spark-shell

BROADCAST HASH JOIN:

spark.conf.get("spark.sql.autoBroadcastJoinThreshold")

spark.conf.get("spark.sql.join.preferSortMergeJoin")

val x = Seq(1,2,3,4,5,6,7,8,9)

val y = Seq(10,2,30,4,50,6,70,8,90)

val df1 = x.toDF("id1")

val df2 = y.toDF("id2")

df1.join(df2,$"id1"===$"id2").show

SHUFFLEHASHJOIN:

spark.conf.set("spark.sql.autoBroadcastJoinThreshold",-1)

spark.conf.set("spark.sql.join.preferSortMergeJoin","false")

spark.conf.get("spark.sql.autoBroadcastJoinThreshold")

spark.conf.get("spark.sql.join.preferSortMergeJoin")

df1.join(df2,$"id1">=$"id2").show

CARTESIANJOIN:

spark.conf.set("spark.sql.autoBroadcastJoinThreshold",2)

spark.conf.get("spark.sql.autoBroadcastJoinThreshold")

spark.conf.get("spark.sql.join.preferSortMergeJoin")

df1.join(df2,$"id1">=$"id2").show