

README

Date	@May 1, 2022
Tag	
Property	@April 30, 2022 10:28 PM

Cluster Information to Prove Our Work

We used three clusters in total. To prove our work, we will list out three clusters name and the configuration below. (Please ignore the single node cluster stopped at this point. I took screenshots of them after task 2.)

- (type, name of cluster)
- (Single Node Cluster, cluster-d729)

The screenshot shows the configuration details for a Dataproc Cluster named 'cluster-d729'. The cluster has a UUID of 'e7690f3c-5a1e-4df6-a42e-0f58f70c52af' and is currently stopped. It is located in the 'us-central1' region and zone 'us-central1-f'. The master node is a single node (1 master, 0 workers) with an n1-standard-4 machine type, 0 GPUs, and 500GB of primary disk space. The cluster uses pd-standard disk type and has 0 local SSDs. Secure boot and VTTPM are disabled. Integrity monitoring is also disabled. The cloud storage staging bucket is 'hw4_spark_ak'. The network is set to 'default' and there are no network tags. Internal IP only is set to 'No'. The configuration tab is selected.

Name	cluster-d729
Cluster UUID	e7690f3c-5a1e-4df6-a42e-0f58f70c52af
Type	Dataproc Cluster
Status	Stopped

MONITORING JOBS VM INSTANCES **CONFIGURATION** WEB INTERFACES

EDIT

Region	us-central1
Zone	us-central1-f
Autoscaling	Off
Dataproc Metastore	None
Scheduled deletion	Off
Master node	Single Node (1 master, 0 workers)
Machine type	n1-standard-4
Number of GPUs	0
Primary disk type	pd-standard
Primary disk size	500GB
Local SSDs	0
Secure Boot	Disabled
VTTPM	Disabled
Integrity Monitoring	Disabled
Cloud Storage staging bucket	hw4_spark_ak
Network	default
Network tags	None
Internal IP only	No

- (3 node 2 worker cluster, cluster-35f7)

Name	cluster-35f7
Cluster UUID	3a086c00-ba10-4df5-a600-bfee48728014
Type	Dataproc Cluster
Status	Running

MONITORING JOBS VM INSTANCES **CONFIGURATION** WEB INTERFACES

 EDIT

Region	us-central1
Zone	us-central1-c
Autoscaling	Off
Dataproc Metastore	None
Scheduled deletion	Off
Master node	Standard (1 master, N workers)
Machine type	n1-standard-4
Number of GPUs	0
Primary disk type	pd-standard
Primary disk size	500GB
Local SSDs	0
Worker nodes	2
Machine type	n1-standard-4
Number of GPUs	0
Primary disk type	pd-standard
Primary disk size	500GB
Local SSDs	0
Secondary worker nodes	0

- (3 node 2 worker cluster, cluster-b400)

Name	cluster-b400
Cluster UUID	d5d454b0-0c95-4bef-9463-6b8b9b46963b
Type	Dataproc Cluster
Status	Running
<hr/>	
MONITORING	JOBs
VM INSTANCES	CONFIGURATION
<hr/>	
 EDIT	
<hr/>	
Region	us-central1
Zone	us-central1-b
Autoscaling	Off
Dataproc Metastore	None
Scheduled deletion	Off
Master node	Standard (1 master, N workers)
Machine type	n1-standard-4
Number of GPUs	0
Primary disk type	pd-standard
Primary disk size	500GB
Local SSDs	0
Worker nodes	2
Machine type	n1-standard-4
Number of GPUs	0
Primary disk type	pd-standard
Primary disk size	500GB
Local SSDs	0
Secondary worker nodes	0
Secure Boot	Disabled
VTPM	Disabled
Integrity Monitoring	Disabled
Cloud Storage staging bucket	hw4_spark_ak

Question 1

The default block size of HDFS is 128 MB. The default replication factor is 2.

