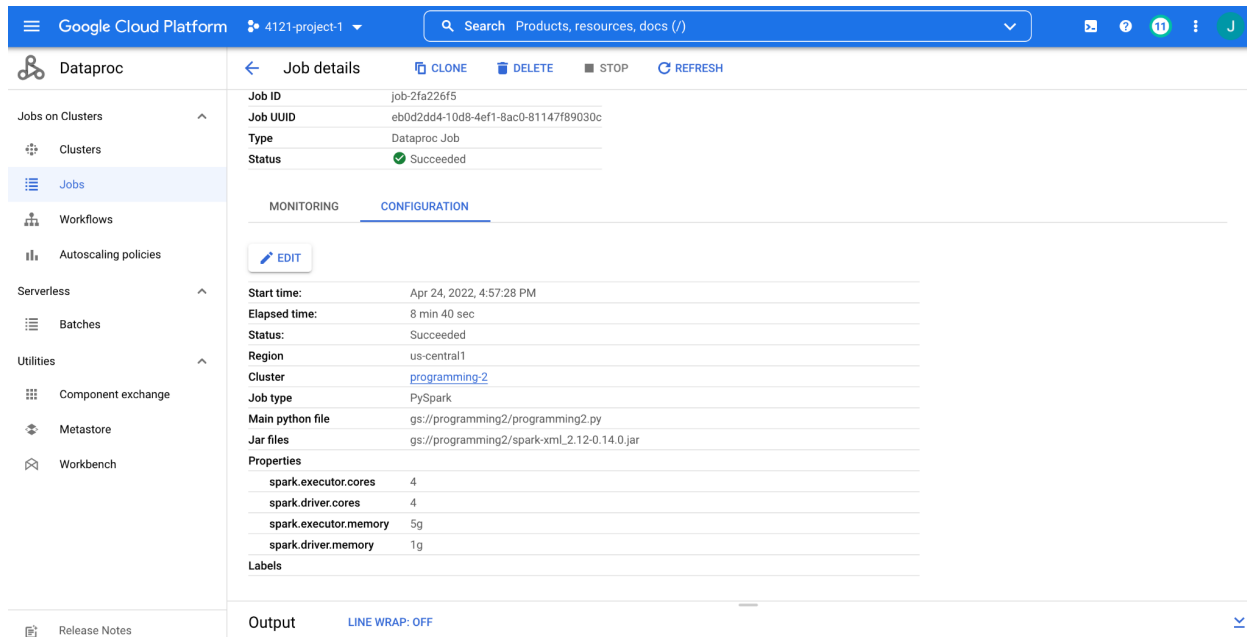


Computer System Programming 2 README

Junhao Zhang jz3430 and Yuning Di yd2617

Question 1: The default block size is 128MB and the default replication factor is 2

Question 2: The completion time is shown in the below graph: 8 min 40 sec



The screenshot shows the Google Cloud Platform interface for a Dataproc job. The left sidebar contains navigation links for Dataproc, Jobs on Clusters, Clusters, Jobs, Workflows, Autoscaling policies, Serverless, Batches, Utilities, Component exchange, Metastore, and Workbench. The main panel displays the 'Job details' for a specific job. The job is titled 'programming-2' and has a status of 'Succeeded'. The 'Elapsed time' is 8 min 40 sec. The 'Main python file' is 'gs://programming2/programming2.py'. The 'Jar files' are 'gs://programming2/spark-xml_2.12-0.14.0.jar'. The 'Properties' section shows 'spark.executor.cores' as 4, 'spark.driver.cores' as 4, 'spark.executor.memory' as 5g, and 'spark.driver.memory' as 1g. The 'Labels' section is empty. The 'Output' section is also empty.

Property	Value
spark.executor.cores	4
spark.driver.cores	4
spark.executor.memory	5g
spark.driver.memory	1g

Question 3: The completion time is 4 min 41 sec. The performance is getting better in terms of the completion time. This is because we are using 3 node clusters with 2 works and thus it's more efficient than the previous problem.

