

Big Data Graded Assignment 3

❖ Step 1 : Upload all files under bucket with public access

The screenshot shows the Google Cloud Storage interface. On the left, there's a sidebar with 'Cloud Storage' selected. The main area displays 'Bucket details' for '13042023bucket'. The bucket's location is 'us (multiple regions in United States)', storage class is 'Standard', public access is 'Subject to object ACLs', and protection is 'None'. Below this, there are tabs for 'OBJECTS', 'CONFIGURATION', 'PERMISSIONS', 'PROTECTION', 'LIFECYCLE', and 'OBSERVABILITY'. The 'OBJECTS' tab is active, showing a list of objects. The list has columns for Name, Size, Type, Created, Storage class, Last modified, and Public access. Two objects are listed: 'hash_file.txt' (548 B, text/plain) and 'hash_time.py' (1.2 KB, application/octet-stream). Both are public to the internet. The interface also includes a search bar at the top and various utility buttons like 'REFRESH', 'HELP ASSISTANT', and 'LEARN'.

Name	Size	Type	Created	Storage class	Last modified	Public access
hash_file.txt	548 B	text/plain	Apr 13, 2023, 7:31:52 AM	Standard	Apr 13, 2023, 7:32:09 AM	Public to internet
hash_time.py	1.2 KB	application/octet-stream	Apr 13, 2023, 7:43:12 AM	Standard	Apr 13, 2023, 7:43:27 AM	Public to internet

❖ Python File

```
> Big Data > GA3 > hash_time.py > ...
1 from pyspark.sql import SparkSession
2
3 #Creating spark session
4 spark= SparkSession.builder.appName("Spark").getOrCreate()
5 #hash file reading from GCP Bucket with spark command
6 df = spark.read.text("gs://13042023bucket/hash_file.txt")
7 #conversion from df to rdd
8 rdd = df.rdd
9 #splitting the rdd using tab separator
10 sep = rdd.map(lambda x :x[0].split("\t"))
11 #collecting time column only using lambda function
12 time = sep.map(lambda x:x[1])
13 #defining hash_time_map function for mapping
14 def hash_time_map(s):
15     val=('X',1)
16     if s!='Time':
17         final = int(s.replace(":", ""))
18         if final >=0 and final<=600:
19             val=("00-06",1)
20         elif final > 600 and final<=1200:
21             val=("06-12",1)
22         elif final > 1200 and final <=1800:
23             val=("12-18",1)
24         elif final>1800 and final <=2400:
25             val=("18-24",1)
26     return val
27 #hash_time_map function call to collect time
28 collect_time = time.map(lambda x: hash_time_map(x))
29 #removing header with sorted time
30 sorted_time = collect_time.filter(lambda x:x[0] != 'X').sortBy(lambda x: x[0])
31 #timezone grouping
32 timezone= sorted_time.reduceByKey(lambda a,b: a+b)
33 #final output
34 print(timezone.collect())
```

❖ Create a cluster which is running

The screenshot shows the Google Cloud Dataproc console for a project named "My First Project". The left sidebar lists various services: Jobs on Clusters (Clusters, Jobs, Workflows, Autoscaling policies), Serverless (Batches), Metastore Services (Metastore, Federation), and Utilities (Component exchange, Release Notes). The main content area is titled "Cluster details" and shows the status of a cluster named "cluster13042023". The cluster is in a "Running" state. Below the status, there are tabs for MONITORING, JOBS, VM INSTANCES, CONFIGURATION, and WEB INTERFACES. The MONITORING tab is active, displaying a graph of YARN memory usage over time. The graph shows a series of peaks and troughs, indicating memory usage fluctuations. The Y-axis is labeled "YARN memory" and ranges from 0GB to 8GB. The X-axis represents time, with a "1 hour" interval selected. A "SAVE AS DASHBOARD" button is visible. To the right of the graph, there is a section for "YARN pending memory" with a 1GB scale.

