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?

➔ In an ETL pipeline, Data Quality refers to the degree to which data is "fit for use" for its intended purpose (e.g., reporting, machine learning, or operations). It is measured by dimensions like accuracy, completeness, consistency, timeliness, and validity.

Cleaning is just fixing mistakes (like deleting a typo). **Data Quality** is the whole process of making sure the data is reliable, arrives on time, and follow the rules from start to finish.

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

➔ If a manager looks at a dashboard that says sales are up 100%, they might hire 10 new people. But if that 100% was actually a mistake caused by data being counted twice, the company will lose money.

The Rule: "Garbage In, Garbage Out." If you put "garbage" data into a system, you will get "garbage" advice out of it.

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

➔ This is when the same information appears more than once in your system (like having the same customer listed twice).

Three simple causes:

1. **System Glitches:** A computer crashes halfway through sending data and sends the same batch again when it restarts.
2. **Bad Math (Joins):** When connecting two lists (like "Customers" and "Orders"), a mistake in the code can cause one customer to be copied for every order they ever made.
3. **Multiple Sources:** You get a list of names from Facebook and a list from Google. If "John Doe" is on both, he becomes a duplicate in your system.

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

➔ **Exact:** The rows are 100% identical. (Copy-paste).

Partial: Most of the info is the same (like the Name and Email), but one small thing is different (like the date they signed up).

Fuzzy: They look like the same person, but the spelling is slightly off. (Example: "Jon Smith" vs "John Smith").

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

➔ Performing validation during the Transformation phase (often called "Pre-load validation") is best practice for several reasons:

- **Integrity of the Warehouse:** It prevents "dirty" data from ever reaching the final production tables, keeping the downstream environment clean.
- **Cost Efficiency:** It is much cheaper and faster to catch an error in flight than to find it later and have to run complex "delete and reload" operations on massive production tables.
- **Immediate Alerting:** If validation happens during transformation, the pipeline can be paused or the bad records can be diverted to a "quarantine" table immediately, alerting engineers before users see the data.

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

➔ **Business rules** are formal statements that define or constrain aspects of a business's operations. In the context of data, they act as the "logic layer" to ensure information is logically sound, consistent, and accurate.

They help validate data accuracy by:

- Preventing Invalid Entries: Setting constraints (e.g., "Quantity cannot be negative").
- Ensuring Completeness: Identifying mandatory fields (e.g., "Every transaction must have a `Txn_Date`").
- Maintaining Consistency: Cross-referencing related data points (e.g., "If `Product_ID` is P11, the `Txn Amount` must be a multiple of the unit price").

Example from dataset:

Look at Txn_ID 205. The `Quantity` is Null, but there is a `Txn Amount` of 2500. A business rule stating "Quantity must be a positive integer and cannot be Null" would flag this record as inaccurate or incomplete.