**As a Data Analyst, you are required to**

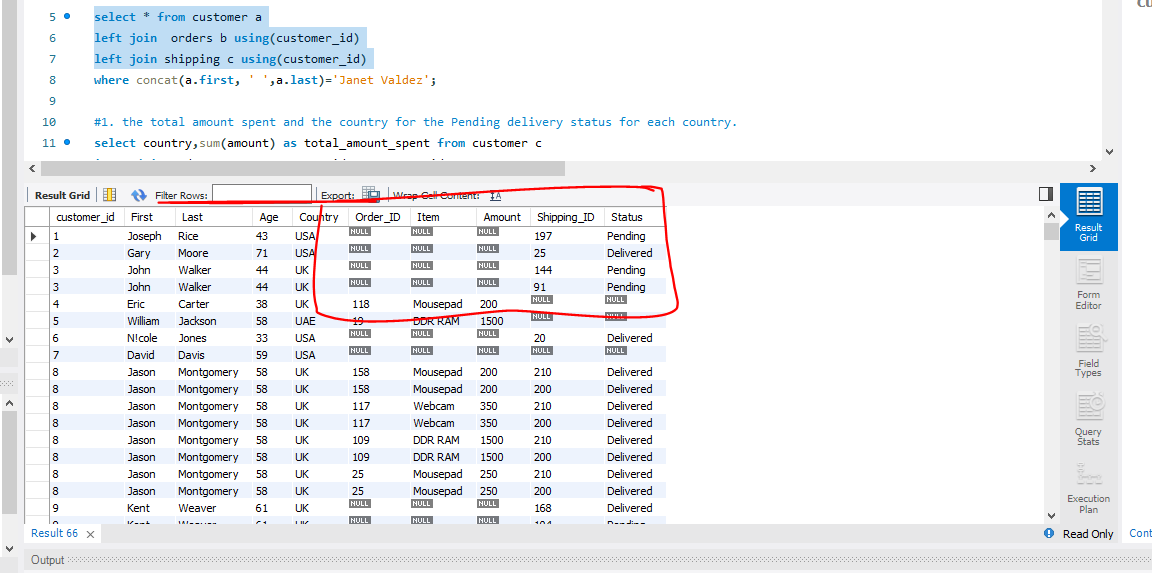
* **Verify the accuracy, completeness, and reliability of source data.**

**Accuracy :**

**Integrity :**

1. select \* from shipping where customer\_id = 153, shipping data for customers is missing from shipping data.

2. Shipping table should have an Order\_id column as customers can have multiple orders.

3

4. Quantity of each item sold is not given.

**Completeness:**

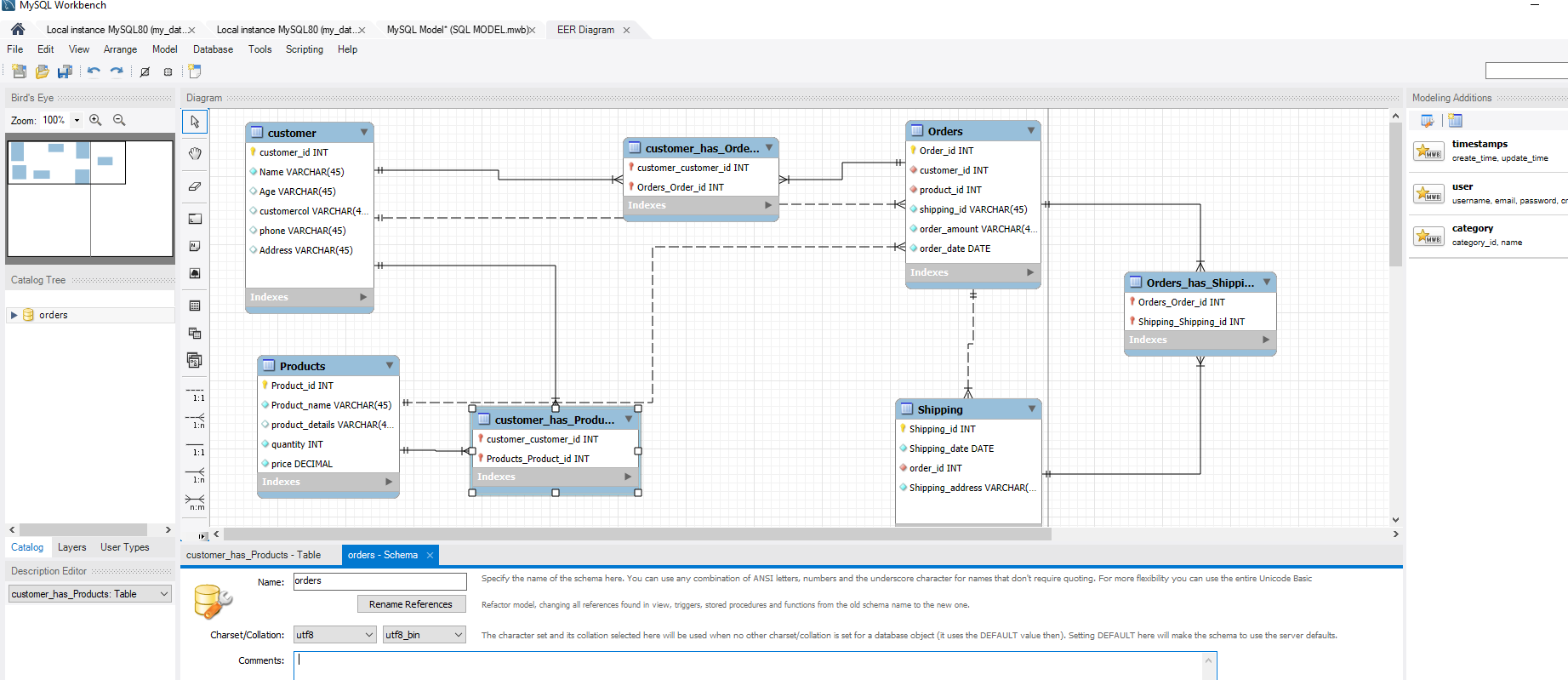
Missing table : Product table should be there, to fetch product tables, and cost price of each product, and quantity sold.

