Task 1: Getting Started

Question 1

What is the default block size on HDFS? **Answer:** 128 MB

What is the default replication factor of HDFS on Dataproc? **Answer:** 3

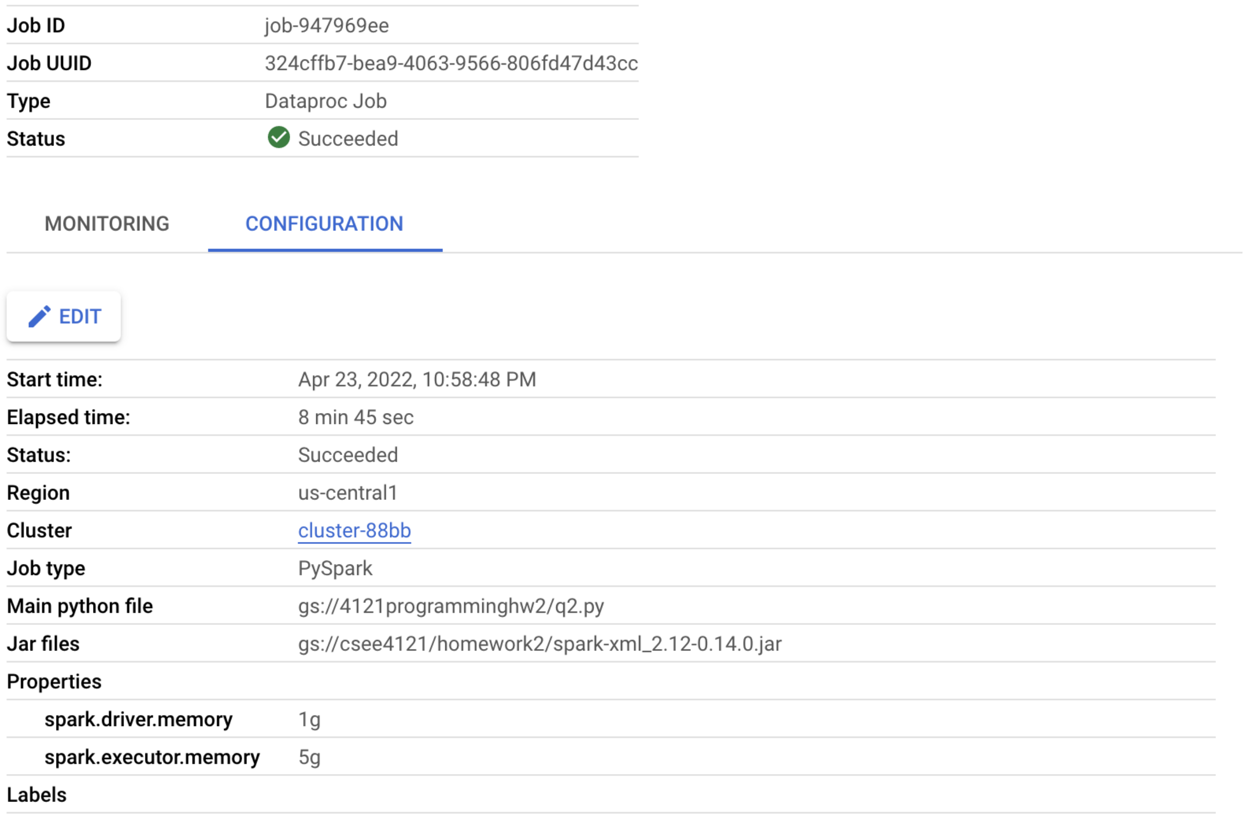
Task 2: Webgraph on Internal Links

Question 2

(Set the **Spark driver memory** to 1GB and the **Spark executor memory** to 5GB, Single Node cluster)

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?

Answer: 8min 45s, see below:



Job ID	job-947969ee
Job UUID	324cffb7-bea9-4063-9566-806fd47d43cc
Type	Dataproc Job
Status	✓ Succeeded

