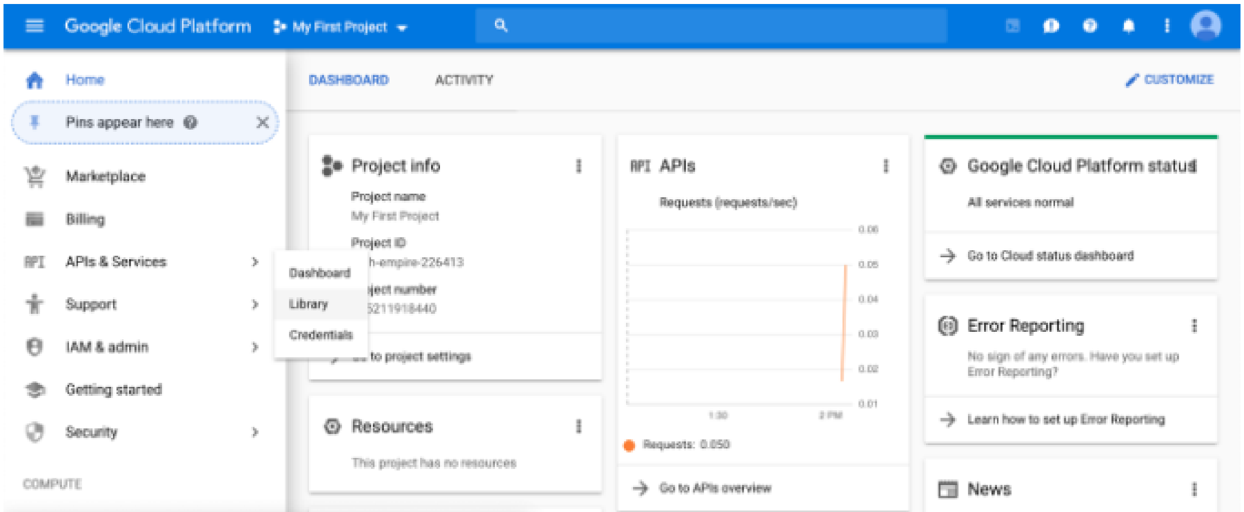
SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK

SERIES OF STEPS PERFORMED IN GOOGLE CLOUD PLATFORM USING CLOUD DATAPRO



Project name
My First Project

Project ID
h-empire-226413

Project number
5211918440

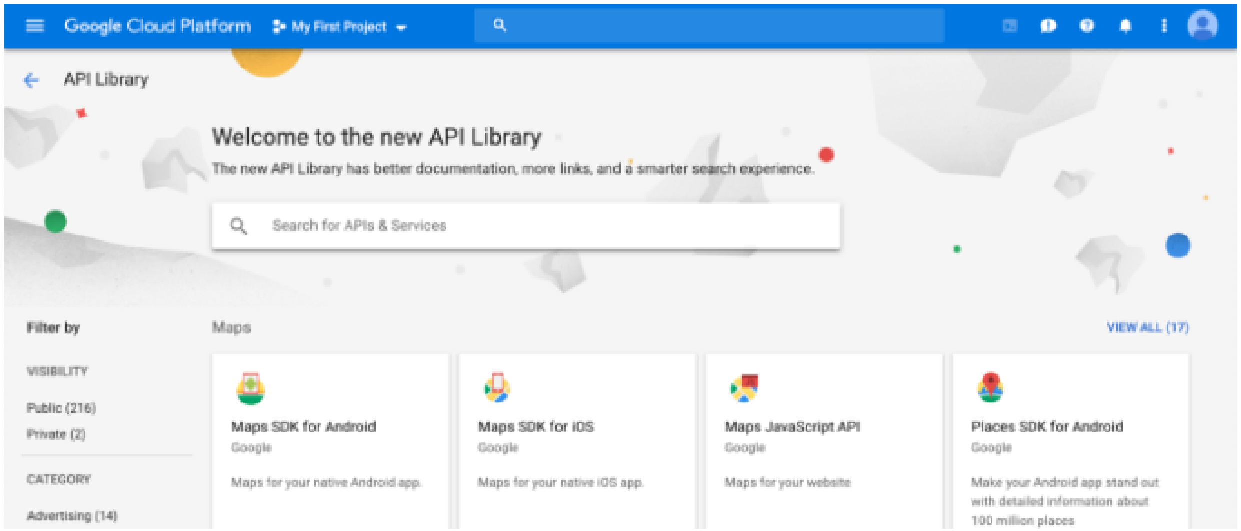
Requests (requests/sec)

Requests: 0.050

All services normal

No sign of any errors. Have you set up Error Reporting?

