

Cluster

Google CloudMy First Project

SearchProducts, resources, docs (/)

Clusters

CREATE CLUSTERREFRESHSTARTSTOPDELETEREGIONS+ 5 RECOMMENDED ALERTS

FilterSearch clusters, press Enter

	Name	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
	cluster-3b15	Running	us-east1	us-east1-c	2	Off	my_bucket_tp_andi	Nov 6, 2022, 12:54:04 PM

Google CloudMy First Project

SearchProducts, resources, docs (/)

Cluster details

SUBMIT JOBSREFRESHSTARTSTOPDELETEVIEW LOGS

Bucket name 'my\_bucket\_tp\_andi' contains underscore, which may cause job failures.

Namecluster-3b15

Cluster UID7021a33b-09e5-42be-8c13-2ef3242907cd

TypeDataproc Cluster

StatusRunning

MONITORING

JOBS

VM INSTANCES

CONFIGURATION

WEB INTERFACES

SAVE AS CUSTOM DASHBOARD

RESET ZOOM

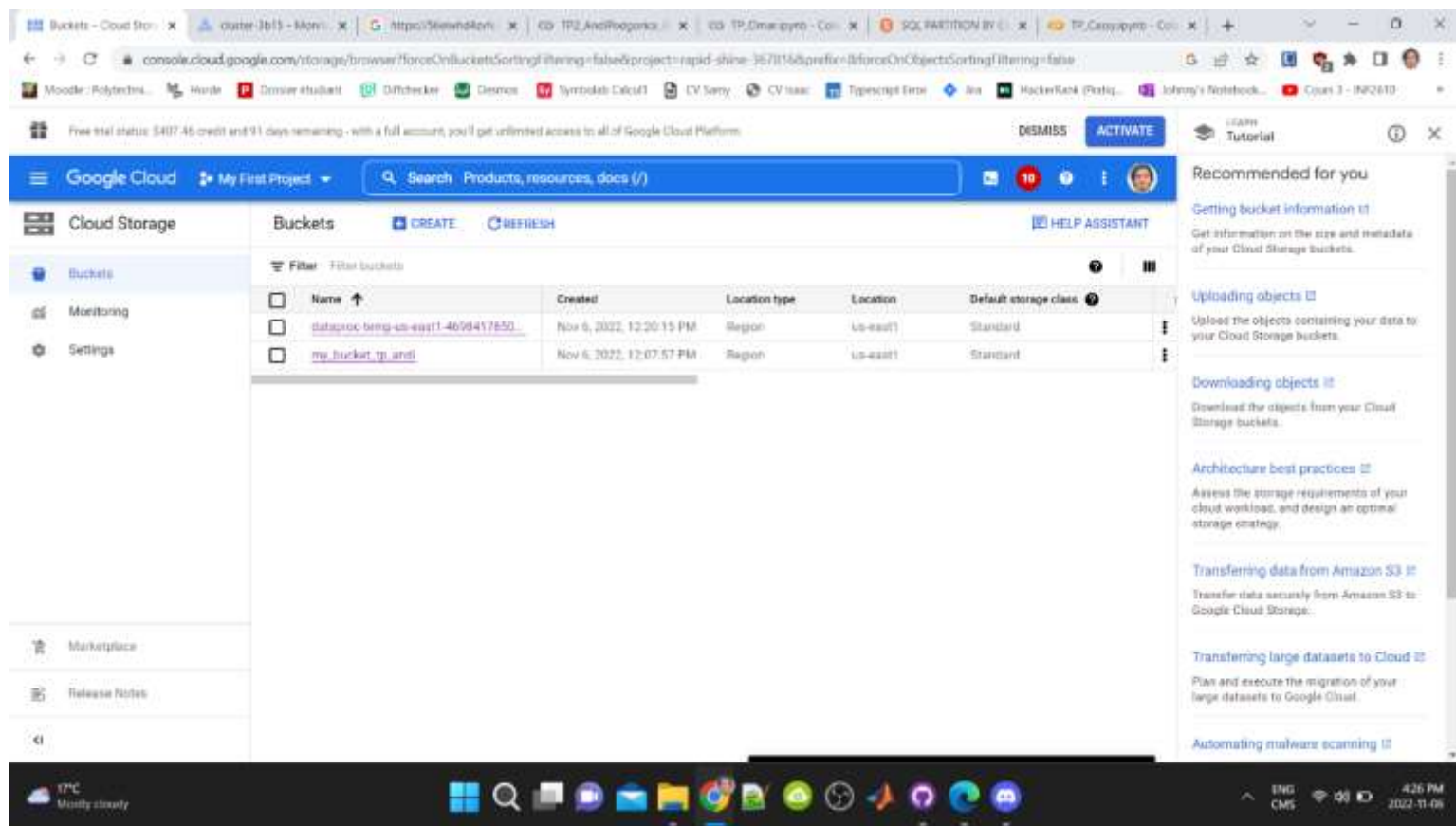
1 hour6 hours12 hours1 day2 days4 days7 days14 days30 daysCustom

