**What is ETL?**

* **EXTRACT data from its original source (local or s3 bucket where we are keeping the data)**
* **TRANSFORM data by deduplicating it, combining it, and ensuring quality**
* **LOAD data into the target database(Postgres, Snowflake, Amazon RDS)**

In this assignment,

I took data from Github

This Data contains yearly records of the product sold ot and the profit.

Now our main task is to take this data and load it into a Warehouse/Database.

**Why I use Airflow?**

Airflow is an orchestration tool which allows the scheduling of task assigned.

Here we want that our ETL script should run regularly at a particular time, so everytime the data will be generated, it will get cleaned and loaded to Postgres.

NOTE: As of now I did it in my local, if I have a particular data provided by the organization then we can perform data cleaning using regex function and Data generation should be done using Kafka.