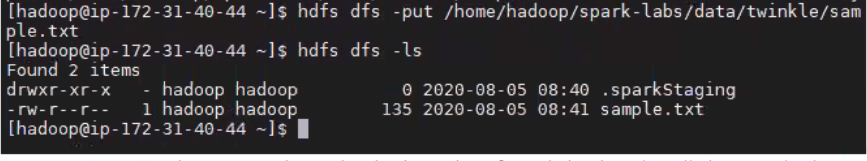
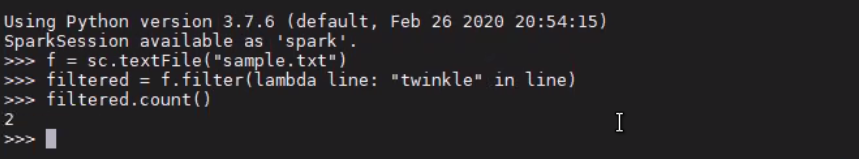
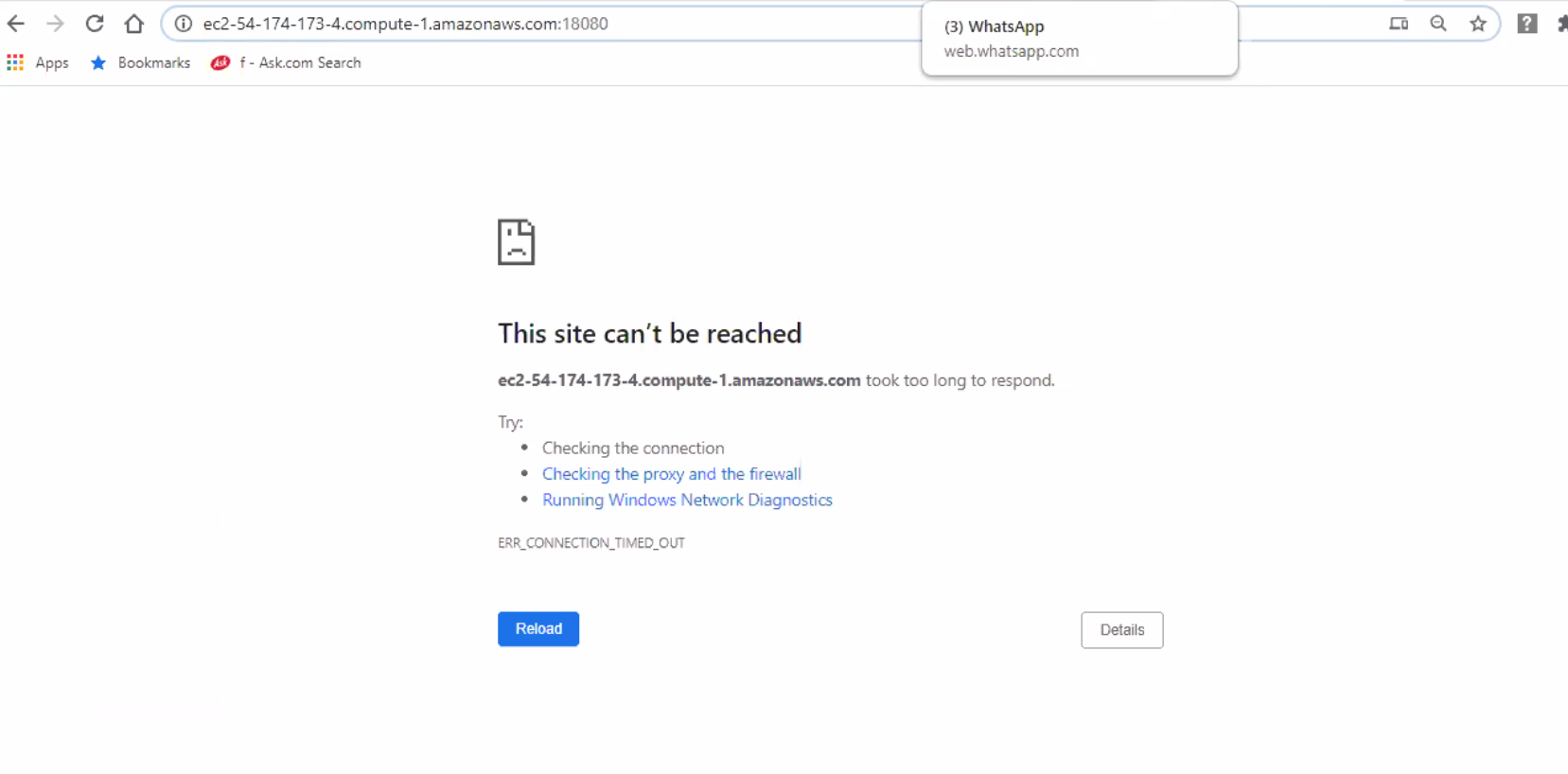
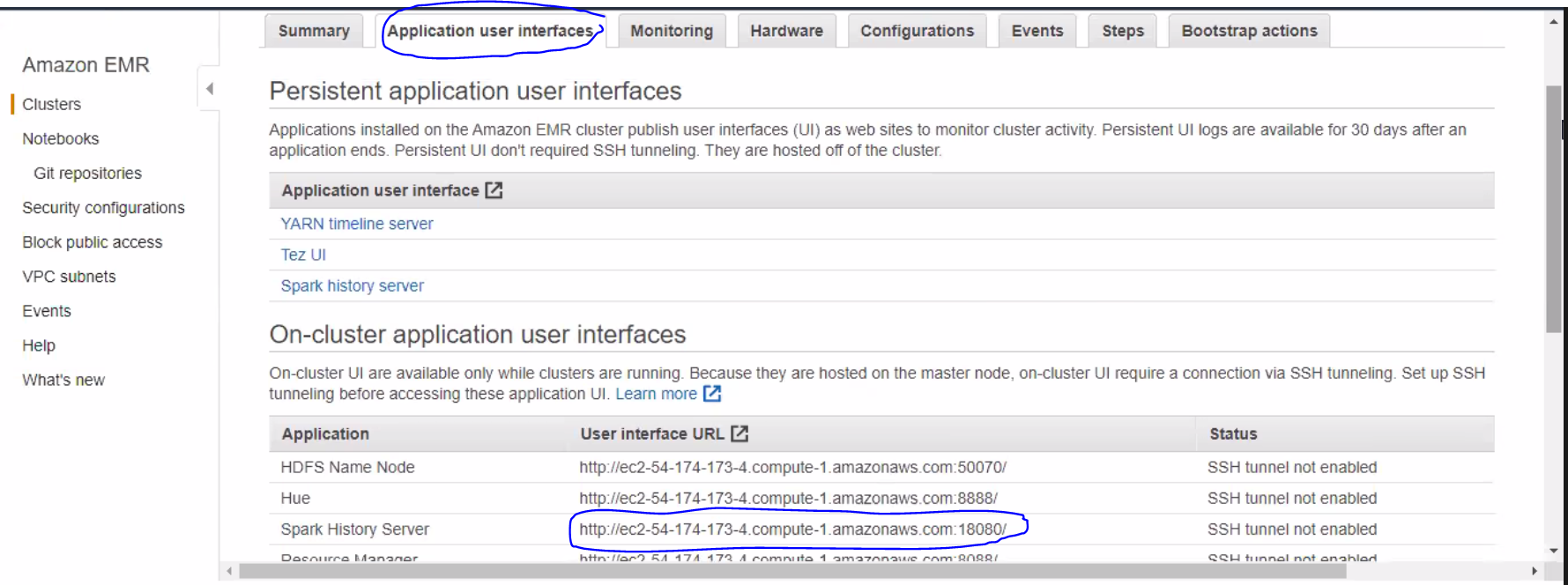


