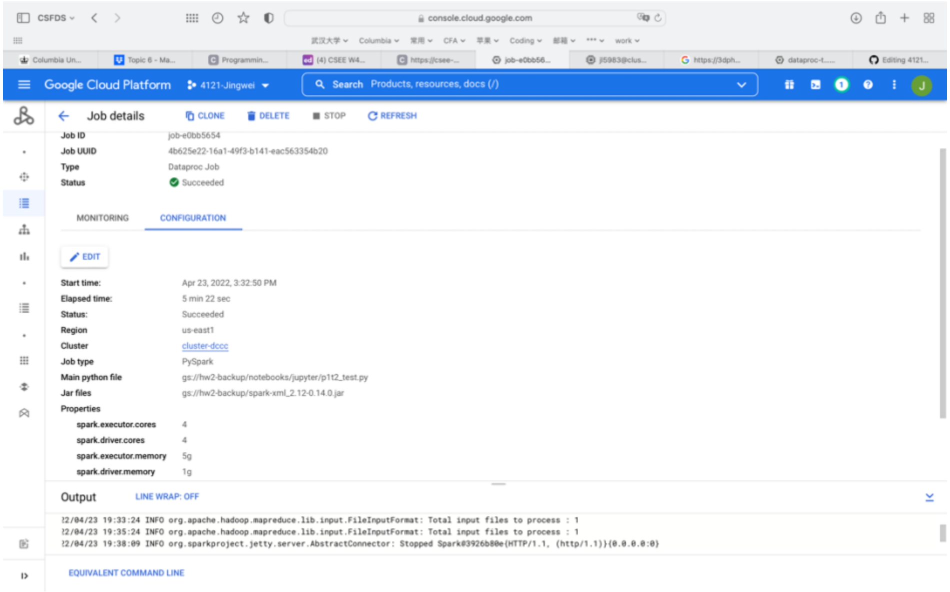
README

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?



Job ID job-e0bb5654
Job UUID 4b625e22-16a1-49f3-b141-eac563354b20
Type Dataproc Job
Status Succeeded

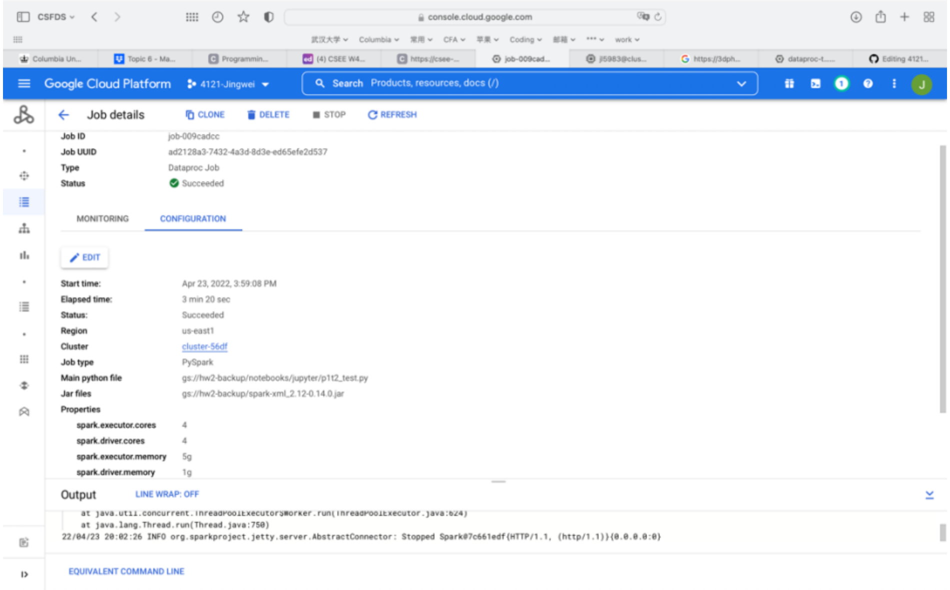
Start time: Apr 23, 2022, 3:32:50 PM
Elapsed time: 5 min 22 sec
Status: Succeeded
Region us-east1
Cluster cluster-dccc
Job type PySpark
Main python file gs://hw2-backup/notebooks/jupyter/p1t2_test.py
Jar files gs://hw2-backup/spark-xml_2.12-0.14.0.jar
Properties
spark.executor.cores 4
spark.driver.cores 4
spark.executor.memory 5g
spark.driver.memory 1g

