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

Source data is in 3 formats namely XLS , CSV and JSON . TO ensure data is consistent we can change all 3 formats into 1 namely CSV so that while reading the data , consistency will be maintained .

Data accuracy – Data types of each field can be examined in this case , For example customer id which is the **Primary key** in customer data can be created with a constraint as Not null . Data types of each field can be marked accordingly , for example AGE should be INT . Name should be VARCHAR

We should check for missing values . IS NULL can help us finding null values . In the erstwhile data set no Null values were present

We can implement some validation checks like format validation for Name , Age etc . Apart from this we can implement range validation . Age should not be more than 3 digits .

Sample SQL query to check Completeness –

SELECT

COUNT(\*) AS Total\_Records,

SUM(CASE WHEN Age IS NULL THEN 1 ELSE 0 END) AS Missing\_Values,

100 \* SUM(CASE WHEN Age IS NULL THEN 1 ELSE 0 END) / COUNT(\*) AS Completeness\_Percentage

FROM

Customers ;

This will calculate if we have any null value and if it is there we can get the percentage as well

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

We have customer dataset with fields – (Customer\_id , First , Last , Age , Country) , Order(order\_id , item , amount , customer\_id ) , shipping (shipping\_id , Status , Customer\_id )

To answer the requirements elicited we will create a join where join will be performed using Customer\_id

Sample query –

SELECT c.country,

SUM(o.amount) AS total\_amount\_spent

FROM customer c

JOIN order\_table o ON c.customer\_id = o.customer\_id

JOIN shipping s ON c.customer\_id = s.customer\_id

WHERE s.status = 'Pending'

GROUP BY c.country;

The intermediary dataset will look like –

Customer\_id , First , Last , Age , Country , order\_id , item , amount , shipping\_id , status , customer\_id : These will be the necessary data components

**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.**

Customer Entity:

This entity represents customers of the bookshop.

Attributes:

CustomerID: A unique identifier for each customer, acting as the primary key.

First : The first name of the customer.

Last : The last name of the customer

Age: The age of the customer.

Country: The country where the customer resides.

Order Entity:

This entity represents orders placed by customers.

Attributes:

OrderID: A unique identifier for each order, acting as the primary key.

Item: The items ordered by the customer.

Amount: The total amount spent on the order.

CustomerID: A foreign key referencing the Customer entity, indicating the customer who placed the order.

Shipping Entity:

This entity represents the shipping details associated with orders.

Attributes:

ShippingID: A unique identifier for each shipping record, acting as the primary key.

Status: The status of the shipping (e.g., pending, shipped, delivered).

CustomerID: A foreign key referencing the Customer entity, indicating the customer associated with the shipping.

Relationships:

Customer-Order Relationship:

One-to-Many relationship from Customer to Order: A customer can place multiple orders, but each order is placed by only one customer.

This relationship is established through the CustomerID attribute in the Order entity, which acts as a foreign key referencing the Customer entity.

Customer-Shipping Relationship:

One-to-Many relationship from Customer to Shipping: A customer can have multiple shipping records associated with their orders, but each shipping record is associated with only one customer.

This relationship is established through the CustomerID attribute in the Shipping entity, which acts as a foreign key referencing the Customer entity.

**Data Flow Processes :**

Level 0 DFD:

Processes:

Data Retrieval Process (DRP): Responsible for retrieving data from the database tables.

Data Processing Process (DPP): Responsible for processing the retrieved data to generate reports.

Reporting Process (RP): Responsible for generating various reports based on processed data.

Data Stores:

Customer Table (CT): Stores information about customers.

Order Table (OT): Stores information about orders.

Shipping Table (ST): Stores information about shipping details.

External Entity:

Reporting User: Represents the end-users who request and consume the reports generated by the system.

Data Flows:

CT Data Flow: Data flow from the Customer Table to the Data Processing Process.

OT Data Flow: Data flow from the Order Table to the Data Processing Process.

ST Data Flow: Data flow from the Shipping Table to the Data Processing Process.

Processed Data Flow: Data flow from the Data Processing Process to the Reporting Process.

Report Data Flow: Data flow from the Reporting Process to the Reporting User.

Level 1 DFD (Data Processing Process):

Processes:

Total Amount Spent Process (TASP): Calculates the total amount spent and country for orders with pending delivery status.

Transaction Analysis Process (TAP): Analyzes total transactions, quantity sold, and amount spent for each customer along with product details.

Maximum Product Purchased Process (MPPP): Determines the maximum product purchased for each country.

Most Purchased Product Process (MPPP): Identifies the most purchased product based on age category.

Country Analysis Process (CAP): Analyzes the country with minimum transactions and sales amount.

Data Stores:

Same as Level 0.

Data Flows:

Processed Data Flow: Data flows from Level 0 processes to Level 1 processes.

TASP Data Flow: Data flow from Level 1 to Total Amount Spent Process.

