IERG4330/IEMS5730 Spring 2022
Homework 3

Release date: Mar 11, 2022
Due date: 11:59:00 pm, Mar 28, 2022
The solution will be posted right after the deadline, so no late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student_____) Date:_____

Name_____ SID_____

Submission notice:

- Submit your homework via the elearning system

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

Q1 [10 marks + 10 bonus marks]: Setup Spark Cluster

a) **[10 marks]** Install Spark 3.2.1 and setup standalone Spark Cluster on your own machine.

Submit the screenshot(s) of your installation process and screenshot(s) of Spark Shell.

b) **[Bonus 10 marks]** Follow the instructions in [8] to setup a multi-node Spark cluster over Hadoop YARN using AWS or Google Cloud.

Submit the screenshot(s) of your installation process and screenshot(s) of Spark Shell.

Q2 [35 marks]: Spark Basic RDD

In this question, you will explore some of the basic RDD concepts in Spark. **You need to submit your Spark application to a Hadoop cluster (either your own Hadoop cluster set up in Q1b or use the Spark Installation in DIC).** You can use any programming language supported by Spark but need to operate on the RDD level.

(a) **[15 marks]** ~~**[10 marks]**~~ Naive implementation of PageRank in Spark. In this part, write your own PageRank program. Then, run your program on the given dataset [2], submit the code and the top 100 nodes.

(b) **[20 marks]** ~~**[10 marks]**~~ Advanced implementation of PageRank in Spark. In this part, you need to take advantage of the pre-partition mechanism to reduce the shuffling overheads. Please adjust the number of partitions (try 3 different cases) and compare their performance. Submit your result and explain your observations.

Q3 [35 marks]: Spark SQL

In this question, we will analyze the report of crime incidents in Washington D.C. . The dataset comes from the District of Columbia's Open Data Catalog. Download the report of crime incidents in 2013 from:

http://opendata.dc.gov/datasets/crime-incidents-in-2013

Upload the data to HDFS. After you explore this CSV file, you can find it consists of around 20 columns. **You need to submit your Spark application to a Hadoop cluster (either your own Hadoop cluster set up in Q1b or use the Spark Installation in DIC).**

(a) **[10 marks]** We are interested in the following information:

(CCN, REPORT_DATE, OFFENSE, METHOD, END_DATE, DISTRICT).

Use Spark to truncate the file and only keep these 6 items of each line of the record. If these fields are empty in some lines, please filter out those lines.

Hints: if these fields are empty in some lines, please filter out those lines.

(b) **[10 marks]** Use Spark queries[3] to count the number of each type offenses and find which time-slot (shift) did the most crimes occur.

(a) **[15 marks]** The dataset below tracks the crime incidents from 2010 to 2018.
http://opendata.dc.gov/datasets?q=crime%20incidents%20
Merge these 9 tables into one and compute the percentage of gun offense for each year. Discuss the effect of Obama's executive actions on gun control.

Q4 **[Mandatory for ESTR4316] [10 Bonus marks for others]**: Setup and run a Spark application over Kubernetes

In this question, you are required to submit and run the Spark WordCount application with the dataset in [4] using a Kubernetes cluster. **You can either use your own Kubernetes cluster or the IE DIC Kubernetes cluster provided by the TAs.** The overall procedure would involve:

1. Use the "$SPARK_HOME/bin/docker-image-tool.sh" tool on your Spark cluster setup from **Q1a/b** to build your own Docker image for the Spark driver and executor program (pods) of WordCount.
2. Publish (Push) your Docker image to a public Docker repository with the help of "$SPARK_HOME/bin/docker-image-tool.sh" in your Spark cluster.
3. Deploy and run your Spark Docker image in a Kubernetes cluster

More detailed instructions are as follows:

(a) **[3 marks]** ~~**[2 marks]**~~ Install Docker on your Spark cluster setup in **Q1a/b** and follow instructions in [5], [9] to build the Docker image(s) containing the WordCount application. The following command may be useful:
`$ ./bin/docker-image-tool.sh -r docker.io/myrepo -t v3.2.1 build`

Please include the screenshot of the output of the command "docker images".

