Table of Contents

Introduction and Assumptions

The analysis is approached based on these assumptions on the reporting aspect:

1. A data visualization tool such as Power BI will be used to show the final report.
2. The requirements posed were analyzed and checked if they can be answered on Power BI. Power BI can easily handle the posed requirements provided the data is accurate. Hence the focus of this exercise is not on the final visualization. Instead, the focus of this assessment has been more on the high-level data understanding, the required data model, and the technical documentation.
3. Ideally, the first step of this analysis would begin with understanding the current business model and processes in detail. For the sake of simplicity, I have inferred the data flow based on the provided data.
4. Please note that the technical specifications provided below is only for a part of the data model process. The complete specifications document would involve the below sections:
   a. Business understanding – covers the terminologies and business processes.
   b. Project timeline and owners
   c. Initial EDA report – based on which the data model and other below steps follow.
   d. Data description – definitions of each column in the source data.
   e. Data Quality – any data quality checks required and defining data constraints
   f. Data Selection and Cleaning – if any subset of data is to be taken and what sort of cleaning needs to be taken care in the ETL.
   g. Data Derivation – which involves any derived fields required for reporting.
   h. Data Modeling – proposed data model.

Technical Specifications of the Data Model – I

Sales Data Overview

Data Modeling Report

Data Team - PEI India

PEI Group – India

Doc version 1.0

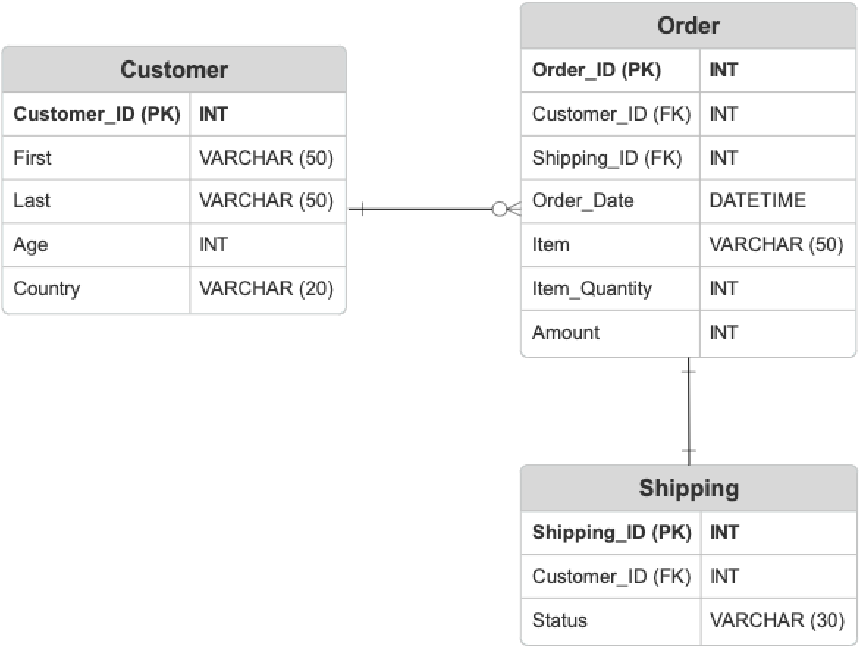
11-07-2024

Revision History

Version	Author	Date	Description
1.0	Sharath V	11-07-2024	Initial version

Schema Overview / ERD

The proposed data model consists of three entities: customer, order and shipping. There are new columns added as per the requirements and findings on the accuracy and completeness of the datasets.

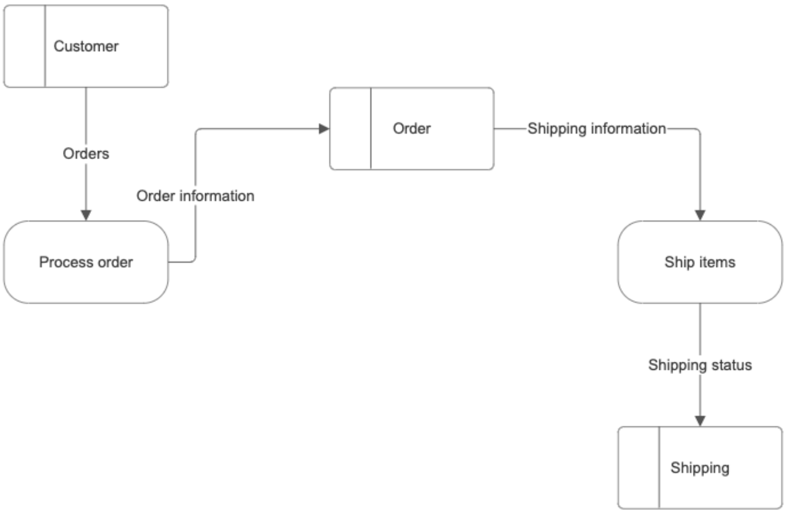


Customer	
Customer_ID (PK)	**INT**
First	VARCHAR (50)
Last	VARCHAR (50)
Age	INT
Country	VARCHAR (20)

Order	
Order_ID (PK)	**INT**
Customer_ID (FK)	INT
Shipping_ID (FK)	INT
Order_Date	DATETIME
Item	VARCHAR (50)
Item_Quantity	INT
Amount	INT

Shipping	
Shipping_ID (PK)	**INT**
Customer_ID (FK)	INT
Status	VARCHAR (30)

Data Flow Diagram



Data Definitions

Table	Column	Business Description	Data Type	Nullable Y/N?	Remarks, if any
Order	Order_ID	Represents a unique customer order	INT	N	
Order	Customer_ID	Represents a unique customer	INT	N	
Order	Shipping_ID	Represents the shipping details per order	INT	N	The proposal is to have the shipping data at order level. Shipping_ID will help with the order to shipping status mapping.
Order	Order_Date	The date on which the order was placed by the customer	DATETIME	N	New column
Order	Item	The product that was ordered by the customer	VARCHAR	N	
Order	Item_Quantity	The quantity of the product that was ordered by the customer	INT	N	New column

Data Mapping

Mapping of the source data into the new schema

New Table	New Column	Source Table	Source Column
Order	Order_ID	Order	Order_ID
Order	Customer_ID	Order	Customer_ID
Order	Shipping_ID	Order	To be checked/added from shipping data
Order	Order_Date	Order	To be checked/added
Order	Item	Order	Item
Order	Item_Quantity	Order	To be checked/added

Areas of Concern

1. The frequency of data retrieval is yet to be ascertained. This will be based on the reporting requirements. If the data must be on real-time, then a live connection is necessary.
2. If there are other components / entities, they need to be considered in the ERD / data model. For example, if product details are available then it must be added.
3. Multiple entries for a single order in shipping data.
4. Missing entries in shipping data for a few orders.
5. If shipping data stores historical shipping status as the order moves from order placed, processed, delivery pending to finally delivered, it is advisable to keep the timestamp for each status. In this way, we can calculate the most recent status during the ETL, and it is helpful to even calculate further KPIs such as the actual time taken from order placing to delivery.