

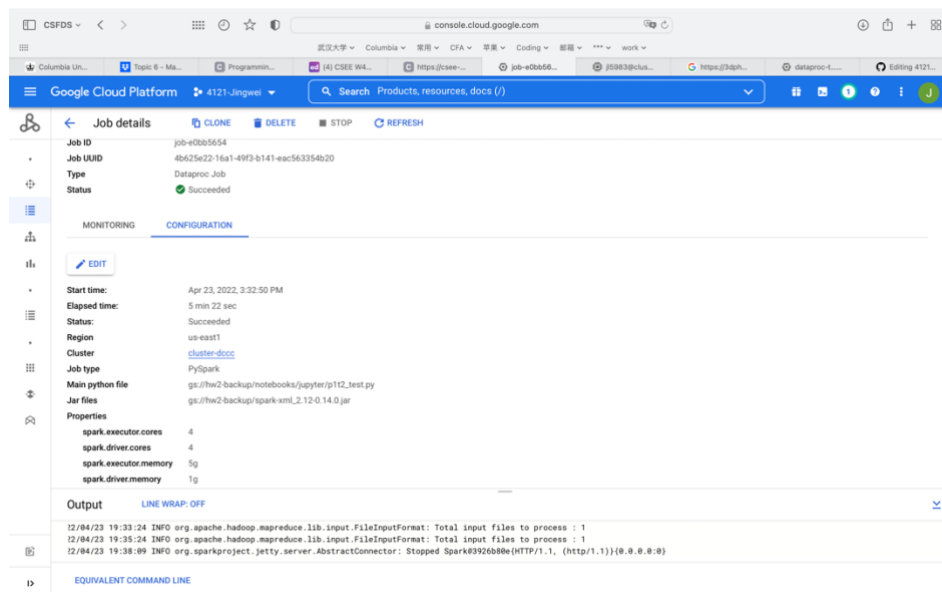
README

Question 1. (4 points) What is the default block size on HDFS? What is the default replication factor of HDFS on Dataproc?

default block-size: 128mb

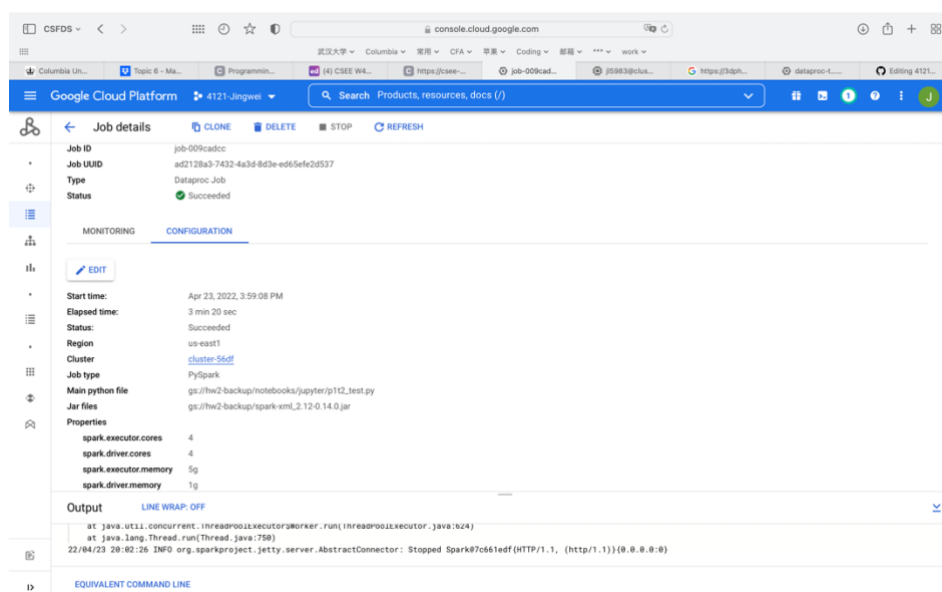
default replication factor: 2

Question 2. (2 points) Use enwiki_test.xml as input and run the program locally on a Single Node cluster using 4 cores. Include your screenshot of the dataproc job. What is the completion time of the task?