YARN memory

YARN pending memory

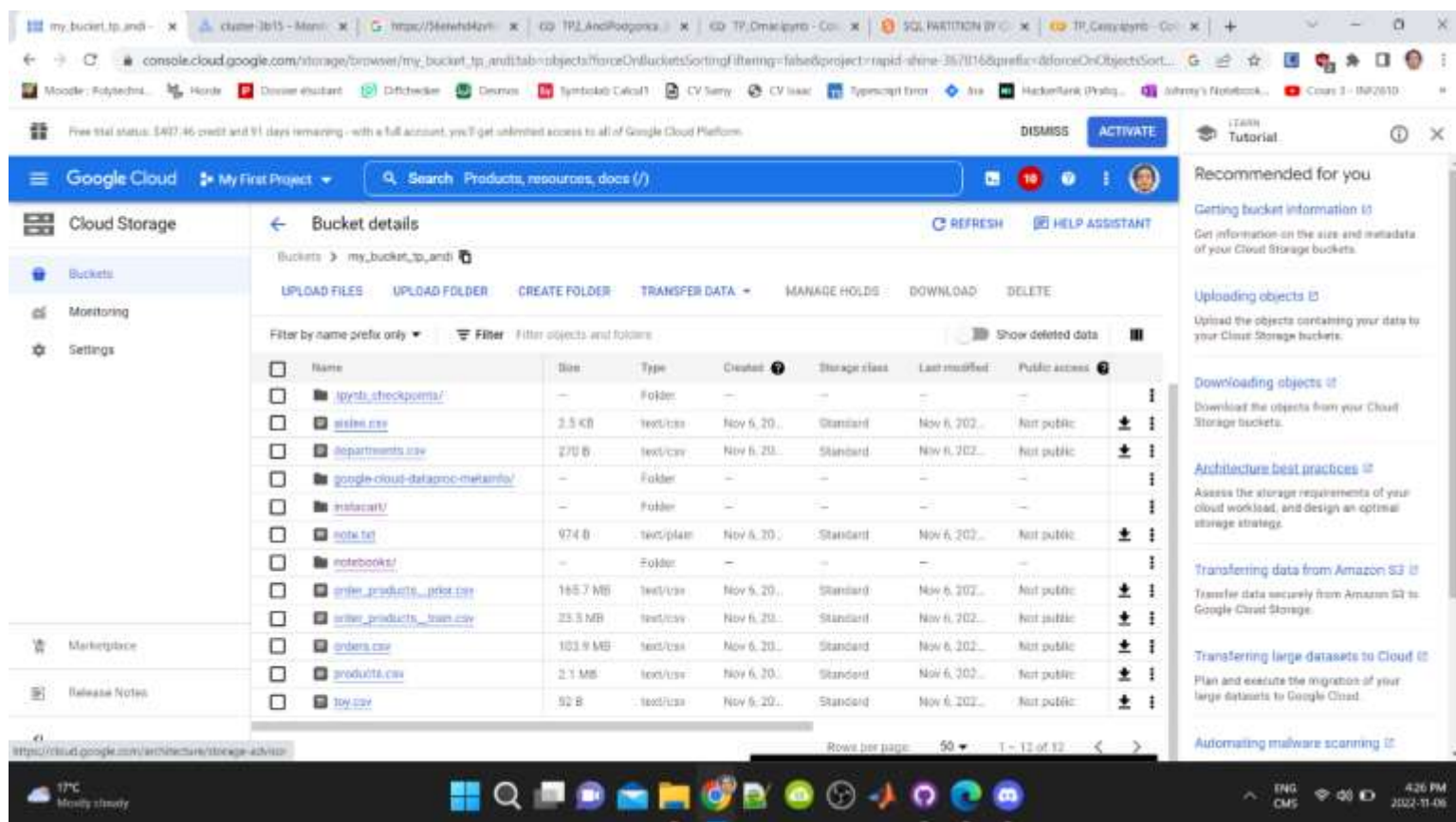
YARN NodeManagers

## Bucket (my\_bucket\_tp\_andi)



This screenshot shows the Google Cloud Storage Buckets page in the console. The left sidebar contains navigation links for Cloud Storage, Buckets, Monitoring, and Settings. The main content area displays a list of buckets with columns for Name, Created, Location type, Location, and Default storage class. Two buckets are listed: 