Distributed Systems Project, Spring 2016 – Assignment 2

Assignment

In this assignment, you are supposed to use Apache Spark to analyze 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 sets. The data sets are simply sets of numbers in a text file. See below for the exact format.

Requirements

We will use two data sets in this assignment. For the first data set (data-1.txt), you need write a program that uses Spark to provide an answer to the following questions. For the first question, you will need to provide 4 numbers and 1 for the second.

1. What are the minimum and maximum values, the average value and the variance?
2. What is the value of the median of the data set?

In addition, you need to 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?

For the second data set (data-2.txt), you need answer the following two questions using Spark. The data set contains the matrix A and you need to compute the following.

1. $A \times A^T \times A$
2. $\text{diag}(A \times A^T)$

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.

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.

Timeline

The assignment is due on February 16th at 10:00. 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_DSP16_EX2.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 1.6.0) has been installed on the Ukko cluster, in CoNe group folder. The complete path is /cs/work/scratch/spark-1.6.0-bin-hadoop2.6. 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, you need log into melkki with the following command:

```
ssh username@melkki.cs.helsinki.fi
```

then log into one of the nodes of Ukko e.g.

```
ssh username@ukko017.hpc.cs.helsinki.fi
```

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-1.6.0-bin-hadoop2.6
export PATH=$PATH:$SPARK_HOME/bin
export PYTHONPATH=$SPARK_HOME/python/:$PYTHONPATH
export PYTHONPATH=$SPARK_HOME/python/lib/py4j-0.9-src.zip:$PYTHONPATH
```

3. Download the example from the course webpage, and run spark-example.py on a Ukko node.

If you can see the following output, it means Spark is correctly running for you now.

```
Avg. = 23.55488086
```

You may also see a lot of other diagnostic output…

4. We will use two datasets: data-1.txt and data-2.txt. 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.

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.

The dataset are accessible at /cs/work/scratch/spark-data.

If you want to start with smaller data sets, we also provide two samples of both data sets, i.e. data-1-sample.txt and data-2-sample.txt. They are in the same directory as original datasets and contain 1000 lines of data.

5. You can use ukko080.hpc.cs.helsinki.fi:8080 to monitor current state of Spark. Note the link is only accessible while you are within Department network.

More Information

You may also find the following links useful:

Spark documentation:
https://spark.apache.org/docs/1.6.0/