Part 1: Spark and Spark SQL

Task 1: Getting Started (10 Points)

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

Ans: The default block size on HDFS is 128MB; The default replication factor is 3.

Question 2: Cluster Screenshot:(block size: 128MB; single node with 4 cores; others are default)

Cluster details SUBMIT JOB REFRESH START STOP DELETE

EDIT

Region	us-central1
Zone	us-central1-c
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
VTPM	Disabled
Integrity Monitoring	Disabled
Cloud Storage staging bucket	csee4121homework
Network	default
Network tags	None
Internal IP only	No
Image version	2.0.37-debian10
Created	Apr 27, 2022, 1:45:05 PM
Optional components	JUPYTER
Properties	Show properties
Advanced security	Disabled
Labels	goog-datap... : cluster-e6...
Encryption type	Google-managed key

Cluster cluster-e60d properties

capacity-scheduler:yarn.scheduler.capacity.root.default.ordering-policy	fair
core:fs.gs.block.size	134217728

Job Properties(Spark driver memory: 1GB; Spark executor memory: 5GB)

Job ID	job-51f5443d
Job UUID	985172af-17c2-4e53-8ac2-2a5ec0a8b6ea
Type	Dataproc Job
Status	Succeeded

MONITORING CONFIGURATION

EDIT

Start time:	Apr 28, 2022, 12:37:25 AM
Elapsed time:	9 min 55 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-e60d
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t2q2.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	

We can see from the above screenshot that the completion time of the task is 9 minutes 55 seconds.

Question 3: Cluster Screenshot: (3 nodes with 2 workers;others are default)

Cluster details SUBMIT JOB REFRESH START

Region	us-central1
Zone	us-central1-c
Autoscaling	Off
Dataproc Metastore	None
Scheduled deletion	Off
Master nodes	High Availability (3 masters, 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

Cluster cluster-1058 properties

capacity-scheduler:yarn.scheduler.capacity.root.default.ordering-policy	fair
core:fs.gs.block.size	134217728

Job Properties(Spark driver memory: 1GB; Spark executor memory: 5GB)

MONITORING CONFIGURATION

EDIT

Start time:	Apr 28, 2022, 1:15:22 AM
Elapsed time:	5 min 24 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-1058
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t2q2.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	

We can see from the above screenshot that the completion time of the task is 5 minutes 24 seconds. This suggests that the performance increases since we have two workers working parallely on this task!

Question 4

Cluster Screenshot:(block size: 64MB; 3 nodes with 2 workers;others default)

Cluster details SUBMIT JOB REFRESH START STOP D

Region	us-central1
Zone	us-central1-a
Autoscaling	Off
Dataproc Metastore	None
Scheduled deletion	Off
Master nodes	High Availability (3 masters, 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

Cluster cluster-7909 properties

hdfs:dfs.blocksize	67108864

Job Properties(Spark driver memory: 1GB; Spark executor memory: 5GB)

Job details CLONE DELETE STOP REFRESH

MONITORING CONFIGURATION

EDIT

Start time:	Apr 28, 2022, 1:54:54 AM
Elapsed time:	5 min 43 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-7909
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t2q2.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	

EQUIVALENT REST

We can see from the above screenshot that the completion time of the task is 5 minutes 43 seconds, which performs slightly worse than the one in Q3 since we have decreased the block size from 128MB to 64MB in hdfs. Thus, more blocks have been created and increased the cost/time of seek that worse the performance.

