

# Nortal - Data Engineer Assignment

At Nortal, you will be responsible for implementing high-scale and large volume data processing and analytics solutions. You will be building data processing pipelines for a variety of data sources, loading and preparing the data into a production environment. Our projects are all about making a positive impact in our client business by developing robust architectures with the latest technologies. Sounds pretty exciting, huh?

So, to make you familiar with your duties, here are three use cases that you can bump into in your daily work:

1. Our client sent to us his data in a csv format and we need to map this data to our internal schema. In this task, we expect that you load the csv file, rename the columns according to the mapping below and cast the columns for the correct format. In addition to the mapping:
  - a. Timestamps are in GMT+3, convert it to GMT+1.
  - b. Generate an id column to the items table by concatenating the transaction id column with the product id column, separated by a dash (-). Ex.: 64532-676

Transactions CSV	Internal Column Name	Data Type
ID	id	Integer
Timestamp	timestamp	Timestamp
Customer ID	guest_id	Integer
Total	total_amount	Float

Items CSV	Internal Column Name	Data Type
Transaction ID	transaction_id	Integer
Product ID	product_id	String

2. Our client didn't send the guests csv but we have the columns Customer ID and Timestamp in the transactions file that can be used to generate the guest table. Since we may have the same Customer ID in different transactions and we don't want duplicate guests, take the first timestamp for that particular guest.

Items CSV	Internal Column Name	Data Type
Customer ID	guest_id	Integer
Timestamp	created_at	Timestamp

3. We also need to ensure data integrity. For this, we expect that:
  - a. All transactions contain items. Transactions without items need to be excluded.
  - b. All items belong to a transaction. Items without a transaction need to be excluded.
  - c. All transactions contain guests. Transactions without guests need to be excluded.

Feel free to use the packages of your choice, it can be pandas, pyspark, koalas or any other that will accomplish the work. We also expect to see the cases covered by simple unit tests.

If you have any questions regarding the cases, feel free to write to us.