

Distributed Data Infrastructures

Fall 2021

Project 2

In the second project, you are asked to perform the following task:

1. Calculate $A \times A^T \times A$ and provide the first row of the resulting matrix as your answer in the same format as the input matrix.

For this project, use the input file (data-2.txt) which contains the matrix A. The data-2.txt is a text file containing a 1000000 x 1000 matrix. The file is stored in the text format, and each line represents a row vector. The row contains 1000 float numbers which are separated by white spaces.

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 project is December 19th 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_EX2.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-2.txt and data-2-sample.txt. The first one (data-2.txt) is the full data set that you should use to provide the answers to the question. The second (data-2-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.