'dataproc-templ-us-east1-469417850' and 'my\_bucket\_tp\_andi'. The right sidebar features a 'Recommended for you' section with links to various guides and tutorials.

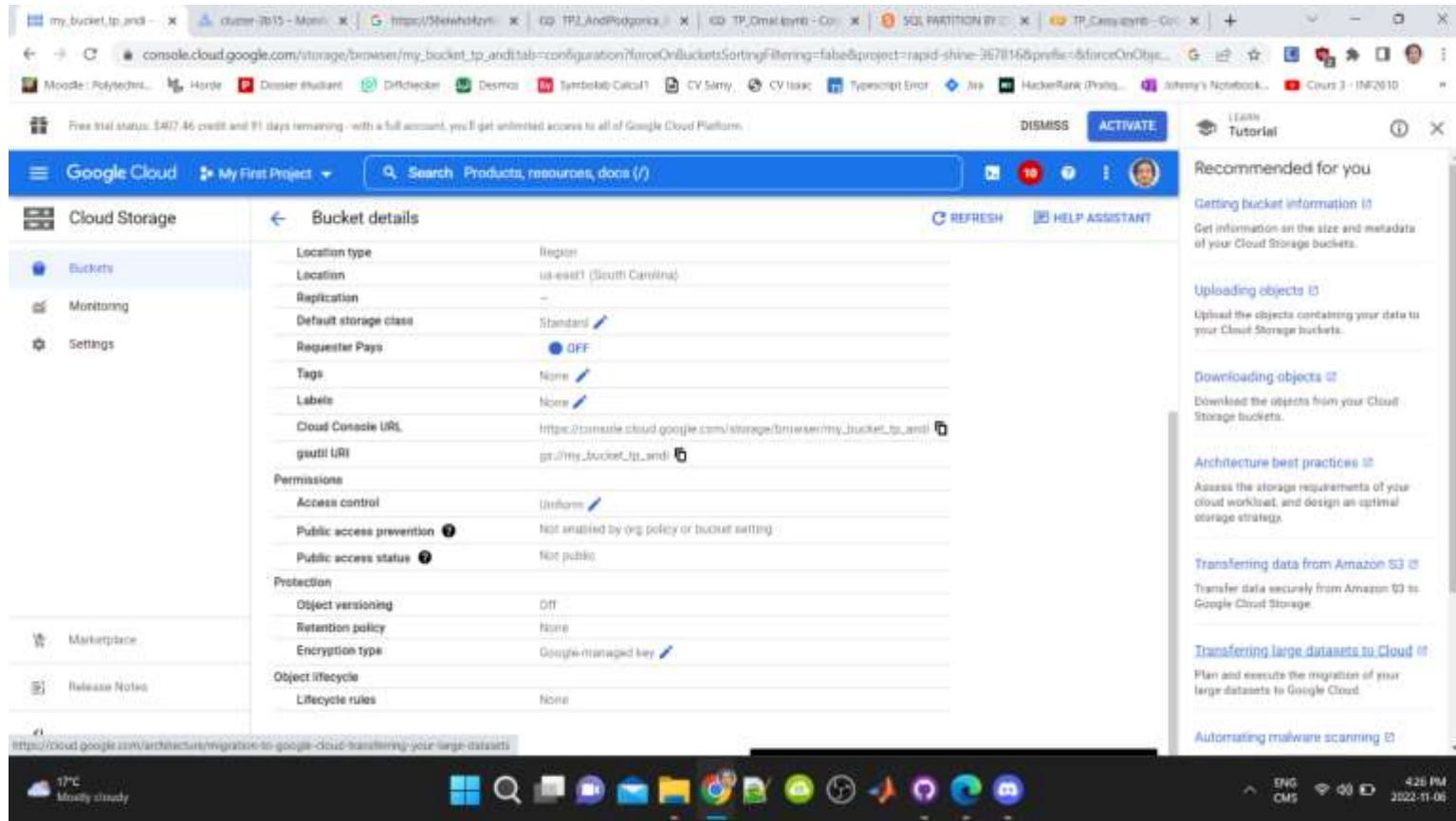
Name	Created	Location type	Location	Default storage class
<a href="#">dataproc-templ-us-east1-469417850</a>	Nov 6, 2022, 12:20:15 PM	Region	us-east1	Standard
<a href="#">my_bucket_tp_andi</a>	Nov 6, 2022, 12:07:57 PM	Region	us-east1	Standard



