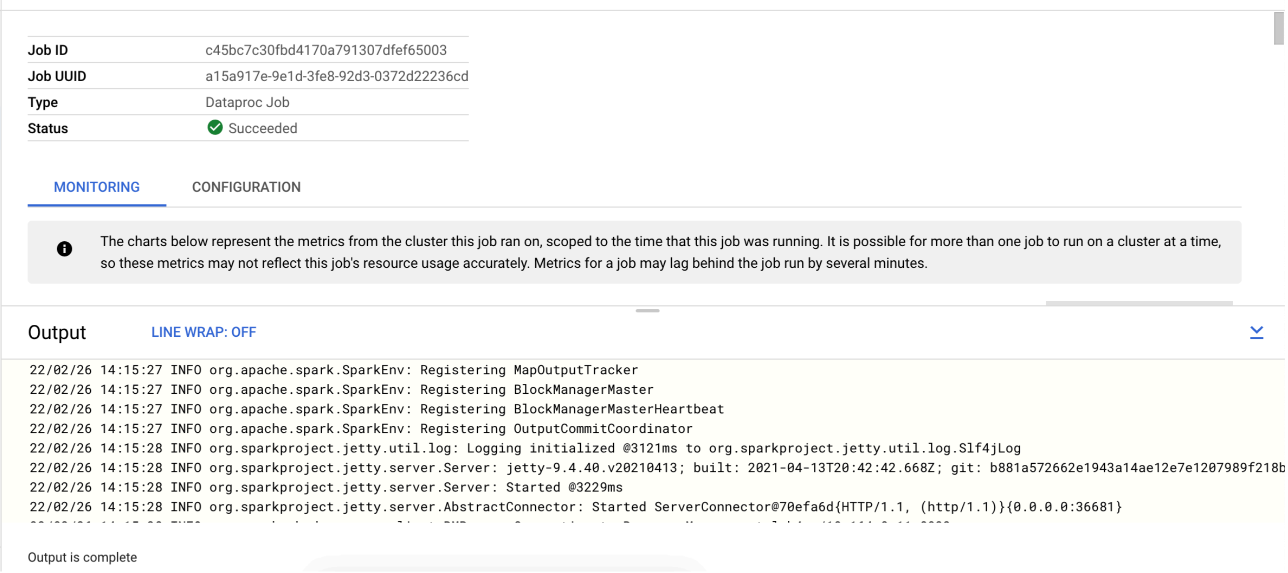
Big Data Lab

Assignment-2

(Vasudev Gupta, ME18B182)

1. **Outputs**

('12-18', 260)
('0-6', 227)
('18-24', 262)
('6-12', 252)

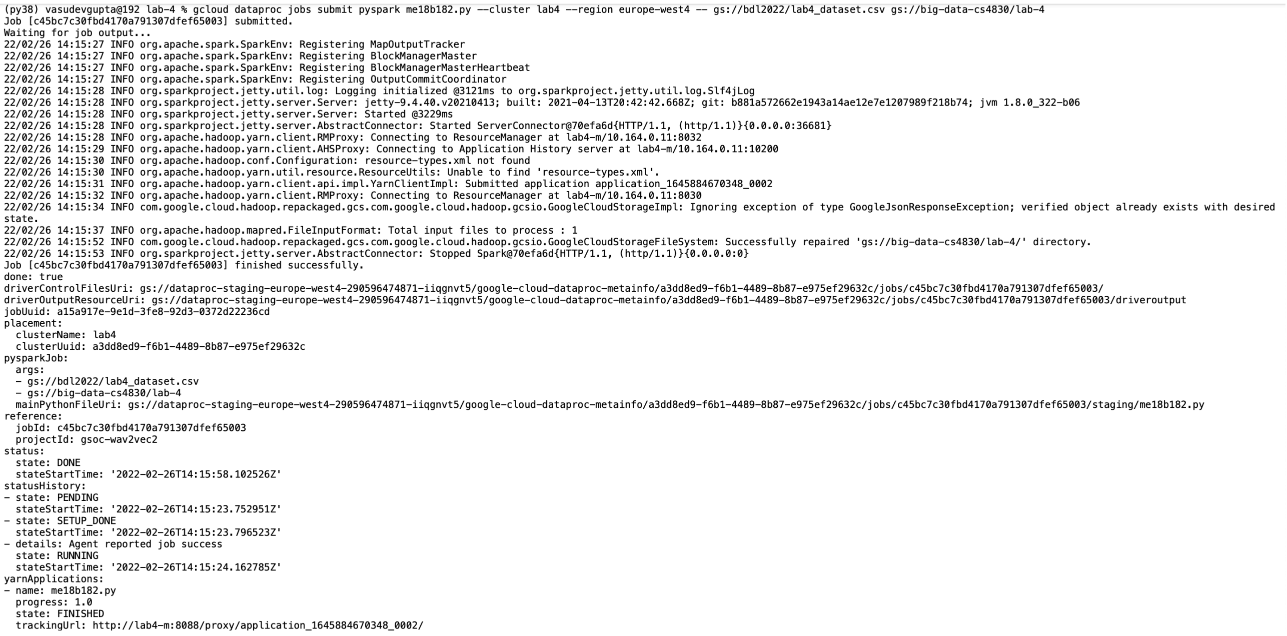


Job ID	c45bc7c30fbd4170a791307dfef65003
Job UUID	a15a917e-9e1d-3fe8-92d3-0372d22236cd
Type	Dataproc Job
Status	✓ Succeeded