* **Based on your findings, define and outline the requirements for anticipated datasets, detailing the necessary data components.**

1. Shipping table has missing data, depending on order id.
2. Order id should be the foreign key of the shipping table.
3. No way to calculate the number of items sold per order. (create new product table with product price) or order table should have quantity column

* **Develop the data models to effectively organise and structure the information and provide a detailed mapping of existing data flows, focussing on the areas of concern.**

We should have a snowflake schema for the data model, rather than the star schema which is being used now.



* **Prepare a story with technical specifications for one part of the data model for a data engineer.**

1. **Data Source Identification :** The pipeline should be able to connect to and extract data from at least two identified source systems (e.g., a CSV file on a shared drive, a REST API endpoint from a supplier).
2. **Data Ingestion Frequency and setting up pipeline to get data(hourly/daily) :** Determine the appropriate frequency for running the pipeline
3. **Technology Stack:** (To be decided based on the existing infrastructure and team expertise, could include):
   * **ETL/ELT Tools:** Apache Airflow, Informatica.
   * **Scripting Languages:** SQL
4. **Data Extraction:** The pipeline successfully extracts relevant product information from the identified sources. This includes fields like product name, description, price, and stock quantity.
5. **Data Transformation:**
   * Data type consistency, price decimal, order id int, customer\_id int.
   * Standard naming conventions.Product Name to product\_name.
   * Basic data cleansing is performed (e.g., trimming whitespace from text fields, handling potential null values – either by setting defaults or flagging for review).
6. **Data Quality Checks:** Implement basic data quality checks during the transformation phase (e.g., checking for negative prices or stock quantities).
7. **Schema Evolution:** Consider how the pipeline will handle changes in the schema of the source data over time

* **Communicate the findings and insights to stakeholders in a visually comprehensive manner.**

1. **UK is the most revenue generating country, and UAE is the least.**
2. **Keyboard is the most sold item.**
3. **Most deliveries are pending.**

For business professionals, understanding these data models offers several key advantages:

* **Improved Sales Analysis:** This model enables comprehensive sales analysis, such as tracking order volumes, identifying top-selling products, and understanding customer lifetime value.
* **Deeper Customer Understanding:** By linking customer information with their orders, businesses can gain valuable insights into purchasing behavior, preferences, and trends.
* **Centralized Product Information:** A product data model acts as a single source of truth for all product-related information, including names, descriptions,prices, and stock levels. This ensures data consistency across all systems and reduces the risk of errors or discrepancies.
* **Better Inventory Management:** By including stock quantity in the model, businesses can track inventory levels in real-time, prevent stockouts or overstocking, and optimize their supply chain.
* **Enhanced Tracking and Visibility:** This model provides the framework for tracking shipments in real-time, giving both the business and the customers better visibility into the delivery process. This proactive approach helps in managing expectations and addressing potential issues promptly.