Dharma Devi 19th May HW

Assignment - 1

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

Tables - Suppliers, Products

SELECT Suppliers.name AS supplier\_name, Products.name AS product\_name FROM Suppliers LEFT JOIN Products ON Suppliers.supplier\_id = Products.supplier\_id;

Assignment -2

**Data Shuffling**:  
 Movement of data across partitions or nodes, typically during operations like join or groupBy.

* **Data Skew**:  
   Uneven distribution of data that causes some partitions to have significantly more data than others, leading to performance bottlenecks.
* **Salting**:  
   A technique used to mitigate data skew by adding randomness (a “salt” key) to skewed keys before shuffling.
* **RDD in PySpark**:  
   Resilient Distributed Dataset — the core low-level distributed data abstraction in PySpark offering fault-tolerant parallel operations.
* **Broadcast Joins**:  
   A join strategy where a small dataset is sent to all nodes to avoid shuffling the larger dataset.
* **Adaptive Query Execution (AQE) in PySpark**:  
   A Spark feature that dynamically optimizes query plans at runtime based on actual data characteristics.
* **reduceByKey()**:  
   A transformation that merges values for each key using an associative and commutative reduce function across partitions.
* **Coalesce in PySpark**:  
   A method to reduce the number of partitions by merging them without a full shuffle, often used for optimizing output.