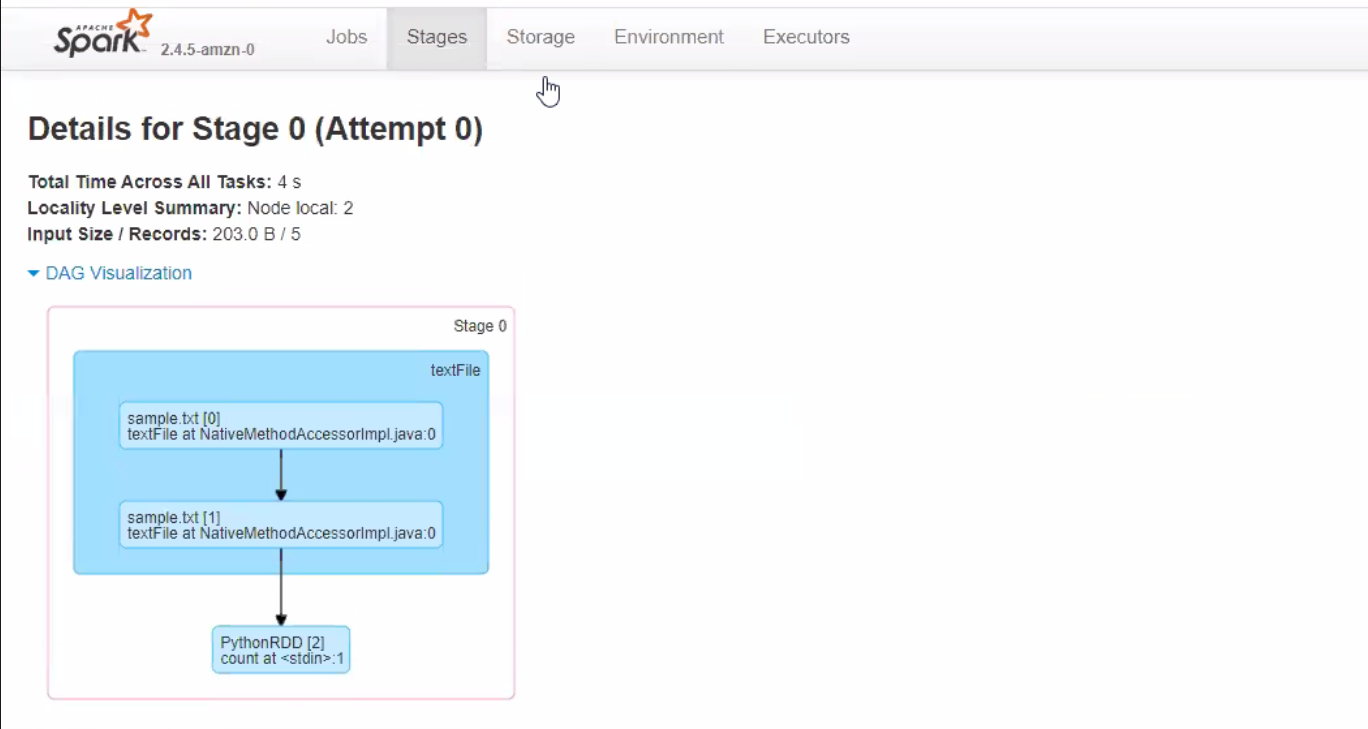
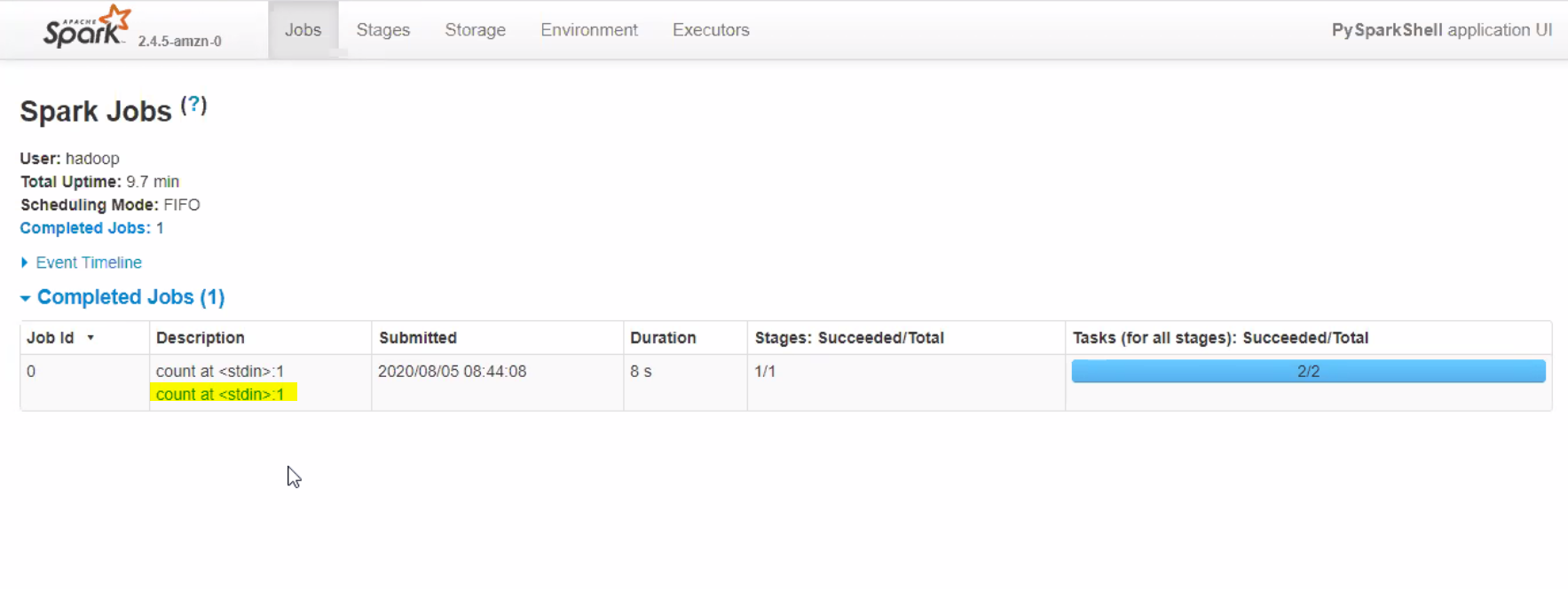
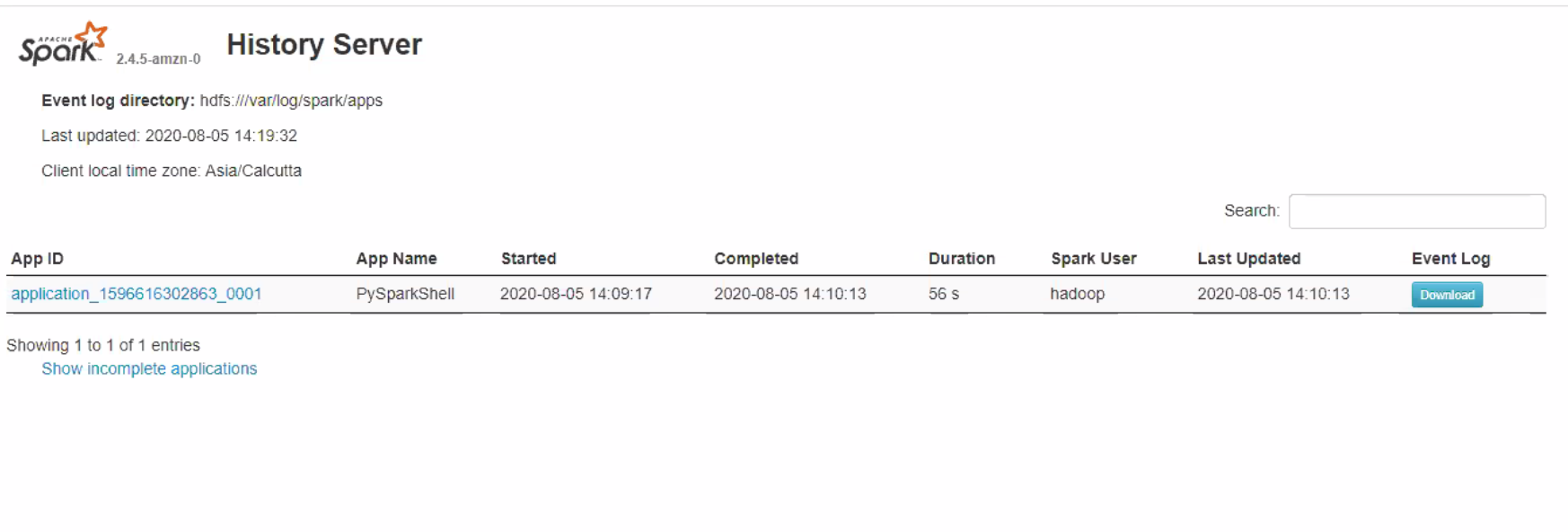
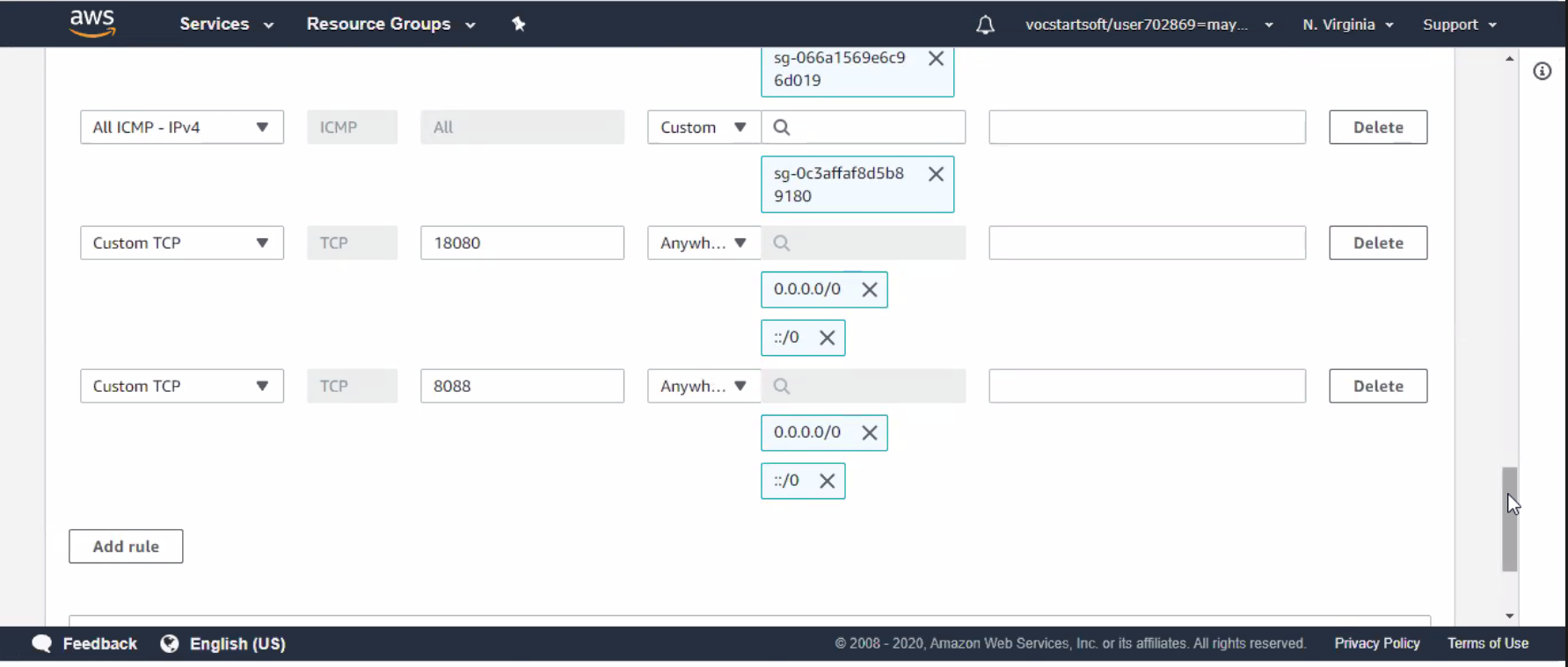
