

Data Loading (ETL) Assignment Question

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: The dataset has the following data quality issues:

1. Duplicate records
 - Order_ID = O101 appears twice.

2. Missing values

- Sales_Amount is NULL for Order 0102.

3. Invalid data type

- Sales_Amount contains text ("Three Thousand") instead of numeric.

4. Inconsistent date formats

- Multiple date formats used in Order_Date.

5. Potential primary key violation

- Duplicate Order_ID breaks uniqueness.

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: a) Is the dataset violating the Primary Key rule?

Yes, it violates the primary key constraint.

b) Which record(s) cause this violation?

- Order_ID = 0101 appears twice with identical data.

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: Which column(s) contain missing values?

- Sales_Amount

a) Affected records

- Order 0102 (Customer C002)

b) Why is loading without handling missing values risky?

- KPIs like Total Sales, Average Sales, and Revenue Trends will be inaccurate.
- NULLs may be treated as:
 - \emptyset → under-reporting revenue
 - excluded → inconsistent aggregation
- Can cause calculation errors in BI tools and SQL queries.

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

- Order 0104 → Sales_Amount = "Three Thousand"
- b) What happens if loaded into SQL with DECIMAL type?
- The record will:

- Fail to load (type conversion error), or
 - Be converted to NULL, depending on DB settings
- This leads to data loss or incorrect 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: a) Date formats present

1. 12-01-2024 → DD-MM-YYYY

2. 15-01-2024 → DD-MM-YYYY

3. 2024/01/18 → YYYY/MM/DD

4. 20-01-2024 → DD-MM-YYYY

5. $25-01-2024 \rightarrow DD-MM-YYYY$

b) Why is this a problem during data loading?

- ETL tools may misinterpret dates
- Sorting, filtering, and time-based analysis break
- Some records may fail parsing or shift dates incorrectly

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 this dataset be loaded directly?

No

b) Justification (at least 3 reasons)

1. Primary key violation (0101 duplicate)

2. Invalid numeric values in $Sales_Amount$

3. Missing sales data

4. Inconsistent date formats
5. High risk of incorrect BI reporting

Q7. Pre-Load Validation Checklist

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

Answer: Before loading, perform:

1. Primary key uniqueness check
2. Null value check (especially **Sales_Amount**)
3. Data type validation for numeric fields
4. Date format standardization check
5. Duplicate row detection
6. Referential integrity checks (if applicable)
7. Range checks for **Sales_Amount** (no negatives, no text)

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

Answer: 1. Remove or resolve duplicate Order_ID records

2. Handle missing Sales_Amount

- Investigate source
- Impute or flag

3. Convert textual sales values

- "Three Thousand" → 3000

4. Standardize date formats

- Convert all to YYYY-MM-DD

5. Validate numeric columns

- Enforce DECIMAL precision

6. Re-run validation checks

7. Load cleaned data into staging

8. Promote to production 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.**

Answer: Assume daily sales data.

a) Full Load or Incremental Load?

Incremental Load

b) Justification

- Daily sales are appended regularly
- Full load is inefficient and risky
- Incremental loads:

- Improve performance
- Reduce data duplication
- Support late-arriving data handling

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

- Total sales may be:
 - Under-reported (NULL or failed records)
 - Over-reported (duplicate 0101)
- b) Records causing misleading insights

1. Duplicate 0101 → double-counted revenue

2. 0102 → missing sales

3. 0104 → invalid sales value ignored or NULL

c) Why BI tools don't auto-detect these issues

- BI tools assume data is clean
- They don't enforce:
 - Primary key constraints
 - Data type rules
- They aggregate whatever is loaded — garbage in, garbage out