Output

```
22/04/23 19:33:24 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/04/23 19:35:24 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/04/23 19:38:09 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@3926b80e{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

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.



Job ID job-009cadcc
Job UUID ad2128a3-7432-4a3d-8d3e-ed65efe2d537
Type Dataproc Job
Status Succeeded

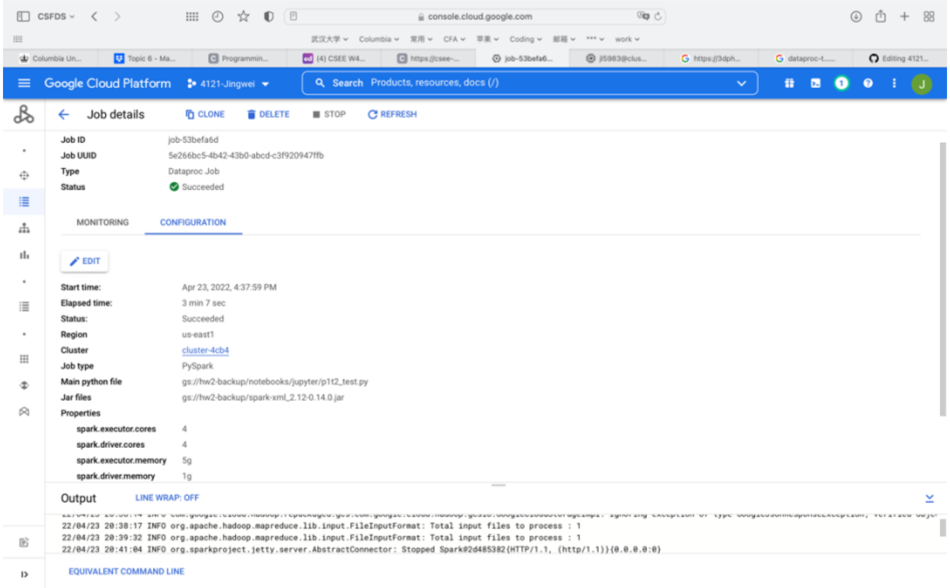
Start time: Apr 23, 2022, 3:59:08 PM
Elapsed time: 3 min 20 sec
Status: Succeeded
Region us-east1
Cluster cluster-56df
Job type PySpark
Main python file gs://hw2-backup/notebooks/jupyter/p1t2_test.py
Jar files gs://hw2-backup/spark-xml_2.12-0.14.0.jar
Properties
spark.executor.cores 4
spark.driver.cores 4
spark.executor.memory 5g
spark.driver.memory 1g

Output

```
	at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
	at java.lang.Thread.run(Thread.java:750)
22/04/23 20:02:26 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7c661edf{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

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.



CSFDS
console.cloud.google.com
武汉大学 Columbia 常用 CFA 苹果 Coding 邮箱 work
Columbia Un... Topic 6 - Ma... Programmin... (4) CSEE W4... https://csee-... job-53befa6... jl5983@clus... https://3dph... dataproc-t..... Editing 4121...
Google Cloud Platform
4121-Jingwei
Search Products, resources, docs (/)
Job details
CLONE
DELETE
STOP
REFRESH
MONITORING
CONFIGURATION
EDIT
Output
LINE WRAP: OFF
EQUIVALENT COMMAND LINE

Job ID job-53befa6d
Job UUID 5e266bc5-4b42-43b0-abcd-c3f920947ffb
Type Dataproc Job
Status Succeeded

