Q10

1. **Broadcast Join:**

Used when one of the datasets is small enough to fit in memory on each worker node. Spark broadcasts the smaller dataset across all nodes, so each partition of the larger dataset can join with the smaller dataset locally without any need for shuffling data across the network

Use Case:

Small dataset joining with a large dataset

2. **Shuffle Merge Join:**

Default join type in Spark for large datasets. Data gets shuffled across the cluster to group matching keys together → all rows with the same key are on the same partition. After that the datasets on each partition are merged to complete the join

Use Case:

Joining two large datasets → None of them can be broadcasted

3. **Sort Merge Join:**

Both datasets are first sorted by the join key. After that, the datasets are merged. It is efficient when both datasets are already sorted or can be efficiently sorted

Use Case:

- When both datasets are already sorted
- Optimal for large datasets → sorting reduces the complexity of the join