PEI

Data Analysis Report

PEI INTERVIEW PROJECT

SARTHAK AGRAWAL

Table of Contents

Table of Figures

Introduction

The company sells computer accessories through e-commerce. They have a global customer base targeting countries USA, UK and UAE.

Business Process

This section provides details about the business considerations related to the given dataset.

Order-Processing Steps

The process from placing an order to delivering it to the customer is as follows:

1. Customer places an order on the website.
2. Orders Management department collects the items and send it to the Shipping Department.
3. The shipping department packs the items, sends them to the customer and tracks the delivery status.

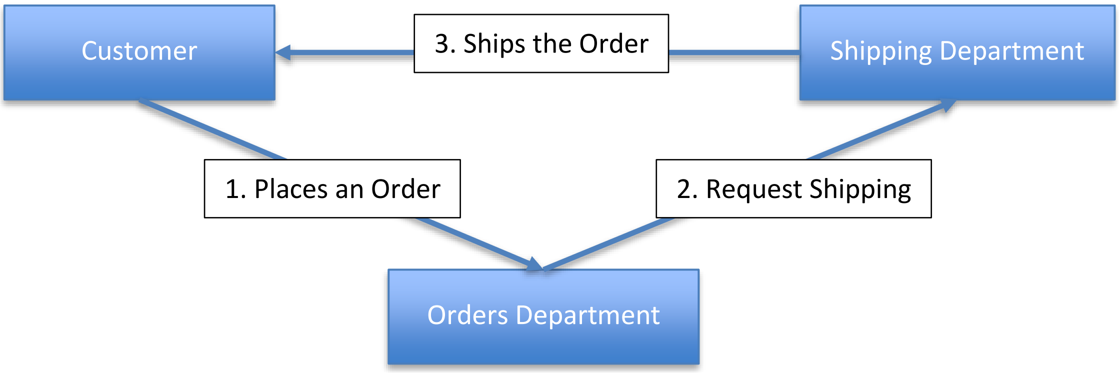

Figure 1: Conceptual Data Flow Diagram

Assumptions

Following business process assumptions have been made based on the provided datasets:

1. The company deals with Computer Accessories only.
2. The company does not operate outside of USA, UK and UAE.
3. The shipping department is an internal entity of the company; hence the data can be shared with other departments internally.
4. All amounts are logged in same currency. All currency conversions (if any) have already been performed and then added to the datasets.

Data Analysis & Reporting Requirements

Based on the document received, following reporting requirements have been identified:

1. List the ***Countries*** and ***Total Amount Spent*** on orders where the Shipping is “*Pending*”.
2. Show ***No. of Transactions***, ***Total Quantity Sold***, and ***Total Amount Spent*** per Customer per Product.
3. ***Top products sold*** in each country.
4. Divide the customers into 2 sections by Age: “Above 30 years” and “Below 30 Years”. List the **most popular products** based on ***No. of Orders***
5. The country with least ***Total Amount Spent*** and least ***No. of Transactions***.

Data Availability

Data Sources

Following datasets are available and their technical specifications are mentioned in the table below:

Table Name	Description	Source	Dataset Format	# Data Points	Size
Customer	Lists details of the customers who have purchased products from the company. ***Caution: Contains PII Information***	Link	Excel (xls)		31 KB
Order	Lists the items purchased by a customer and the transaction value.	Link	Text (csv)		6 KB
Shipping	Lists the customer and the status of delivery.	Link	Text (JSON)		20 KB

Data Dictionary

Customer

Field Name	Description	Assumed Data Type	Assumed Constraints
Customer_ID	Unique Id assigned to each customer.	Integer	Primary Key, Unique, Not Null
First	First Name of the Customer	String	Not Null
Last	Last Named of the Customer	String	Not Null
Age	Age of the Customer (assumed) as of today.	Integer	Not Null
Country	Country where the customer lives in	String	Not Null

