Assignment

The main intention of this assignment is to help you in understanding when to use the different join types, when the different join strategies are applied and when the Spark AQE feature comes into picture.

1. Upload 2 datasets of your choice one big and one small(less than 10mb)
   or
   choose the existing sample datasets in the path /public/trendytech of the external lab

2. Create 2 Dataframes, one on each file and perform a join using Dataframes approach as well as spark SQL style. Do check the sparkUI to see the join strategy used.

3. Disable the broadcast join by changing the threshold and perform a join again. Now check the Join strategy used.

4. Give a hint for shuffle hash join and invoke the join again and check the spark UI for the join strategy used.

5. By default the AQE was disabled, enable the AQE and also disable the broadcast join. Now perform a join again and see if there is a change in the number of shuffle partitions.

6. You need to explain left outer and semi join with relevant use cases. Demonstrate it by running in the notebook.