Google Cloud | My First Project | Search (/) for resources, docs, products, and more

Dataproc | Cluster details | SUBMIT JOB | REFRESH | START | STOP | DELETE | VIEW LOGS

Jobs on Clusters

- Clusters
- Jobs
- Workflows
- Autoscaling policies

Serverless

- Batches

Metastore Services

- Metastore
- Federation

Utilities

- Component exchange
- Release Notes

Cluster details

Consider using Auto Zone rather than selecting a zone manually. See <https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone> [MORE](#)

Name: cluster13042023
Cluster UUID: 017d5f84-1e5d-441b-8a18-b9aceb6fd3f6
Type: Dataproc Cluster
Status: Running

MONITORING | JOBS | VM INSTANCES | CONFIGURATION | WEB INTERFACES

SAVE AS DASHBOARD | RESET ZOOM | 1 hour | 6 hours | 12 hours | 1 day | 2 days | 4 days | 7 days | 14 days | 30 days | Custom

YARN memory

YARN pending memory

❖ Get gsutil url from file

The screenshot shows the Google Cloud Cloud Storage console for a project named "My First Project". The left sidebar lists various services: Buckets, Monitoring (NEW), Settings, Marketplace, and Release Notes. The main content area is titled "Object details" and shows the details of a file named "hash_time.py" in a bucket named "13042023bucket". The file is in a "LIVE OBJECT" state. Below the status, there are tabs for LIVE OBJECT and VERSION HISTORY. The LIVE OBJECT tab is active, displaying a table of object details. The table includes fields for Type, Size, Created, Last modified, Storage class, Custom time, Public URL, Authenticated URL, gsutil URI, Permissions, and Protection. The gsutil URI is highlighted in blue. The table also shows the Public access level as "Public to internet" and the Version history as "None".

Google Cloud | My First Project | Search (/) for resources, docs, products, and more

Cloud Storage | Object details

Buckets > 13042023bucket > hash_time.py

LIVE OBJECT | VERSION HISTORY

DOWNLOAD | EDIT METADATA | EDIT ACCESS | DELETE

Overview

Type	application/octet-stream
Size	1.2 KB
Created	Apr 13, 2023, 7:43:12 AM
Last modified	Apr 13, 2023, 7:43:27 AM
Storage class	Standard
Custom time	—
Public URL	https://storage.googleapis.com/13042023bucket/hash_time.py
Authenticated URL	https://storage.cloud.google.com/13042023bucket/hash_time.py
gsutil URI	gs://13042023bucket/hash_time.py

Permissions

Public access	Public to internet
---------------	--------------------

Protection

Version history	—
Retention policy	None

❖ Create & submit a job with correct output

The screenshot displays the Google Cloud Dataproc console interface. The top navigation bar includes the Google Cloud logo, the project name "My First Project", a search bar, and user profile icons. The left sidebar contains a navigation menu with categories like "Jobs on Clusters", "Serverless", "Metastore Services", and "Utilities". The "Jobs" option under "Jobs on Clusters" is selected.

The main content area is titled "Job details" and shows the following job information:

Job ID	job-51e51fcd
Job UUID	1df4f07f-0f7f-46a5-92f5-4ac9e91ef7ef
Type	Dataproc Job
Status	✔ Succeeded

Below the job details, there are tabs for "MONITORING" and "CONFIGURATION". A message states: "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."

The "Output" section is expanded, showing a log of messages:

- Spark jobs take ~60 seconds to initialize resources. (DISMISS)
- 23/04/13 02:41:45 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
- 23/04/13 02:41:45 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
- 23/04/13 02:41:45 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1681351778996_0002
- 23/04/13 02:41:46 INFO org.apache.hadoop.yarn.client.RMPProxy: Connecting to ResourceManager at cluster13042023-m/10.128.0.3:8030
- 23/04/13 02:41:50 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object [({'00-06', 4), ('06-12', 8), ('12-18', 12), ('18-24', 6)]
- 23/04/13 02:42:07 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@f2c6a3b(HTTP/1.1, {http/1.1}){0.0.0.0:0}

At the bottom, there is a link for "EQUIVALENT COMMAND LINE".