Question 5

Cluster Screenshot:(1 node with 2 workers; others are default)

Cluster details SUBMIT JOB REFRESH START STOP DELETE

MONITORING JOBS VM INSTANCES CONFIGURATION WEB INTERFACES

EDIT

Region	us-central1
Zone	us-central1-a
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

Cluster cluster-ba6a properties

capacity-scheduler:yarn.scheduler.capacity.root.default.ordering-policy	fair
core:fs.gs.block.size	134217728

Job Properties(without killing one worker)(Spark driver memory: 5GB; Spark executor memory: 5GB):

Job details CLONE DELETE STOP REFRESH

Job ID	job-5d0d54db
Job UUID	7bd61483-2130-45f3-8e34-1e26b59f487b
Type	Dataproc Job
Status	Succeeded

MONITORING CONFIGURATION

EDIT

Start time:	Apr 28, 2022, 2:48:20 AM
Elapsed time:	1 hr 16 min
Status:	Succeeded
Region	us-central1
Cluster	cluster-ba6a
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t2q5.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	5g
Labels	

We can see from the above screenshot that the completion time of the task is 1 hour and 16 minutes without killing one worker.

Job Properties(after killing one worker)(Spark driver memory: 5GB; Spark executor memory: 5GB):

Job details CLONE DELETE STOP REFRESH

Job ID	job-1def1185
Job UUID	517321f5-845e-426a-b267-b830b5ffecd5
Type	Dataproc Job
Status	Succeeded

MONITORING CONFIGURATION

EDIT

Start time:	Apr 28, 2022, 4:41:48 AM
Elapsed time:	2 hr 26 min
Status:	Succeeded
Region	us-central1
Cluster	cluster-ba6a
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t2q5.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	5g
Labels	

EQUIVALENT REST

After killing one worker, we can see that the job has been finished. However, the completion time has been doubled compared to the one without killing one.

Question 6

Cluster Screenshot:(1 node with 2 workers;replication factor = 1; others are default)

Cluster details SUBMIT JOB REFRESH START STOP

MONITORING JOBS VM INSTANCES CONFIGURATION WEB INTERFACES

EDIT

Region	us-central1
Zone	us-central1-a
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

Cluster cluster-d6b3 properties

capacity-scheduler:yarn.scheduler.capacity.root.default.ordering-policy	fair
core:fs.gs.block.size	134217728

Cluster cluster-d6b3 properties

hdfs:dfs.namenode.service.handler.count	10
hdfs:dfs.namenode.servicerpc-address	cluster-d6b3-m:8
hdfs:dfs.replication	1

Job Properties(without killing one worker)(Spark driver memory: 5GB; Spark executor memory: 5GB):

Job details CLONE DELETE STOP REFRESH

Job ID	job-d9a73f15
Job UUID	8f810748-f3ff-432b-9457-2813ed35433f
Type	Dataproc Job
Status	Succeeded

MONITORING CONFIGURATION

EDIT

Start time:	Apr 28, 2022, 3:19:36 AM
Elapsed time:	1 hr 15 min
Status:	Succeeded
Region	us-central1
Cluster	cluster-d6b3
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t2q5.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	5g
Labels	

We can see from the above screenshot that the completion time of the task is 1 hour and 15 minutes without killing one worker using replication factor = 1, which decreases 1 minutes compared to Q5 without killing one worker.

Question 7(Cluster Screenshot:(block size: 64MB; 1 node with 2 workers; others are default))

Cluster details SUBMIT JOB REFRESH START

MONITORING JOBS VM INSTANCES CONFIGURATION WEB I

EDIT

Region	us-central1
Zone	us-central1-f
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

Cluster cluster-ce31 properties

hadoop-env:HADOOP_DATANODE_OPTS	-Xmx512m
hdfs:dfs.blocksize	67108864

Job Properties(without killing one worker)(Spark driver memory: 5GB; Spark executor memory: 5GB):

Job details CLONE DELETE STOP REFRESH

Job ID	job-46541f6b
Job UUID	fadb75c9-bf50-49aa-8b33-12c154494d0e
Type	Dataproc Job
Status	Succeeded

MONITORING CONFIGURATION

EDIT

Start time:	Apr 28, 2022, 8:33:31 AM
Elapsed time:	1 hr 24 min
Status:	Succeeded
Region	us-central1
Cluster	cluster-ce31
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t2q5.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	5g

We can see from the above screenshot that the completion time of the task is 1 hour and 24 minutes without killing one worker using block size = 64MB, which has a worse performance compared to Q5. Since we have decreased the block size from 128MB to 64MB in hdfs, more blocks have been created and increased the cost/time of seeking more on this larger data set that makes performance worse.