This screenshot shows the Google Cloud Storage Bucket details page for 'my\_bucket\_tp\_andi'. The left sidebar is the same as the previous screenshot. The main content area displays the bucket's details, including a list of objects with columns for Name, Size, Type, Created, Storage class, Last modified, and Public access. The right sidebar is also the same as the previous screenshot.

Name	Size	Type	Created	Storage class	Last modified	Public access
<a href="#">ipynb_checkpoints/</a>	—	Folder	—	—	—	—
<a href="#">index.csv</a>	2.5 KB	text/csv	Nov 6, 2022	Standard	Nov 6, 2022	Not public
<a href="#">departments.csv</a>	270 B	text/csv	Nov 6, 2022	Standard	Nov 6, 2022	Not public
<a href="#">google-cloud-dataproc-metadata/</a>	—	Folder	—	—	—	—
<a href="#">install.sh</a>	—	Folder	—	—	—	—
<a href="#">note.txt</a>	974 B	text/plain	Nov 6, 2022	Standard	Nov 6, 2022	Not public
<a href="#">notebooks/</a>	—	Folder	—	—	—	—
<a href="#">order_products_prior.csv</a>	165.7 MB	text/csv	Nov 6, 2022	Standard	Nov 6, 2022	Not public
<a href="#">order_products_train.csv</a>	23.3 MB	text/csv	Nov 6, 2022	Standard	Nov 6, 2022	Not public
<a href="#">orders.csv</a>	103.9 MB	text/csv	Nov 6, 2022	Standard	Nov 6, 2022	Not public
<a href="#">products.csv</a>	2.1 MB	text/csv	Nov 6, 2022	Standard	Nov 6, 2022	Not public
<a href="#">toy.csv</a>	92 B	text/csv	Nov 6, 2022	Standard	Nov 6, 2022	Not public

