

# README

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

This is a submission of programming homework 2 by Hyo Won Kim (hk3175) and Alex Kita (ak4729).

## 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)

Name	cluster-d729			
Cluster UUID	e7690f3c-5a1e-4df6-a42e-0f58f70c52af			
Type	Dataproc Cluster			
Status	Stopped			
<hr/>				
MONITORING	J OBS	VM INSTANCES	<b>CONFIGURATION</b>	WEB INTERFACES
<hr/>		<a href="#"> EDIT</a>		
<hr/>				
Region	us-central1			<hr/>
Zone	us-central1-f			<hr/>
Autoscaling	Off			<hr/>
Dataproc Metastore	None			<hr/>
Scheduled deletion	Off			<hr/>
Master node	Single Node (1 master, 0 workers)			<hr/>
Machine type	n1-standard-4			<hr/>
Number of GPUs	0			<hr/>
Primary disk type	pd-standard			<hr/>
Primary disk size	500GB			<hr/>
Local SSDs	0			<hr/>
Secure Boot	Disabled			<hr/>
VTPM	Disabled			<hr/>
Integrity Monitoring	Disabled			<hr/>
Cloud Storage staging bucket	<a href="#">hw4_spark_ak</a>			<hr/>
Network	default			<hr/>
Network tags	None			<hr/>
Internal IP only	No			<hr/>

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

Name	cluster-35f7
Cluster UUID	3a086c00-ba10-4df5-a600-bfee48728014
Type	Dataproc Cluster
Status	<span>Running</span>
<hr/>	
MONITORING	JOBSS
VM INSTANCES	<b>CONFIGURATION</b>
<hr/>	
<span>EDIT</span>	
<hr/>	
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	<span>Running</span>
<hr/>	
MONITORING	J OBS
VM INSTANCES	CONFIGURATION
WEB INTERFACES	
<hr/>	
<a href="#"> EDIT</a>	
<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	<a href="#">hw4_spark_ak</a>

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

**MONITORING**

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 02:21:10 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 02:22:29 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
22/05/01 02:23:33 INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat: Total input files to process : 1
java.lang.InterruptedException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:519)
    at com.google.common.util.concurrent.FluentFuture.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)
22/05/01 02:25:57 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark5b04000B{HTTP/1.1, (http://1.1)}{0.0.0.0:0}
```

Output is complete

EQUIVALENT COMMAND LINE

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

**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 (GCP) interface for a Dataproc job. The job ID is job-c3986184, with a UUID of 9cc25f5-382e-4280-b8e3-5f5e92aa1207. The status is Succeeded. The monitoring tab displays three line charts: YARN memory usage (25GB), YARN pending memory (20GB), and YARN NodeManagers (2). The configuration tab shows the job's configuration details, including start and end times, region (us-central1), cluster (cluster-357), job type (PySpark), main Python file (gs://hw4\_spark\_uk/notebooks/jupyter/hw2-q3-test.py), and properties like spark.executor.memory (5g) and spark.driver.memory (1g). The output section shows the command run and its execution log.

```

at java.util.concurrent.Executors$RunnableExecutor.runWorker(Executors.java:610)
at java.util.concurrent.Executors$RunnableExecutor$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}

```

**Job details**

**MONITORING**

**Configuration**

**Output**

**EQUIVALENT COMMAND LINE**

```

at java.util.concurrent.Executors$RunnableExecutor.runWorker(Executors.java:610)
at java.util.concurrent.Executors$RunnableExecutor$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}

```

**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\_uk/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**

# 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.

The screenshot shows the Google Cloud Platform Dataproc Job details page. The job ID is job-c18c2766, and the status is Succeeded. The monitoring section displays charts for YARN memory, YARN pending memory, and YARN NodeManagers. The log output shows Java stack traces and INFO logs related to Spark and Hadoop execution. The log output ends with the message "INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@7c661edf{HTTP/1.1}{(0.0.0.0:0)}".

