**SALES DATA ANALYSIS**

**PROJECT OVERVIEW:**

PEI group provided 3 datasets which must be cleaned, modelled and joined to get insights for the business.

Below image shows the higher-level architecture of the project.

All the 3 datasets to be cleaned and loaded to the storage database. The database will have two layers

1. Exact data layer
2. Aggregated data layers.

Aggregated data layer will be powering the tableau/power bi which will provide the insights

A computer screen shot of a diagram

Description automatically generated

**Table structure and data Issues:**

There was total 3 tables provided by the DE team. And they are Customer Master Table, Orders Table and Shipping table.

*Customer* Master table: It has major details of the customer demographics

*Orders table:* It has the information about the order details and can be treated as our fact table or base table.

*Shipping table:* it has the details of shipping information about the customer.

We have done the analysis of the three datasets which were given to us.

There were many gaps in the datasets provided by the team.

Below are the key findings:

1. Customer master table has many customer\_ids(customer numbers) who neither placed an order or their orders are shipped.

The number of such customers is approximately 35. Please find the below examples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer\_ID | First | Last | Age | Country |
| 7 | David | Davis | 59 | USA |
| 16 | David | Benson | 61 | USA |
| 19 | James | Johnson | 67 | UK |
| 25 | Raymond | Taylor | 36 | UAE |

b. Data discrepancy related to order and shipping: There are few customers whose details are missing in orders table, they are appearing in shipping table. Below examples for shipping table

|  |  |  |
| --- | --- | --- |
| Shipping\_ID | Status | Customer\_ID |
| 1 | Pending | 173 |
| 2 | Pending | 155 |
| 5 | Delivered | 72 |
| 9 | Pending | 199 |

Similarly, we found cases where there were details about the customers in orders table but not in shipping table and pasted few examples of such cases in orders table below

|  |  |  |  |
| --- | --- | --- | --- |
| Order\_ID | Item | Amount | Customer\_ID |
| 1 | Keyboard | 400 | 139 |
| 2 | Mouse | 300 | 250 |
| 3 | Monitor | 12000 | 239 |
| 4 | Keyboard | 400 | 153 |

A**ctual Data Model based on the provided data:**

We have created physical data model using the data provided to us.

We couldn’t identfy the relatinship between the orders and shipping table which has to be the foremost requirment.

Customer table doesnt provide many data cuts related to customer demographics which might not help in understanding the customer behaviour.

A diagram of a product

Description automatically generated with medium confidence

**Communication to DE team:**

1. Apart from data discrepancy, we are unable to model the data because we are unable to draw the relationship between orders and shipping table.

We must use customer table as our base table which will connect both orders and shipping table which is not an optimized way to connect.

1. If we get the relationship between orders and shipping table, then we can go with starschema model which is optimized way to draw the relationship and querying will be faster.
2. Indexing can be done with starschema model. Which will improve faster query response.
3. If we can track the load time for each table, which will help the business in analysing the KPIs based on the reporting period.

Few assumptions made as per our analysis:

1. As we can see customer master table have customer details who don’t have any records in shipping or orders meaning it might be some historic data. If we can get the historic data, then we can report other sales KPIs like customer cohort or return customer ratio over the timeline.
2. If we draw any relationship between the shipping table and orders table, it has to be based on the shipping\_id. Each shipping\_id can have multiple orders or single order.
3. Shipping date, order date is missing. We have created proposed data model where we have identified few datapoints which can be tracked and helpful in our analysis.

**Proposed data Model:**

We have created the below data model, and it can be enhanced furthermore based on the conceptual data model.

We have added many data points in all the tables and the data load time column which will help in understanding when the data has been loaded.

The datatype of each column can be changed accordingly, and many lookup tables can be added as dimensional tables to this model.

Order date and shipping date give overall picture of when the orders are ordered and when it’s shipped. We can track customer satisfaction related to the delivery timeline.

**A screenshot of a computer

Description automatically generated**

**Analysis of the data:**

We have built a tableau dashboard by using the existing data. There are multiple key takeaways which we have mentioned in the dashboard itself as insights sheet.

Below is the image of the dashboard.

A screenshot of a data analysis dashboard

Description automatically generated