spark sql data bigger than node memory when coalesce(1)

Asked 2 years, 10 months ago Modified 2 years, 10 months ago Viewed 77 times



I'm working on spark 1.6.1

1 I have a dataframe that is distributed and is for sure bigger than any nodes i have in my cluster.



What will happen if i bring all in a node?



df.coalesce(1)

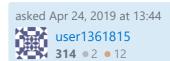
Will the job fail?

Thanks

apache-spark apache-spark-sql coalesce

Share Follow





1 Short answer: Yes. – eliasah Apr 24, 2019 at 14:30

To add to eliasah's answer, it will fail with OutOfMemory Exception. There are many ways to circumvent the need of coalesce. What are you trying to achieve by bringing all the data to the driver? – Sai Apr 24, 2019 at 16:37

1 Answer

Active Oldest Votes



It will fail for sure as data will not fit in memory. If you want to return single file as a output, you can merge HDFS files later using HDFS getMerge.



You can use utility to merge multiple files into one file from below mentioned git project https://github.com/gopal-tiwari/hdfs-file-merge



Share Follow

answered Apr 24, 2019 at 16:26

Gopal Tiwari