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 the context of ETL (Extract, Transform, Load) means

Ensuring that data is accurate, complete, consistent, reliable, valid, and ready for analysis before it is loaded into the target system (like a data warehouse)

Data quality means the data is trustworthy and usable for business decisions.

Accuracy – Data correctly represents real-world values
Completeness – No important fields are missing
Consistency – Same data is uniform across systems
Validity – Data follows rules and formats
Uniqueness – No duplicate records

Why Data Quality is More Than Just Data Cleaning

Many people think data quality = removing null values or duplicates.

Data cleaning fixes problems.
Data quality ensures systems and processes prevent errors in the first place.

Example:
Cleaning → Removing duplicate customers
Data Quality → Creating validation rules so duplicates never enter the system

It Involves Business Rules
Data quality checks business logic.

example
Order amount cannot be negative
Date of joining cannot be in the future

QUESTION 2

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

ANSWER

If the input data is wrong, incomplete, or inconsistent, the dashboard will confidently show wrong insights — and that’s dangerous.

Incorrect KPIs

Revenue may be underreported (missing data)

If sales data has missing or duplicate records:

Revenue may be overreported (duplicate data)

Wrong Trends & Forecasts

Wrong values

Incorrect dates

Missing months

Poor Customer Insights

Wrong age

Incorrect gender

Duplicates

Missing segments

Delayed or Outdated Data (Timeliness Issue)

Dashboard shows last month’s data

But decision is being made for today.

QUESTION 3 **What is duplicate data? Explain three causes in ETL pipelines**

ANSWER The same record appears more than once in a dataset when it should only exist once.

Exact duplicate (all fields same)
Partial duplicate (same customer but slightly different spelling, email, etc.)

Improper Incremental Load Logic
If incremental load is not configured correctly, the same data may be loaded again during the next ETL run.

Multiple Source Systems
Website forms
Sometimes data comes from:
Billing system
CRM system

example	“Rahul Mehta”
	“R. Mehta”
	“Rahul M.”

Missing Primary Key or Constraints in Target Table

If the target table does not have:
Primary key
Unique constraint

QUESTION 4

Differentiate between exact, partial, and fuzzy duplicates

ANSWER

1) Exact Duplicates

Records where all column values are exactly the same.

ID	Name	City
101	mitesh shimpi	Mumbai
101	mitesh shimpi	Mumbai

2) Partial Duplicates

Records where some key fields are same, but other fields differ.

ID	Name	City
101	mitesh shimpi	Mumbai
101	mitesh shimpi	Pune

3) Fuzzy Duplicates

Records that look similar but are not exactly the same due to spelling or formatting differences.

Exact duplicates are records where all fields are identical. Partial duplicates share key fields but differ in other attributes. Fuzzy duplicates are records that represent the same entity but contain slight spelling or formatting differences, requiring advanced matching techniques to identify.

QUESTION 5

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

ANSWER

In ETL, the transformation stage is where data is cleaned, standardized, and validated before it reaches the target system

Prevents Bad Data from Entering the System

If validation is done after loading, incorrect data is already stored in:

Data warehouse

BI dashboards

Reports

Reduces Rework and Maintenance Cost

Fixing data after loading means:

Updating tables

Re-running ETL jobs

Refreshing dashboards

protects Business Trust

Invalid sales amount loaded

Dashboard shows inflated revenue

Management makes a decision

Allows Proper Error Handling

During transformation, you can:

Reject invalid records

Send bad records to an error table

Log validation failures

QUESTION 6

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

ANSWER

How Business Rules Help in Validating Data Accuracy

Business rules are logical conditions based on how the business actually operates.

They help ensure that data is not just technically correct, but logically correct according to business reality.

What is Data Accuracy?

The data correctly represents real-world facts.

Business rules act like a reality check for the data.

