**E-commerce Order Fulfillment & Delivery Analytics Assignment Report**

**-- Sayan Das**

**1. Introduction**

In this assignment, we integrated SAP ERP data into a SQL-based data warehouse to enable robust reporting and analytics. The goal was to extract data from multiple SAP modules—covering sales orders, shipments, customer information, and carrier details—and transform it into a relational data model that supports advanced analysis in Power BI. This approach allows the e-commerce company ("ShopX") to monitor order processing times, delivery performance, and carrier efficiency, thereby optimizing fulfillment and improving customer satisfaction.

**2. ETL Process Overview**

**Data Extraction**

Data was extracted from an Excel file containing SAP ERP sheets:

* **KNA1:** Customer master data
* **LFA1:** Vendor (carrier) data
* **VBAK:** Sales order header
* **VBAP:** Sales order items
* **LIKP:** Delivery (shipment) header
* **LIPS:** Delivery (shipment) items
* **VTTK & VTTP:** Shipment status details

**Data Transformation**

The transformation process was designed to build a robust relational data model with the following tables:

* **customers:** Created from KNA1, this table stores customer details.
* **sap\_customers:** A duplicate of customers, to retain raw SAP customer information if needed.
* **orders:** Derived from VBAK, representing order header information.
* **order\_items:** Derived from VBAP, capturing the item-level details of each order.
* **shipments:** Created from LIKP, holding shipment header data.
* **shipment\_items:** Derived from LIPS, containing details about individual shipment items.
* **carriers:** Populated from LFA1, storing carrier/vendor details.
* **delivery\_status:** Created by merging VTTK and VTTP, this table tracks shipment statuses and related timestamps.
* **delivery\_analytics:** An aggregated table that computes performance metrics such as delivery time and on-time status by joining orders and shipments.

Key transformation tasks included:

* Renaming columns to standardized lower-case names with underscores.
* Converting date fields to proper datetime formats.
* Merging related data sources (e.g., joining VBAK with VBAP, LIKP with LIPS, VTTK with VTTP) to create comprehensive tables.
* Validating data quality through primary key uniqueness and referential integrity checks.

**Data Validation**

Before loading the data into MySQL, we performed several validation checks:

* **Primary Key Uniqueness:** Ensured that key fields (or composite keys) are unique for tables like orders, shipments, and delivery\_status. (A warning was issued for duplicate shipment numbers in delivery\_status, which can be addressed by aggregating or removing duplicates based on business rules.)
* **Foreign Key Consistency:** Verified that foreign keys (e.g., customer\_id in orders) exist in the corresponding parent tables.

**Data Load**

The validated data was loaded into a MySQL database, using SQLAlchemy, with table names stored in lower-case. This ensures consistency and avoids issues with case sensitivity in SQL queries.

**3. Representative Code Snippet**

Below is a representative code snippet from the Python notebook that demonstrates the extraction, transformation, validation, and loading process:

import pandas as pd

from sqlalchemy import create\_engine

import numpy as np

# Define Excel file path and load sheets

excel\_file = 'SAP-DataSet.xlsx'

xls = pd.ExcelFile(excel\_file)

print("Available sheets:", xls.sheet\_names)

# Extract data from SAP sheets

df\_kna1 = pd.read\_excel(excel\_file, sheet\_name='KNA1')

df\_lfa1 = pd.read\_excel(excel\_file, sheet\_name='LFA1')

df\_vbak = pd.read\_excel(excel\_file, sheet\_name='VBAK')

df\_vbap = pd.read\_excel(excel\_file, sheet\_name='VBAP')

df\_likp = pd.read\_excel(excel\_file, sheet\_name='LIKP')

df\_lips = pd.read\_excel(excel\_file, sheet\_name='LIPS')

df\_vttk = pd.read\_excel(excel\_file, sheet\_name='VTTK')

df\_vttp = pd.read\_excel(excel\_file, sheet\_name='VTTP')

# Transform: Example for Orders table

orders\_df = df\_vbak.copy()

orders\_df.rename(columns={

    'Sales Document': 'order\_id',

    'Order Date': 'order\_date',

    'Customer ID': 'customer\_id',

    'Order Type': 'order\_type',

    'Sales Organization': 'sales\_organization',

    'Distribution Channel': 'distribution\_channel',

    'Division': 'division',

    'Order Status': 'order\_status'

