SparkR::sql, DBI::dbExecute, and sparklyr::spark_sql

Xiande Yang

2025-08-02

SparkR, DBI, and sparklyr

Databricks deprecated SparkR using which I wrote more than 150k lines of code. This really made me upset. However, I had to change to sparklyr.

sparkR::sql, sparklyr::spark_sql, dplyr::tbl(sc, dplyr::sql()), and DBI::dbExecute.

sparkR::sql() can be used for parallel tasks which is essential for big data.

sparklyr::spark_sql() is the equivalent one of sparkR::sql().

dplyr::tbl(sc, dplyr::sql()) is good for select operation in SQL or dplyr equivalent operations excellent for pipe operation %>% but it does not work for create or replace table etc.

DBI::dbExecute() works equivalent to sparkR:sql() or sparklyr::spark_sql() however, it is for sequential tasks but not for parallel work. So, when I ran code parallelly in Databricks, it has the error of race condition.

Hence, finally, I changed all these code use sparklyr::spark_sql.

Bye bye SparkR!! I love you but somebody else does not like you and make you a legacy package.