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.

```
22/02/26 14:15:27 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/02/26 14:15:27 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/02/26 14:15:27 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/02/26 14:15:27 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/02/26 14:15:28 INFO org.sparkproject.jetty.util.log: Logging initialized @3121ms to org.sparkproject.jetty.util.log.Slf4jLog
22/02/26 14:15:28 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b
22/02/26 14:15:28 INFO org.sparkproject.jetty.server.Server: Started @3229ms
22/02/26 14:15:28 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@70efa6d{HTTP/1.1, (http/1.1)}{0.0.0.0:36681}
```

Output is complete



```
(py38) vasudevgupta@192 lab-4 % gcloud dataproc jobs submit pyspark me18b182.py --cluster lab4 --region europe-west4 -- gs://bdl2022/lab4_dataset.csv gs://big-data-cs4830/lab-4
Job [c45bc7c30fbd4170a791307dfef65003] submitted.
Waiting for job output...
22/02/26 14:15:27 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/02/26 14:15:27 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/02/26 14:15:27 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/02/26 14:15:27 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/02/26 14:15:28 INFO org.sparkproject.jetty.util.log: Logging initialized @3121ms to org.sparkproject.jetty.util.log.Slf4jLog
22/02/26 14:15:28 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_322-b06
22/02/26 14:15:28 INFO org.sparkproject.jetty.server.Server: Started @3229ms
22/02/26 14:15:28 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@70efa6d{HTTP/1.1, (http/1.1)}{0.0.0.0:36681}
22/02/26 14:15:28 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at lab4-m/10.164.0.11:8032
22/02/26 14:15:29 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at lab4-m/10.164.0.11:10200
22/02/26 14:15:30 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/02/26 14:15:30 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/02/26 14:15:31 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1645884670348_0002
22/02/26 14:15:32 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at lab4-m/10.164.0.11:8030
22/02/26 14:15:34 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired
state.
22/02/26 14:15:37 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1
22/02/26 14:15:52 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFileSystem: Successfully repaired 'gs://big-data-cs4830/lab-4/' directory.
22/02/26 14:15:53 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@70efa6d{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [c45bc7c30fbd4170a791307dfef65003] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-europe-west4-290596474871-iiqgnvt5/google-cloud-dataproc-metainfo/a3dd8ed9-f6b1-4489-8b87-e975ef29632c/jobs/c45bc7c30fbd4170a791307dfef65003/
driverOutputResourceUri: gs://dataproc-staging-europe-west4-290596474871-iiqgnvt5/google-cloud-dataproc-metainfo/a3dd8ed9-f6b1-4489-8b87-e975ef29632c/jobs/c45bc7c30fbd4170a791307dfef65003/driveroutput
jobUuid: a15a917e-9e1d-3fe8-92d3-0372d22236cd
placement:
  clusterName: lab4
  clusterUuid: a3dd8ed9-f6b1-4489-8b87-e975ef29632c
pysparkJob:
  args:
  - gs://bdl2022/lab4_dataset.csv
  - gs://big-data-cs4830/lab-4
  mainPythonFileUri: gs://dataproc-staging-europe-west4-290596474871-iiqgnvt5/google-cloud-dataproc-metainfo/a3dd8ed9-f6b1-4489-8b87-e975ef29632c/jobs/c45bc7c30fbd4170a791307dfef65003/staging/me18b182.py
reference:
  jobId: c45bc7c30fbd4170a791307dfef65003
  projectId: gsoc-wav2vec2
status:
  state: DONE
  stateStartTime: '2022-02-26T14:15:58.102526Z'
statusHistory:
- state: PENDING
  stateStartTime: '2022-02-26T14:15:23.752951Z'
- state: SETUP_DONE
  stateStartTime: '2022-02-26T14:15:23.796523Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2022-02-26T14:15:24.162785Z'
yarnApplications:
- name: me18b182.py
  progress: 1.0
  state: FINISHED
  trackingUrl: http://lab4-m:8088/proxy/application_1645884670348_0002/
```

2.

HDFS	HDFS is a network based file system that can store very large datasets by abstracting the location of the data using a mapping table. To store huge amounts of data, it employs a cluster (combined storage) of tiny storage devices. If more storage capacity is required, HDFS does not need the complete hardware be modified, but rather needs a suitable number of next storage nodes (machines) be added to the cluster. To accommodate a network machine failure, HDFS keeps three copies of the same machine (one primary machine + two duplicates). The two versions are stored on two different computers.
Hive	Hive was designed primarily to let users to interact with data in HDFS using SQL-like queries. Hive does this by converting SQL queries into MapReduce queries, which are subsequently run in Yarn on HDFS data. Hive also has data analysis and summarization capabilities.
Pig	Pig, like Hive, is a data processing and manipulation framework. Pig abstracts MapReduce and turns it to a scripting language, whereas Hive abstracts MapReduce and converts it to SQL.
Yarn	Yarn is in charge of cluster resource management and acts as a compute abstraction layer. It functions similarly to a cluster's operating system. Yarn also provides information on the amount of resources available in the cluster as well as the number of resources that do not have any tasks executing in them. Yarn has a non-functional requirement that unsuccessful tasks be automatically re-tried. As a result, it is responsible for work completion in the case of machine failure. Yarn's other non-functional role is to offer scalability, which means that if more storage is needed, Yarn is responsible for running the same operation (with the same code) while increasing the number of servers in the cluster.