**Sakina banu**

**Assignment 1: SQL Query – Retrieve Suppliers and Products**

**Question:**Write a query to retrieve the names of all suppliers and their corresponding products, including suppliers that have no products listed.

**Tables Involved:**

* Suppliers
* Products

**SQL Query:**

SELECT

Suppliers.SupplierName,

Products.ProductName

FROM

Suppliers

LEFT JOIN

Products ON Suppliers.SupplierID = Products.SupplierID;

**Explanation:**

* The query uses a LEFT JOIN to ensure all suppliers are included in the result, even if they do not have any products listed.
* If a supplier has no products, the ProductName will appear as NULL.
* This approach ensures completeness in reporting supplier data alongside available product information.

A screenshot of a computer

AI-generated content may be incorrect.

**Assignment 2**

**Interview based questions**

1. **What is Data Shuffling?**  
   *“Data shuffling is the process of redistributing data across partitions or nodes during operations like joins, groupBy, or repartition. It can be expensive because it involves disk and network I/O, so minimizing it is key for performance.”*
2. **What is Data Skew?**  
   *“Data skew occurs when certain keys have significantly more records than others, leading to uneven partitioning and slow tasks during distributed processing. I usually handle it using salting or broadcast joins.”*
3. **What is Salting in PySpark?**  
   *“Salting is a technique to mitigate data skew by adding a random prefix or suffix to keys, which helps distribute the data more evenly across partitions during transformations like joins or aggregations.”*
4. **What is an RDD in PySpark?**  
   *“RDD stands for Resilient Distributed Dataset—it's the low-level abstraction in Spark for distributed collections. It supports fault tolerance and in-memory computation, though in most cases, I now prefer using DataFrames for better optimization.”*
5. **What are Broadcast Joins?**  
   *“A broadcast join sends a small lookup table to all worker nodes, allowing Spark to perform map-side joins and avoid costly shuffling. I often use it when joining a large dataset with a small dimension table.”*
6. **What is Adaptive Query Execution (AQE)?**  
   *“AQE is a Spark feature that dynamically adjusts query plans based on runtime stats—like switching join strategies or optimizing skewed joins. It’s really useful in unpredictable data environments.”*
7. **What does reduceByKey() do in PySpark?**  
   *“reduceByKey() groups data by key and applies a reduction function. It performs local aggregation before the shuffle, making it more efficient than groupByKey for large datasets.”*
8. **What is Coalesce in PySpark and when do you use it?**  
   *“Coalesce reduces the number of partitions without a full shuffle. I use it when writing output to minimize the number of small files or when optimizing performance after filtering or sampling.”*