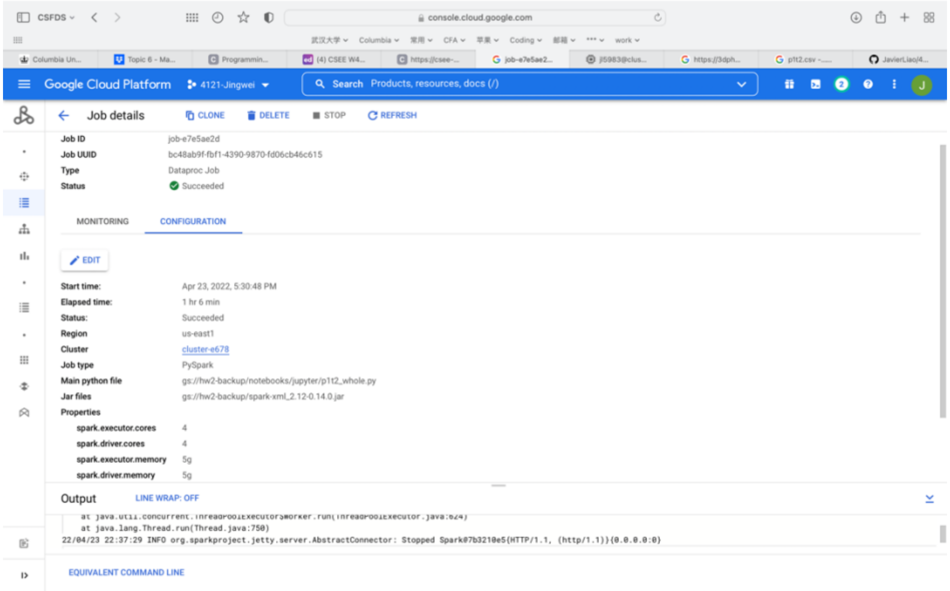
Start time: Apr 23, 2022, 4:37:59 PM
Elapsed time: 3 min 7 sec
Status: Succeeded
Region us-east1
Cluster cluster-4cb4
Job type PySpark
Main python file gs://hw2-backup/notebooks/jupyter/p1t2_test.py
Jar files gs://hw2-backup/spark-xml_2.12-0.14.0.jar
Properties
spark.executor.cores 4
spark.driver.cores 4
spark.executor.memory 5g
spark.driver.memory 1g

```
22/04/23 20:38:17 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/04/23 20:39:32 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/04/23 20:41:04 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@2d485382{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

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:



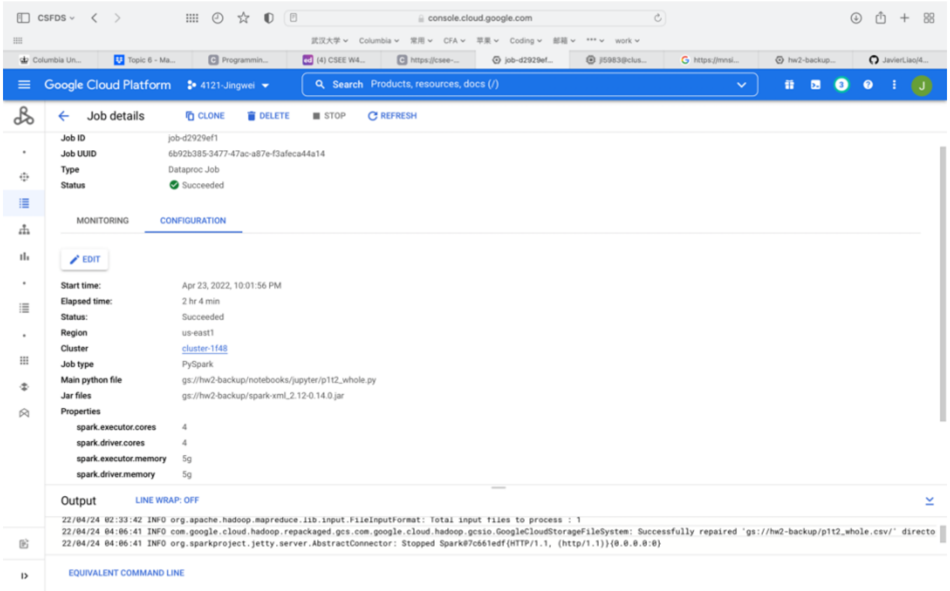
Job ID job-e7e5ae2d
Job UUID bc48ab9f-fbf1-4390-9870-fd06cb46c615
Type Dataproc Job
Status Succeeded

Start time: Apr 23, 2022, 5:30:48 PM
Elapsed time: 1 hr 6 min
Status: Succeeded
Region us-east1
Cluster cluster-e678
Job type PySpark
Main python file gs://hw2-backup/notebooks/jupyter/p1t2_whole.py
Jar files gs://hw2-backup/spark-xml_2.12-0.14.0.jar
Properties
spark.executor.cores 4
spark.driver.cores 4
spark.executor.memory 5g
spark.driver.memory 5g

Output

```
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)
22/04/23 22:37:29 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7b3210e5{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

Kill in the process:



Job ID job-d2929ef1
Job UUID 6b92b385-3477-47ac-a87e-f3afeca44a14
Type Dataproc Job
Status Succeeded

