Newday Data Engineer Interview Homework

Introduction

This homework is intended to understand your capability in Spark and Scala/Python. The problem is posed quite loosely to allow you to solve it in the way you are most comfortable. We will be looking at the approach used, the code style, structure, and quality as well as testing.

Data

Download the movie lens open data set (ml-1m.zip) from
http://files.grouplens.org/datasets/movielens/ml-1m.zip

Brief

The job needs to do the following:

1. Read in movies.dat and ratings.dat to spark dataframes.
2. Creates a new dataframe, which contains the movies data and 3 new columns max, min and average rating for that movie from the ratings data.
3. Create a new dataframe which contains each user's (userId in the ratings data) top 3 movies based on their rating.
4. Write out the original and new dataframes in an efficient format of your choice.

You should also include the command to build and run your code with spark-submit.

Delivery

Please provide a link to a public repo.