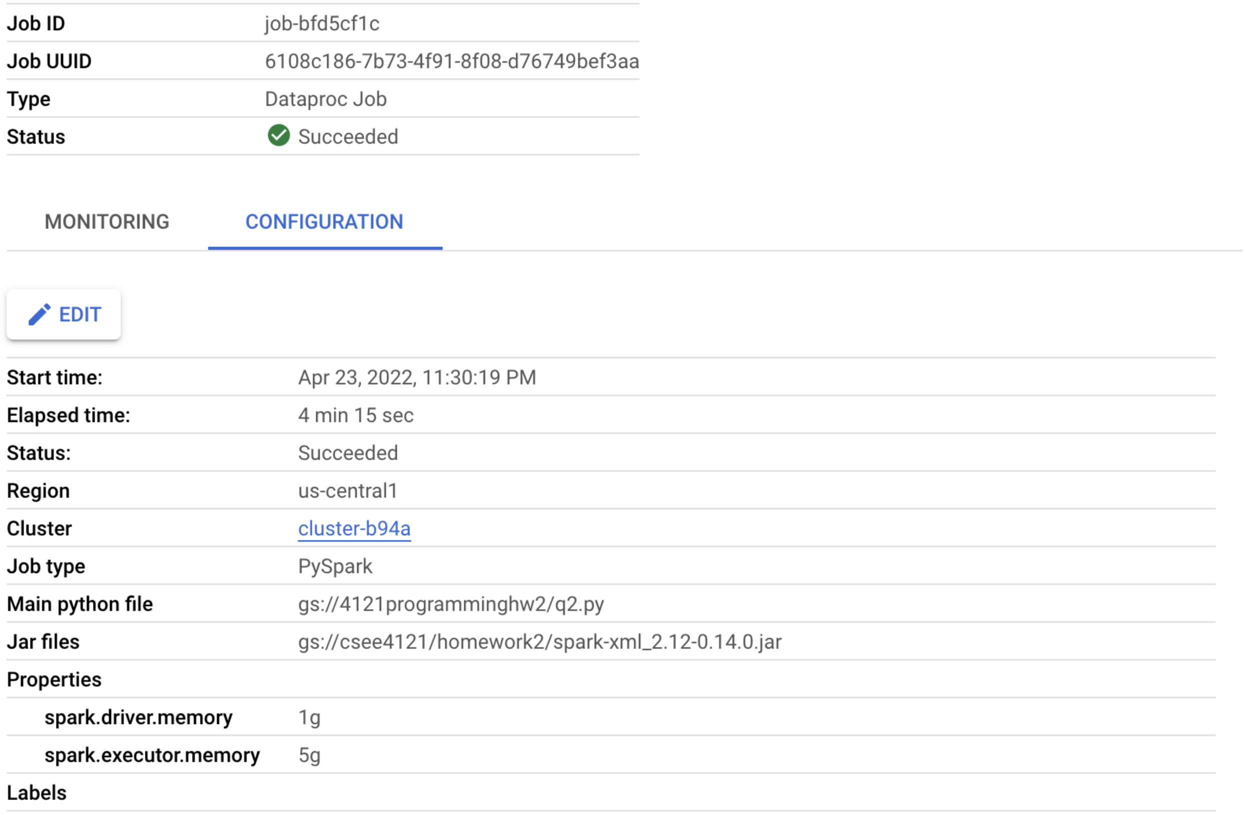
Start time:	Apr 23, 2022, 10:58:48 PM
Elapsed time:	8 min 45 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-88bb
Job type	PySpark
Main python file	gs://4121programminghw2/q2.py
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties	
spark.driver.memory	1g
spark.executor.memory	5g
Labels	

Question 3

(Set the **Spark driver memory** to 1GB and the **Spark executor memory** to 5GB, 3 node cluster)

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.

Answer: 4 min 15 seconds. As expected, the performance improves when we increase the number of nodes. It is roughly twice as fast compared to the previous question.



