**Name:** **Priyanshu Sahu**
**Batch:** **Data Analytics Nov live Batch**

# Assignment 21: Data Quality and Validation in ETL

Question 1: Define Data Quality in the context of ETL pipelines. Why is it more than just data cleaning?

Answer:

**Data Quality** in ETL means ensuring that data is:

- Accurate
- Complete
- Consistent
- Valid
- Reliable
- Timely

It ensures that the data loaded into the data warehouse is trustworthy for reporting and decision-making.

**Name:** Priyanshu Sahu
**Batch:** Data Analytics Nov live Batch

## Why it is more than data cleaning?

Data Cleaning only:

- Removes null values
- Removes duplicates
- Fixes formatting

But Data Quality includes:

- Applying business rules
- Checking referential integrity
- Validating data types
- Ensuring accuracy
- Maintaining consistency across systems

So, Data Cleaning is just one part of Data Quality.

**Name:** Priyanshu Sahu
**Batch:** Data Analytics Nov live Batch

Question 2: Explain why poor data quality leads to misleading dashboards and incorrect decisions.

Answer:

If data is incorrect:

Example from Sales_Transactions dataset:

- Txn_ID 206 → Txn_Amount = Null
- Txn_ID 205 → Quantity = Null
- Duplicate transactions (201 & 208)

**Problems:**

- Total sales will be incorrect
- Revenue calculation wrong
- Wrong business decisions
- Fake performance insights

Example:

If duplicates are not removed → Sales will show ₹8000 instead of ₹4000 for Rahul Mehta.

Question 3: What is duplicate data? Explain three causes in ETL pipelines.

Answer:

**Duplicate Data:**

Duplicate data means same record stored multiple times.

Example:

Txn_ID 201 and 208 are same:

C101 + P11 + 2025-12-01 + 4000

**Three Causes in ETL:**

1. **Multiple Data Loads**
   - Same file loaded twice.
2. **Missing Primary Key**

- No unique constraint on business key.

3. **System Integration Issues**
    - Data coming from multiple sources without matching rules.

Question 4: Differentiate between exact, partial, and fuzzy duplicates

Answer:

**Exact Duplicate:**

Records that are **100% identical in every column**.

| Customer_ID | Name | Email | City |
|---|---|---|---|
| 101 | Rahul Verma | rahul@gmail.com | Kanpur |
| 101 | Rahul Verma | rahul@gmail.com | Kanpur |

Both rows are completely same.

**Name:** Priyanshu Sahu
**Batch:** Data Analytics Nov live Batch

## Partial Duplicates:

Records that are **not fully identical**, but some key fields are same.

| Customer_ID | Name | Email | City |
|---|---|---|---|
| 101 | Rahul Verma | rahul@gmail.com | Kanpur |
| 101 | Rahul V. | rahul@gmail.com | Lucknow |

Email same, but name slightly different and city different

## Fuzzy Duplicates:

Records that look similar but are **not exactly same** (spelling differences, formatting differences).

| Name | Phone |
|---|---|
| Rahul Verma | 9876543210 |

**Name:** Priyanshu Sahu
**Batch:** Data Analytics Nov live Batch

| Name | Phone |
|---|---|
| RAHUL VERMA | 9876543210 |
| Rahul Vearma | 9876543210 |

Question 5: Why should data validation be performed during transformation rather than after loading?

Answer:

**Reasons:**

1. Saves storage
2. Prevents bad data entering warehouse
3. Improves performance
4. Reduces rework

If we load bad data first → then clean later → it wastes time & resources.

So best practice:

Extract → Transform (validate) → Load clean data

Question 6: Explain how business rules help in validating data accuracy. Give an example.

Answer:

Business Rule = Condition defined by company.

**Example:**

Rule 1:
Quantity cannot be NULL

But in Sales_Transactions dataset:

Txn_ID 205 → Quantity = NULL (Invalid)

Rule 2:
Txn_Amount must not be NULL

Txn_ID 206 → Txn_Amount = NULL

**Name:** Priyanshu Sahu
**Batch:** Data Analytics Nov live Batch

Rule 3:

Customer_Name should not be 'N/A'

Txn_ID 206 → Customer_Name = N/A

These rules ensure accurate and meaningful data

Question 7: Write an SQL query on Sales_Transactions  to list all duplicate keys and their counts using the business key (Customer_ID + Product_ID + Txn_Date + Txn_Amount )

Answer:

```
SELECT Customer_ID, Product_ID, Txn_Date, Txn_Amount, COUNT(*) AS Duplicate_Count

FROM Sales_Transactions

GROUP BY  Customer_ID,Product_ID, Txn_Date, Txn_Amount HAVING COUNT(*) > 1;
```

**Name:** Priyanshu Sahu
**Batch:** Data Analytics Nov live Batch

Question 8: Enforcing Referential Integrity

Assume the following Customers_Maste table:

Identify Sales_Transactions.Customer_ID values that violate referential integrity when joined with Customers_Master and write a query to detect such violations.

Answer:

**Enforcing Referential Integrity**

Customers_Master contains:
C101, C102, C103, C104

In Sales_Transactions:
C105 and C106 are present but not in Customers_Master.

Therefore, C105 and C106 violate referential integrity.

SQL to detect violations:

```
SELECT DISTINCT s.Customer_ID
```

```
FROM Sales_Transactions s
LEFT JOIN Customers_Master c
ON s.Customer_ID = c.CustomerID
WHERE c.CustomerID IS NULL;
```

This query identifies invalid Customer_ID values.

**ETL Implementation Explanation:**

In ETL, referential integrity is enforced during the transformation stage using a lookup between Sales_Transactions and Customers_Master.

- If Customer_ID exists → record is loaded into Fact table.
- If Customer_ID does not exist → record is rejected or stored in an error table.

This ensures only valid customer records are loaded into the data warehouse.

**Name:** Priyanshu Sahu
**Batch:** Data Analytics Nov live Batch