Put the file sample.txt in HDFS

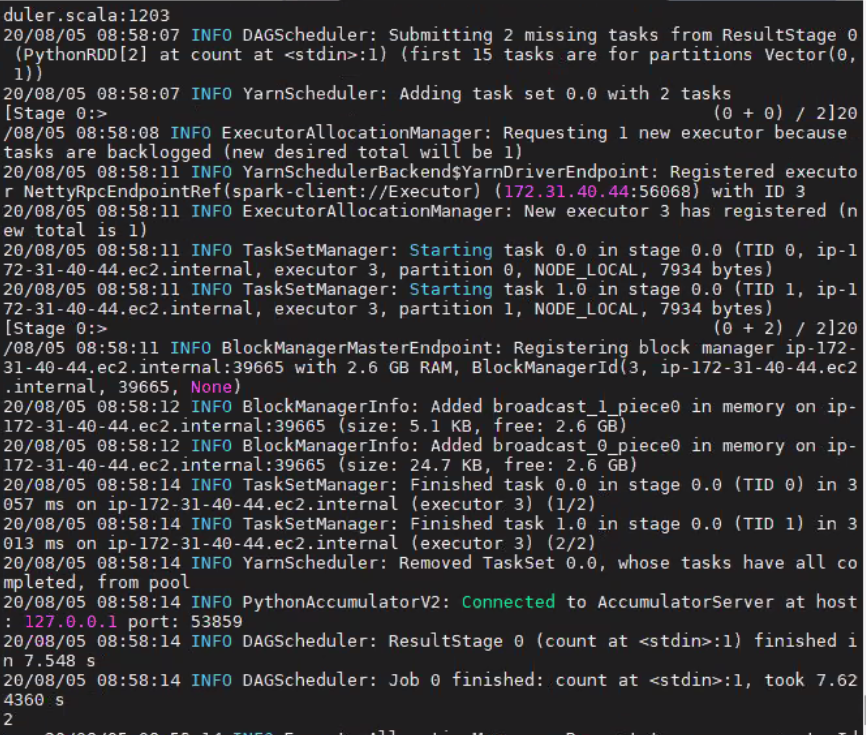
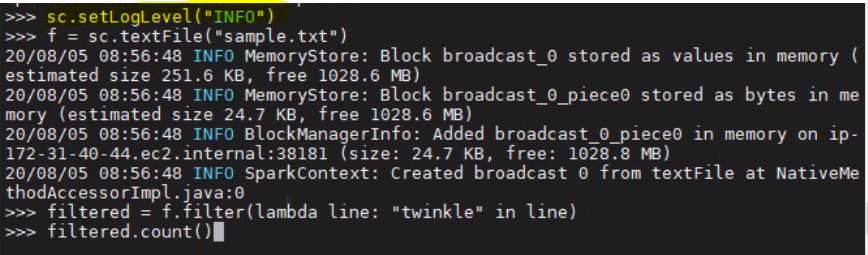


Then enter pyspark

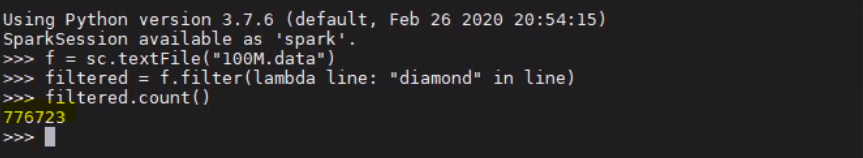
