Task 07 ETL Process on Sales Data using Python 📌

You'll be given a csv dataset file and you'll be extracting data from it through pandas or pyspark. Then there will be the transformation phase and after performing that you'll try to load the final data to a database.

[There are some problems with the data and you need to remove them during the transformation phase to make data useful for everyone:

- An order id should always exist as an integer

- A product id cannot be 0

- We never had a product priced more than 1500 Rs. so any item with amount greater than 1500 Rs is an anomaly and it should be treated as 1500 Rs

- A status of an item can never be null or None, if it is then its an anomaly and item rows to be considered as fake orders and should not be kept in final data

- There must be no duplication in final data.

Directions to solve the problem:

- You'll be learning how to CREATE a table in database with pre-defined schema

- Will be figuring out how in the best way you can remove anomalies from the data

- How to read the data with pandas or pyspark