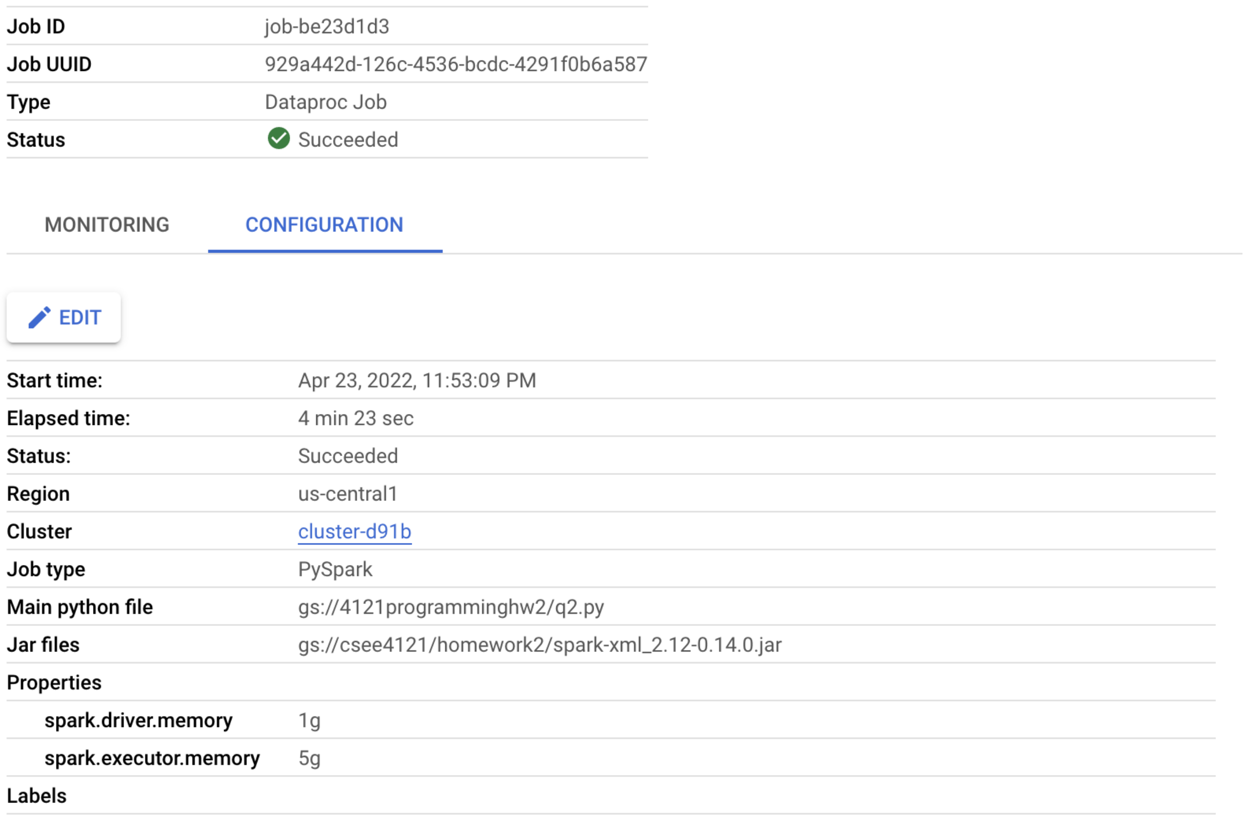
Job ID	job-bfd5cf1c
Job UUID	6108c186-7b73-4f91-8f08-d76749bef3aa
Type	Dataproc Job
Status	✓ Succeeded

Start time:	Apr 23, 2022, 11:30:19 PM
Elapsed time:	4 min 15 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-b94a
Job type	PySpark
Main python file	gs://4121programminghw2/q2.py
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties	
spark.driver.memory	1g
spark.executor.memory	5g
Labels	

Question 4

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.

Answer: 4 min 23 s. The performance is a little bit worse since there are more blocks to manipulate (the HDFS file is split into 64MB sized blocks rather than 128MB sized blocks) which results in a higher completion time.



Job ID	job-be23d1d3
Job UUID	929a442d-126c-4536-bcdc-4291f0b6a587
Type	Dataproc Job
Status	✓ Succeeded

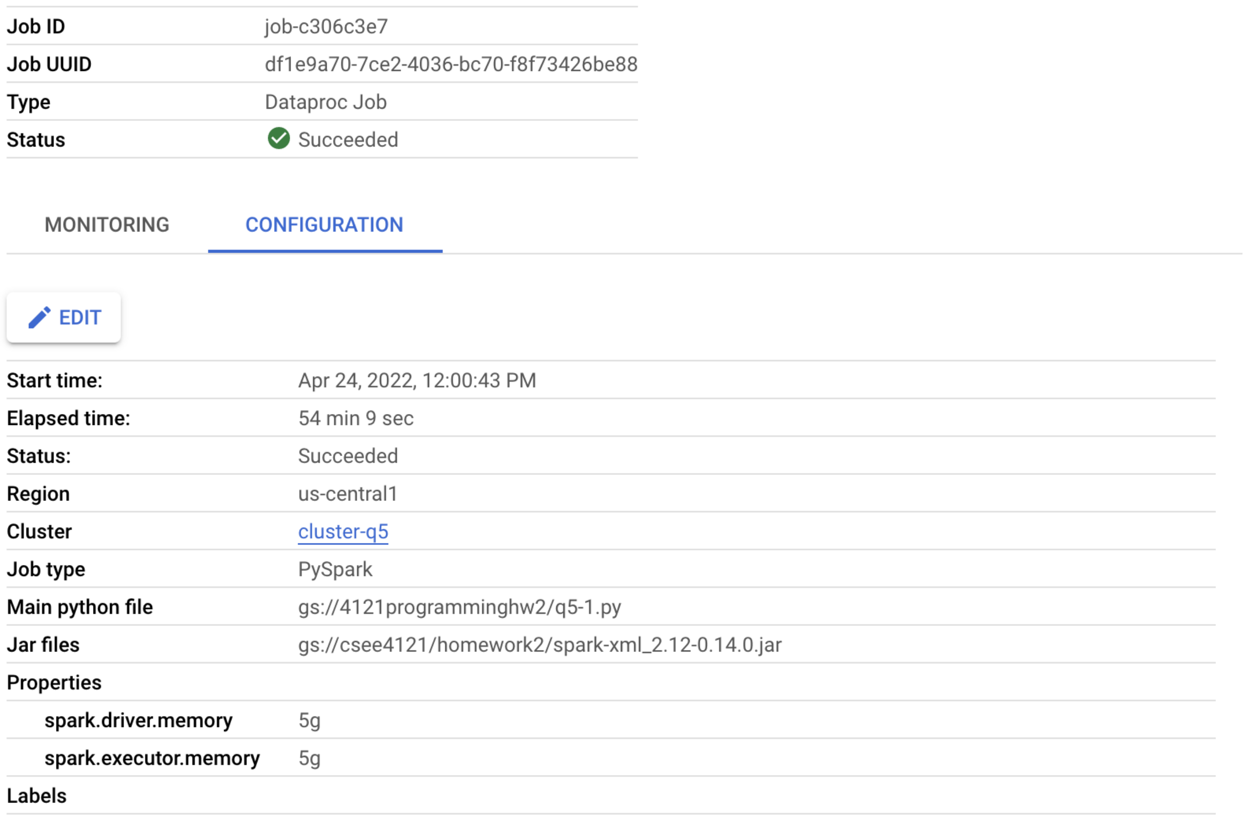
Start time:	Apr 23, 2022, 11:53:09 PM
Elapsed time:	4 min 23 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-d91b
Job type	PySpark
Main python file	gs://4121programminghw2/q2.py
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties	
spark.driver.memory	1g
spark.executor.memory	5g
Labels	

