Qinyun Lin – SEC01 (NUID 001582464)

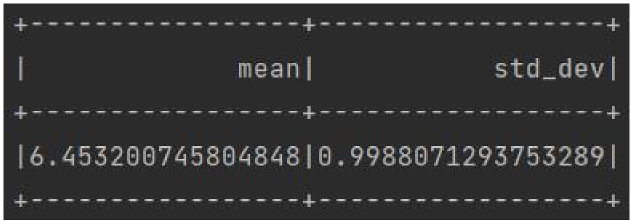
Big Data System Engineering with Scala
Spring 2022
Assignment No. 6

Task

1. **Read Movie dateset into spark**
2. **Calculate the mean rating and standard deviation for all movies**

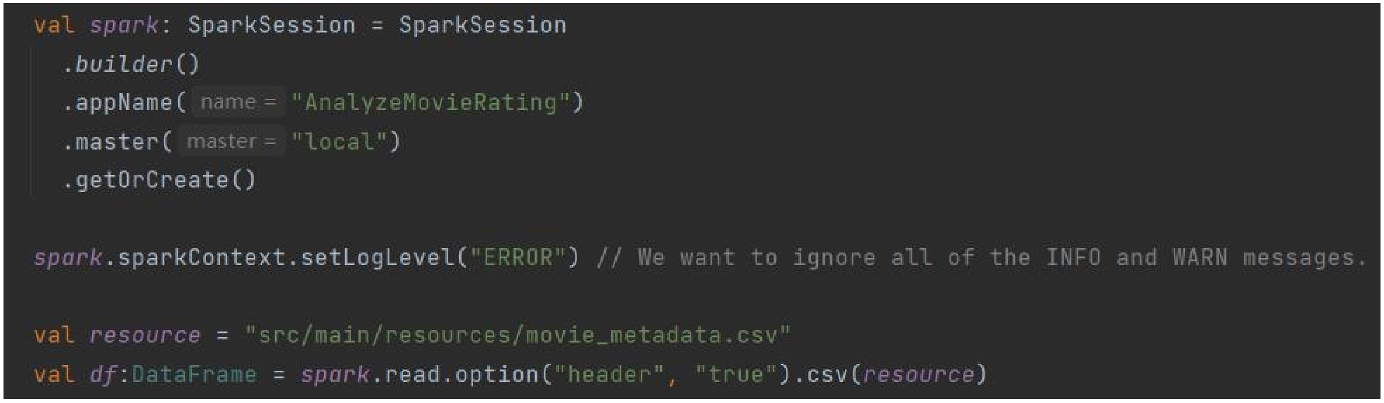
Solution

The data I used is the one in the repository.
There are 1609 rows of records in the data, I calculate the “Ratings” based on the “imdb_score” column. The result I got is shown below:



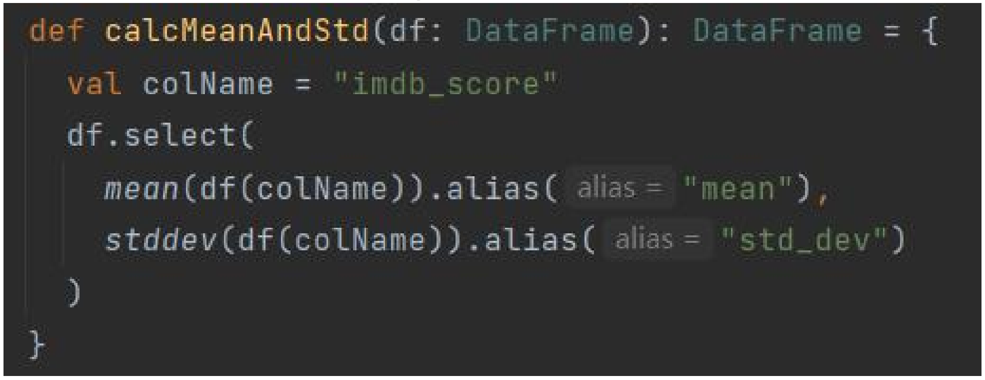
```
+----------------+-----------------+
|            mean|          std_dev|
+----------------+-----------------+
|6.453200745804848|0.9988071293753289|
+----------------+-----------------+
```

