User Story for Data Engineers

Data to be ingested :

1. Customer Table with Customer ID column as primary key, First name , last name, age and country. With below data checks:
   1. Customer ID has to be Unique.
   2. All the columns should not have any null values.
   3. First name and Last name should have exclusively only A-Z or a-z no special characters or numbers.
2. Order Table with Order ID column as primary key , Item , Amount and Customer ID. With the below data checks:
   1. Order ID has to be unique.
   2. All columns should not have null values.
   3. Amount has to be a Integer or a Float.
   4. Customer ID should have values which are available in Customer Table's Customer ID.
3. Shipping Table with Shipping ID as primary key , status and Customer ID. With the below data checks:
   1. Shipping ID has to be Unique.
   2. All columns should not have null values.
   3. Customer ID should have values which are available in Customer Table's Customer ID.

User Story for Quality Engineers

1. Go through the user story of Data Engineers.
2. Compare the data against the source and note down the differences.
3. All the data types has to be verified to have correct data type. Any discrepancy to be noted.
4. Context has to be checked meaning the orders can be placed by customers listed in customer table, and only those items can be orders which are listed in customer table (Items table not given currently)
5. All the data has to be checked against the latest snapshot of data ensuring only the most recent data stays in the reporting layer.

Methods Suggested:
If the data ingestion is a one time activity then ensuring the above is sufficient. In case if its a pipeline where the data has to ingested regularly then all the above checks are to be created as **DBT tests** in case if DBT is used for data ingestion and transformation, if we dont have DBT setup available and if we are using Python for data pipelining and Ingestion then we can use the code available in the shared sanity check file (https://github.com/BharathNandan036/PEI/blob/main/PEI_Data_Sanity_Checks.ipynb) to ensure the correct data is ingested in the reporting layer.
Additionally since the customer data has demographic and personal identifying information we have use any data encryption algorithm so that there is no data leakage and also ensuring only the correct teams get correct access. These data encryption can be done using Python or DBT based on the tools available.