COMP 415/515: Distributed Computing Systems

Assignment-3

Due: May 2, 2020, 11:59 pm (Late submissions are <u>not</u> accepted.)

This is an individual assignment. You are not allowed to share your codes with each other.

Distributed Data Processing with Apache Spark

This assignment is about learning and practicing with the distributed **data processing framework Apache Spark** to process large-scale data in a distributive way. Due to in-memory processing, it is faster than the Hadoop MapReduce framework. In our first course assignment, you deployed a distributed storage system in the cloud. Suppose that you want to analyze the files stored or log files of the users on your distributed storage system. The simplest way is that you should collect all the files on one server and just write a simple program to analyze the files. However, as you noticed from Assignment 1 Part-2 that it is quite probable that you do not have enough storage to store all the files on the central server. In addition to this, by doing analysis this way you are utilizing a lot of bandwidth, incurring latency, and requiring unnecessary central storage. To address these issues, one can use Apache Spark to analyze the data in a distributive way i.e., without moving the data to a central server. By using Spark, each node in the storage system will analyze its data and send back the result to a central/parent node. In this way, you do not need data to be transferred to a central place to be analyzed.

The main purpose of this assignment is to make you familiar with the **distributive data processing framework (Apache Spark)**. You can run Spark in **Python**, **Java**, **R** and **Scala**. You are free to choose any language, but it is recommended that you write in Python as it is easier and the tutorials provided include related examples.

In this assignment, you are asked to run Spark on a single machine to analyze the file "book.txt" and files in "numbers.zip". However, in the optional part of this assignment, you are required to deploy it on 5 AWS nodes.

Tasks:

1. (10 points)

Using Spark, write a program to count the number of words in "book.txt".

<u>Example:</u>

Input: "*A distributed system is a collection of autonomous computing elements that appears to its users as a single coherent system.*"

Output: system: 2, distributed: 1, …..

2. (20 points)

Using Spark, write a program to count how many times each letter appeared in the “book.txt.”

3. (20 points)

Using Spark, write a program to replace the words to lowercase letters and write it the file “words_lower.txt.”

4. (20 points)

Using Spark, write a program to replace spaces with “-” in the “book.txt” and write it to “words-.txt”.

5. (20 points)

Using Spark, compute the sum of the numbers given in “numbers.txt” in the numbers.zip file.

Additionally, you are given files numbers2.txt, numbers4.txt, numbers8.txt, numbers16.txt, and numbers32.txt.

Compute the sum of the numbers in the individual files and plot a bar-graph. On the x-axis plot the size of the file and on y-axis plot the time taken by the Spark to compute the result.

Optional Part (30 points)

Deploy the Spark on 5 AWS nodes and store “book.txt” and “numbers.txt” on these nodes and repeat the task 1, task 2, and task 5.

Report (10 points)

Write a 1-page report on Spark and mention its main features & use cases. For instance, what kind of data can be processed in it. What are RDDs?

If you completed the optional part, then also write the step by step procedure to deploy Spark on the nodes.

Submission and Demonstration:

Please submit your assignment including a report on the Blackboard (lastname_KU_id.zip). The TA would announce the demo sessions. Attending the demo session is required for your assignment to be graded. Please read this assignment document carefully BEFORE the implementation. Take the report part seriously and draft your report as accurately and complete as possible.

Resources:

The online materials on **Apache Spark** are provided on the course site.

https://sites.google.com/a/ku.edu.tr/comp515/home/lecture-notes/apachespark

Installation of Spark:

Please follow the following steps to install Spark on AWS (Linux) or on your Linux machine:

1. cd /tmp && curl -O

   Dowload: https://repo.anaconda.com/archive/Anaconda3-2019.03-Linux-x86_64.sh
2. bash Anaconda3-2019.03-Linux-x86_64.sh
   - Press enter or yes on each command prompt
3. Logout and login again to the AWS server/local machine
4. conda update anaconda-navigator
5. conda install pyspark
6. sudo apt install default-jre
7. sudo apt install default-jdk
8. sudo apt install openjdk-8-jdk
9. sudo update-alternatives --config java
   - (choose the option of java 8)
10. Installation is complete (You can run the following code to test your installation. You will see the squares of [1,2,3,4] in the output).

```
import pyspark

from pyspark import SparkContext

sc = SparkContext()

nums = sc.parallelize([1,2,3,4])

squared = nums.map(lambda x: x*x).collect()

for num in squared:

    print('%i ' % (num))
```

Good Luck!