User Story for Data Engineers

Data to be ingested :

1. dim_customer Table with customer_id (string) column as primary key, First_name , last_name, age and country. With below schema:
   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.
   4. Age should be an integer value and > 0

2. dim_Item table with Item_id (string) (numbers only) as primary key, item_name with the below schema:
   1. Item_id string numeric only and should be unique
   2. Item_name string and alphabets only, trim spaces and standardise no duplicates.

3. fact_order Table with Order ID column as primary key , Item_id , Amount and Customer_ID. With the below schema:
   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 (positive decimal numbers only)
   4. Customer_ID should have values which are available in Customer Table's Customer_ID column and is foreign key to join customer table.
   5. Item_id should have values from Item table and is a foreign key to join to item table.

4. Shipping Table with Shipping ID as primary key , status and Customer ID. With the below schema:
   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 column and is foreign key to join customer table.

User Story for Quality Engineers

1. Dim_customer :
   1. Verify if the schema matches to the specifications.
   2. Check if first name and last name column contains only alphabets
   3. Check if the customer_id is unique
   4. Check if age is a positive number (> 0)

2. Dim_item:
   1. Item_id has to be unique with string data type and only numbers.
   2. Item_name should be standardised eg: there should not be 2 entries for “Hard Disk” and “Harddisk”

3. Fact_order:
   1. Check if the foreign keys (customer_id and item_id) joins with dim_customer and dim_item.
   2. Order_id has to be unique
   3. Amount should be only positive decimal value.

4. Fact_shipping:
   1. Check if the status is ENUM have only “Pending” and “Delivered” values

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.