Question 5

(Set the **Spark driver memory** to 5GB and the **Spark executor memory** to 5GB to answer Question 5-7)

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.

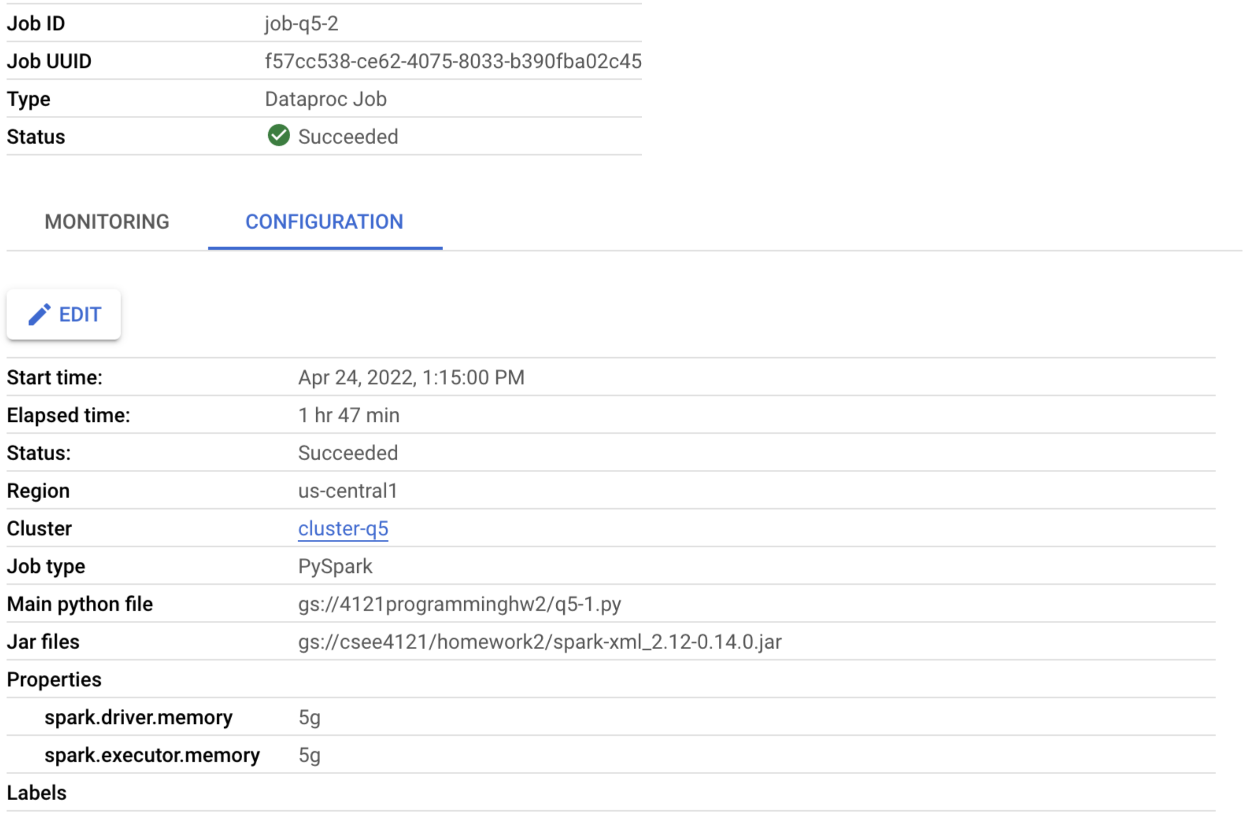
Answer: Q5-1: 54min 9s (baseline)
Even though the driver memory was increased from 1GB to 5GB, it still takes a very long time to complete the job with enwiki_whole.xml due to the size of the data.