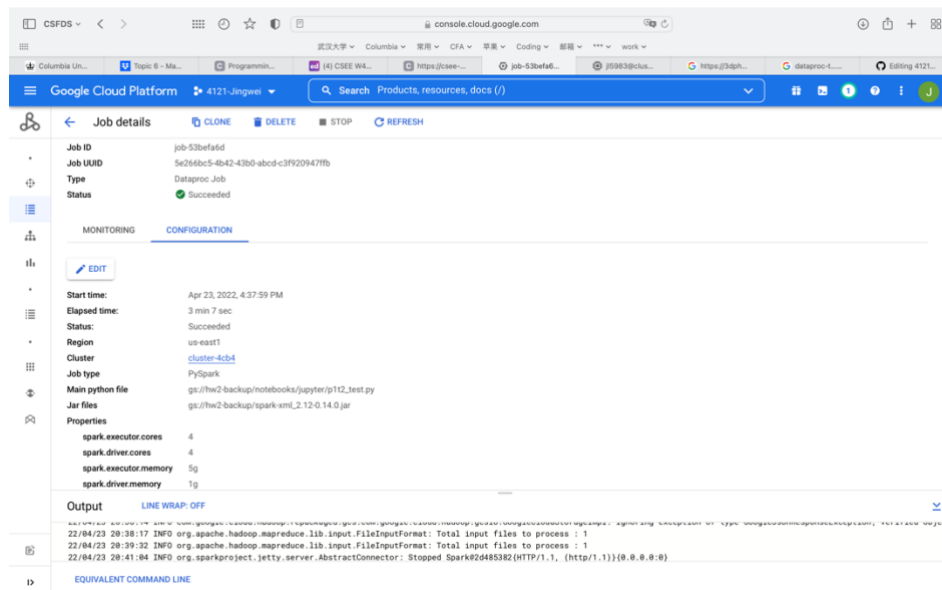
The completion time is 5mins 22secs.

Question 3. (2 points) Use enwiki_test.xml as input and run the program under HDFS inside a 3-node cluster (2 worker nodes). Include your screenshot of the dataproc job. Is the performance getting better or worse in terms of completion time? Briefly explain.



The performance is getting better. The extra 2 workers provide with more resources to finish the job. In a single node cluster, the master node not only need to manage the job, but also need to work on the job. But in a 3-node cluster, workers can work on the job and master node only need to manage those workers. This rapid the process.

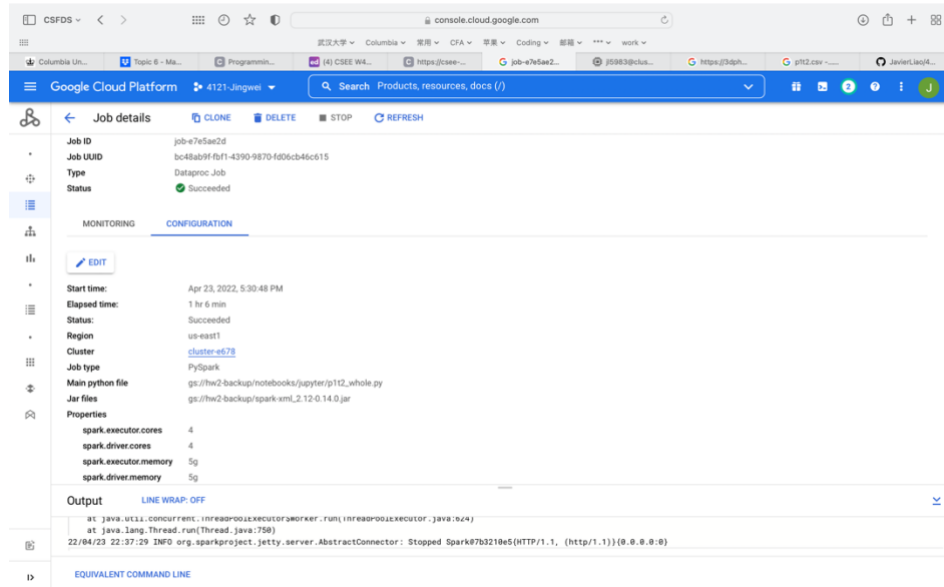
Question 4. (2 points) For this question, change the default block size in HDFS to be 64MB and repeat Question 3. Include your screenshot of the dataproc job. Record run time, is the performance getting better or worse in terms of completion time? Briefly explain.