1. Read Movie dateset into spark



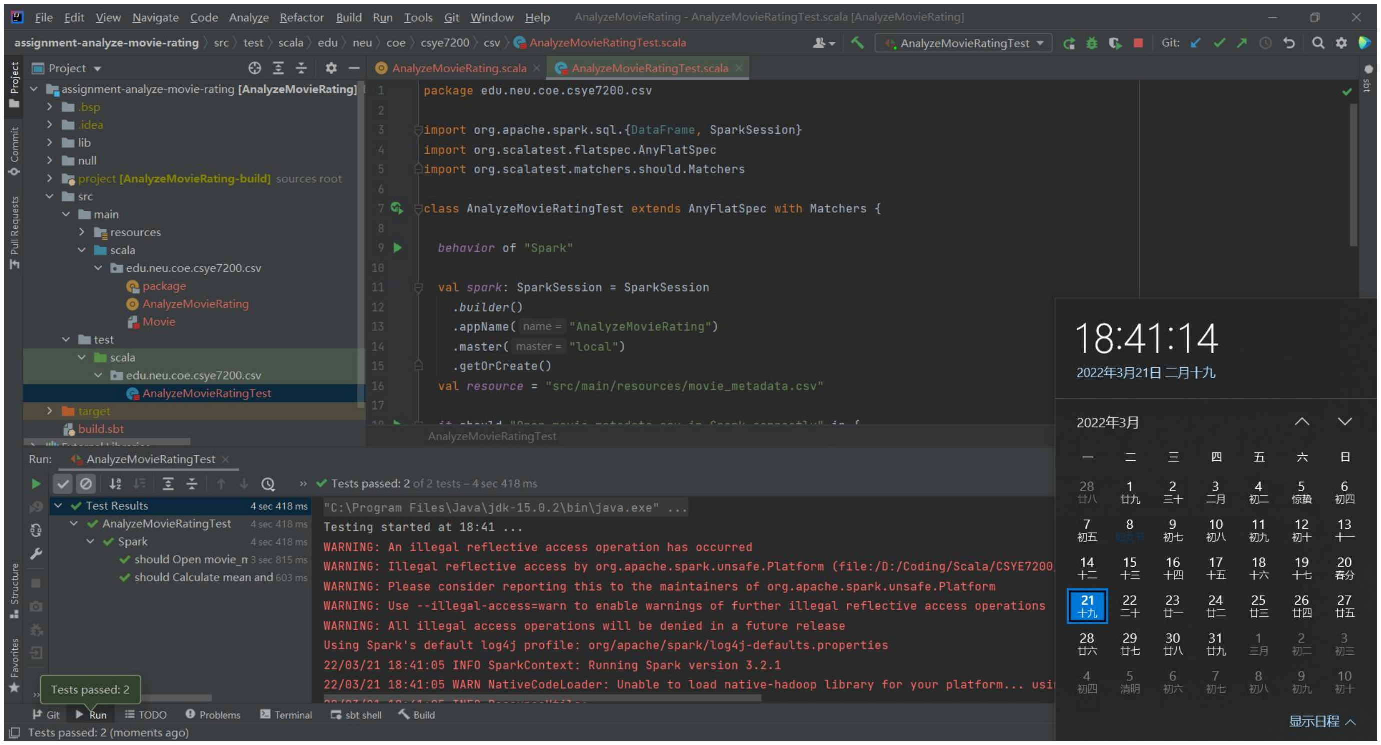
```
val spark: SparkSession = SparkSession
  .builder()
  .appName( name = "AnalyzeMovieRating")
  .master( master = "local")
  .getOrCreate()

spark.sparkContext.setLogLevel("ERROR") // We want to ignore all of the INFO and WARN messages.

val resource = "src/main/resources/movie_metadata.csv"
val df:DataFrame = spark.read.option("header", "true").csv(resource)
```

2. Calculate the mean rating and standard deviation for all movies



```
def calcMeanAndStd(df: DataFrame): DataFrame = {
  val colName = "imdb_score"
  df.select(
    mean(df(colName)).alias( alias = "mean"),
    stddev(df(colName)).alias( alias = "std_dev")
  )
}
```

Unit Test Screenshot


File Edit View Navigate Code Analyze Refactor Build Run Tools Git Window Help
AnalyzeMovieRating - AnalyzeMovieRatingTest.scala [AnalyzeMovieRating]
assignment-analyze-movie-rating 〉 src 〉 test 〉 scala 〉 edu 〉 neu 〉 coe 〉 csye7200 〉 csv 〉 AnalyzeMovieRatingTest.scala
AnalyzeMovieRatingTest
Git:
Project
assignment-analyze-movie-rating [AnalyzeMovieRating]
.bsp
.idea
lib
null
project [AnalyzeMovieRating-build] sources root
src
main
resources
scala
edu.neu.coe.csye7200.csv
package
AnalyzeMovieRating
Movie
test
scala
edu.neu.coe.csye7200.csv
AnalyzeMovieRatingTest
target
build.sbt
AnalyzeMovieRating.scala
AnalyzeMovieRatingTest.scala
Run:
AnalyzeMovieRatingTest
Git
Run
TODO
Problems
Terminal
sbt shell
Build
Structure
Favorites
Commit
Pull Requests

```scala
package edu.neu.coe.csye7200.csv

import org.apache.spark.sql.{DataFrame, SparkSession}
import org.scalatest.flatspec.AnyFlatSpec
import org.scalatest.matchers.should.Matchers

class AnalyzeMovieRatingTest extends AnyFlatSpec with Matchers {

  behavior of "Spark"

  val spark: SparkSession = SparkSession
    .builder()
    .appName( name = "AnalyzeMovieRating")
    .master( master = "local")
    .getOrCreate()
  val resource = "src/main/resources/movie_metadata.csv"

  it should "Open movie metadata csv in Spark correctly" in {
```

Tests passed: 2 of 2 tests – 4 sec 418 ms

```
Test Results                         4 sec 418 ms
  AnalyzeMovieRatingTest             4 sec 418 ms
    Spark                            4 sec 418 ms
      should Open movie_m            3 sec 815 ms
      should Calculate mean and      603 ms
```

```
"C:\Program Files\Java\jdk-15.0.2\bin\java.exe" ...
Testing started at 18:41 ...
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/D:/Coding/Scala/CSYE7200,
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
22/03/21 18:41:05 INFO SparkContext: Running Spark version 3.2.1
22/03/21 18:41:05 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... usi
```

Tests passed: 2

Tests passed: 2 (moments ago)

18:41:14
2022年3月21日 二月十九

2022年3月

一	二	三	四	五	六	日
28 廿八	1 廿九	2 三十	3 二月	4 初二	5 惊蛰	6 初四
7 初五	8 妇女节	9 初七	10 初八	11 初九	12 初十	13 十一
14 十二	15 十三	16 十四	17 十五	18 十六	19 十七	20 春分
21 十九	22 二十	23 廿一	24 廿二	25 廿三	26 廿四	27 廿五
28 廿六	29 廿七	30 廿八	31 廿九	1 三月	2 初二	3 初三
4 初四	5 清明	6 初六	7 初七	8 初八	9 初九	10 初十

显示日程

Project Source

https://github.com/MrNiro/CSYE7200/tree/Spring2022/assignment-analyze-movie-rating