Distributed Data Infrastructures

Fall 2021

Assignment

In this assignment, you are going to use Apache Spark to analyze a large data set. Spark is an open-source big data framework that we have discussed in the course. The assignment is divided into two projects, each of which has a set of questions that need to be answered.

Project 1

In the first project, you need to write a program that uses Spark to provide an answer to the following questions.

1. What are the minimum and maximum values, the average, and the variance?

2. What is the value of the median of the data set? You should provide the exact median, not an approximation, and sorting the complete data set or using the .median() or .quantile() functions are not acceptable solutions.

3. How would you compute the mode for the data set? Explain how you would solve the problem and justify why this data set is not very well suited for computing the mode. What kind of data set would be better?

The source file for this project has the following format.

```
3.01316363
16.41347991
11.73966247
74.71116433
29.53299636
5.91881846
21.12204071
...
```

The file has one billion rows and each row contains only one float number.

Documentation

In the documentation, you should explain how your code solves the problems and how it uses Spark. You also need to provide the answers to the above questions.

Grading

Grading is based on the correctness of the program and the answers, quality of the program code, and associated documentation.

Guidelines

The assignment is individual work. You can of course discuss any problems you encounter with other students, but sharing code is not allowed and if found, will be considered plagiarism.

On the course Moodle, there is a discussion forum for asking questions about the project. There will also be a Q&A session at the end of each class session during the project. You can also ask questions in Slack and any relevant answers from Slack will also be posted by us on Moodle.

Deliverables

Program source code with documentation. You can return the code as a python script. The document should explain how you have solved the problems and provide answers to the questions from project description. Even if your code does not work or does not work correctly, explain in the documentation how you have tried to solve the problem.

Timeline

The deadline for the is December 1st, 23:00. No extensions will be given.

Return

Store all the files in a directory that has the same name as your username. Zip this directory, name the zip-file "username_DDI21_EX1.zip", and return the zip-file via Moodle. Please indicate clearly your name and student ID in every source code file.

Dataset and sample code

We created a group directory for the course on Ukko2 "/wrk/group/grp-ddi-2021" which will have all the relevant datasets, sample code and modules necessary for this assignment.

For this assignment, there are two files, data-1.txt and data-1-sample.txt. The first one (data-1.txt) is the full data set that you should use to provide the answers to the questions. The second (data-1-sample.txt) is a subset of the bigger data set that you should use when developing your programs so that they run faster. Make sure everything works smoothly with the sample data sets before trying out the real data sets. Please ignore the data-2.txt for now.