Sales Data Analysis & Modelling Documentation –

**Project Name**: Sales Insights  
**Tools Used**: Snowflake

**Step 1 : Data Ingestion** –

**Load Method** –

1. Files staged in Snowflake Internal Stage
2. Data loaded via COPY INTO command into respective raw tables.

**Raw Tables Created –**

| **Table Name** | **Description** |
| --- | --- |
| **Customer\_Raw** | **Customer personal information** |
| **Order\_Raw** | **Product sales transactions** |
| **Json\_Shipping\_Raw** | **Semi structured shipping data from json file** |
| **Shipping\_Raw** | **Shipping status of orders** |

**Customer\_Raw table structure -**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Customer\_ID | INT | Unique ID for each customer |
| First | STRING | Customer's first name |
| Last | STRING | Customer's last name |
| Age | INT | Age of the customer |
| Country | STRING | Country of residence |

**Order\_Raw Table Structure**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Order\_ID | INT | Unique ID for each order |
| Item | STRING | Item Name |
| amount | FLOAT | Total amount spent |
| Customer\_ID | INT | Foreign key to customer |

**Json\_Shipping\_Raw Structure**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Data | VARIANT | Semi- structured Data from Json file |

**Shipping\_Raw Structure**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Shipping\_ID | INT | Unique shipping record |
| Status | STRING | Shipping status |
| Customer\_ID | INT | Foreign key to customer |

**Step 2: Data Quality Checks –**

**Goal: Validate accuracy, completeness, and reliability of ingested data**

**2.1 Completeness Checks –**

1) Null Values per Table

2) Missing foreign keys (Referential Integrity)

**2.2 Accuracy Checks –**

1) To check no negative order amounts in Order table

2) Special characters in names in Customer table

**2.3 Reliability Checks –**

1) Count of duplicate records in Shipping table

2) Validate default changing fields (Age is not constant for customers)

**Findings Summary Table -**

| **Check** | **Issue Found** | **Affected Table** |
| --- | --- | --- |
| Missing customer IDs | No | Order\_Raw, Shipping\_Raw |
| Missing foreign keys | No | Order\_Raw, Shipping\_Raw |
| Validate default changing fields | Yes | Customer\_Raw |
| Negative order amounts | No | Order\_Raw |
| Special characters in names | Yes | Customer\_Raw |
| Duplicate shipping entries | Yes (based on customer\_id) | Shipping\_Raw |

**Suggestions to improve Raw Data Quality for below points –**

1. Instead of age value in Customers Raw Data, date of birth value will give more reliable value of Age.
2. In Shipping data, Only Customer ID is given but one customer can have multiple Orders in Order table so tracking of shipping status specific to order may not be reliable. It’s better to add Order ID instead of /along with customer ID in shipping data.

**Step 3: Domain Model & Data Story for Data Engineers**

**Data Modeling Assumptions and Transformation:**

1. All necessary columns required for building the dimensional and fact tables are present in the source files.
2. Every Order\_ID has a valid Customer\_ID in Customer\_Raw. If not, such records are excluded from the fact table.
3. Each customer may appear in the shipping table more than once, but only the latest shipping status per customer is considered for reporting.
4. Names (first\_name, last\_name) are cleaned to remove special characters — it's assumed that these are not critical to customer identity.
5. If no matching shipping record is found for a customer in the Shipping\_Raw, the order is marked with a default "Yet to Ship" status.
6. One customer can place many orders (1:M), but each order is placed by exactly one customer.

**Target Schema: Star Schema**

**3.1 Dimension Tables**

| Table | Description |
| --- | --- |
| Dim\_Customer | Cleaned customer data with derived columns |
| Dim\_Shipping | Latest shipping status per customer |

**Dim\_Customer table structure:**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Customer\_ID | INT | From Customer\_Raw, Primary Key |
| First\_Name | STRING | From Customer\_Raw |
| Last\_Name | STRING | From Customer\_Raw |
| Age | INT | From Customer\_Raw |
| Country | STRING | From Customer\_Raw |

**Dim\_Shipping table structure:**

| Column Name | Data Type | Description |
| --- | --- | --- |
| Customer\_ID | INT | Foreign key to Dim\_Customer |
| Latest\_Shipping\_Id | INT | Latest shipping record ID |
| Latest\_Status | STRING | Status of most recent shipment (Pending/Shipped) |

**3.2 Fact Table**

| Table | Description |
| --- | --- |
| Fact\_Order | Contains transactional data + enriched dimensions |

**Fact\_Order table structure -**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Order\_ID | INT | From Order\_Raw |
| Status | STRING | Latest status from Dim\_Shipping, else ‘Yet to Ship’ |
| Customer\_ID | INT | Foreign key to customer |
| Amount | FLOAT | From Order\_Raw |
| Item | STRING | From Order\_Raw |

**Entity Relationship Diagram -**

Dim\_Customer (1) ────< Fact\_Order ──── (1) Dim\_Shipping

1. Dim\_Customer table has one to many relationships with Fact\_Order table
2. Dim\_Shipping has one to one relationship with Fact\_Order table

**Step 4: QA Checklist (Test Scenarios)**

| **Test Case** | **Expected Result** |
| --- | --- |
| No NULLs in primary keys | True |
| Every fact row has a matching customer | True |
| Special characters removed from names | True |
| Correct shipping status default logic | "Yet to Ship" where no match |
|  |  |
|  |  |