Distributed Data Infrastructures, Fall 2020, Project 1

Assignment

In this project, you are supposed to use Apache Spark to analyze a large data set. Spark is an open source big data framework which we have discussed in the course. We provide you with two data sets and you should run Spark on Ukko2 to answer one question about each data set.

Requirements

You need write a program that uses Spark to provide an answer to the following questions.

1. For the first data set (data-1.txt), provide the value of the median of the data set. You should provide the exact median, not an approximation, and sorting the complete data set is not an acceptable solution.
2. The second data set (data-2.txt) contains the matrix A. Your task is to calculate A x A^T x A and provide the resulting matrix as your answer in the same format as the input matrix.

The data-1.txt is in 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.

The data-2.txt is 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 as 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 Requirements section. 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 assignment is due on November 22th at 23:00. No extensions will be given.

Return

Store all the files in a directory that has same name as your username. Zip this directory, name the zip-file "username_DDI20_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-2020" which will have all the relevant datasets, sample code and modules necessary for this assignment.

For both questions, there are two versions of the data sets, data-(1|2).txt and data-(1|2)-sample.txt. The first one (data-(1|2).txt) is the full data set that you should use to provide the answers to the questions. The second (data-(1|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.

Running the experiment

1. Log in to Ukko2

First, you will need to connect to Ukko2 (it will ask you to login twice).

```
# Access to Ukko2 (goes through Melkkipass server)
ssh <username>@ukko2.cs.helsinki.fi
```

2. Running the script

Run your scripts from your working directory. The work directory uses HDFS for faster I/O operations. In order to access your working directory, use the environment variable $WRKDIR. Your work directory will be located at /wrk/users/<username>

```
cd $WRKDIR
$ /wrk/users/<username>
```

3. Load the Spark Module

Spark cluster is already installed and running. However, you will need to load the Spark module in order to submit jobs to it. The module is available at the course project and can be loaded with the following instructions:

```
# Add Spark module to the list of modules to be used
$ module use /wrk/group/grp-ddi-2020/MyModules
# Check if Spark module is available to be used. You should see it at the top of your list
$ module avail
# Load the module itself
$ module load Spark
# Check if the module is loaded correctly. You should be able to see it in the loaded list
$ module list
```

4. Submitting a job to Spark

Now that the module has been configured, you can submit a job to the Spark cluster. For that:

a. Make a local copy of the sample code (/wrk/group/grp-ddi-2020/sampleScript.py) in your work directory.
b. Change the master IP in the script. The current master IP address can be found on DDI 2020 Moodle page.
c. After making necessary changes to the python script, you can submit it to spark cluster via spark-submit <script_name.py>

```
# Submit job to Spark cluster
$ spark-submit <script_name.py>
```

Use the sample script, sampleScript.py, provided in spark-data as base for your submission. The script provides sample code for counting and adding dataset for question 1 (data-1.txt). The configuration section of the script contains several Spark control parameters which can be played with, including address to Spark master node.

5. Access to the Spark Management GUI

You can access the Spark Management GUI through SSH tunneling. Make sure that address to master node is correct from the webUI.

- If you are within the University network, you can also access the Spark visualizer running at master at http://<master-IP>:8080.
- In case you are connected via *eduroam*, you will need to pipe the address to your local computer (localhost) to a specific port (8081). You can then access the UI from http://localhost:8081 in your browser. For that, you will need to create a SSH tunnel that binds a local port in your machine to connect to the Spark master UI port and you will need the IP address of the Spark master node. Spark management UI is available at port 8080.

```
# SSH Tunnel connecting localhost port 8081 to the Spark master node UI at port 8080
ssh -N -f -L localhost:8081:<spark master node IP>:8080 <username>@ukko2.cs.helsinki.fi
```

Open your web-browser and access the following link to see the Spark Management UI: http://localhost:8081

6. Troubleshooting

- In case Spark throws a "FileNotFoundexception" at spark-submit, it means that master does not have permission to access your dataset and script. To solve the problem, give group permission to your path to the dataset and script (chmod g+xr <directory_name>)

References

Spark documentation link: https://spark.apache.org/docs/2.2.0/

Ukko2 guide: https://wiki.helsinki.fi/display/it4sci/Ukko2+User+Guide