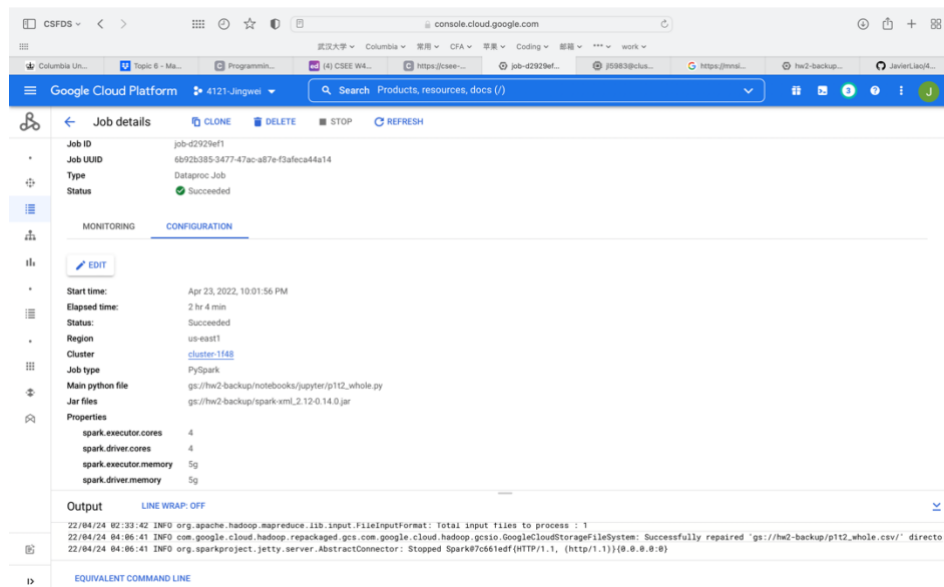
The performance is the best. However, in a common sense when we reduce the block size, it should get slower as we have much smaller space to cache. The faster result may be due to the small scale of the job.

Question 5. (2 points) Use `enwiki_whole.xml` as input and run the program under HDFS inside the Spark cluster you deployed. Record the completion time. Now, 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. Include your screenshot of the dataproc jobs.

Not Kill:

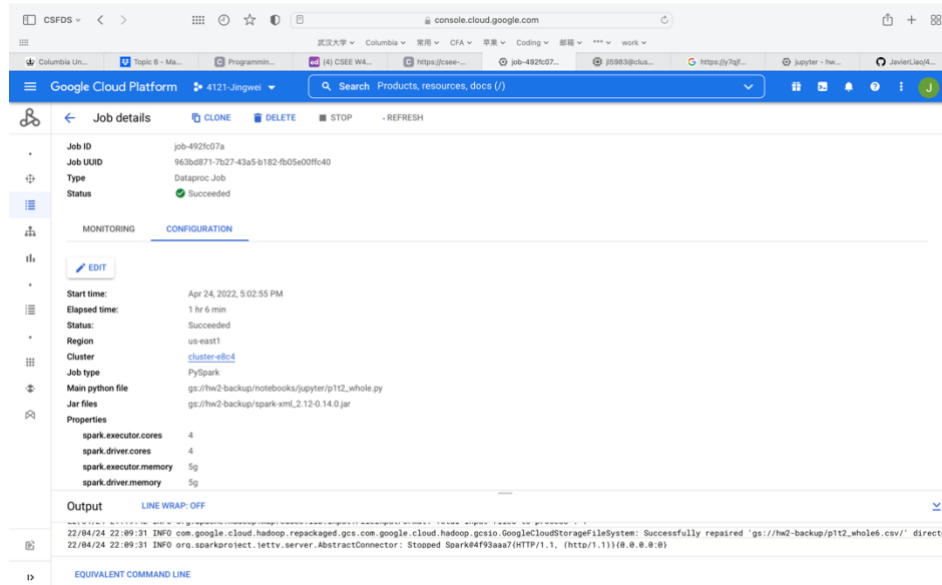


Kill in the process:



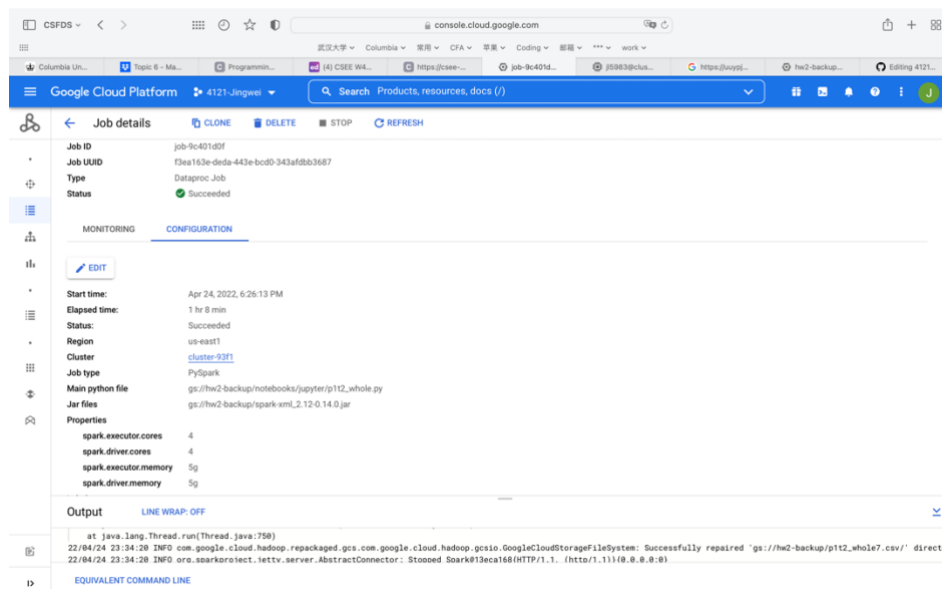
When we kill one of the workers, the job is still finished but the completion time doubled from about 1 hour to about 2 hours. When we kill one of the worker, there is only one left and naturally the time doubled.

Question 6. (2 points) Only for this question, change the replication factor of enwiki_whole.xml to 1 and repeat Question 5 without killing one of the worker nodes. Include your screenshot of the dataproc job. Do you observe any difference in the completion time? Briefly explain.



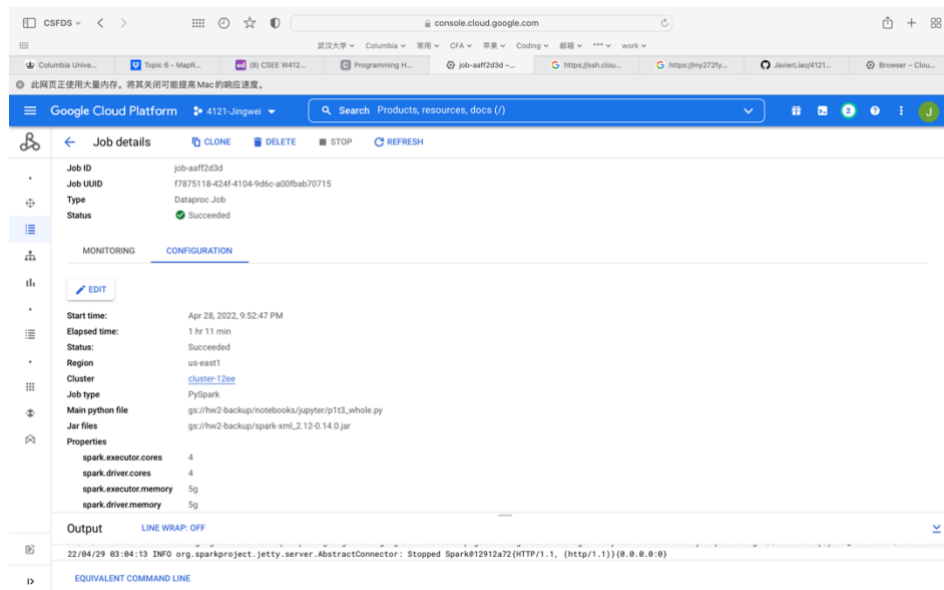
The completion time doesn't differ a lot. As we didn't encounter any break-down or network issue, the replication factor may not affect the completion time significantly.

Question 7. (2 points) Only for this question, change the default block size in HDFS to be 64MB and repeat Question 5 without killing one of the worker nodes. Record run time, include your screenshot of the dataproc job. Is the performance getting better or worse in terms of completion time? Briefly explain.



The completion time still doesn't differ a lot.

Question 8. (2 points) Use your output from Task 2 with enwiki_whole.xml as input, run Task 3 using a 3 node cluster (2 worker nodes). Include your screenshot of the dataproc job. What is the completion time of the task?



The completion time is 1hour 11mins

Questions 11 (2 points) How many hours did you spend in this assignment?

20hours