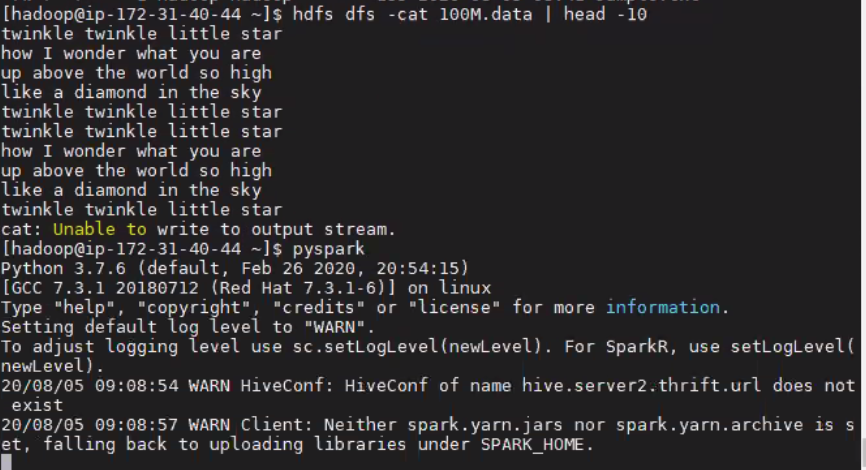
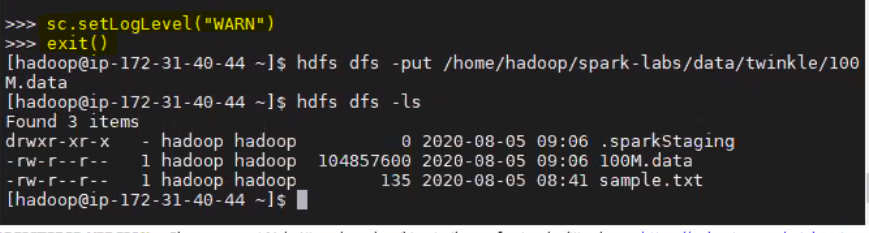


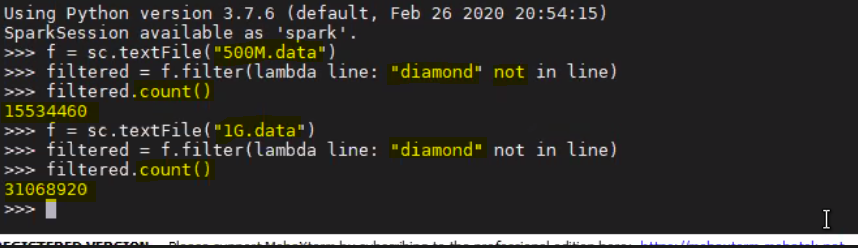
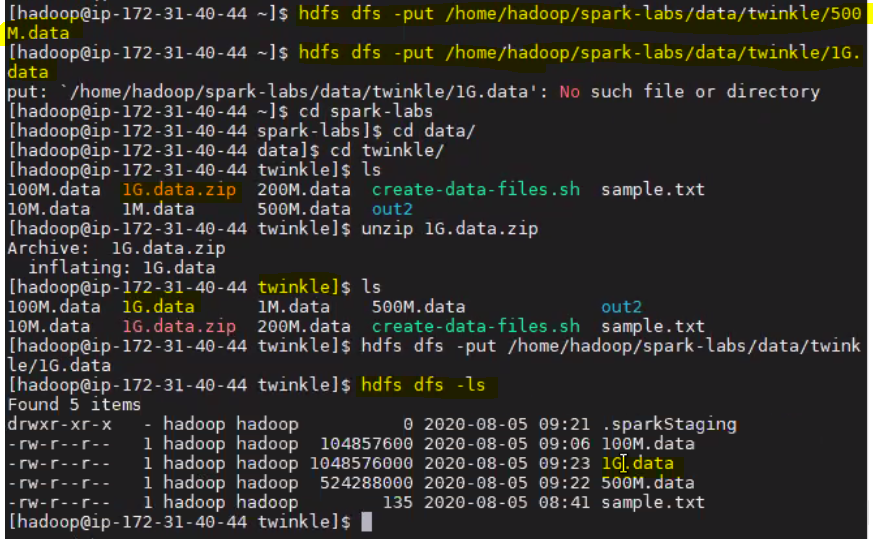
