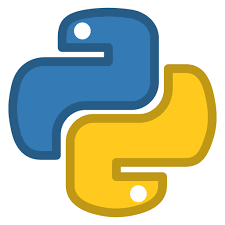
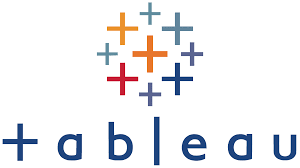
**Day 4**

**ETL – Pipeline**





The dataset, sourced from<https://data.gov.in/>, comprises information from State Road Transport Undertakings (SRTUs) and is available in three different formats: profit and loss data in CSV, financial data in JSON, and physical data in XML. The objective is to create an ETL pipeline as outlined in the diagram and perform data visualization.

The process involves:

1. Extraction: Data is fetched from various sources, hypothetically stored on a cloud platform like Google Drive. The extraction involves gathering data from structured and semi-structured sources.

1. Transformation: Data is brought into a Python Colab environment for standardization. This includes aligning state names and correcting column names. Once uniformity is achieved, the datasets are combined (Joined).

1. Loading: The transformed data is loaded back onto the cloud, in our case Google Drive. This data can then be utilized for analytics. Additionally, a connection to Tableau is established for visualization and deriving insights from the data.

Steps to Load data in tableau from google drive:

1. In tableau connect>>to a server open more..
2. Click on google drive.
3. Choose the google account you want to connect and enter the password.
4. Allow tableau to connect google drive.
5. Select the file from the pop up window appearing in tableau.
6. The data will be loaded from the file.