Question 2

completion time: 5 min 6 sec

Job details for job-a387a3c1

Job ID: job-a387a3c1
Job UUID: a98e4a04-ce85-4b06-bf66-115f5c245139
Type: Dataproc Job
Status: Succeeded

MONITORING **CONFIGURATION**

The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for more than one job to run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

RESET ZOOM **1 hour** **6 hours** **12 hours** **1 day** **2 days** **4 days** **7 days** **14 days** **30 days** **✓ 10:20 PM - 10:26 PM**

YARN memory 15GiB **YARN pending memory** 15GiB **YARN NodeManagers** 1

Output **LINE WRAP: OFF**

```
22/05/01 02:21:18 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists.
22/05/01 02:21:14 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 02:23:29 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 02:23:29 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedIOException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:519)
    at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
    at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
    at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:98)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)
22/05/01 02:25:57 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@5b04000B{HTTP/1.1, {http/1.1}}{0.0.0.0:0}
```

Output is complete

EQUIVALENT COMMAND LINE

Job details for job-a387a3c1

Job ID: job-a387a3c1
Job UUID: a98e4a04-ce85-4b06-bf66-115f5c245139
Type: Dataproc Job
Status: Succeeded

MONITORING **CONFIGURATION**

EDIT

Start time: Apr 30, 2022, 10:20:56 PM
Elapsed time: 5 min 6 sec

Status: Succeeded
Region: us-central1
Cluster: cluster-d729
Job type: PySpark
Main python file: gs://hw4_spark_ak/notebooks/jupyter/hw2-q2-test.py
Jar files: gs://csee4121/homework2/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:2

Output **LINE WRAP: OFF**

Output is complete

EQUIVALENT COMMAND LINE

Question 3

completion time: 2 min 55 sec

The completion time is shorter than the single node cluster. That is because we have two worker node on top of one master so we can parallelize our tasks.

The screenshot shows the Google Cloud Platform DataProc Job details page for a job named 'job-c3986184'. The job ID is 'job-c3986184' and the job UUID is '9cc255f5-382e-4280-b8e3-5f5e92aa1207'. The status is 'Succeeded'. The monitoring tab displays three line charts: YARN memory, YARN pending memory, and YARN NodeManagers. The YARN memory chart shows a step increase from 15GiB to 20GiB, then to 25GiB. The YARN pending memory chart shows a step increase from 15GiB to 20GiB. The YARN NodeManagers chart shows a constant value of 2. Below the charts, the log output shows:

```

at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:1157)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:750)
22/05/01 01:37:16 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 01:38:33 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 01:39:55 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@13eca168(HTTP/1.1, {http://1.1}){0.0.0.0:0}

```

The output is complete.

EQUIVALENT COMMAND LINE

```

$ spark-submit --py-files gs://hw4_spark_ak/notebooks/jupyter/hw2-q3-test.py

```

The configuration tab shows the following parameters:

Parameter	Value
Start time	Apr 30, 2022, 9:37:02 PM
Elapsed time	2 min 55 sec
Status	Succeeded
Region	us-central1
Cluster	cluster-357
Job type	PySpark
Main python file	gs://hw4_spark_ak/notebooks/jupyter/hw2-q3-test.py
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties	spark.executor.memory: 5g spark.driver.memory: 1g
Labels	question:3

EQUIVALENT REST

Output

```

22/05/01 01:39:55 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@13eca168(HTTP/1.1, {http://1.1}){0.0.0.0:0}

```

Question 4

completion time: 2 min 50 sec

The completion time is not too different from question 3 even though we change block size from 128MB to 64MB.

Google Cloud Platform Job details for job-c18c2766

Job ID: job-c18c2766
Job UUID: 5c93da05-3d2c-4418-bb89-34e884413bc9
Type: DataProc Job
Status: Succeeded

MONITORING **CONFIGURATION**

The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for more than one job to run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

RESET ZOOM

1 hour 6 hours 12 hours 1 day 2 days 4 days 7 days 14 days 30 days ✓ 9:59 PM - 10:04 PM

YARN memory **YARN pending memory** **YARN NodeManagers**

Output LINE WRAP: OFF

```

at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:750)
22/05/01 02:08:46 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #0,5,main]) interrupted:
java.lang.InterruptedException
at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:98)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:750)
22/05/01 02:08:46 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 02:02:02 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 02:03:21 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7c561edf{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
```

Output is complete

EQUIVALENT COMMAND LINE

Google Cloud Platform Job details for job-c18c2766

Job ID: job-c18c2766
Job UUID: 5c93da05-3d2c-4418-bb89-34e884413bc9
Type: DataProc Job
Status: Succeeded