The screenshot displays the Google Cloud Platform interface for a Dataproc job. The left sidebar shows the navigation menu with 'Jobs' selected. The main panel shows the 'Job details' for a specific job, including its ID, UUID, type, and status (Succeeded). Below this, there are tabs for 'MONITORING' and 'CONFIGURATION'. The 'CONFIGURATION' tab is active, showing an 'EDIT' button and various job parameters such as start time, elapsed time, status, region, cluster, job type, main python file, jar files, properties (spark.executor.cores, spark.driver.cores, spark.executor.memory, spark.driver.memory), and labels. The 'Output' section at the bottom is partially visible.

Job Details	
Job ID	job-f3f69c81
Job UUID	beb0dd76-20e4-48e7-a264-a2567a7596d7
Type	Dataproc Job
Status	Succeeded

Configuration	
Start time	Apr 24, 2022, 5:39:00 PM
Elapsed time	4 min 41 sec
Status	Succeeded
Region	us-central1
Cluster	programming-3
Job type	PySpark
Main python file	gs://programming2/programming2.py
Jar files	gs://programming2/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
Labels	

Question 4: The completion time is 4 min 44 sec. We can see that even though we decreased the block size from 128MB to 64 MB, the completion time has not increased too much, only a few seconds longer than before. This makes sense because the decrease in block size will increase the completion time. On the other hand, the change in block size is not big enough to generate a significant effect on the performance and that's why the time difference between question 4 and question 3 is not too big.

Google Cloud Platform 4121-project-1 Search Products, resources, docs (/)

Dataproc Job details CLONE DELETE STOP REFRESH

Jobs on Clusters Clusters **Jobs** Workflows Autoscaling policies Serverless Batches Utilities Component exchange Metastore Workbench

Job ID: job-80908d1f
 Job UUID: 761117c7-2fe9-4d86-95ae-7a379166bc14
 Type: Dataproc Job
 Status: ✔ Succeeded

MONITORING CONFIGURATION

[EDIT](#)

Start time: Apr 24, 2022, 5:11:10 PM
 Elapsed time: 4 min 44 sec
 Status: Succeeded
 Region: us-central1
 Cluster: [programming-4](#)
 Job type: PySpark
 Main python file: gs://programming2/programming2.py
 Jar files: gs://programming2/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

Labels

Release Notes Output [LINE WRAP: OFF](#)

Question 5: We can see from the below two graphs that the job still finishes even though 1 worker is killed. The completion time of the normal cluster is around 1 hour and the completion time of the other is around 2 hour 2 min. This makes sense because the normal cluster has 2 workers. The other has only 1 worker. When we kill 1 worker, the efficiency will decrease and the time it takes to finish the job should be 2 times longer than the normal one, which matches the result.

1. Normal cluster with 2 workers

Google Cloud Platform 4121-project-1 Search Products, resources, docs (/)

Dataproc

Jobs on Clusters Clusters **Jobs** Workflows Autoscaling policies

Serverless Batches Utilities Component exchange Metastore Workbench

Release Notes

Job details CLONE DELETE STOP REFRESH

Job ID: job-318720a6
Job UUID: 1e78442f-e338-4285-85f3-9f4cfa95ba92
Type: Dataproc Job
Status: Succeeded

MONITORING CONFIGURATION

EDIT

Start time: Apr 24, 2022, 5:49:14 PM
Elapsed time: 1 hr 57 sec
Status: Succeeded
Region: us-central1
Cluster: [programming-3](#)
Job type: PySpark
Main python file: gs://programming2/programming2.py
Jar files: gs://programming2/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

Labels

EQUIVALENT REST

Output LINE WRAP: OFF

2. Cluster with 1 worker killed

Google Cloud Platform 4121-project-1 Search Products, resources, docs (/)

Dataproc

Jobs on Clusters Clusters **Jobs** Workflows Autoscaling policies

Serverless Batches Utilities Component exchange Metastore Workbench

Release Notes

Job details CLONE DELETE STOP REFRESH

Job ID: job-9906faf6
Job UUID: caa85f23-df66-41e2-b240-691869f559a2
Type: Dataproc Job
Status: Succeeded

MONITORING CONFIGURATION

EDIT

Start time: Apr 24, 2022, 7:01:37 PM
Elapsed time: 2 hr 2 min
Status: Succeeded
Region: us-central1
Cluster: [programming-3](#)
Job type: PySpark
Main python file: gs://programming2/programming2.py
Jar files: gs://programming2/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

Labels

EQUIVALENT REST

Output LINE WRAP: OFF

Question 6: The completion time is 1 hour 3 min and it's slower than question 5 but the difference in completion time is not too big. Recall from the lecture that by increasing the replication factor, the resource manager is more likely to find a node that stores the data. So if we decrease the replication factor from 2 to 1, then the resource manager is less likely to find a

node that stores the data and that's why the performance gets slower in terms of completion time.

