DATA LOADING (ETL)

Dataset:

Order_ID	Customer_ID	Sales_Amount	Order_Date
O101	C001	4500	12-01-2024
O102	C002	Null	15-01-2024
O103	C003	3200	2024/01/18
O101	C001	4500	12-01-2024
O104	C004	Three Thousand	20-01-2024
O105	C005	5100	25-01-2024

Q1. Data Understanding

Identify all data quality issues present in the dataset that can cause problems during data loading.

Answer: All Data Quality Issues:

1. **Duplicate Primary Key** → O101 appears twice.
2. **Missing value** → sales_Amount is NULL for O102
3. **Invalid data type** → "Three Thousand" in Sales_Amount (O104).
4. **Inconsistent date formats** → 12-01-2024 and 2024/01/18.
5. Potential numeric type inconsistency in Sales_Amount coulmn.

Q2. Primary Key Validation

Assume Order_ID is the **Primary Key.**

a) Is the dataset violating the Primary Key rule?
b) Which record(s) cause this violation?

Answer: Primary Key Validation:

a) Yes, Primary Key rule is violated
b) Record causing violation:
- O101 (appears twice)

Primary keys must be unique. Databases are not emotionally flexible about this.

Q3. Missing Value Analysis

Which column(s) contain **missing values?**

a) List the affected records.
b) Explain why loading these records without handling missing values is risky.

Answer: Column with missing value:

- Sales_Amount

Affected record:

- O102

Why risky?

- Total sales calculations become incorrect
- BI reports show wrong revenue
- Aggregations like SUM may ignore or miscalculate values

Q4. Data Type Validation

Identify records where **Sales_Amount** violates expected data type rules.

a) Which record(s) will fail numeric validation?
b) What would happen if this dataset is loaded into a SQL table with Sales_Amount as DECIMAL?

Answer:

a) Records failing numeric validation:

- O102 (NULL)
- O104 ('Three Thousand")

b) If loaded as DECIMAL

- Load may fail
- Or invalid values may be converted to NULL
- Causes incorrect KPI totals

Q5. Date Format Consistency

The Order_Date column has multiple formats.

a) List all date formats present in the dataset
b) Why is this a problem during data loading?

Answer: Date formats present:

- DD-MM-YYYY → 12-01-2024
- YYYY/MM/DD → 2024/01/18

Problem:

- Database may misinterpret dates
- Sorting and filtering will break
- Can cause load errors

Q6. Load Readiness Decision

Based on the dataset condition:

a) Should this dataset be loaded directly into the database? (Yes/No)
b) Justify your answer with at **least three reasons.**

Answer:

a) **Should it be loaded directly?**
   No
b) **Reasons:**
   1. Duplicate primary key
   2. Missing sales value
   3. Invalid numeric data
   4. Inconsistent date formats

This dataset is not "load-ready." It's 'cry-for-help-ready."

Q7. Pre-Load Validation Checklist

List the exact **pre-load validation checks** you would perform on this dataset before loading.

Answer: Pre-Load Validation Checklist:

1. Check primary key uniqueness
2. Validate non-null constraints
3. Validate numeric fields
4. Standardize date format
5. Remove duplicates
6. Check data type compatibility
7. Validate referential integrity (Customer_ID)

Q8. Cleaning Strategy

Describe the **step-by-step cleaning actions** required to make this dataset load-ready.

Answer: Cleaning Strategy (Step-by-Step):

1. Remove duplicate record (O101)
2. Convert “Three Thousand” → 3000
3. Handle NULL in O102
- Either fill with the correct value or remove record
4. Convert all dates to one format (YYYY-MM-DD recommended)
5. Revalidate data types
6. Perform final validation check

Q9. Loading Strategy Selection

Assume this dataset represents **daily sales data.**

a) Should a **Full Load** or **Incremental Load** be used?
b) Justify your choice.

Answer:

a) Use Incremental Load
b) Why?
- It represents daily sales data
- Only new daily records should be added
- Faster and more efficient
- Reduces system load

Full load daily would be dramatic and unnecessary. Like rewriting your entire life story evry morning.

Q10. BI Impact Scenario

Assume this dataset was loaded **without cleaning** and connected to a BI dashboard.

a) What incorrect results might appear in Total Sales KPI?
b) Which records specifically would cause misleading insights?
c) Why would BI tools not detect these issues automatically?

Answer:

a) Incorrect Total Sales KPI:
- Duplicate O101 doubles revenue
- NULL sales ignored
- “Three Thousand” may not be counted

b) Misleading records:
- O101 (duplicate)
- O102 (NULL)
- O104 (Invalid text value)

c) Why BI tools don’t detect automatically?

- BI tools trust the loaded data
- They aggregate, they don’t validate business logic
- Garbage in → very confident garbage out