EDIT

Start time: Apr 30, 2022, 10:00:33 PM
Elapsed time: 2 min 50 sec
Status: Succeeded
Region: us-central1
Cluster: cluster-35f7
Job type: PySpark
Main python file: gs://dee4_spark_ak/notebooks/jupyter/hw2-q4-test.py
Jar files: gs://dee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties

- spark.executor.memory 5g
- spark.driver.memory 1g
- dfs.block.size 64m

Labels question:4

EQUIVALENT REST

Output LINE WRAP: OFF

Output is complete

Question 5

Completion time: 1 hr 9 min

It still finished the job even though we kill one of the worker node. However, it took much longer than Question 6 or 7 which did not kill the worker node.

This screenshot shows the Google Cloud Platform Job details page for a Dataproc job named 'job-c896fca2'. The job has a status of 'Succeeded'. The monitoring section displays three line charts: 'YARN memory' (25GB), 'YARN pending memory' (800GB), and 'YARN NodeManagers' (2). Below the charts, the log output shows a single error message:

```
22/05/01 21:54:38 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 3 on cluster-35f7-w-0.c.spark-348122.internal: Container marked as failed: container_1651364765876_0006_01
22/05/01 22:27:21 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 23:03:21 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark02d485382{HTTP/1.1, {http://0.0.0.0:0}}
```

This screenshot shows the Google Cloud Platform Job details page for the same job. The configuration section includes fields for Start time (May 1, 2022, 5:54:15 PM), Elapsed time (1 hr 9 min), Status (Succeeded), Region (us-central1), Cluster (cluster-35f7), Job type (PySpark), Main python file (gs://hw4_spark_ak/notebooks/jupyter/hw2-q5-whole.py), Jar files (gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar), Properties (spark.driver.memory: 5g, spark.executor.memory: 5g), and Labels (question: 5). The log output is identical to the previous screenshot.

Question 6

Completion time: 35 min 7 sec

By not killing the worker node, it significantly improved the completion time. Changing replication factor to 1 does not affect the performance significantly.

This screenshot shows the Google Cloud Platform DataProc Job details page for a job named 'job-dc252cb7'. The left sidebar lists various DataProc components: Clusters, Jobs (selected), Workflows, Autoscaling policies, Serverless, Batches, Utilities, Component exchange, Metastore, and Workbench. The main area displays monitoring charts for YARN memory, YARN pending memory, and YARN NodeManagers, along with a log output window showing successful execution of a Python script. A note at the top of the log indicates that metrics may not reflect the job's resource usage accurately due to multiple jobs running on a cluster.

Job ID: job-dc252cb7
Job UUID: aec681b1-8169-4e8e-afee-d64f0d5c451d
Type: Dataproc Job
Status: Succeeded

MONITORING CONFIGURATION

The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for more than one job to run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

RESET ZOOM 1 hour 6 hours 12 hours 1 day 2 days 4 days 7 days 14 days 30 days ✓ 10:52 PM - 11:27 PM

YARN memory YARN pending memory YARN NodeManagers

Output LINE WRAP OFF

```
22/05/02 02:52:29 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: inread (/inread/getfileinto #1,5,main)) interrupted:  
java.lang.InterruptedException  
at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)  
at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)  
at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)  
at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)  
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)  
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)  
at java.lang.Thread.run(Thread.java:750)  
22/05/02 02:52:29 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1  
22/05/02 03:09:08 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1  
22/05/02 03:27:19 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@4f93aaa7(HTTP/1.1, (http/1.1)){0.0.0.0:0}
```

Output is complete

EQUIVALENT COMMAND LINE

This screenshot shows the Google Cloud Platform DataProc Job details page for the same job ('job-dc252cb7'). The left sidebar is identical to the previous screenshot. The main area now shows configuration settings under the 'CONFIGURATION' tab. The 'EDIT' button is highlighted. Configuration parameters include Start time (May 1, 2022, 10:52:22 PM), Elapsed time (35 min 7 sec), Status (Succeeded), Region (us-central1), Cluster (cluster-35f7), Job type (PySpark), Main python file (gs://hw4_spark_ak/notebooks/jupyter/hw2-q6-whole.py), Jar files (gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar), Properties (spark.driver.memory: 5g, spark.executor.memory: 5g, dfs.replication: 1), and Labels (question: 6). The log output window shows the same successful execution of the Python script.

