# Data_Loading(ETL)

**Q1. Data Understanding Identify all data quality issues present in the dataset that can cause problems during data loading.**
The dataset has duplicates, NULL values, inconsistent data types, inconsistent date formats, and primary key violations, which can cause problems during data loading.

**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?**
a) Yes, The dataset violates the primary key constraint (Unique), one of the record appears twice.
b) Order_ID = O101 appears more than once

**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**
a) The sales_Amount column contains missing data for Order_ID O102, and
b) Loading this record without proper handling can lead to errors, data inconsistency, and incorrect analysis.

**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 as DECIMAL?**
a) Order_ID O104 will fail numeric validation because Sales_Amount contains text ("Three Thousand").
b) It will cause data type conversion errors when loading into a DECIMAL column.

**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?**
a) The Order_Date column contains multiple date formats - (DD-MM-YYYY and YYYY/MM/DD)
b) Having multiple date formats in the Order_Date column can cause data load failures, incorrect date values, and unreliable analysis, so dates must be standardized before loading.

**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**
a) No, the dataset should not be loaded directly into the database.
b) Due to 1) Duplicate keys 2) Missing values 3) Incorrect data types and 4) Inconsistent date formats, the dataset must be cleaned and validated before loading into the database.

**Q7. Pre-Load Validation Checklist List the exact pre-load validation checks you would perform on this dataset before loading.**
Before loading the dataset into the database, the following pre-load validation checks should be performed:

1. Primary Key Validation
   Ensure Order_ID is unique and non-NULL (check for duplicates like O101).

2. Missing Value Check
   Identify and handle NULL values in critical columns such as Sales_Amount.
3. Data Type Validation
   Verify that Sales_Amount contains only numeric values and no text entries.
4. Date Format Consistency Check
   Standardize all Order_Date values into a single date format (e.g., YYYY-MM-DD).
5. Duplicate Record Detection
   Detect and remove exact duplicate rows to avoid double counting.
6. Domain / Range Validation
   Ensure Sales_Amount falls within a valid business range (e.g., non-negative values).

**Q8. Cleaning Strategy Describe the step-by-step cleaning actions required to make this dataset load-ready.**

Cleaning Strategy (Step-by-Step)

1. Remove Duplicate Records
   - Identify duplicate Order_ID values.
   - Keep only one valid record for Order_ID = O101.
   - Remove exact duplicate rows.
2. Handle Missing Values
   - Identify NULL values in Sales_Amount (Order_ID O102).
   - Impute with a valid value (mean/median), or
   - Replace with 0 if business-approved, or
   - Reject the record for correction.
3. Fix Invalid Data Types
   - Convert text values in Sales_Amount (e.g., "Three Thousand") into numeric form (3000).
4. Standardize Date Formats
   - Convert all Order_Date values to a single format (e.g., YYYY-MM-DD).
   - Validate date correctness.
5. Validate Primary Key Constraint
   - Re-check Order_ID for uniqueness and non-NULL values after cleaning.
6. Apply Domain and Range Checks
   - Ensure Sales_Amount is non-negative and within a valid business range.
7. Final Schema Validation
   - Verify column names, data types, and order match the target database table.

**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.**

a) Incremental Load should be used.

b) Justification

1. Daily Data Updates:
   Since sales data is generated every day, only new records need to be loaded.
2. Improved Performance:
   Incremental loading processes only new or changed data, reducing load time and system resource usage.
3. Avoids Data Duplication: Prevents reloading previously processed sales records.

**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?**

a) incorrect results:

1) Inflated Total Sales: Duplicate records (Order_ID O101) would be counted twice, increasing total sales incorrectly.

2) Underestimated Total Sales: Records with NULL Sales_Amount (O102) would be ignored in aggregation, lowering total sales.

3) Missing or Incorrect Totals: Text values like "Three Thousand" (O104) may be excluded or cause calculation errors.

b)

| Issue | Order_ID | Reason |
|---|---|---|
| Duplicate record | O101 | Double-counted sales |
| Missing value | O102 | Sales not included |
| Invalid data type | O104 | Excluded or miscalculated |
| Inconsistent date | 2024/01/18 | Incorrect time-based analysis |

c) BI Tools Assume Clean Data:

BI dashboards rely on the data source and do not enforce data quality rules.

Limited Validation:

BI tools aggregate values but do not validate business logic or constraints.

Silent Failures:

Invalid or NULL values are often ignored without warnings.