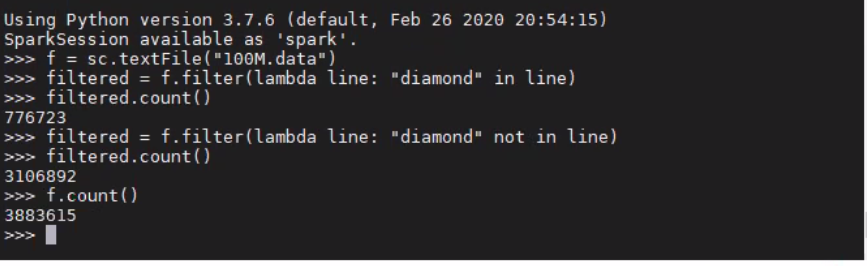


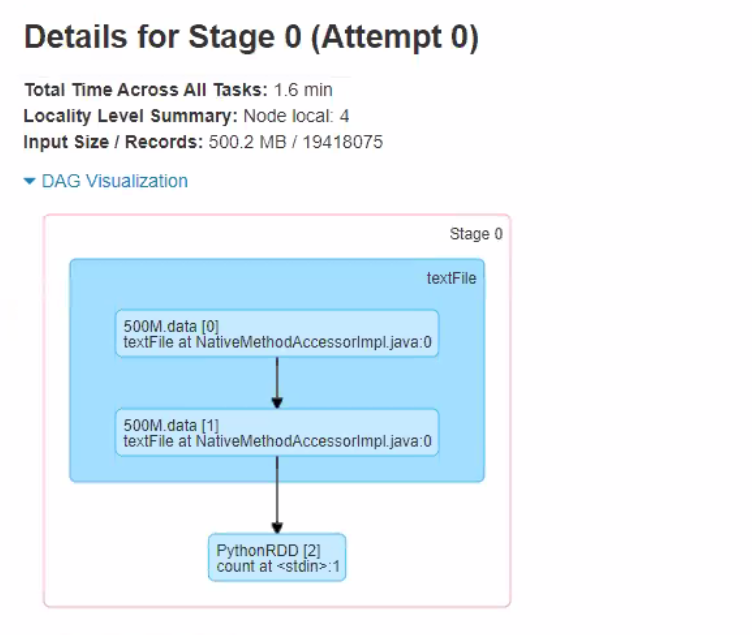


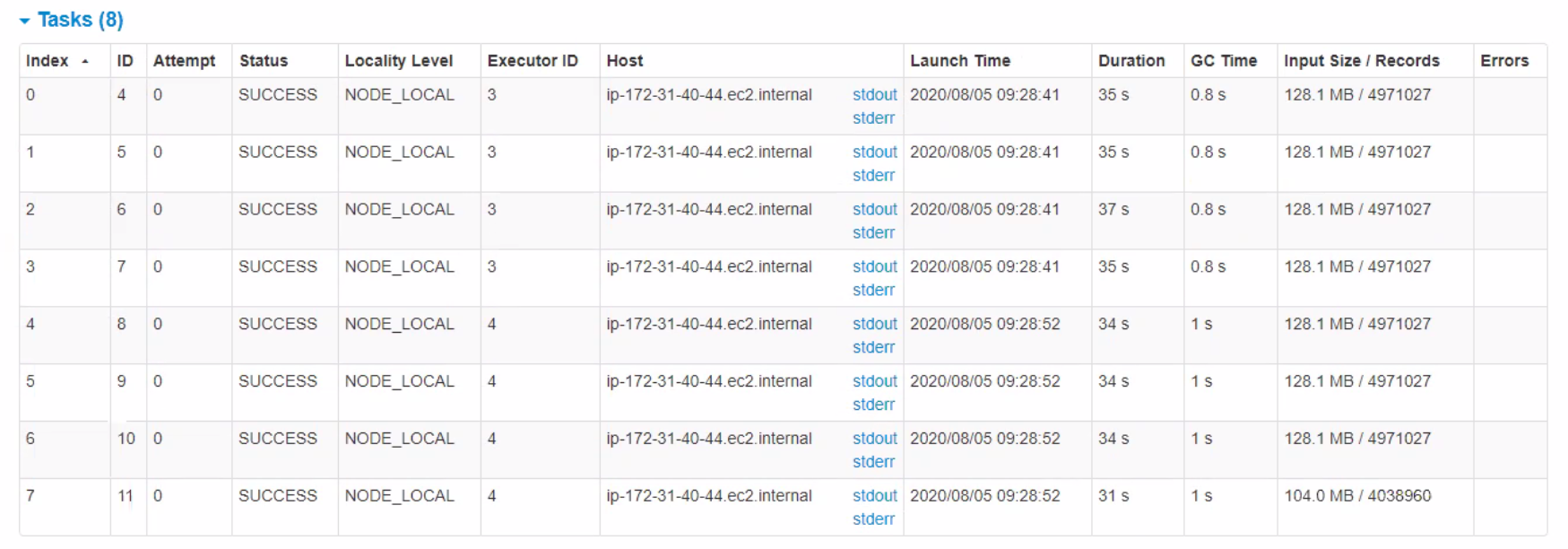
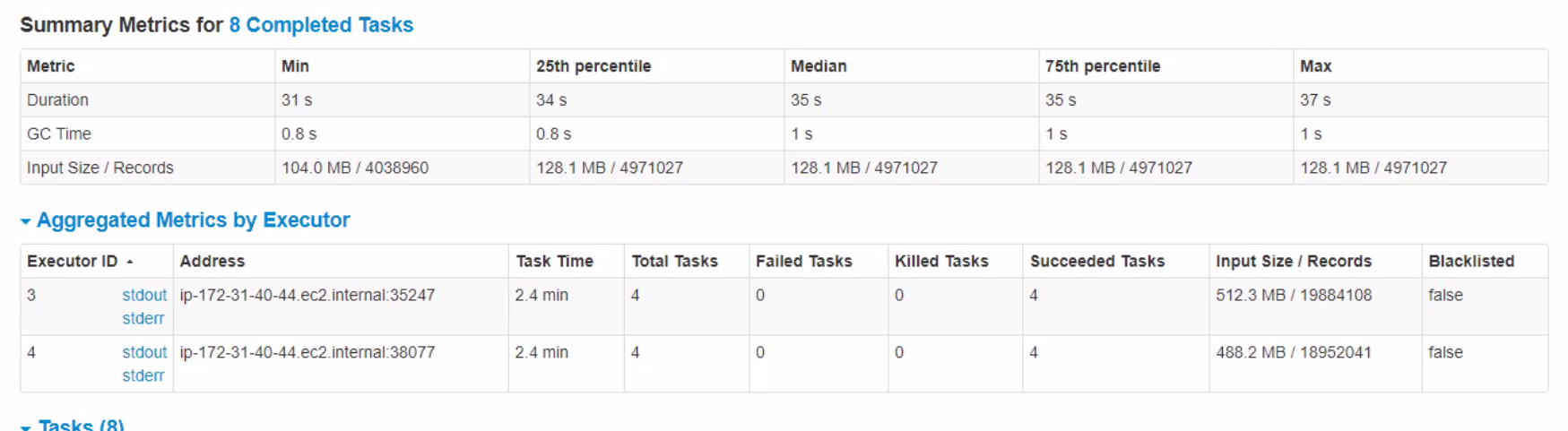