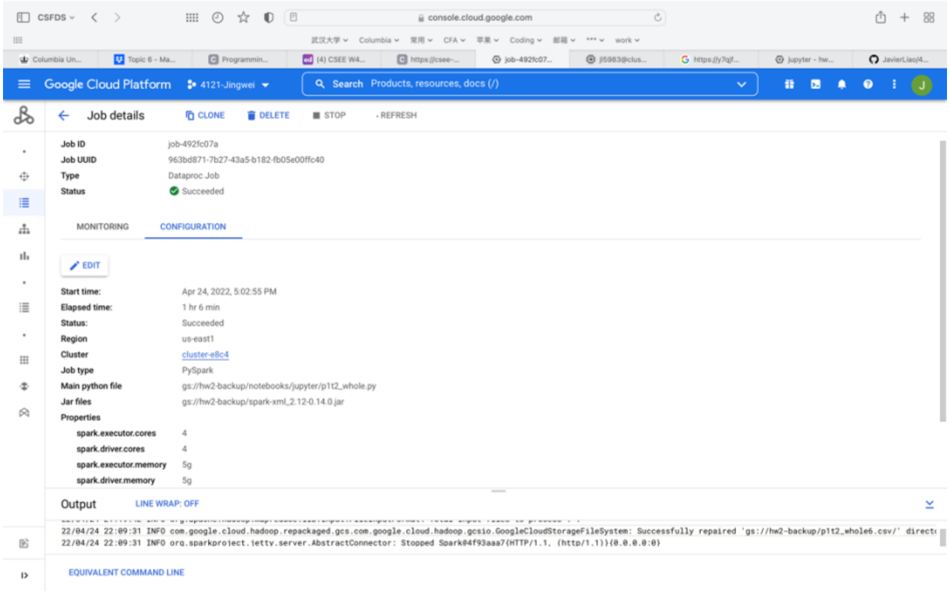
Start time: Apr 23, 2022, 10:01:56 PM
Elapsed time: 2 hr 4 min
Status: Succeeded
Region us-east1
Cluster cluster-1f48
Job type PySpark
Main python file gs://hw2-backup/notebooks/jupyter/p1t2_whole.py
Jar files gs://hw2-backup/spark-xml_2.12-0.14.0.jar
Properties
spark.executor.cores 4
spark.driver.cores 4
spark.executor.memory 5g
spark.driver.memory 5g

Output

```
22/04/24 02:33:42 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/04/24 04:06:41 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired 'gs://hw2-backup/p1t2_whole.csv/' directo
22/04/24 04:06:41 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7c661edf{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

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.



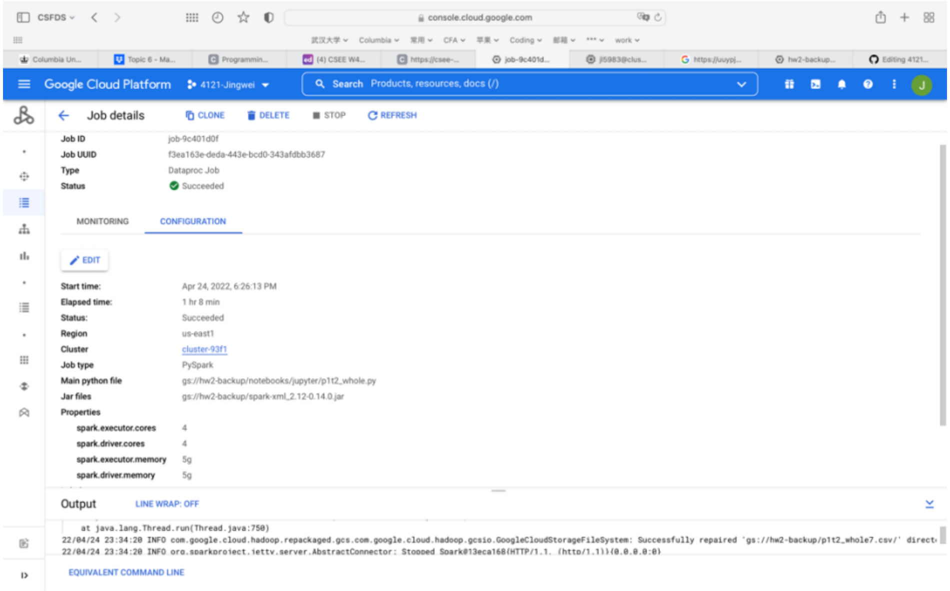
Job ID job-492fc07a
Job UUID 963bd871-7b27-43a5-b182-fb05e00ffc40
Type Dataproc Job
Status Succeeded

Start time: Apr 24, 2022, 5:02:55 PM
Elapsed time: 1 hr 6 min
Status: Succeeded
Region us-east1
Cluster cluster-e8c4
Job type PySpark
Main python file gs://hw2-backup/notebooks/jupyter/p1t2_whole.py
Jar files gs://hw2-backup/spark-xml_2.12-0.14.0.jar
Properties
spark.executor.cores 4
spark.driver.cores 4
spark.executor.memory 5g
spark.driver.memory 5g

```
22/04/24 22:09:31 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired 'gs://hw2-backup/p1t2_whole6.csv/' direct
22/04/24 22:09:31 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@4f93aaa7{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

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.