(b) **[3 marks]** ~~**[2 marks]**~~ Push the Docker image(s) to a public Docker repository (e.g., Docker Hub, Amazon Elastic Container Registry, Google Container Repository). Make sure you can pull this Docker image from your Kubernetes cluster/ IE DIC Kubernetes cluster. The following command may be useful:
`$ ./bin/docker-image-tool.sh -r docker.io/myrepo -t v3.2.1 push`

In your homework submission, show the URL of your public Docker repository.

~~**[3 marks]**~~ If you setup your own Kubernetes cluster, follow instructions in [6] to configure Kubernetes RBAC roles and service accounts. This is to grant the Spark

driver program the necessary permission to launch Spark executor pods in the Kubernetes Cluster. The following command may be useful:

```
$ kubectl create serviceaccount spark --namespace=<your_namespace>
$ kubectl create clusterrolebinding spark-role --clusterrole=edit
--serviceaccount=default:spark --namespace=<your_namespace>
```

(c) **[4 marks]** ~~**[3 marks]**~~ Follow instructions in [7], [9] and launch Spark WordCount program to the Kubernetes cluster using "spark-submit". The following command may be useful:

```
bin/spark-submit \
    --master <kubernetess_master_address> \
    --deploy-mode cluster \
    --name <application_name> \
    --class <application_main_class_(for Java/ Scala apps)> \
    --conf spark.app.name=<spark_application_name> \
    --conf spark.kubernetes.authenticate.diver.serviceAccountName=spark \
    --conf spark.kubernetes.container.image=http://docker.io/myrepo/spark:v3.2.1 \
local:///opt/spark/examples/jars/spark-examples_2.12–3.2.1.jar
```

Tips:

1. You are recommended to use Amazon S3/ Google Cloud Storage as a distributed file system to read the dataset and store the output/ results of your program.
2. If you use the DIC Kubernetes cluster (MASTER_URL=k8s://https://172.16.5.98:6443), you need to connect to IE VPN before running spark-submit. See [9] for details.

Q5 **[10 Bonus marks]**: Fault tolerance mechanisms for Spark over Kubernetes

In this part, you need to figure out the fault-tolerance mechanism for Spark over Kubernetes.

(a) Re-run your WordCount application from **Q4d**, but try to suddenly kill a Spark executor pod using the command line "kubectl delete pod xxxxxx" in the middle of its execution. Describe and explain your observations. You should use "kubectl get pods" to check the status of each pod with timestamps and use "kubectl log Spark-driver-pod" to figure out the fault-tolerance mechanism from logs.

(b) Re-run your WordCount application from **Q4d**, but try to suddenly kill the Spark driver pod in the middle of its execution. Describe and explain your observations. Please check the status of each pod with timestamps.

Reference:

[1] Spark
https://spark.apache.org/docs/latest/
[2] SNAP Google Web Data
https://snap.stanford.edu/data/web-Google.html
[3] Spark SQL programming guide
http://spark.apache.org/docs/1.6.0/sql-programming-guide.html
[4] Word Counting dataset (User:bigdata Password:spring2021bigdata)
http://mobitec.ie.cuhk.edu.hk/ierg4330Spring2021/homework/shakespeare_basket.zip
[5] Submitting Applications to Kubernetes:
https://spark.apache.org/docs/latest/running-on-kubernetes.html#docker-images
[6] Configure Kubernetes RBAC roles and service accounts:
https://spark.apache.org/docs/latest/running-on-kubernetes.html#rbac
[7] Launch Spark Pi in cluster mode:
https://spark.apache.org/docs/latest/running-on-kubernetes.html#cluster-mode
[8] Running Spark on YARN
https://spark.apache.org/docs/latest/running-on-yarn.html
[9] Spark over K8s tutorial
https://mobitec.ie.cuhk.edu.hk/ierg4330/tutorials/07_spark_over_k8s/