How Business Rules Improve Accuracy

Validate Logical Conditions

Order amount must be greater than 0

Discount cannot exceed 50%

Date of joining cannot be in the future

Ensure Data Matches Business Policy

Loan can only be approved if credit score > 700

Customer age must be 18 or above

Salary must be within company pay scale range

Detect Outliers or Abnormal Values

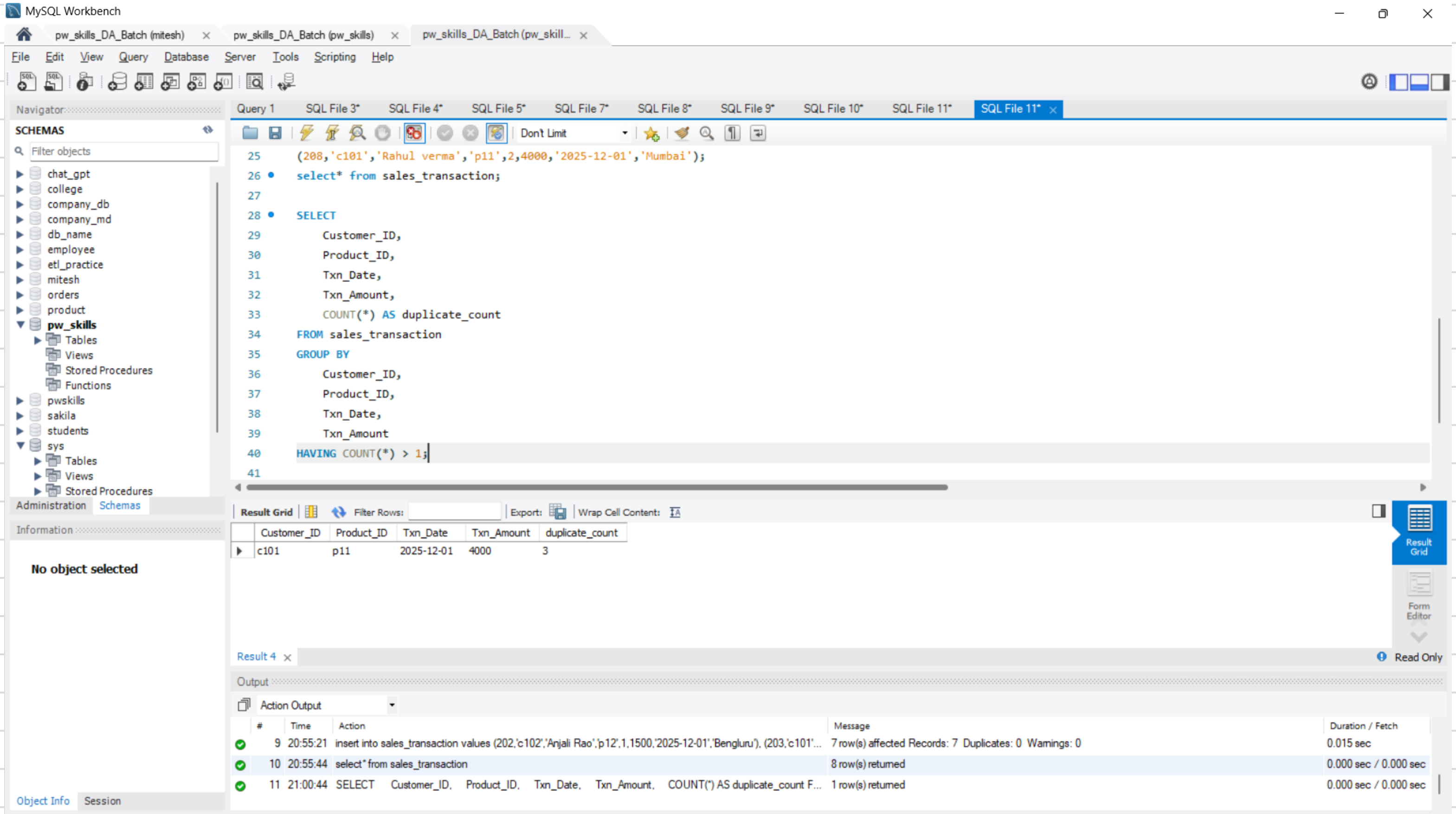
If normal daily sales are ₹1 lakh and suddenly one entry shows ₹50 crore,
a business rule can flag it for review.

Order_ID	Product_Price	Quantity	Total_Amount
101	500	2	200

QUESTION 7

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

ANSWER



```
(208,'c101','Rahul verma','p11',2,4000,'2025-12-01','Mumbai');
select* from sales_transaction;

SELECT
    Customer_ID,
    Product_ID,
    Txn_Date,
    Txn_Amount,
    COUNT(*) AS duplicate_count
FROM sales_transaction
GROUP BY
    Customer_ID,
    Product_ID,
    Txn_Date,
    Txn_Amount
HAVING COUNT(*) > 1;
```

Customer_ID	Product_ID	Txn_Date	Txn_Amount	duplicate_count
c101	p11	2025-12-01	4000	3

```
#   Time      Action                                                                                       Message                                                   Duration / Fetch
9   20:55:21  insert into sales_transaction values (202,'c102','Anjali Rao','p12',1,1500,'2025-12-01','Bengluru'), (203,'c101'...   7 row(s) affected Records: 7  Duplicates: 0  Warnings: 0   0.015 sec
10  20:55:44  select* from sales_transaction                                                               8 row(s) returned                                         0.000 sec / 0.000 sec
11  21:00:44  SELECT    Customer_ID,    Product_ID,    Txn_Date,    Txn_Amount,    COUNT(*) AS duplicate_count F...   1 row(s) returned                                         0.000 sec / 0.000 sec
```

QUESTION 8

Enforcing Referential Integrity

Assume the following customers_master table:

ANSWER

sorry i did not get this question answer