}, inplace=True)

orders\_df['order\_date'] = pd.to\_datetime(orders\_df['order\_date'], errors='coerce')

# Data Validation: Check primary key uniqueness

def check\_primary\_key\_uniqueness(df, key\_columns, table\_name):

    duplicates = df.duplicated(subset=key\_columns)

    if duplicates.any():

        print(f"WARNING: Duplicates in {table\_name} for key columns {key\_columns}:")

        print(df[duplicates][key\_columns])

    else:

        print(f"Primary key check passed for {table\_name}.")

check\_primary\_key\_uniqueness(orders\_df, ['order\_id'], 'orders')

# Load Data into MySQL

username = 'root'

password = '12345'

host = 'localhost'

port = '3306'

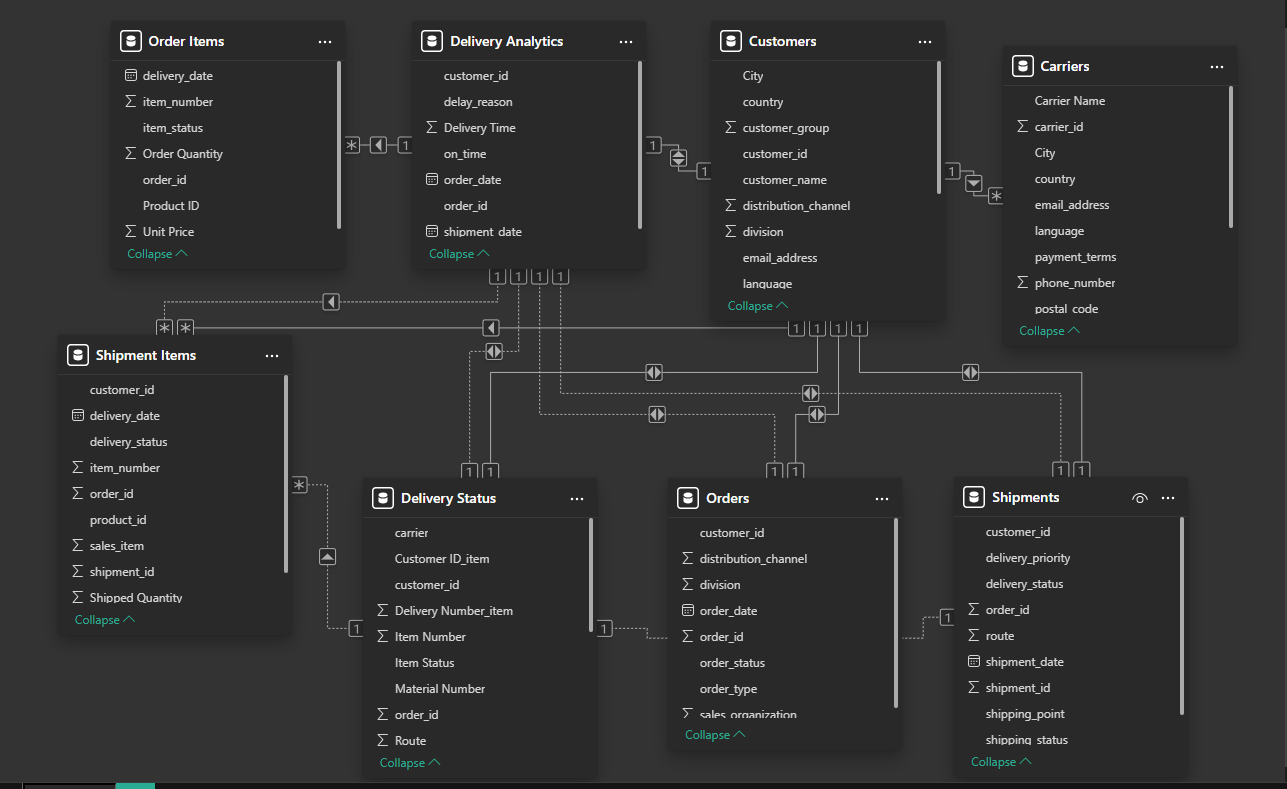
database = 'case1'

engine = create\_engine(f'mysql+pymysql://{username}:{password}@{host}:{port}/{database}')

orders\_df.to\_sql('orders', con=engine, if\_exists='replace', index=False)

print("Data loaded to MySQL database successfully.")

