Distributed Systems – Big exercise 4

Assignment

In this assignment, your task is to use Apache Spark to analyse a large data set. Spark is an open source big data framework which we have seen in the course. We provide the environment where to run Spark and also the data set. The data set is simply a set of numbers in a text file. See below for more details.

Requirements

For this exercise we use a data set provided by us. You need write a program that uses Spark to provide an answer to the following questions.

1. For the provided data set, calculate minimum, maximum, average, and variance?

For this question the answer is four numbers.

2. Explain how you would 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 a data set would be better?

3. Calculate three histograms, where the number of bins are a) 10, b) 100, and c) 1000.

In a) you should produce an output file that has ten lines, and one integer on each line. The integer on the first line tells how many numbers in the data set are from the range $0 < x \leq 10$. The second line tells how many numbers are from the range $10 < y \leq 20$, and so on until the last row which tell how many numbers are from between $90 < z \leq 100$.

In b) the task is the same, but output will have 100 rows, and rows describe the ranges beginning from $0 < x \leq 1$; $1 < y \leq 2$; $2 < z \leq 3$, and so on until $99 < q \leq 100$.

In c) the histogram is of even finer granularity, and will have 1000 lines.

Deliverables

Program source code with documentation. The document should explain how you have solved the problems and provide answers to the questions from Requirements section.

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.

Deadline

dThe assignment is due on December 18th at 23:55. No extensions will be given.

Return

Store all the files in a directory that has same name as your username. Put this directory into a zip-file by the name “`username_DS16_BE4.zip`” and return the zip-file via Moodle. Please indicate clearly your name and student ID in every source code file.

Set Up

Spark (version 2.0.2) has been installed on the Ukko cluster. The master node is **ukko007** and it is dedicated for that purpose. Hence you can not login there, and neither should you. However, you need to reference it in your code. An example can be seen in the provided example code. Here is a very brief instruction to help you start quickly. Additional help will be provided in the Q&A sessions as needed.

1. First, log into one of the nodes of Ukko e.g.

```
ssh username@ukko042.hpc.cs.helsinki.fi
```

And as before, Ukko nodes are accessible only from within the CS network. You need to use e.g. melkki in between if you're working from outside the CS department.

2. In order to use our Spark installation, you need to add the related paths to your profile dot-file (~/.profile). Add the following:

```
# SPARK
export SPARK_HOME=/cs/work/scratch/spark-2.0.2-bin-hadoop2.6
export PATH=$PATH:$SPARK_HOME/bin
export PYTHONPATH=$SPARK_HOME/python/:$PYTHONPATH
export PYTHONPATH=$SPARK_HOME/python/lib/py4j-0.10.3-src.zip:$PYTHONPATH
```

Dot-files are meant to be hidden, so a normal file listing doesn't show if the file exists. If you don't have `.profile` in your home directory, then create one (e.g. with `$ touch ~/.profile`). A re-login to the system is required for changes to take place.

3. Download the example from the course webpage (link also at the end of this document), and run spark-example.py on a Ukko node. Note that the sample code uses Python 2 syntax. For the exercise you may use either python 2 or 3. The default python version for PySpark is still 2.x, and that is probably easier to get started with in this exercise. If you can see the following output, it means Spark is correctly running for you now.

```
Avg. = 50.19059339
```

You may also see a lot of other diagnostic output…

4. The data set we use (data-1.txt) is in the following format:

```
3.01316363
16.41347991
11.73966247
74.71116433
```

```
29.53299636
5.91881846
...
```

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

The dataset is located at `/cs/work/scratch/spark-data/data-1.txt`

If you want to start experimenting with a smaller data set, you can use a sample of the full set data we provide by the name `data-1-sample.txt` in the same directory. Developing will be faster with a small data set, since one execution round with the full data will take a while. The sample file contains only the first 1000 lines of the full data set. Final answers have to be calculated from the full data set!

5. You can use ukko007.hpc.cs.helsinki.fi:8080 to monitor current state of Spark.

Once again, note the link is only accessible while you are within the Department network.

More Information

You may also find the following links useful:

Spark documentation:

https://spark.apache.org/docs/2.0.2/

Example code:

https://www.cs.helsinki.fi/u/owaltari/distsys/spark-example.py