Job ID	job-c306c3e7
Job UUID	df1e9a70-7ce2-4036-bc70-f8f73426be88
Type	Dataproc Job
Status	✓ Succeeded

Start time:	Apr 24, 2022, 12:00:43 PM
Elapsed time:	54 min 9 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-q5
Job type	PySpark
Main python file	gs://4121programminghw2/q5-1.py
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties	
spark.driver.memory	5g
spark.executor.memory	5g
Labels	

Q5-2: The job still runs after killing one of the worker nodes but it takes 1h 47min to complete, which is twice as long compared to Q5-1. This makes sense because there is one worker completing the job (as opposed to two workers from before).



Job ID	job-q5-2
Job UUID	f57cc538-ce62-4075-8033-b390fba02c45
Type	Dataproc Job
Status	✓ Succeeded

MONITORING CONFIGURATION

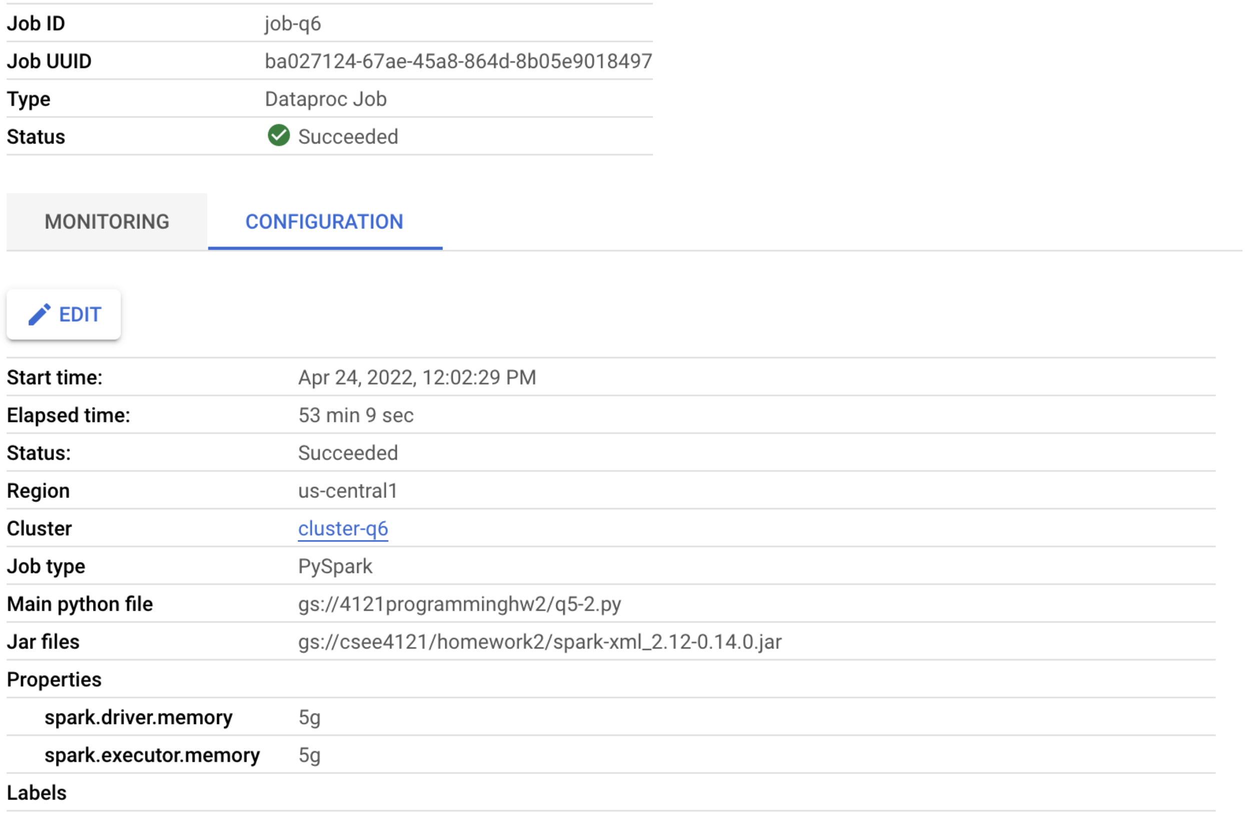
EDIT

Start time:	Apr 24, 2022, 1:15:00 PM
Elapsed time:	1 hr 47 min
Status:	Succeeded
Region	us-central1
Cluster	cluster-q5
Job type	PySpark
Main python file	gs://4121programminghw2/q5-1.py
