New York University
Computer Science Department
Courant Institute of Mathematical Sciences

Course Title: Cloud Computing
Course Number: csci-ga.3033-026
Instructor: Jean-Claude Franchitti
Session: 4

Assignment #4

I. Due

March 17, 2022 at the beginning of class.

II. Objectives

1. Understand how to leverage MapReduce using Spark on the Big Clouds.
2. You will learn how to process Big Data and streaming Big Data on the Big Clouds.

III. References

1. Slides and handouts posted on the course Web site for Session 4
2. Relevant course textbook sections

IV. Software Required

1. Microsoft Word.
2. Win Zip as necessary.
3. Cloud platforms related services as applicable.

V. Assignment

Part I: Get familiar with Spark; execute and document the following exercises

To run the notebooks below, you may bring up Jupyter in your own Linux data science vm (e.g., https://youvm-IP-address:8000 and sign in with vm userid and password as needed and as per assignment #2).

Another alternative is to run the course textbook tutorial container:

docker run -it -p 8888:8888 dbgannon/tutorial

Both of these alternatives have spark built in.

The various notebooks are available in the tutorial container or can be uploaded to the data science vm from the tutorial package installed earlier.

- spark-euler is a simple illustration of Spark used to demonstrate a trivial map-reduce computation. Download notebook file.

- spark provides the second demonstration of Spark for a simple k-means clustering algorithm. Download notebook file.

- sql-magic demonstrates how SQL commands may be executed in Spark. This notebook illustrate the use of a special set of commands that allows embedding of SQL in an IPython notebook directly. Download notebook file.

Part II: Experiment with Spark on AWS; execute and document the following exercise

- aws-emr provides a small tutorial on how to deploy Jupyter in a Spark cluster on an AWS Elastic Map Reduce cluster. The Notebook illustrates this with a small example of exploring Wikipedia data. Download notebook file.

Part III: Experiment with Google Datalab; execute and document the following exercises

- datalab1 illustrates Google's Datalab. This notebook is an exploration of contagious disease records from the the U.S. CDC, specifically looking at Rubella cases over a period of time. Download notebook file.

- datalab2 examines weather station data and spots an anomaly in one station's reporting. Download notebook file.

Part IV: Experiment with Spark on Azure; execute and document the following exercise

- sparkml Azure's HDInsight plus Spark are used here to look at food inspection records.Download notebook file.

Part V: Experiment with Streaming Big Data on AWS; execute and document the following exercise

- kinesis uses AWS Kinesis together with Spark to detect anomalies in data from the Chicago Array of Things instrument streams. Download notebook file. The data file is in Kinesis-spark-Aot.

Part VI: Build additional experiments to be able to provide an illustration of big data processing and streaming big data processing on all four Big Clouds. The experiments covered in Parts I-IV above should count towards the complete set of experiments provided as part of your assignment solution. Please note that the notebooks.azure.com site does not have Spark installed so you may have to find another Azure solution for running spark on big data.

Part VII: Build additional experiments to document the use of MapReduce using Hadoop/Yarn.

VIII. Deliverables

1. Electronic:

   Your assignment file must be submitted via NYU Brightspace. The file must be created and sent by the beginning of class. After the class period, the homework is late. The email clock is the official clock.

2. Cover page and other formatting requirements:

   The cover page supplied on the next page must be the first page of your assignment file.

   Fill in the blank area for each field.

   NOTE:

   The sequence of the electronic submission is:

   1. **Cover sheet**
   2. **Assignment Answer Sheet(s)**

3. Grading guidelines:

   Assignment Layout (15%)

   - Assignment is neatly assembled on 8 1/2 by 11 layout.
   - Cover page with your name (last name first followed by a comma then first name), username and section number with a signed statement of independent effort is included.
   - File name is correct.

   Answers to Individual Questions (85%):

   - Answers to all questions are complete and correct.
   - Assumptions provided as required.

   (100 points total, all questions weighted equally)

VIII. Sample Cover Sheet:

Name _____ Date: _____
(last name, first name)
Section: _____

Assignment 4

Total in points (100 points total): _____

Professor's Comments: