Section 4: Quick Conceptual Check

1. When would you choose Spark over Pandas?

You choose Spark when working with large-scale datasets that don't fit into memory. Spark runs computations in a distributed manner, making it ideal for big data processing across multiple machines.

2. What is a broadcast join in Spark and when should it be used?

A broadcast join sends a small table to all worker nodes to avoid data shuffling. Use it when one table is very small compared to the other - it significantly speeds up the join.

3. Difference between LEFT JOIN and INNER JOIN with an example.

- LEFT JOIN: Returns all rows from the left table and matched rows from the right.
- INNER JOIN: Returns only rows that match in both tables.

Example: LEFT JOIN customers with orders includes customers who haven't ordered; INNER JOIN would exclude them.

4. What is a DAG in Airflow and how do you monitor it?

A DAG (Directed Acyclic Graph) is a workflow structure defining task execution order. In Airflow, you monitor it using the Airflow web UI - you can view DAG status, task logs, retries, and schedule.

5. How would you explain "partitioning" and "shuffling" to a beginner?

- Partitioning: Divides data into smaller chunks across nodes for faster parallel processing.
- Shuffling: Data is moved across partitions (expensive) during operations like groupBy, join, or distinct.