

COMP 6231: Distributed Systems Design

Assignment 4 on RDD and Apache Spark

Summer 2022, sections BB

July 18, 2022

Contents

1	General Information	2
2	Introduction	2
3	Ground rules	2
4	Overview	3
4.1	Implementation Platform	3
4.2	Software Engineering Best Practices	3
5	Your Assignment	3
5.1	Queries	3
6	What to Submit	4
7	Grading Scheme	5

1 General Information

Date posted: Monday, July 18th, 2022.

Date due: Wednesday, August 3rd, 2022, by 23:59.¹.

Weight: 5% of the overall grade.

2 Introduction

This assignment targets using Resilient Distributed Dataset (RDD) in Apache Spark. In this assignment you use a small RDD and implement some queries using Apache Spark.

3 Ground rules

You are allowed to work on a team of 2 students at most (including yourself). Each team should designate a leader who will submit the assignment electronically. See Submission Notes for the details.

ONLY one copy of the assignment is to be submitted by the team leader. Upon submission, you must book an appointment with the marker team and demo the assignment. All members of the team must be present during the demo to receive the credit. Failure to do so may result in zero credit.

This is an assessment exercise. You may not seek any assistance from others while expecting to receive credit (**You must work strictly within your team**). Failure to do so will result in penalties or no credit.

¹see submission notes

4 Overview

At a high level, every Spark application consists of a driver program that runs the user's main function and executes various parallel operations on a cluster. The main abstraction Spark provides is a resilient distributed dataset (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. You may read more in ref. 2.

4.1 Implementation Platform

The assignment does not require using any specific programming language. It only enforces using Apache spark. You may use python, java, scala, C#, or any other programming environment by which you may use spark. Tutorials will be provided during the labs.

4.2 Software Engineering Best Practices

Use software engineering best practices for both architecture (structure) and implementation; Consider performance, scalability, and maintenance; Using RDD is expected,

5 Your Assignment

In this assignment, you use the movie-lens data-set available in ref 1. Use the small data-set. Your assignment is to load the above data-set into spark and develop an Apache Spark application to perform some queries, as specified in section 5.1.


You are expected to use a design as you would normally do in a “real-life” project.

Document your application; explain how you understand and solve the problem, explain your approach to the question, possible caveats and limitations.



5.1 Queries

1. How many movies of genre “drama” are there?
2. How many unique movies are rated, how many are not rated?

3. Who gave the most ratings, how many rates did the person make?
4. Compute min, average, max rating per movie. 
5. Output data-set containing users that have rated a movie but not tagged it.
6. Output data-set containing users that have rated AND tagged a movie.
7. Output data-set showing the number of movies per genre, per year².

6 What to Submit

The whole assignment is submitted by the due date under the corresponding assignment box. It has to be completed by ALL members of the team in one submission file.

Submission Notes

Clearly include the names and student IDs of all members of the team in the submission. Indicate the team leader.

IMPORTANT: You are allowed to work on a team of 2 students at most (including yourself). Any teams of 3 or more students will result in 0 marks for all team members. If your work on a team, ONLY one copy of the assignment is to be submitted. You must make sure that you upload the assignment to the correct assignment box on Moodle. No email submissions are accepted. Assignments uploaded to the wrong system, wrong folder, or submitted via email will be discarded and no resubmission will be allowed. Make sure you can access Moodle prior to the submission deadline. The deadline will not be extended.

Naming convention for the uploaded file: Create one zip file, containing all needed files for your assignment using the following naming convention. The zip file should be called a#_studids, where # is the number of the assignment, and studids is the list of student ids of all team members, separated by (_). For example, for the first assignment, student 12345678 would submit a zip file named a1_12345678.zip. If you work on a

²Movies associated with multiple genres, are counted multiple times.

team of two and your IDs are 12345678 and 34567890, you would submit a zip file named `a1_12345678_34567890.zip`.

Submit your assignment electronically on Moodle based on the instruction given by your instructor as indicated above: <https://moodle.concordia.ca>

Please see course outline for submission rules and format, as well as for the required demo of the assignment. A working copy of the code and a sample output should be submitted for the tasks that require them. A text file with answers to the different tasks should be provided. Put it all in a file layout as explained below, archive it with any archiving and compressing utility, such as WinZip, WinRAR, tar, gzip, bzip2, or others. You must keep a record of your submission confirmation. This is your proof of submission, which you may need should a submission problem arises.

7 Grading Scheme

Using RDD	10 marks
Best Practices	10 marks
Queries	70 marks
Documentattion	10 marks

Total: 100 marks.

References

1. The movie-lens data-set:
<http://files.grouplens.org/datasets/movielens/ml-latest-small.zip>
2. RDD Programming Guide:
<https://spark.apache.org/docs/latest/rdd-programming-guide.html>
3. Spark by Examples:
<https://sparkbyexamples.com/>