SE 4455B Lab #4: Hadoop with Docker - Due Tuesday, Apr. 4, 11:55 pm

Please ensure that all group members are able to complete this assignment. Please submit the required information only from one attempt among group members.

# Objectives:

1. Setting up a Hadoop cluster on Linux.
2. Running a sample application and validating results.
3. Create a mapper and reducer task for a data summarization task.
4. Run a data summarization task using a Hadoop cluster on Linux
5. Control basic operating parameters of a job.

# Grading Rubric

selins-MacBook-Pro:~ isacar$ ssh root@138.197.149.23

root@slackchat:~# ssh -p 995 [root@mim.eng.uwo.ca](mailto:root@mim.eng.uwo.ca)

purchases.txt is located in ~/storedata/purchases.txt in clack chat, and /usr/storedata/purchases.txt in mim.

|  |  |  |
| --- | --- | --- |
| Task | Assessment | Mark |
| **a.**  **> see what docker images exist**  **sudo docker images**  **> Install a Docker image**  sudo docker pull sequenceiq/hadoop-docker:2.7.1  selinuxenabled && echo enabled || echo disabled  **(if it’s disabled, continue if not:**  echo 0 > /selinux/enforce  to disable.)  **> Start docker container:**  **sudo docker run -it sequenceiq/hadoop-docker:2.7.1 /etc/bootstrap.sh -bash**  **> GO TO THE HADOOP PREFIX DIRECTORY! #must**  cd $HADOOP\_PREFIX  **> run the sample Hadoop task**  **> run map reduce**  bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar grep input output 'dfs[a-z.]+'  **> check the output**  bin/hdfs dfs -cat output/\*  <http://stackoverflow.com/questions/20821584/hadoop-2-2-installation-no-such-file-or-directory> | Needs improvement | 60 |
| b. Create mapper and reducer to rank stores based on purchase amounts.  c. Run this task and produce a ranked list of stores  <https://www.youtube.com/watch?v=MYo8EZwDRUA&list=PLAwxTw4SYaPkXJ6LAV96gH8yxIfGaN3H-&index=58>  **> see that input files exist**  # bin/Hadoop fs –ls  ~ Hadoop streaming lets you write mappers and reducers in any lang – so they’re in python  ~NOW TO TEST CODE!!!! RUN RUN RUN!  $ Call the mapper  ./mapper.py  $enter in different test fields i.e.  field1 f2 f3 f4 f5 f6  fieldA fB fC fD fE fF  $ctl+d to simulate end of input, will output the results of mapper:  f3 f5  fC fE  ~~!!! OR BETTER – create a small set data file and pipe it to the mapper!  $find the directory where the data is stored | Meets expectations | 80 |
| d. Create a map/reduce task to rank the payment methods by payment amount .  e. Setup port forwarding to monitor task progress from an external client | Exceeds expectations | 100 |

# Schedule:

Due on Tuesday, Apr. 4, 11:55 pm. Late submissions will be accepted until Apr. 6, 11:55pm (subject to 10% penalty per day).

# Submit commit ID on Owl before the deadline

1. Create a directory called ‘lab4’ in a regular user account (not root).
2. Initialize a git repository inside this directory.
3. Copy the Hadoop logs for all runs to this directory.
4. Copy the command history that covers lab 4 activities. E.g. . ‘**history > lab4/commands**’
5. Copy the data output produced by all Hadoop jobs. E.g. using ‘**hadoop fs ...**’ commands to copy output to a local file and then copy this file from Docker container to the ‘**lab4**’ directory on server.
6. Commit all the files in ‘lab4’ folder and submit the commit ID on Owl.
7. Demonstrate your lab to a TA before April 5.

# Procedure

Before you start, please ensure that all group members have regular (non-root) accounts on the server and that everyone can log in.

Note: Servers are not setup to run ‘sudo’ command by regular users. Hence, any command that requires root privileges (E.g. commands that start with ‘sudo’) must be run by separately logging in as root.

Some of these tasks require you to investigate possible solutions and implement a working solution. Please see the class notes for possible starting points.

1. Obtain the Hadoop container from <https://github.com/sequenceiq/hadoop-docker> and start a container based on this image.
2. Run the sample task from the ‘Howto’ document that accompanied this image.
3. Familiarize with basic Hadoop job control tasks including file copy operations from local file system HDFS and vice versa, starting a job etc.

Move files from local to there:

On my cpu:

scp /Users/isacar/Downloads/purchases.txt root@138.197.149.23:~/storedata

on slackchat:

scp /src/purchases.txt root@mim.eng.uwo.ca:/usr/storedata

or….

scp ~/storedata/ purchases.txt root@mim.eng.uwo.ca

https://www.youtube.com/watch?v=MYo8EZwDRUA&list=PLAwxTw4SYaPkXJ6LAV96gH8yxIfGaN3H-&index=58

1. Investigate methods to run a streaming task where mapper and reducer task take input from stdin and produce output on stdout. Get the ‘purchases.txt.gz’ dataset from Owl and create a map/reduce task to add all purchases by each store and execute this job.
2. Create another map/reduce task to rank the payment methods based on the payment amount.
3. Set up port forwarding to monitor progress of the Hadoop cluster.