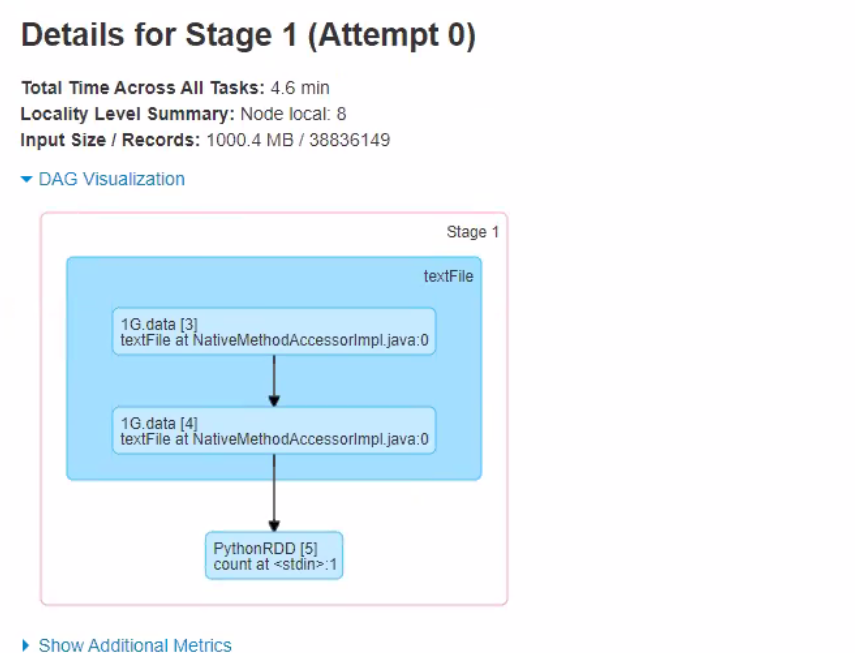
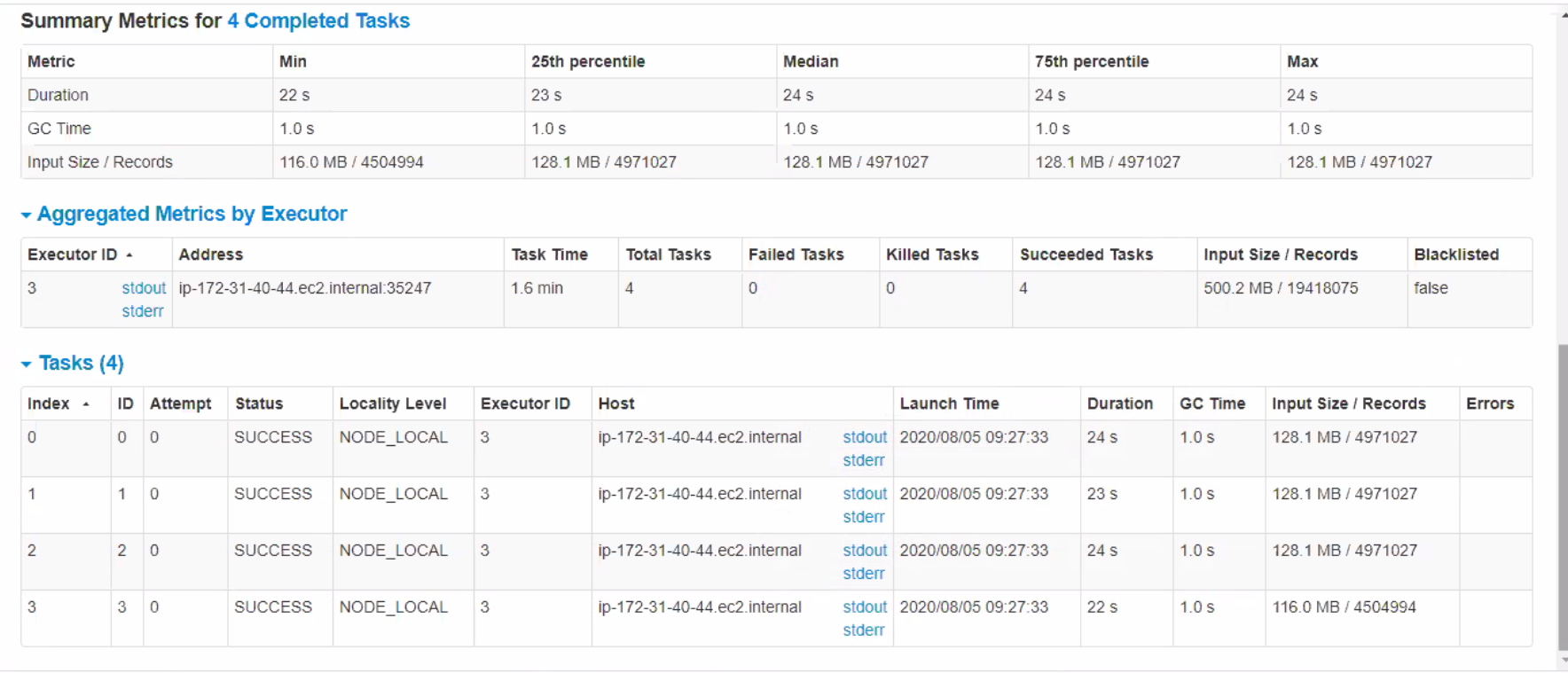


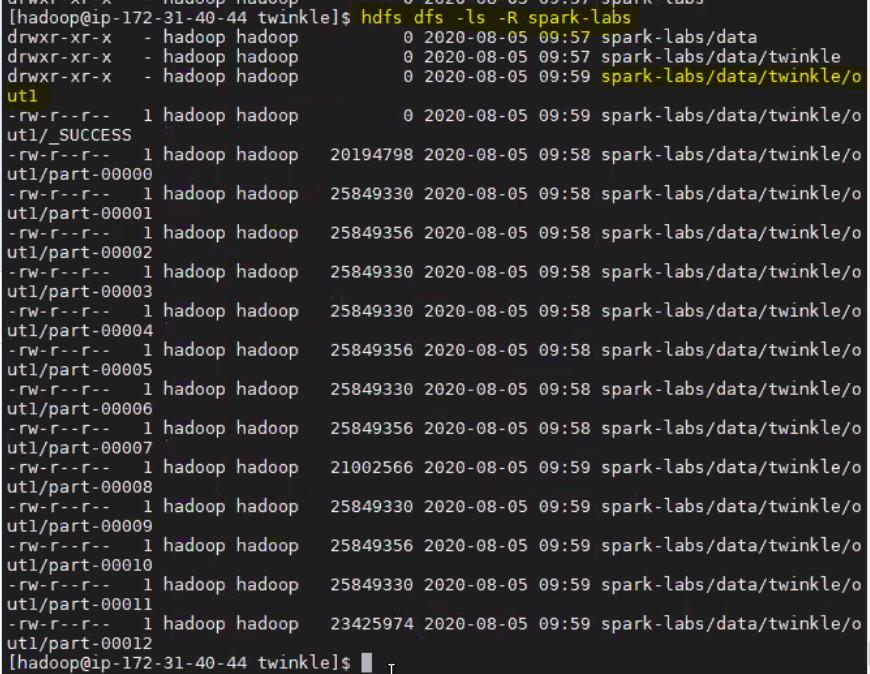