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties	
spark.driver.memory	5g
spark.executor.memory	5g
Labels	

Question 6

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.

Answer: 53min 9s. With a replication factor of 1, only one copy of the block is kept in the cluster. This means fewer data blocks that the system needs to manage. As expected, we see slightly better performance compared to the baseline in Question 5. Note the downside of this is that the data may be more prone to getting corrupted if there is a database failure.



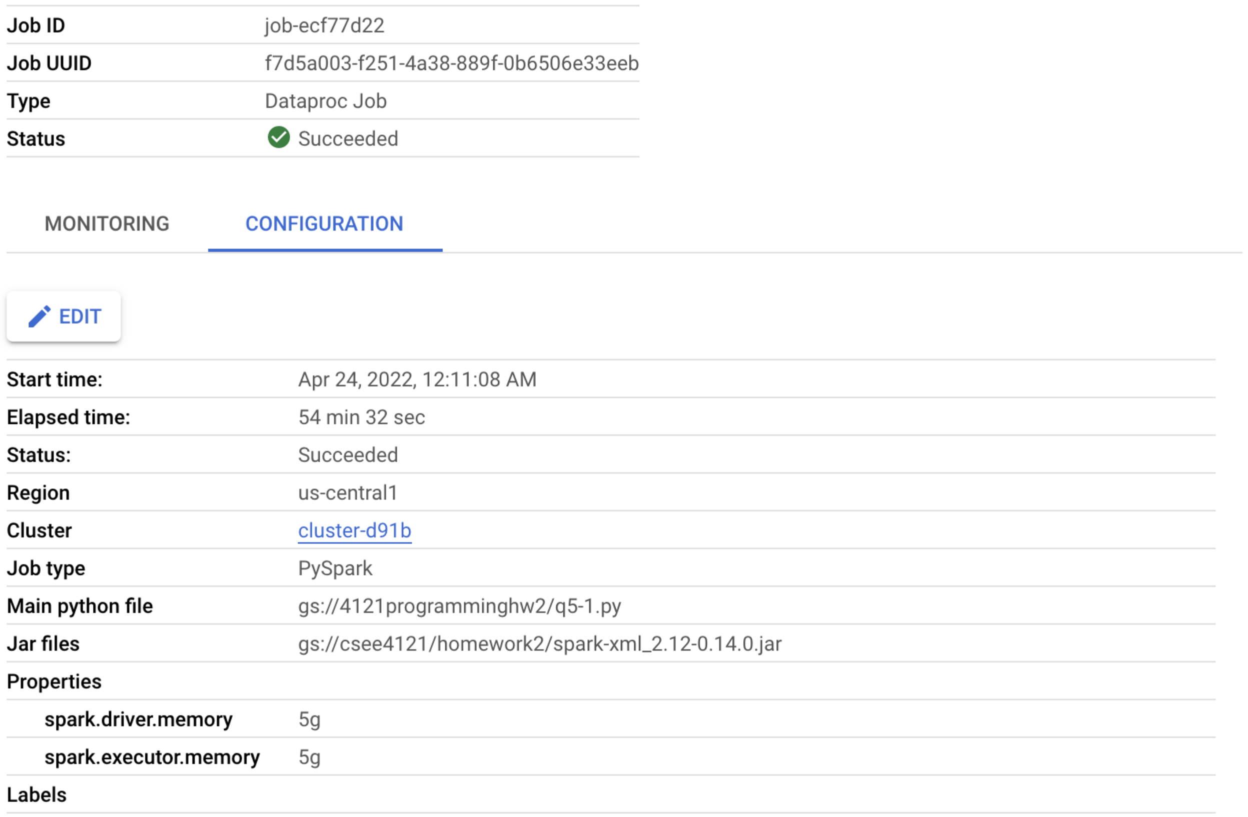
Job ID	job-q6
Job UUID	ba027124-67ae-45a8-864d-8b05e9018497
Type	Dataproc Job
Status	✓ Succeeded

Start time:	Apr 24, 2022, 12:02:29 PM
Elapsed time:	53 min 9 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-q6
Job type	PySpark
Main python file	gs://4121programminghw2/q5-2.py
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties	
spark.driver.memory	5g
spark.executor.memory	5g
Labels	

Question 7

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.