The screenshot shows the Google Cloud Platform interface for a Dataproc job. The left sidebar contains navigation links for Jobs on Clusters, Clusters, Jobs, Workflows, Autoscaling policies, Serverless, Batches, Utilities, Component exchange, Metastore, and Workbench. The main content area is titled 'Job details' and includes buttons for CLONE, DELETE, STOP, and REFRESH. The job information is as follows:

Property	Value
Job ID	job-f42cd414
Job UUID	610a1a41-f172-4728-9b8b-1be2499749fa
Type	Dataproc Job
Status	Succeeded

Below the job information, there are tabs for MONITORING and CONFIGURATION. The CONFIGURATION tab is active, showing an EDIT button and the following details:

Property	Value
Start time	Apr 24, 2022, 9:36:57 PM
Elapsed time	1 hr 3 min
Status	Succeeded
Region	us-central1
Cluster	programming
Job type	PySpark
Main python file	gs://programming2/programming2.py
Jar files	gs://programming2/spark-xml_2.12-0.14.0.jar

Under the Properties section, the following configuration is listed:

Property	Value
spark.executor.cores	4
spark.driver.cores	4
spark.executor.memory	5g
spark.driver.memory	5g

At the bottom, there is a Labels section and a link to EQUIVALENT REST.

Question 7: The performance is getting a little bit worse after changing the block size 64MB since the completion time is longer than question 5. As mentioned in the previous problem, decreasing the block size will decrease the efficiency and slow the performance and that's why the time to complete question 7 is a little bit longer than question 5. Since the change in block size is not too big, the time difference between question 7 and 5 is not too big as well.

Google Cloud Platform 4121-project-1 Search Products, resources, docs (/)

Dataproc

Jobs on Clusters Clusters **Jobs** Workflows Autoscaling policies

Serverless Batches Utilities Component exchange Metastore Workbench

Release Notes

Job details CLONE DELETE STOP REFRESH

Job ID: job-18b5835d
Job UUID: 8e4c9813-a9a1-4fe7-9717-21601a345de4
Type: Dataproc Job
Status: Succeeded

MONITORING CONFIGURATION

EDIT

Start time: Apr 24, 2022, 11:03:55 PM
Elapsed time: 1 hr 5 min
Status: Succeeded
Region: us-central1
Cluster: [programming](#)
Job type: PySpark
Main python file: gs://programming2/programming2.py
Jar files: gs://programming2/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

Labels

EQUIVALENT REST

Question 8: The completion time is 34 min 17 sec.

Google Cloud Platform 4121-project-1 Search Products, resources, docs (/)

Dataproc

Jobs on Clusters Clusters **Jobs** Workflows Autoscaling policies

Serverless Utilities Component exchange Metastore Workbench

Release Notes

Job details CLONE DELETE STOP REFRESH

Job ID: job-e20809a1
Job UUID: 309d788b-df8c-48de-8d14-e8137f235bef
Type: Dataproc Job
Status: Succeeded

MONITORING CONFIGURATION

EDIT

Start time: May 1, 2022, 10:37:57 PM
Elapsed time: 34 min 17 sec
Status: Succeeded
Region: us-central1
Cluster: [t3](#)
Job type: PySpark
Main python file: gs://programming2/p1t3.py
Jar files: gs://programming2/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

Labels

EQUIVALENT REST

Output LINE WRAP: OFF

Question 9:

There are 2675032 articles in the database that have a rank greater than 0.5.

Question 10: It is feasible and efficient. Using TCP sockets, we can guarantee that the streaming data we received is complete. And we can also identify different data in the same

directory (folder). For file streams, the official document says “Once moved, the files must not be changed. So if the files are being continuously appended, the new data will not be read”.

However, the TCP socket can be always listened to receive continuous appended data.

Sadly, we have to bear flow control of TCP which means that the throughput might be not very large.

Question 11: For this extra credit assignment, we spent one night on it. We spent three days on this whole assignment,