Job ID: job-dc252cb7
Job UUID: aec681b1-8169-4e8e-afee-d64f0d5c451d
Type: Dataproc Job
Status: Succeeded

MONITORING CONFIGURATION

EDIT

Start time: May 1, 2022, 10:52:22 PM
Elapsed time: 35 min 7 sec
Status: Succeeded
Region: us-central1
Cluster: cluster-35f7
Job type: PySpark
Main python file: gs://hw4_spark_ak/notebooks/jupyter/hw2-q6-whole.py
Jar files: gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties: spark.driver.memory: 5g, spark.executor.memory: 5g, dfs.replication: 1
Labels: question: 6

Output LINE WRAP OFF

```
22/05/02 03:09:08 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1  
22/05/02 03:27:19 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@4f93aaa7(HTTP/1.1, (http/1.1)){0.0.0.0:0}
```

Output is complete

Question 7

Completion time: 34 min 43 sec.

Even though we decrease the block size, we do not observe the worse the completion time. In fact, it got the better completion compared than Q5 and Q6.

Job details for job-36e62461

MONITORING

YARN memory: 25GB (Actual), 20GB (Target), 15GB (Min)

YARN pending memory: 800GB (Actual), 600GB (Target)

YARN NodeManagers: 1.5 (Actual), 2 (Target)

Output

```
22/05/01 02:59:48 INFO org.apache.hadoop.com.Configuration: resource-types.xml not found
22/05/01 02:59:48 INFO org.apache.hadoop.util.resource.ResourceFile: Unable to find 'resource-types.xml'.
22/05/01 02:59:48 INFO org.apache.hadoop.yarn.api.impl.YarnClientImpl: Submitted application application_1651364765876_0005
22/05/01 02:59:50 INFO org.apache.hadoop.yarn.client.proxy.ProxyAllocation: To queue queue at cluster-35f7-m/10.128.0.11:8880
22/05/01 02:59:52 INFO com.google.cloud.hadoop.mapreduce.gcs.GcsGoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists.
22/05/01 02:59:54 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 03:16:21 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 03:34:21 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@74100b16{HTTP/1.1}{0.0.0.0:0}
```

Release Notes

Output is complete

Job details for job-36e62461

CONFIGURATION

EDIT

Properties

dfs.block.size	64m
spark.driver.memory	5g
spark.executor.memory	5g

Labels

question:7

Output

```
22/05/01 02:59:54 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 03:16:21 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 03:34:21 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@74100b16{HTTP/1.1}{0.0.0.0:0}
```

Release Notes

Output is complete

Question 8

Completion time: 58 min 55 sec

The screenshot shows the Google Cloud Platform DataProc Job details page. The left sidebar lists various services: Jobs on Clusters, Clusters, Jobs (selected), Workflows, Autoscaling policies, Serverless, Batches, Utilities, Component exchange, Metastore, and Workbench. The main area displays job information: Job ID (job-1d6806c1), Job UUID (d6fd5184-783c-4821-b3d2-1518f04e977b), Type (Dataproc Job), and Status (Succeeded). Below this are tabs for MONITORING and CONFIGURATION. The MONITORING tab shows three charts: YARN memory (20GB), YARN pending memory (1.5TB), and YARN NodeManagers (2). The CONFIGURATION tab shows the equivalent command line used to run the job.

```

at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromFutureExecutor(ExecutorHelper.java:48)
at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:750)
22/05/02 18:27:58 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark3@766858{HTTP/1.1}/{http/1.1}(0.0.0.0:0)

```

Job ID	job-1d6806c1
Job UUID	d6fd5184-783c-4821-b3d2-1518f04e977b
Type	Dataproc Job
Status	✓ Succeeded

MONITORING		CONFIGURATION
EDIT		
Start time:	May 2, 2022, 1:29:16 PM	
Elapsed time:	58 min 55 sec	
Status:	Succeeded	
Region	us-central1	
Cluster	cluster-35f	
Job type	PySpark	
Main python file	gs://hw4_spark_ak/notebooks/jupyter/hw2-q8-whole.py	
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar	
Properties		
spark.executor.memory	5g	
spark.driver.memory	5g	
Labels	question : 8	