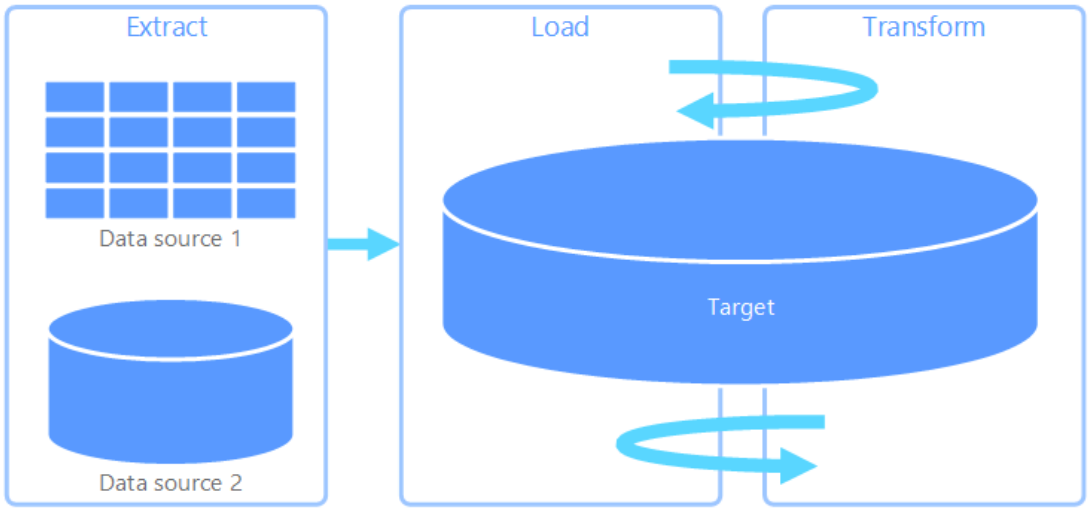
## Homework | Cycle 30

- What is the purpose of a data engineer?
- why we use ELT model in Cloud?
- what is the significance of ADF?

❖ **Purpose of a data engineer**

- Data Engineer has to Extract data from different places it could be either from different databases, excel or also from some share point locations, then load the extracted data into cloud (ADLS GEN2).
- Next,he is going to store the data in a folder called “Bronze” and he will place the files in PARQUET format in sub folders.
- Next step is to combine the extracted data which also includes to perform cleanup activity in the data or to check if there are any null values are present.
- Once data is combined, he will perform SCD type transformation and keep the data in a dedicated SQLPOOL called as Datawarehouse.
- To get this all done Data Engineer has to develop a pipeline and once the pipeline is developed. He will perform ETL activities as mentioned in the above points.
- If there are any failures happening on them, Data Engineer has to fix those things.

❖ **Use of ELT model in Cloud**




❖

- Extract, load, transform (ELT) differs from ETL solely in where the transformation takes place
- In the ELT pipeline, the transformation occurs in the target data store. Instead of using a separate transformation engine, the processing capabilities of the target data store are used to transform data.
- scaling the target data store also scales the ELT pipeline performance
- Azure Synapse Analytics uses powerful MPP (Massively Parallel Processing) to run large-scale SQL transformations after loading data.
- Raw data is first loaded into cheap, scalable storage like Azure Data Lake Storage (ADLS Gen2), and transformations happen later.
- Azure Data Factory (ADF) allows code-free orchestration of Extract, Load, and Transform steps in a single pipeline.
- The ELT model fits well with the Bronze (raw), Silver (cleaned), and Gold (curated) architecture used in Azure.
- Azure Databricks supports advanced Spark-based transformations on semi-structured or unstructured data.
- Azure services like Synapse and Databricks can auto-scale to handle big data transformations efficiently.
- Azure's architecture separates data storage (ADLS) and compute (Synapse, Databricks) for better cost and performance optimization.

❖ **what is the significance of ADF**

Azure Data Factory is Microsoft developed cloud based tool.

It is a Orchestration tool which we use to perform ETL (Extract, Transform, Load) operations.

And below are the 5 important components in ADF

- **Pipeline**: It is combinations 1 or more activities
- **Activities**: Activities are actual things which do work like copy data from source to destination, which includes either copy file data or table data. Similarly it can also perform delete, foreach etc.
- **Linked Services:** linked services contain the connection details or connection strings of sensitive details (username, passwords, url, account key..etc) of source and destination systems.
- **DataSets**: Source and Target systems file paths will be stored Datasets.
- **Integration Runtime**: It is nothing, but a machine will be provided by Microsoft to move data which means to create or make a link or connection from source to cloud to bring the data. We have 3 types of integration runtime are available for listed below.
- autoresolve ir
- self hosted ir
- azure ssis ir