```
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:759)
22/05/01 02:00:46 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:518)
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:759)
22/05/01 02:00:46 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@7c661edf{HTTP/1.1}{(0.0.0.0:0)}
```

The screenshot shows the Google Cloud Platform DataProc interface. On the left, there's a sidebar with navigation links for 'Jobs on Clusters' (Clusters, Jobs, Workflows, Autoscaling policies), 'Serverless' (Batches), 'Utilities' (Component exchange, Metastore), and 'Workbench'. The main area is titled 'Job details' for a job with ID 'job-c18c2766'. It shows the job succeeded with a start time of April 30, 2022, at 10:00:33 PM, an elapsed time of 2 min 50 sec, and was run in the 'us-central1' region on cluster 'cluster-357'. The configuration tab is selected, showing properties like spark.executor.memory (5g), spark.driver.memory (1g), and dfs.block.size (64m). Below the configuration is an 'EQUIVALENT REST' section with an output panel containing the message '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 DataProc Job details page for a job named 'job-c896fca2'. The job ID is 'job-c896fca2', Job UUID is 'f6cc2d4f-9f09-4ff5-9b91-df39f4694add', Type is 'DataProc Job', and Status is 'Succeeded'. The 'MONITORING' tab is selected, displaying three line charts: 'YARN memory', 'YARN pending memory', and 'YARN NodeManagers'. The 'YARN memory' chart shows usage from 10GB to 25GB. The 'YARN pending memory' chart shows pending memory from 400GB to 800GB. The 'YARN NodeManagers' chart shows two NodeManagers from 1.5 to 2. Below the charts, the 'Output' section shows log entries:

```

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 Spark@2d485382{HTTP/1.1}{0.0.0.0:0}

```

The 'EQUIVALENT COMMAND LINE' section shows the command used to run the job.

This screenshot shows the Google Cloud Platform DataProc Job details page for the same job ('job-c896fca2'). The 'CONFIGURATION' tab is selected, showing various configuration parameters:

- 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
- Labels:** question:5

The 'EQUIVALENT REST' section shows the equivalent API endpoint. The 'Output' section indicates the job is complete.

# 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 completed job. The left sidebar lists various cluster components like Clusters, Jobs, Workflows, and Utilities. The main area displays job metadata (Job ID: job-dc252cb7, Job UUID: aec681b1-8169-4e8e-afee-d64f0d5c451d, Type: Dataproc Job, Status: Succeeded) and monitoring metrics for YARN memory, YARN pending memory, and YARN NodeManagers. The log output section shows standard Java and Hadoop logs indicating the job's execution and completion.