Answer: 54 min 32 s. Just like in Question 4, the performance is a little bit worse since there are more blocks to manipulate which results in a higher completion time.



Job ID	job-ecf77d22
Job UUID	f7d5a003-f251-4a38-889f-0b6506e33eeb
Type	Dataproc Job
Status	✓ Succeeded

Start time:	Apr 24, 2022, 12:11:08 AM
Elapsed time:	54 min 32 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-d91b
Job type	PySpark
Main python file	gs://4121programminghw2/q5-1.py
Jar files	gs://csee4121/homework2/spark-xml_2.12-0.14.0.jar
Properties	
spark.driver.memory	5g
spark.executor.memory	5g
Labels	

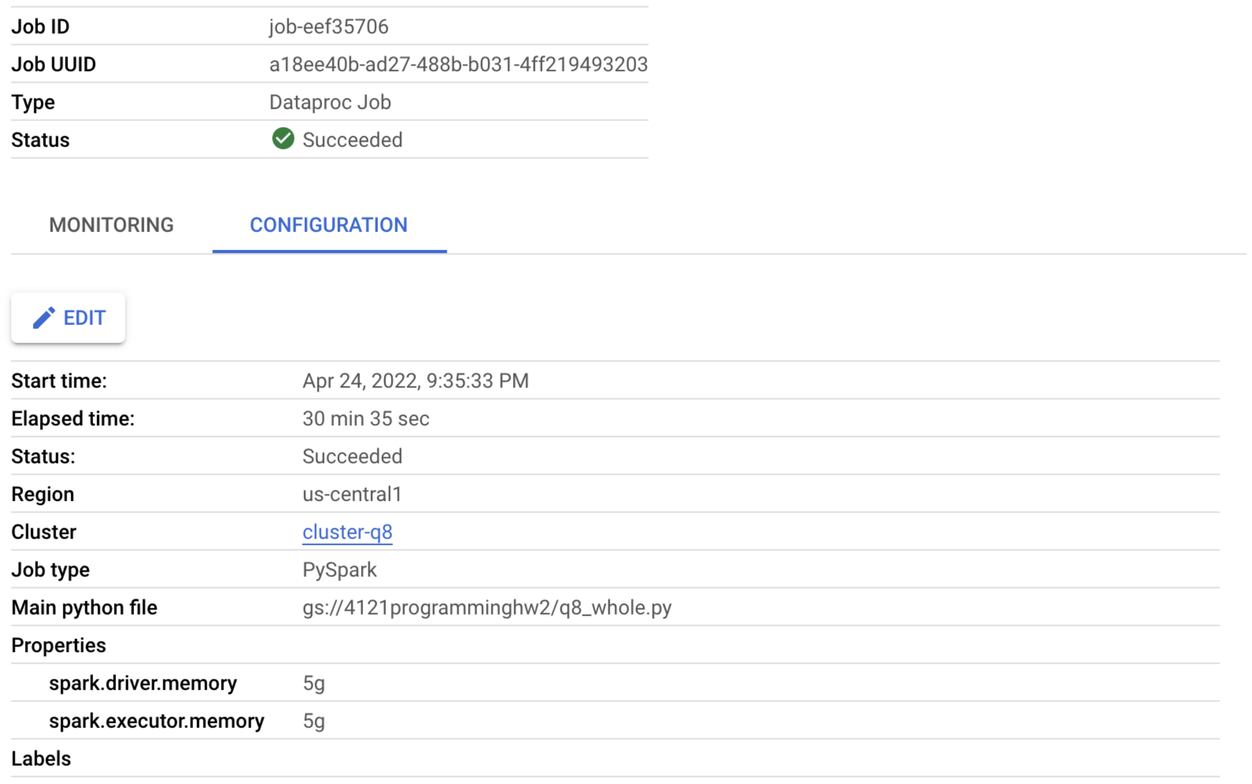
Task 3: Spark PageRank

Question 8

Set the Spark driver memory to 5GB and the Spark executor memory to 5GB whenever you run your PageRank program. Write a script to first run Task 2, and then run Task 3 using the csv output generated by Task 2, and answer the following questions. Always use 10 iterations for the PageRank program. When running Task 2, use `enwiki_whole.xml` as input.

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?