Order

Field Name	Description	Assumed Data Type	Assumed Constraints
Order_ID	Unique Id assigned to each order.	Integer	Primary Key, Unique, Not Null
Item	Item purchased in the order.	String	Not Null
Amount	Amount spent in the order	Decimal	Not Null
Customer_ID	Customer who placed this order	Integer	Foreign Key, Not Null

Shipping

Field Name	Description	Assumed Data Type	Assumed Constraints
Shipping_ID	Unique Id assigned to each shipping activity.	Integer	Primary Key, Unique, Not Null
Status	Status of the shipping (Delivered / Pending)	String	Not Null, only 2 possible values
Customer_ID	Last Named of the Customer	String	Foreign Key, Not Null

Entity-Relationship Diagram

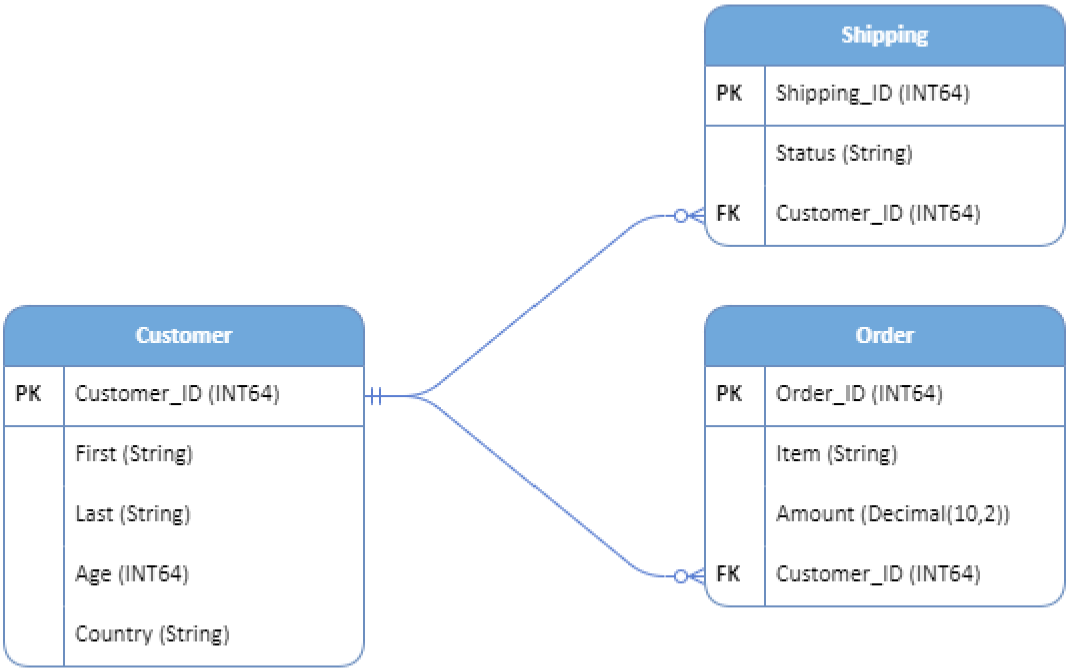

Figure 2: Entity-Relationship Diagram

Understanding the Data

Key Metrics Identified

Summarizing the requirements into key metrics that the Business wants to focus on:

- **Total Amount Spent**: Sum of Amount from Orders table
- **Total Quantity Sold**: Count of Items (not distinct) from Orders table
- **Total Transactions**: Count of Order_ID from Orders table
- **No. of Orders**: Count of Order_ID from Orders

Data Flow Diagram

The diagram below shows how data is generated by the customer when they place an order, then processed 3 times while interacting with three data stores owned by the company.