## Bucket details

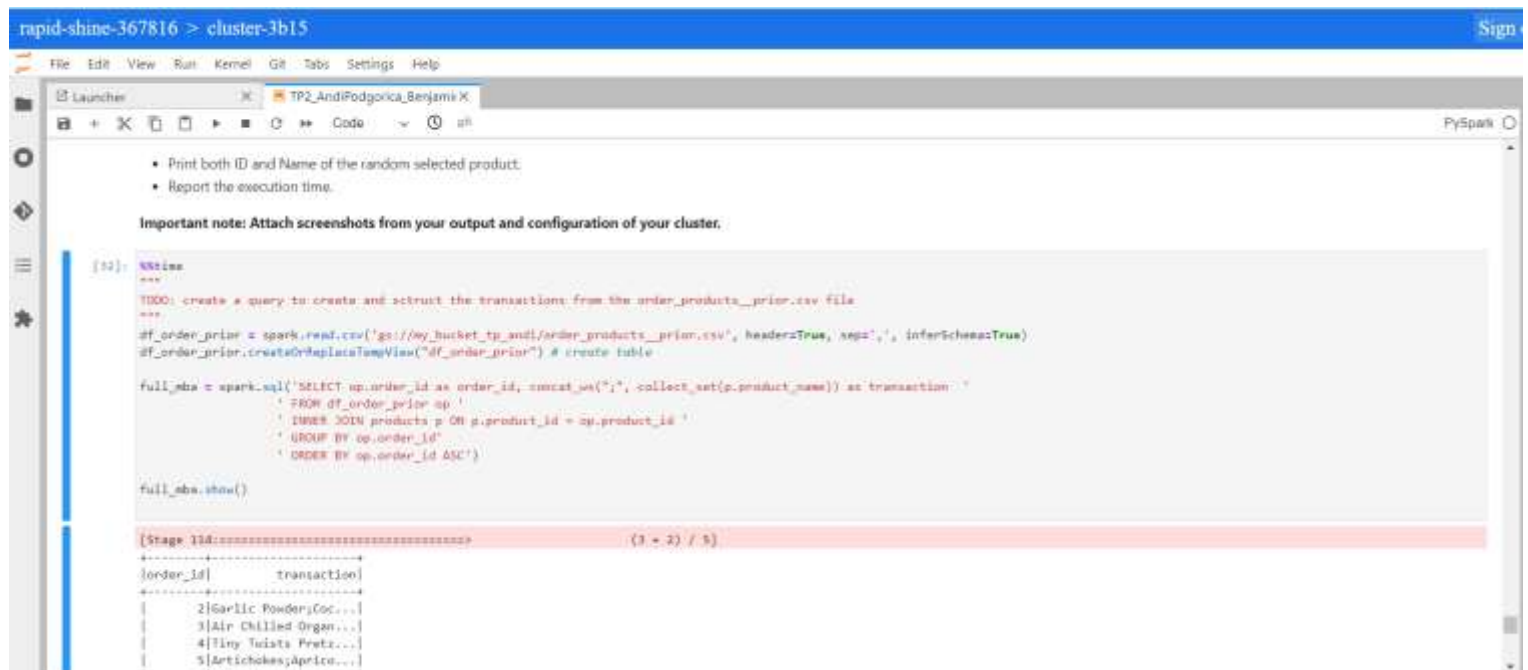


The screenshot shows the Google Cloud console interface for a Cloud Storage bucket named 'my\_bucket\_tp\_and'. The left sidebar contains navigation links for Cloud Storage, Buckets, Monitoring, and Settings. The main area displays the 'Bucket details' page with various configuration options.

Property	Value
Location type	Region
Location	us-east1 (South Carolina)
Replication	—
Default storage class	Standard
Requester Pays	OFF
Tags	None
Labels	None
Cloud Console URL	<a href="https://console.cloud.google.com/storage/browser/my_bucket_tp_and">https://console.cloud.google.com/storage/browser/my_bucket_tp_and</a>
gsutil URL	<a href="gs://my_bucket_tp_and">gs://my_bucket_tp_and</a>
Permissions	
Access control	Uniform
Public access prevention	Not enabled by org policy or bucket setting
Public access status	Not public
Protection	
Object versioning	Off
Retention policy	None
Encryption type	Google-managed key
Object lifecycle	
Lifecycle rules	None

On the right side, there is a 'Recommended for you' section with links to various guides and best practices.

## 4. MBA for the full dataset cluster



The screenshot shows a Jupyter Notebook interface with a code cell containing Spark SQL queries. The output of the code is displayed below the cell.

```
[12]: %sql
---
TODD: create a query to create and extract the transactions from the order_products_prior.csv file
---
df_order_prior = spark.read.csv('gs://my_bucket_tp_and/order_products_prior.csv', header=True, sep=',', inferSchema=True)
df_order_prior.createOrReplaceTempView("df_order_prior") # create table

full_mba = spark.sql('SELECT op.order_id as order_id, concat_ws(";", collect_set(p.product_name)) as transaction
FROM df_order_prior op
INNER JOIN products p ON p.product_id = op.product_id
GROUP BY op.order_id
ORDER BY op.order_id ASC')

full_mba.show()
```

The output shows the execution of the query, resulting in a table with two columns: 'order\_id' and 'transaction'. The first row of data is shown below:

order_id	transaction
2	Garlic Powder;Coc...

```
.....*
[order_id]      transaction]
.....*
```

```
2|Garlic Powder;Coc...
3|Air Chilled Organ...
4|Tiny Twists Pretz...
5|Artichokes;Aprico...
6|Clean Day Lavende...
7|Orange Juice;Pine...
8|Original Hawaiian...
9|French Baguettes,...
10|Organic Cilantro,...
11|Mango Pineapple S...
12|All Natural Bone...
13|Handmade Vodka Fr...
14|Total Greek Strai...
15|Organic Extra Vir...
16|Sea Salt Made Wit...
18|Globe Eggplant;Sa...
19|Organic Whole Whi...
20|Milla Wafers;Red ...
21|Organic Firm Tofu...
22|Cream Cheese;Pres...
```

```
.....*
only showing top 20 rows
```

```
CPV times: user 10.2 ms, sys: 4.00 ms, total: 20.3 ms
Wall time: 22.7 s
```