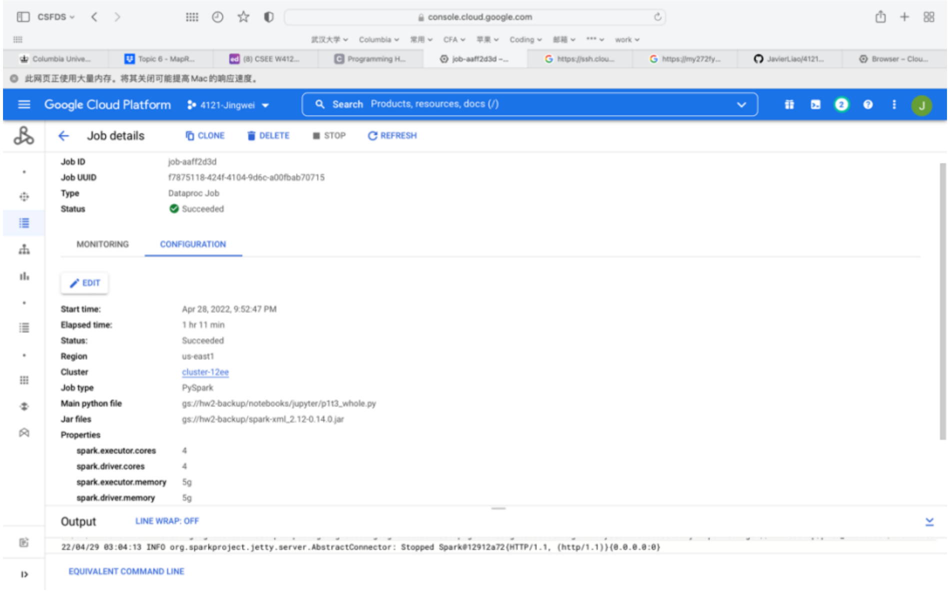
Job ID job-9c401d0f
Job UUID f3ea163e-deda-443e-bcd0-343afdbb3687
Type Dataproc Job
Status Succeeded

Start time: Apr 24, 2022, 6:26:13 PM
Elapsed time: 1 hr 8 min
Status: Succeeded
Region us-east1
Cluster cluster-93f1
Job type PySpark
Main python file gs://hw2-backup/notebooks/jupyter/p1t2_whole.py
Jar files gs://hw2-backup/spark-xml_2.12-0.14.0.jar
Properties
spark.executor.cores 4
spark.driver.cores 4
spark.executor.memory 5g
spark.driver.memory 5g

```
    at java.lang.Thread.run(Thread.java:750)
22/04/24 23:34:20 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired 'gs://hw2-backup/p1t2_whole7.csv/' direct
22/04/24 23:34:20 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@13eca168{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

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?


CSFDS
console.cloud.google.com
武汉大学 Columbia 常用 CFA 苹果 Coding 邮箱 *** work
Columbia Unive... Topic 6 - MapR... (8) CSEE W412... Programming H... job-aaff2d3d -... https://ssh.clou... https://my272fy... JavierLiao/4121... Browser – Clou...
此网页正使用大量内存。将其关闭可能提高 Mac 的响应速度。
Google Cloud Platform
4121-Jingwei
Search Products, resources, docs (/)
Job details
CLONE
DELETE
STOP
REFRESH
MONITORING
CONFIGURATION
EDIT
LINE WRAP: OFF
EQUIVALENT COMMAND LINE

Job ID	job-aaff2d3d
Job UUID	f7875118-424f-4104-9d6c-a00fbab70715
Type	Dataproc Job
Status	Succeeded

Start time:	Apr 28, 2022, 9:52:47 PM
Elapsed time:	1 hr 11 min
Status:	Succeeded
Region	us-east1
Cluster	cluster-12ee
Job type	PySpark
Main python file	gs://hw2-backup/notebooks/jupyter/p1t3_whole.py
Jar files	gs://hw2-backup/spark-xml_2.12-0.14.0.jar
Properties	
spark.executor.cores	4
spark.driver.cores	4
spark.executor.memory	5g
spark.driver.memory	5g

Output

```
22/04/29 03:04:13 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@12912a72{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

The completion time is 1hour 11mins