CST 5321: ADVANCED OPERATING SYSTEM

FALL 2020: WRITING ASSIGNMENT 1

This assignment will introduce you to distributed operating system where you are going to use different big data applications.

Learning Outcomes

After completing this programming assignment, students should be able to:

- Deploy and configure Apache Spark and HDFS in Google Cloud Dataproc.
- Write Spark applications using Python and launch them in the cluster.
- Describe how your program performs given different scenarios.

Steps

Here is the list of things that you have to do:

1. Read the paper Apache Spark: A Unified Engine for Big Data Processing M. Zaharia, et al., CACM 2016
2. Redeem your GCP coupon

Here is the URL you will need to access in order to request a Google Cloud Platform coupon. You will be asked to provide your school email address and name. An email will be sent to you to confirm these details before a coupon is sent to you.

Student Coupon Retrieval Link

- You will be asked for a name and email address, which needs to match the domain. A confirmation email will be sent to you with a coupon code.
- You can only request ONE code per unique email address.

3. Go through the following tutorial and create a dataproc cluster with single node (by default each node has 4 CPU)

https://cloud.google.com/dataproc/docs/tutorials/gcs-connector-spark-tutorial#python

Another tutorial that might also help: https://towardsdatascience.com/step-by-step-tutorial-pyspark-sentiment-analysis-on-google-dataproc-fef9bef46468

4. Download the following 2 files from canvas and upload them to your storage bucket in the cloud.
   Word-count.py
   Pg100.txt

5. Run the word-count.py with pg100.txt in your cluster. (see question 2 below).

6. Answer the following questions:

- **Question 1.** What is the default block size on HDFS? What is the default replication factor of HDFS on Dataproc? **15 points**
- **Question 2.** Using pg100.txt as input and run the word-count.py program on a Single Node cluster using 4 cores (Step 5 above). What is the completion time of the task? Please take a snapshot of your VM instances monitoring page while running. **25 points**
- **Question 3.** Using pg100.txt as input and run the word-count.py program under HDFS inside a 2-node cluster (1 master, 1 worker nodes). Is the performance getting better or worse in terms of completion time? Briefly explain. **25 points**
- **Question 4.** Using pg100.txt as input and run the word-count.py program under HDFS inside a 3 node cluster (1 master, 2 worker nodes). Is the performance getting better or worse in terms of completion time? Briefly explain. **25 points**

- **Question 5.** For this question, change the default block size in HDFS to be 64MB and repeat Question 4. Record run time, is the performance getting better or worse in terms of completion time? Briefly explain. **10 points**
- **Extra Credit: Question 6.** Run the settings in Question 4, kill one of the worker nodes immediately. You could kill one of the worker nodes by go to the **VM Instances** tab on the Cluster details page and click on the name of one of the workers. Then click on the STOP button. Record the completion time. Does the job still finish? Do you observe any difference in the completion time? Briefly explain your observations. **5 points**

- **Submission deadline**: By 9:00 PM, 09/02/2020.
- **What to submit**:
  - A report answering the questions with screenshots.
- **How to submit**:
  - Through Canvas.