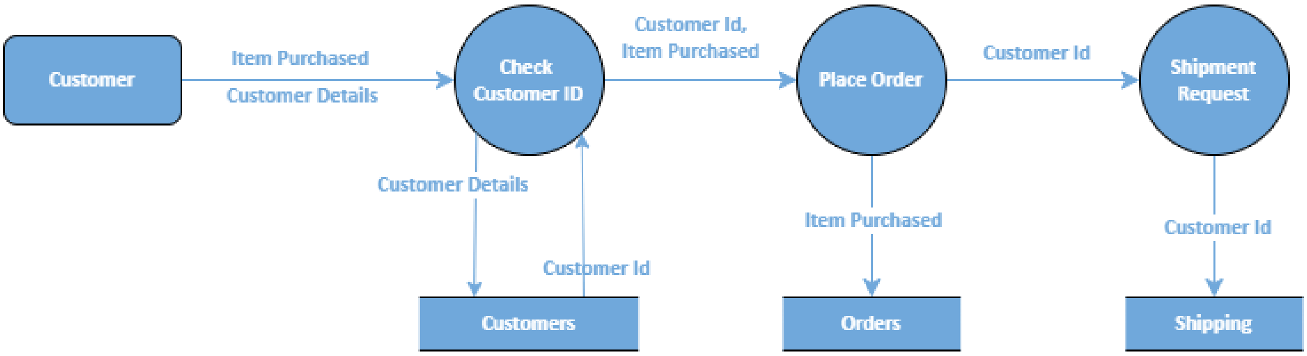

Figure 3: Data Flow Diagram

Anomalies Observed in Schema

- Shipping information does not contain order information.
  - Since orders are shipped to customers, the shipment information should be related to both Customer and Order tables.
- Item should not be present in Orders table.
  - Item can exist without an order; hence it should have its own identity (or, Primary Key).
  - One table cannot contain multiple Primary Keys; hence, Item details should be moved to a separate table, including the Amount.
- Amount is an ambiguous field here. We do not know if it is Total Order Amount, or Unit Price Amount.

Exploratory Data Analysis

Data-based Assumptions

1. Customers have placed only one item per order. This is because quantity ordered information is not provided.
2. The Shipping Status only checks for 2 values, Pending and Delivered. Information about packing, shipping initiation and shipping location (or, In Progress) will not be considered in the analysis.
3. There are only 8 products that the company sells. All analyses will be done accordingly.
4. The *Amount* field in *Order* table has been considered as both Unit Price and Total Order Amount.
5. No date columns are provided, so time-series analysis is not possible.
6. All anomalies such as missing values and misspellings will be considered as Manual Errors / Data Entry errors.
7. No more data can be requested now. All requirements to be addressed after sharing the first final version of the report and may be considered based on communications with Stakeholders.

Data Cleaning

S.No.	Cleanup Task	Observations
1	Check for misspellings in fields: Country, Status, Item	• There are no misspelled data in *Country*, *Status* and *Item* columns. • *Country*, *Status* and *Item* columns have consistent datatypes. • *Country*, *Status* and *Item* columns follow Atomicity, i.e. one data point contains only one value **Resolution:** Not Required
2	Check missing data in all fields	• There are 250 rows in each table, and all columns have 250 non-null values • There are no NULLs, zeroes, or <Blanks> in all three tables. **Resolution:** Not Required
3	Check the tables should not contain any duplicates.	• Assumed primary keys *Customer_ID, Order_ID* and *Shipping_ID* follow all data consistency constraints. • There is only one item purchased per order. **Resolution:** Not Required
4	Check for Outliers based on Amounts & No. of Orders Valid range will be considered from [Quartile_1 - (IQR * 1.5)] to [Quartile_3 + (IQR * 1.5)]	• The outliers have been identified based on the statistical method which considers outliers as values outside the range from [Quartile_1 - (IQR * 1.5)] to [Quartile_3 + (IQR * 1.5)]. This method especially works with data that does not follow normal curve. • From customers table, we don't see any customer outside the Age range (0, 114), hence there are no outliers. Since age cannot be negative, we can assume the range starts from 0. • From orders table, the max of Amount value is outside the valid data range (0, 3300). On further analysis, we found that there are significant number of orders with the amount above 3300. Hence, no action is required. • shipping table does not contain numeric data, except the identifiers. **Resolution:** Not Required

Data Validations Performed

S.No.	Validation Task	Observations
1	Do all entries in a column have same datatype. **Validation Successful**	• All primary keys (Order_ID, Customer_ID and Shipping_ID) have no missing values and have int64 data type. • All Foreign Keys (Customer_ID in Shipping and Order table) have no missing value and have int64 datatype. • All numeric columns, i.e. Age and Amount have no missing values and have int64 data type.

		• All string columns (First, Last, Country, Status, Item) do not have missing values and have object type.
2	Do all entries lie in the same/expected range. **Validation Successful**	• As established, we can see that Customer_ID, Order_ID and Shipping_ID primary key columns have no missing values, no duplicates and min and max values are 1 and 250 respectively. Hence, these columns contain all integers between 1 and 250, both inclusive. • Customer_ID foreign key has values from 4 to 250 in Orders table and 1 to 248 in Shipping table. Hence, there is one primary key for each foreign key here. • Age column ranges from 18 to 80, which indicates the orders were placed by customers with appropriate age. • Status column has only 2 unique values as expected. • Item column has 8 unique values as expected. • **Special characters '!', '1', '0' & '@' were found in First and Last columns. These will need to be replaced with appropriate characters.**
3	Do all entries have the same format. **Validation Unsuccessful**	

While importing data into reporting tool, please update the following:

- Replace '!' and '1' with 'l' in First and Last name columns.
- Replace '0' with 'o' in First and Last name columns.
- Replace '@' with 'a' in First and Last name columns.

Data Anomalies Checked

S.No.	Exploration Task	Observations
1	Have all orders have been shipped?	There are 94 orders for which shipment has not been started.
2	Number of customers who placed multiple orders should be same as customers who have multiple shipments.	There are 67 customers who have multiple shipments. There are 60 customers who have placed multiple orders.
3	Is there any shipment for which order doesn't exist?	There are 98 shipments for which order could not be found.
4	Is there any item sold at 2 or more different amount?	Mousepad has been sold at 2 different prices 25 times each.
5	Is there any order that is linked to multiple shipments?	There are 92 orders that are linked to multiple shipments.

***CAUTION:** Above mentioned anomalies will greatly affect the quality of reports generated from the given data.*

Data Reporting

All data reporting has been done using Power BI because:

- Power BI is readily available, free of charge on all Windows PC
- Power BI can be used to export data and visualizations into a file and shared through Git as requested.
- Power BI is most popular Business Intelligence tool according to Gartner report.

Data Preparation Steps:

- Downloading and importing the files into Power BI
- Replacing special characters with correct letters
    - Replace ‘!’ and ‘1’ with ‘l’ in First and Last name columns.
    - Replace ‘0’ with ‘o’ in First and Last name columns.
    - Replace ‘@’ with ‘a’ in First and Last name columns.
- Check if the data types have been imported correctly.
- Setup table relationships using Foreign Keys
- Creating measures to calculate No. of Orders, Total Amount Spent, Total Quantity.
- Create visualizations that best suit the reporting requirements.
- Add captions to explain the results.
- Publish and share the dashboard with Stakeholders.

Analysis Results

1. Total Amount spent in each country for PENDING shipments are as follows:

Country	Total Amount Spent on Pending Shipment
UK	128800
USA	52850
UAE	34750

2. The total number of transactions, total quantity sold, and total amount spent for each customer, along with the product details has been listed on the dashboard in a table format due to high number data points.
3. The maximum product purchased in **UAE** is **Keyboard**, while **Mousepad** is the most purchased product in **UK** and **USA**
4. The most popular product Age Group <30 is **Mousepad** and for Age Group > 30 is **Keyboard**.
5. **UAE** is the country with least *Total Amount Spent* at 49950 and least *No. of Orders* at 40.