```
22/05/02 02:52:29 WARN org.apache.hadoop.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#04f93aa7{HTTP/1.1}{(http/1.1)}{0.0.0.0:0}
```

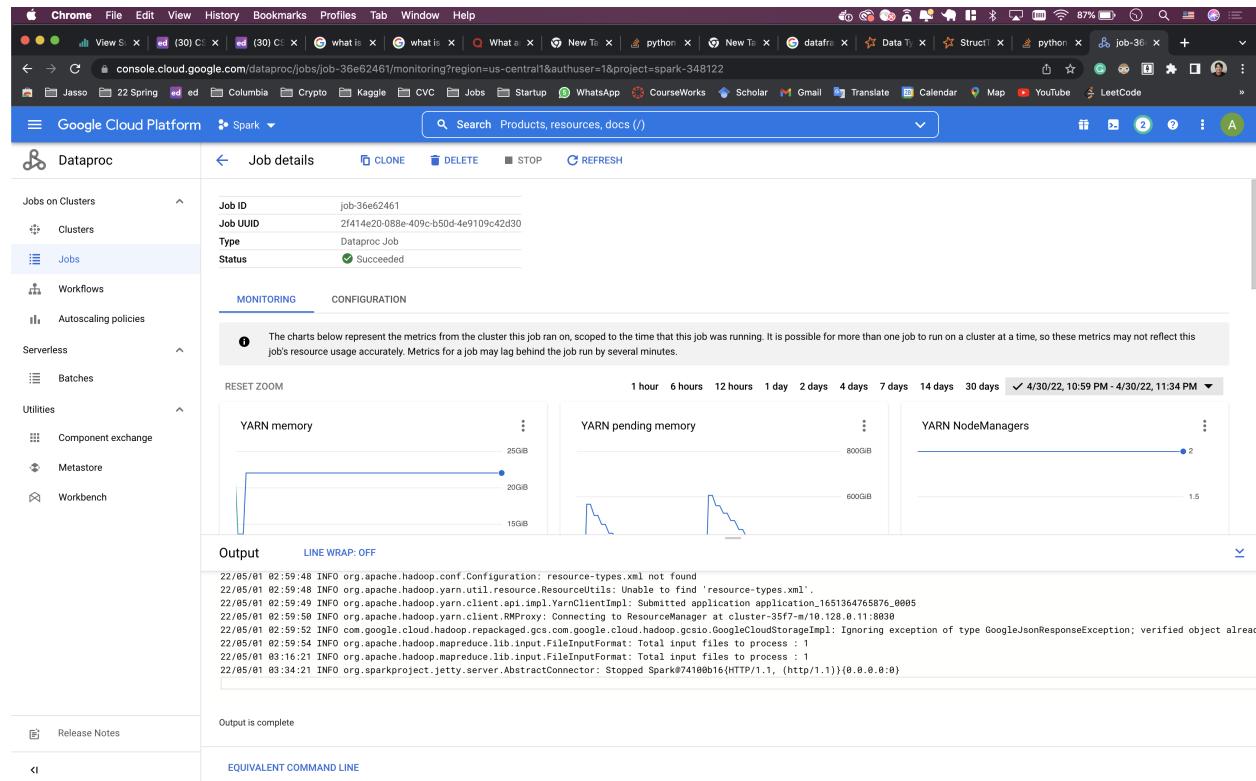
This screenshot shows the Google Cloud Platform Dataproc Job details page for the same completed job, but with different configuration settings. The configuration tab is selected, showing detailed configuration parameters such as Start time (May 1, 2022, 10:52:22 PM), Elapsed time (35 min 7 sec), Status (Succeeded), Region (us-central1), Cluster (cluster-35f7), Job type (PySpark), Main python file (gs://hw4\_spark\_ak/notebooks/jupyter/hw2-q6-whole.py), and Jar files (gs://cseee4121/homework2/spark-xml\_2.12-0.14.0.jar). The properties section includes spark.driver.memory (5g), spark.executor.memory (5g), and dfs.replication (1). The log output section is identical to the previous screenshot.

```
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#04f93aa7{HTTP/1.1}{(http/1.1)}{0.0.0.0:0}
```

# 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 ID: job-36e62461  
 Job UUID: 2f414e20-088e-409c-b50d-4e9109c42d30  
 Type: DataProc Job  
 Status: Succeeded

**MONITORING**

**CONFIGURATION**

**Properties**

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

**Labels**: question: 7

**Output** (LINE WRAP: OFF)

```
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@7410#b16{HTTP/1.1, {http/1.1}}{0.0.0.0:0}
```

Output is complete

# Question 8

Completion time: 58 min 55 sec

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

**MONITORING**

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 ✓ 1:29 PM - 2:28 PM

**YARN memory**: 25GB (Actual: 20GB)

**YARN pending memory**: 1.5TB (Actual: 1.7TB)

**YARN NodeManagers**: 2 (Actual: 1.5)

**Output** (LINE WRAP: OFF)

```
at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
at org.apache.hadoop.util.concurrent.ExecutorHelper$LogThrowableCommitterExecutor.execute(ExecutorHelper.java:48)
at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.execute(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 Spark@3e766858{HTTP/1.1, {http/1.1}}{0.0.0.0:0}
```

Output is complete

EQUIVALENT COMMAND LINE

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

---

MONITORING	CONFIGURATION
------------	---------------

---

<span>EDIT</span>
-------------------

---

Start time:	May 2, 2022, 1:29:16 PM
Elapsed time:	58 min 55 sec
Status:	Succeeded
Region	us-central1
Cluster	<a href="#">cluster-35f7</a>
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

---

## Question 9

More than 1,148,000 articles in enwiki-whole.xml have a rank greater than 0.5.

## Question 10

The data server design via TCP is feasible. This can be implemented using `readStream.format("socket")`. In terms of efficiency, it can be inefficient for some aspects but efficient in other aspects. Because we have to wait for the socket connection request, the latency and the time to complete the task would be increased. However, once the socket connection is secured, it enables us to avoid writing and reading directly from the disk and we gain efficiency. For larger tasks, especially, data server design via TCP is much more efficient.

# **Question 11**

We spent at least 30 hours on this assignment. Thank you for grading!