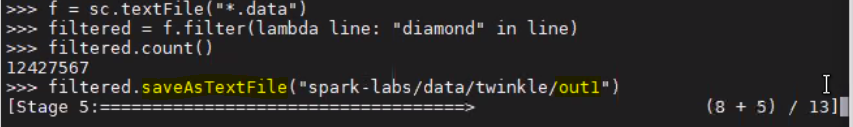
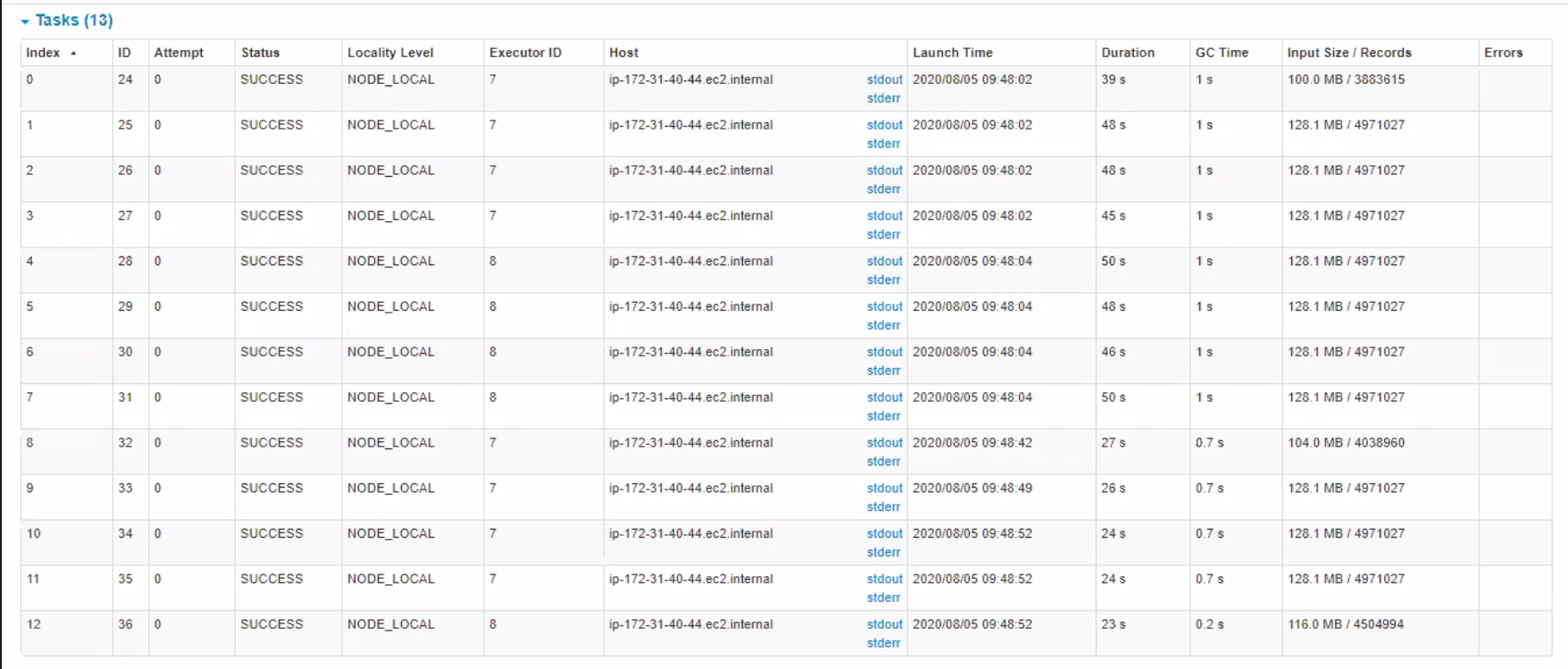
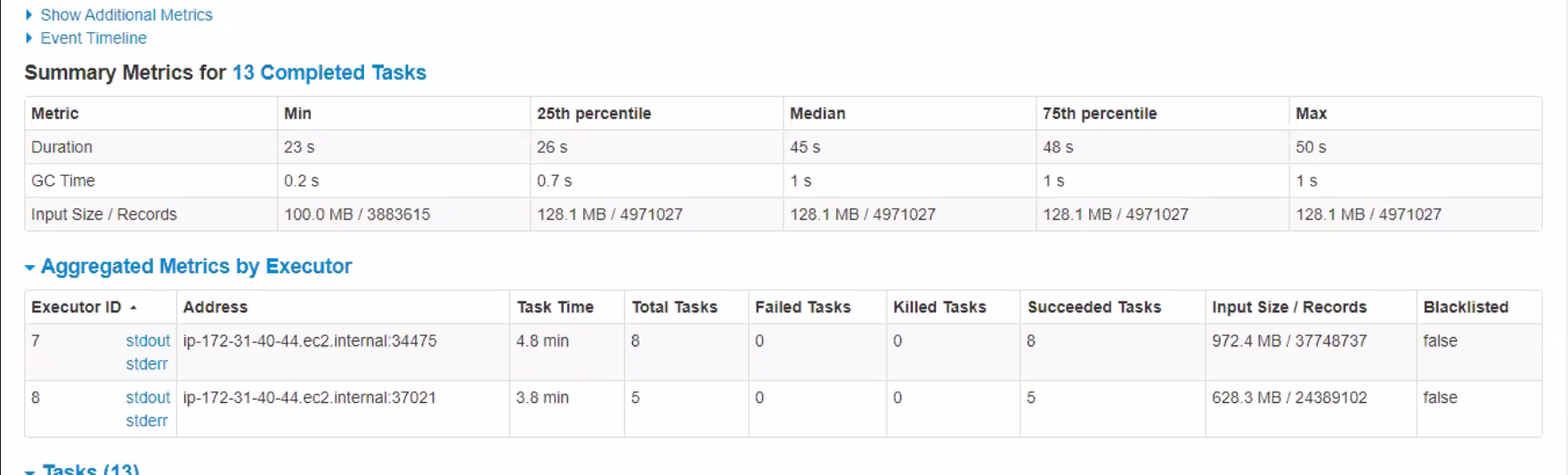
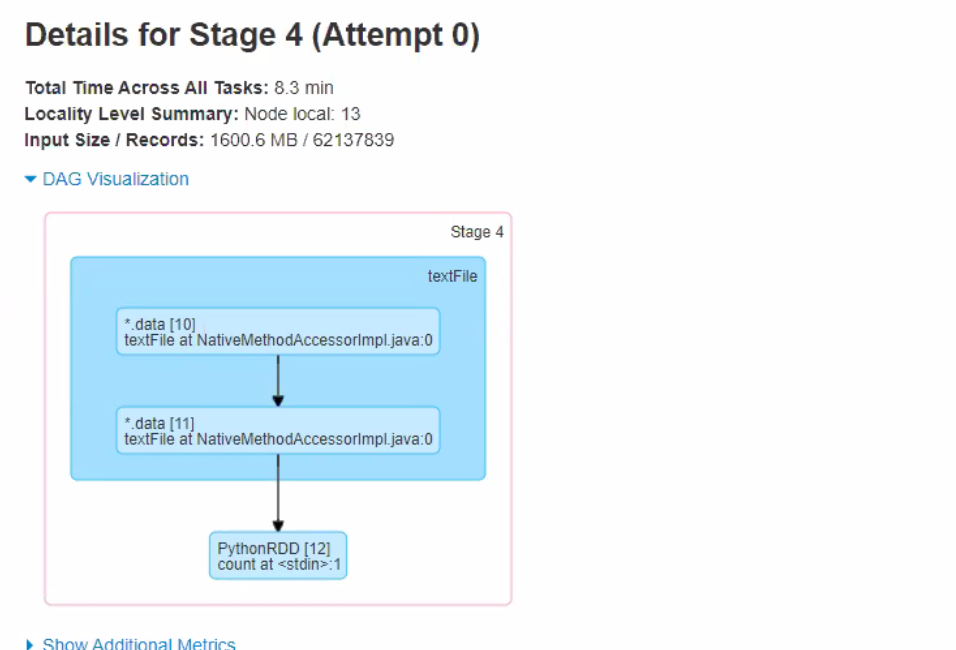
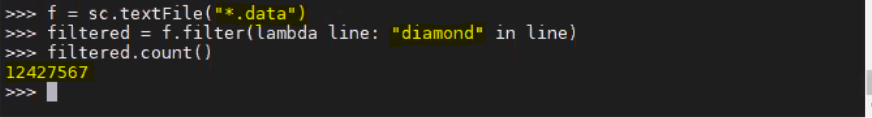












To store the result in single file use following command

**filtered.coalesce(1).saveAsTextFile(“<Any directory name you want>”)**