Question 8(Cluster Screenshot:(3 nodes with 2 workers))

Cluster details SUBMIT JOB REFRESH START STOP DE

EDIT

Region	us-central1
Zone	us-central1-b
Autoscaling	Off
Dataproc Metastore	None
Scheduled deletion	Off
Master nodes	High Availability (3 masters, 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

Job Properties(Spark driver memory: 5GB; Spark executor memory: 5GB):

EDIT

Start time:	Apr 28, 2022, 2:49:17 PM
Elapsed time:	2 hr 19 min
Status:	Succeeded
Region	us-central1
Cluster	cluster-ea57
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t3q8.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	5g
Labels	

From the above screenshot, we can see that it takes 2 hours and 19 minutes to complete the task using the given job properties.

Q9

Job Properties(cores: 4; Spark driver memory: 5GB; Spark executor memory: 5GB):

Job details CLONE DELETE STOP REFRE

ck to Job List

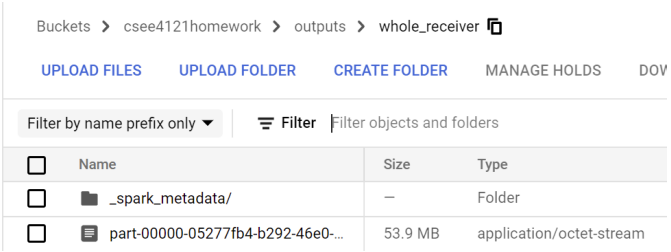
Job ID	job-42f64cc3
Job UUID	0096709d-8909-401f-a02c-fa6a744806a9
Type	Dataproc Job
Status	Succeeded

MONITORING CONFIGURATION

EDIT

Start time:	May 1, 2022, 2:40:26 AM
Elapsed time:	2 hr 24 min
Status:	Succeeded
Region	us-central1
Cluster	cluster-e936
Job type	PySpark
Main python file	gs://csee4121homework/notebooks/jupyter/p1t3q8.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	5g
Labels	

Document received by stream receiver:




```
]:
print("Number of articles in the database has a rank greater than 0.5:",\
      spark.read.csv("gs://csee4121homework/outputs/whole_receiver",sep = "\t").count())

Number of articles in the database has a rank greater than 0.5: 1762190
```

After receiver has been killed, we can see from the generated csv file catched by the stream that there are 1,762,190 articles in the database generated by the pagerank algorithm has a rank greater than **0.5.**

Q10

```
In [20]: spark = SparkSession.builder.getOrCreate()
         print("Number of articles in the database has a rank greater than 0.5 for part 2 task 1:",\
               spark.read.csv("gs://csee4121homework/outputs/whole_receiver",sep = "\t").count())
         print("Number of articles in the database has a rank greater than 0.5 for part 2 task 2:",\
               spark.read.csv("gs://csee4121homework/outputs/whole_receiver_from_emitter",sep = "\t").repartition(1).count())

         Number of articles in the database has a rank greater than 0.5 for part 2 task 1: 1762190
         Number of articles in the database has a rank greater than 0.5 for part 2 task 2: 1762190
```

I think such data server design is feasible and efficient since it may thus allows the multiples emitters to send data to receiver via TCP sockets instead of changing receiver's receiving path to catch data in different location. For example, if we have two pagerank algorithm running and generating output at two different locations. When we only have one receiver, we can use two emitters to emit those two results to receiver, and receiver would be able to catch those stream one by one via TCP sockets. However, without an emitter, we might need two receivers to receive those two files simultaneously.

Question 11

I spent five days and around 10 hours per day working on this assignment.