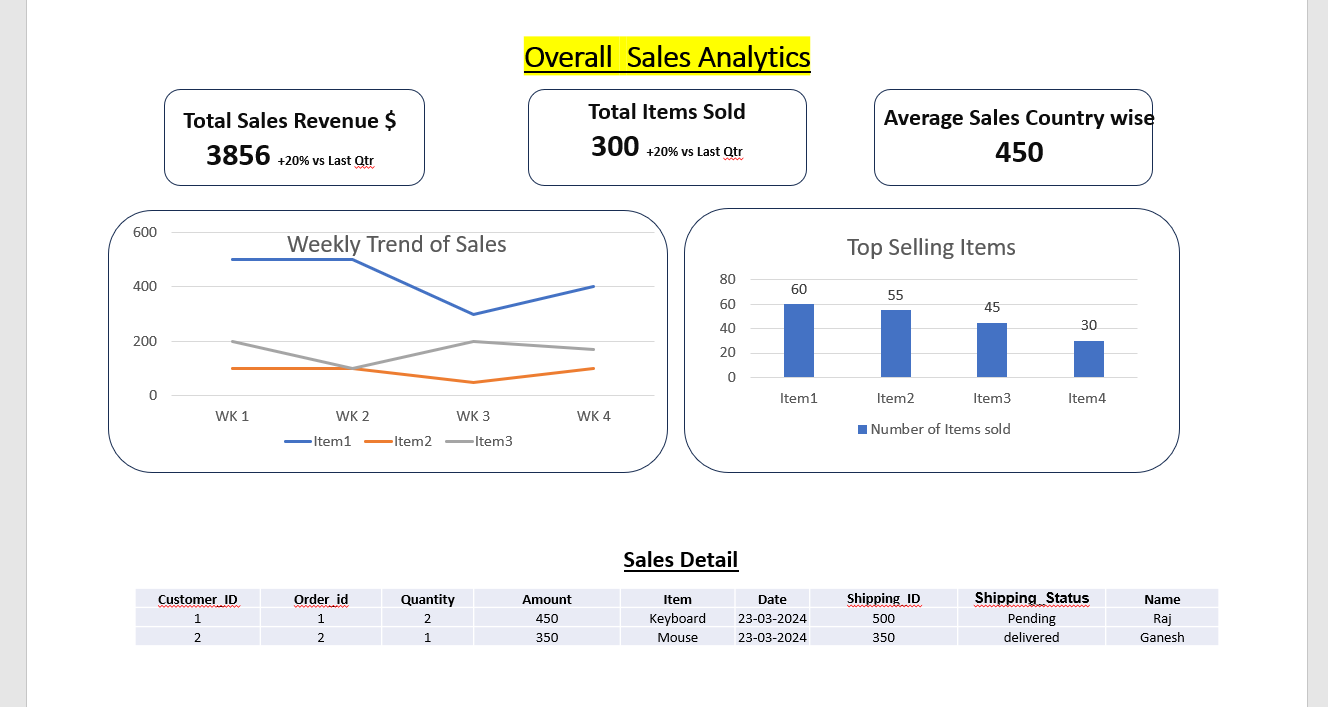
TAP Data Flow: Data flow from Level 1 to Transaction Analysis Process.

MPPP Data Flow: Data flow from Level 1 to Maximum Product Purchased Process.

Most Purchased Product Data Flow: Data flow from Level 1 to Most Purchased Product Process.

CAP Data Flow: Data flow from Level 1 to Country Analysis Process.

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



**What will be your insights to other peer teams of Data Engineers, Data Scientists and other technical and non-technical stakeholders?**

**Data Engineers:**

Ensure that the data pipelines are robust and efficient in extracting, transforming, and loading (ETL) data from the customer, order, and shipping tables.

Optimize database performance to handle complex queries efficiently, especially for aggregations and joins required for reporting.

Collaborate on data modeling and schema design to support reporting requirements effectively.

Implement data quality checks and monitoring to ensure the accuracy and completeness of the data used for reporting.

**Data Scientists:**

Provide clean and well-structured datasets for predictive analytics and machine learning models.

Collaborate on feature engineering and selection to identify relevant predictors for customer behavior analysis, sales forecasting, or product recommendation models.

Share insights from the BI reports to validate assumptions and refine models based on real-world trends and patterns observed in the data.

**Technical Stakeholders (Developers, Architects):**

Discuss any technical challenges or limitations encountered during data extraction, processing, or reporting.

Provide guidance on implementing best practices for data management, security, and governance.

Seek feedback on tools and technologies used for BI reporting and explore opportunities for integration or automation to enhance efficiency.

**Non-Technical Stakeholders (Management, Business Analysts):**

Present actionable insights and recommendations derived from the BI reports to drive strategic decision-making.

Provide training and support to enable stakeholders to interpret and leverage BI reports effectively.

Solicit feedback on reporting requirements and iterate on dashboards or visualizations to better meet business needs.

**Cross-Functional Collaboration:**

Foster collaboration and knowledge sharing across teams to leverage expertise and insights from different domains.

Encourage cross-functional discussions to identify synergies and opportunities for innovation in data-driven initiatives.

Establish regular communication channels and meetings to align on project goals, priorities, and progress updates.