

Yang Song– SEC01 (NUID 001003647)

Big Data System Engineering with Scala

Spring 2022

Assignment No. 6



Task

Read movie rating dataset and calculate the mean rating and the standard deviation for all movies and create test cases

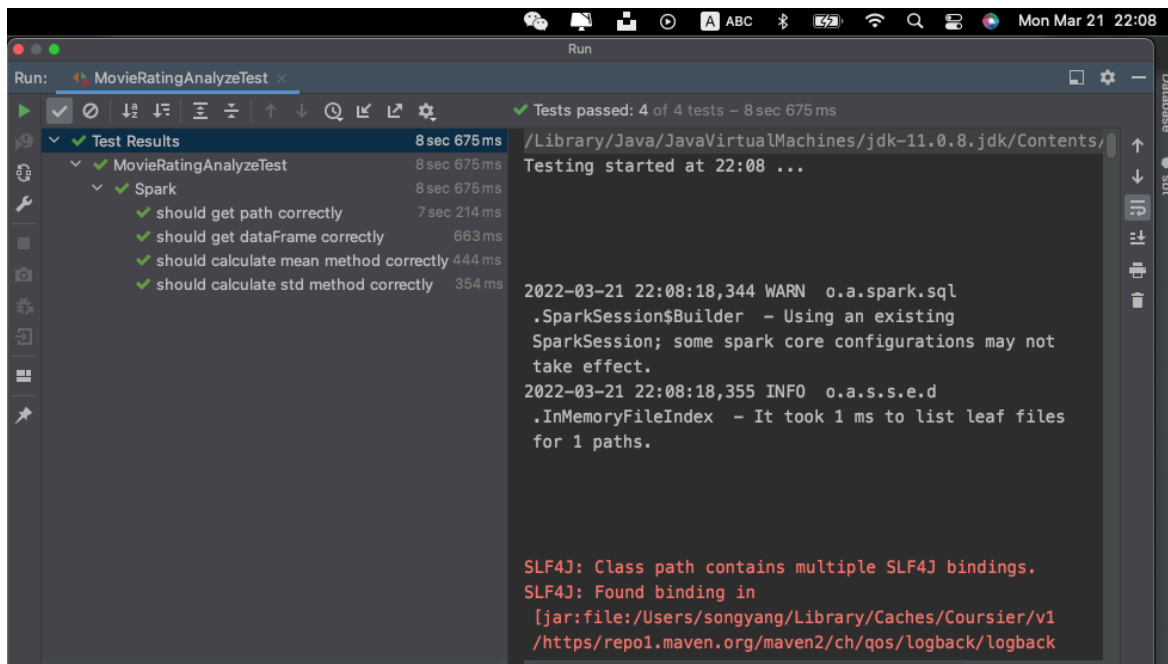
Solution/ Unit test (screenshot)

All Scala files are in the spark-csv module. I create a new class called "MovieRatingAnalyze.scala" and its related test class "MovieRatingAnalyzeTest.scala". I use the "movie_metadata.csv" from the resources directory of this module.

The mean result and the standard deviation result:

```
+-----+ +-----+
| avg(imdb_score) | | stddev_samp(imdb_score) |
+-----+ +-----+
| 6.453200745804848 | | 0.9988071293753289 |
+-----+ +-----+
```

The unit test result:



```
Run: MovieRatingAnalyzeTest
Tests passed: 4 of 4 tests - 8 sec 675 ms

Test Results
  MovieRatingAnalyzeTest 8 sec 675 ms
    Spark 8 sec 675 ms
      should get path correctly 7 sec 214 ms
      should get dataFrame correctly 663 ms
      should calculate mean method correctly 444 ms
      should calculate std method correctly 354 ms

2022-03-21 22:08:18,344 WARN o.a.spark.sql
.SparkSession$Builder - Using an existing
SparkSession; some spark core configurations may not
take effect.
2022-03-21 22:08:18,355 INFO o.a.s.s.e.d
.InMemoryFileIndex - It took 1 ms to list leaf files
for 1 paths.

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in
[jar:file:/Users/songyang/Library/Caches/Coursier/v1
/https/repo1.maven.org/maven2/ch/qos/logback/logback
```

Project Source

<https://github.com/Shelleyaaa/CSYE7200/tree/Spring2022/spark-csv>