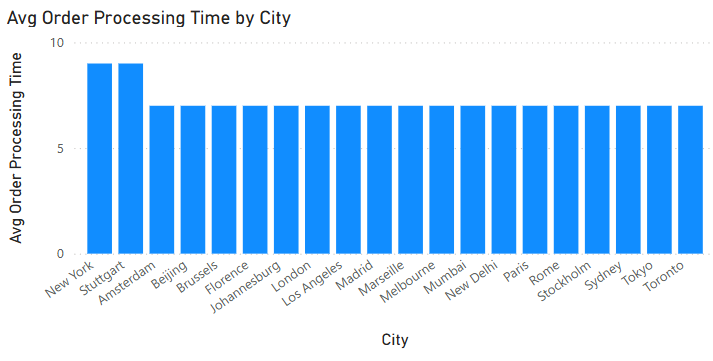
**SQL Schema**



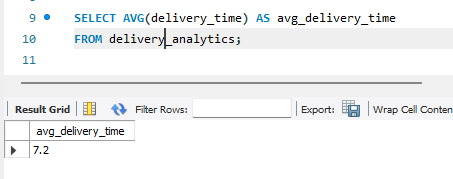
**4. Power BI Reporting and Analytics**

After loading the data into the MySQL database, the next step is to build Power BI reports and dashboards to visualize key performance metrics. The following KPIs and graphs are planned for the Power BI dashboard:

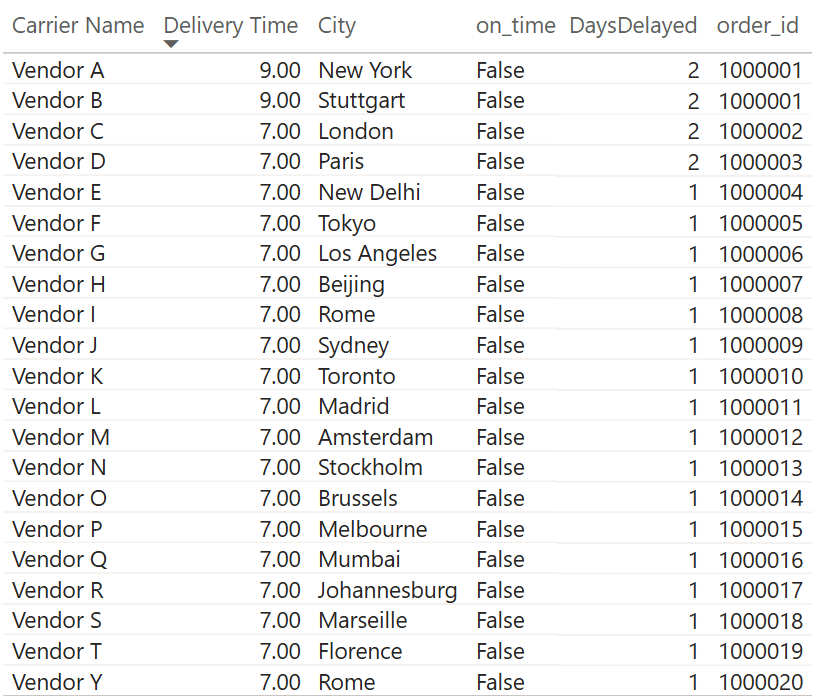
* **Order Processing Time: Difference between delivery date and order date**
  + *Minimum time – 7* days
  + *Maximum time – 9 days*