Answer: 30 min 35 s



Job ID	job-eef35706
Job UUID	a18ee40b-ad27-488b-b031-4ff219493203
Type	Dataproc Job
Status	✓ Succeeded

Start time:	Apr 24, 2022, 9:35:33 PM
Elapsed time:	30 min 35 sec
Status:	Succeeded
Region	us-central1
Cluster	cluster-q8
Job type	PySpark
Main python file	gs://4121programminghw2/q8_whole.py
Properties	
spark.driver.memory	5g
spark.executor.memory	5g
Labels	

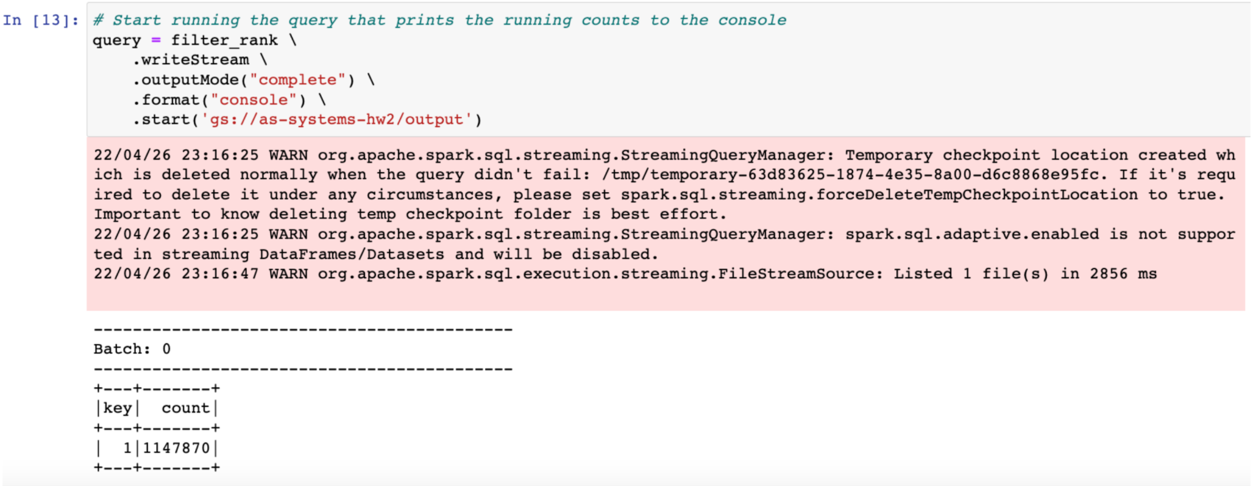
Part 2: Spark Streaming (Extra Credit)

Task 1: Stream Receiver

Question 9

Start a PageRank program you wrote in Part 1 Task 3 whose input is the link graph generated from “enwiki_whole.xml” and store the output to a directory inside HDFS. Set your stream receiver to read the files generated by the PageRank program. Kill the receiver when the PageRank task is finished. How many articles in the database has a rank greater than **0.5**? You can SSH into the master node and start with ‘hdfs dfs -mv’ or ‘mv’ to help you write your program.

Answer: There are 1147870 articles with pagerank greater than 0.5.



```
In [13]: # Start running the query that prints the running counts to the console
         query = filter_rank \
             .writeStream \
             .outputMode("complete") \
             .format("console") \
             .start('gs://as-systems-hw2/output')
```

```
22/04/26 23:16:25 WARN org.apache.spark.sql.streaming.StreamingQueryManager: Temporary checkpoint location created wh
ich is deleted normally when the query didn't fail: /tmp/temporary-63d83625-1874-4e35-8a00-d6c8868e95fc. If it's requ
ired to delete it under any circumstances, please set spark.sql.streaming.forceDeleteTempCheckpointLocation to true.
Important to know deleting temp checkpoint folder is best effort.
22/04/26 23:16:25 WARN org.apache.spark.sql.streaming.StreamingQueryManager: spark.sql.adaptive.enabled is not suppor
ted in streaming DataFrames/Datasets and will be disabled.
22/04/26 23:16:47 WARN org.apache.spark.sql.execution.streaming.FileStreamSource: Listed 1 file(s) in 2856 ms
```

```
-------------------------------------------
Batch: 0
-------------------------------------------
+---+-------+
|key|  count|
+---+-------+
|  1|1147870|
+---+-------+
```

Task 2: Stream Emitter

Question 10

Spark Streaming can also be used to send data via TCP sockets. The Emitter in this case will wait on a socket connection request from the receiver, and upon accepting the connection request it will start sending data. Do you think such data server design is feasible and efficient? Briefly explain.

Answer: The design is feasible and efficient. That's because if A transmit the data to B before B manages to connect to the server, partial data that are transferred will be missing. Another possible reason is it is possible to send the data to wrong receivers if there is no TCP socket.

Question 11

How many hours did you spend in this assignment?

Answer: 32 hours