****On-Time Delivery Rate:** **Percentage of deliveries that arrive on time**

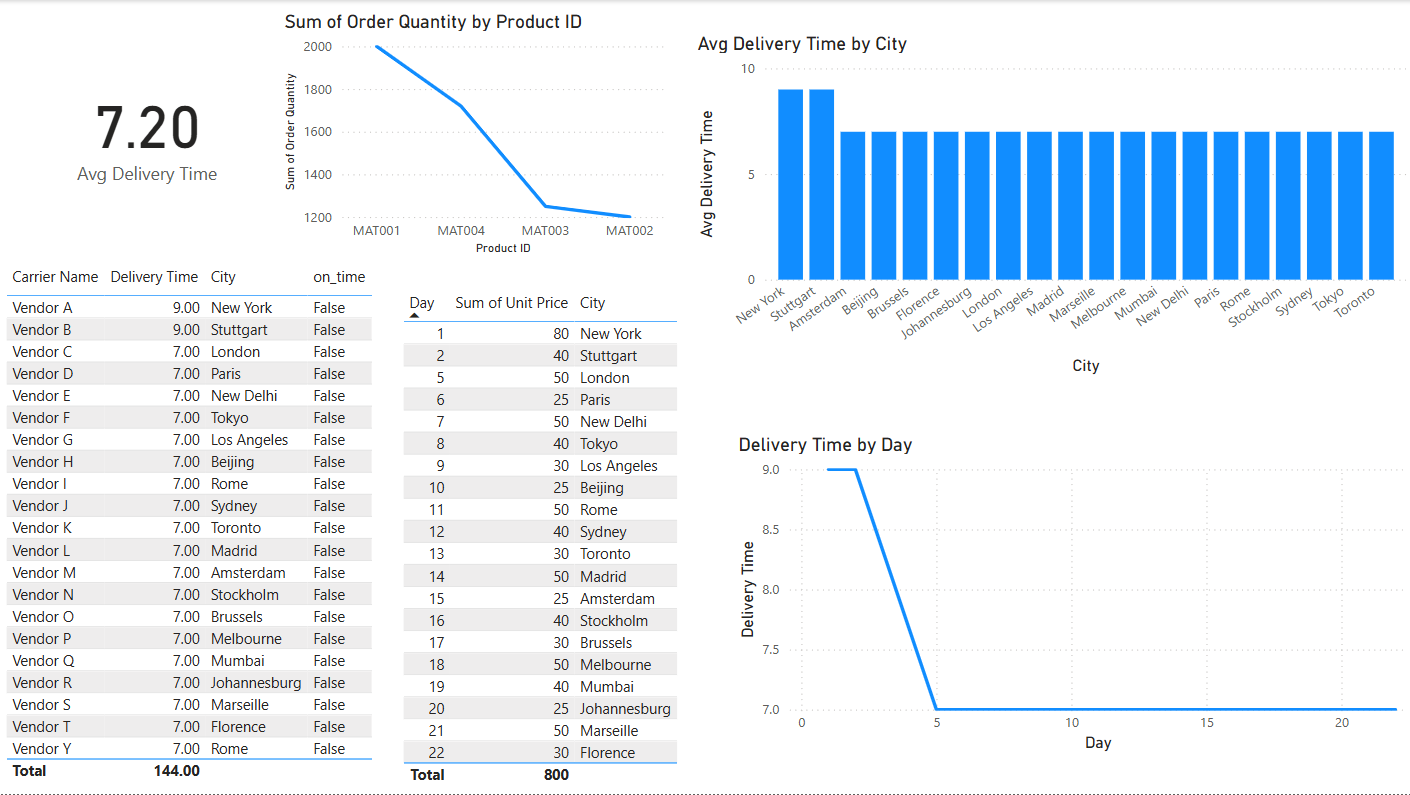
* + *None of the packages was delivered on time.*
* **Average Delivery Time:**

****

* **Carrier Performance: Carrier's performance by aggregating the average delivery time**

**

**Full Report**

****

**5. Conclusion**

In summary, we have:

* **Extracted** raw SAP ERP data from an Excel file containing multiple sheets.
* **Transformed** the data into a structured relational model that includes tables for customers, orders (header and items), shipments (header and items), carriers, delivery\_status, and delivery\_analytics.
* **Validated** the data for completeness, uniqueness, and referential integrity.
* **Loaded** the clean, transformed data into a MySQL database.
* **Prepared** the environment for Power BI reporting by creating key measures and KPIs (such as Order Processing Time, On-Time Delivery Rate, Average Delivery Time, and Carrier Performance).

We found that orders were always delivered late and Vendor A,B,C and D were worse than rest.  
ShopX needs to work on improving their delivery time.