This project has no resources



Welcome to the new API Library

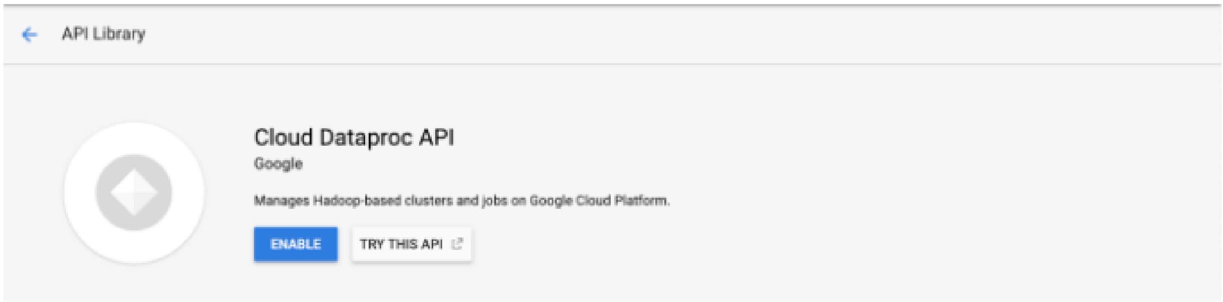
The new API Library has better documentation, more links, and a smarter search experience.

Maps SDK for Android
Google
Maps for your native Android app.

Maps SDK for iOS
Google
Maps for your native iOS app.

Maps JavaScript API
Google
Maps for your website

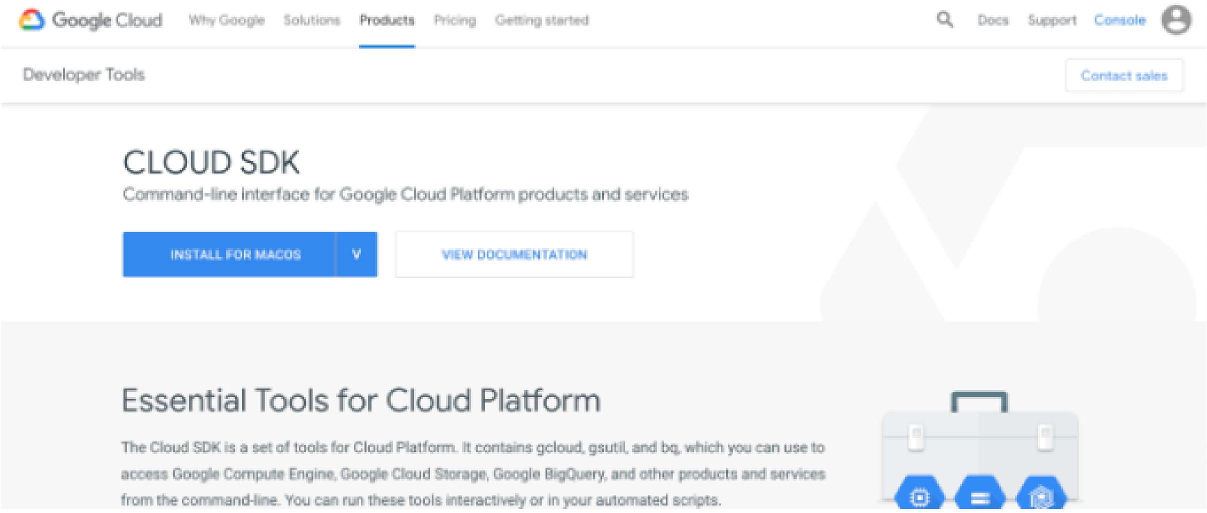
Places SDK for Android
Google
Make your Android app stand out with detailed information about 100 million places



Cloud Dataproc API
Google

Manages Hadoop-based clusters and jobs on Google Cloud Platform.

SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK



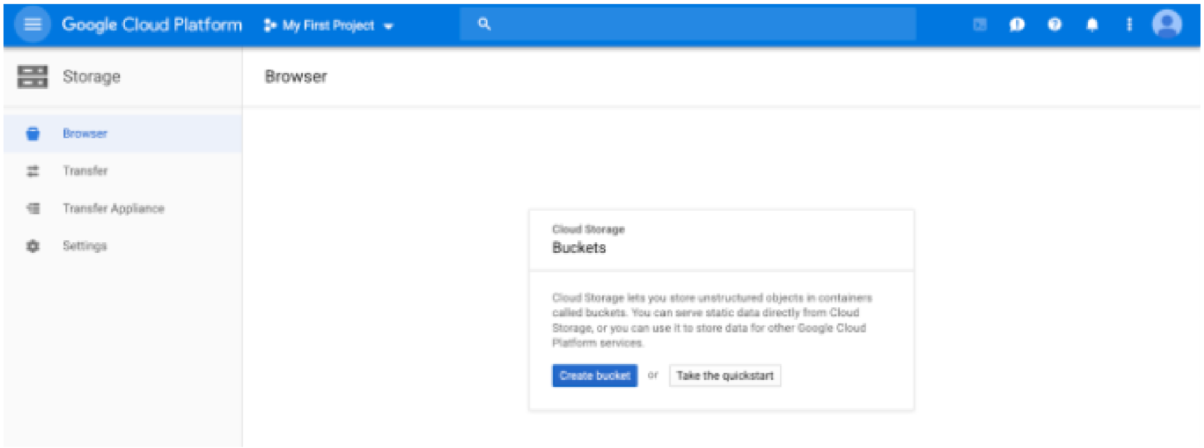
CLOUD SDK

Command-line interface for Google Cloud Platform products and services

Essential Tools for Cloud Platform

The Cloud SDK is a set of tools for Cloud Platform. It contains gcloud, gsutil, and bq, which you can use to access Google Compute Engine, Google Cloud Storage, Google BigQuery, and other products and services from the command-line. You can run these tools interactively or in your automated scripts.

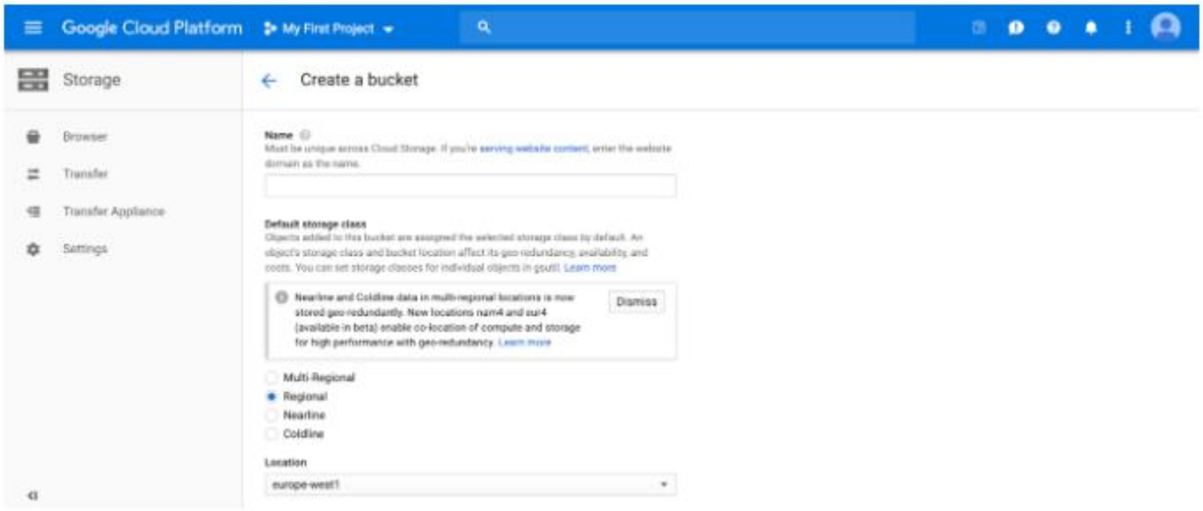
Install Google Cloud SDK by following instructions on https://cloud.google.com/sdk/



Cloud Storage

Buckets

Cloud Storage lets you store unstructured objects in containers called buckets. You can serve static data directly from Cloud Storage, or you can use it to store data for other Google Cloud Platform services.



Create a bucket

Name

Must be unique across Cloud Storage. If you're serving website content, enter the website domain as the name.

Default storage class

Objects added to this bucket are assigned the selected storage class by default. An object's storage class and bucket location affect its geo-redundancy, availability, and costs. You can set storage classes for individual objects in gsutil. Learn more

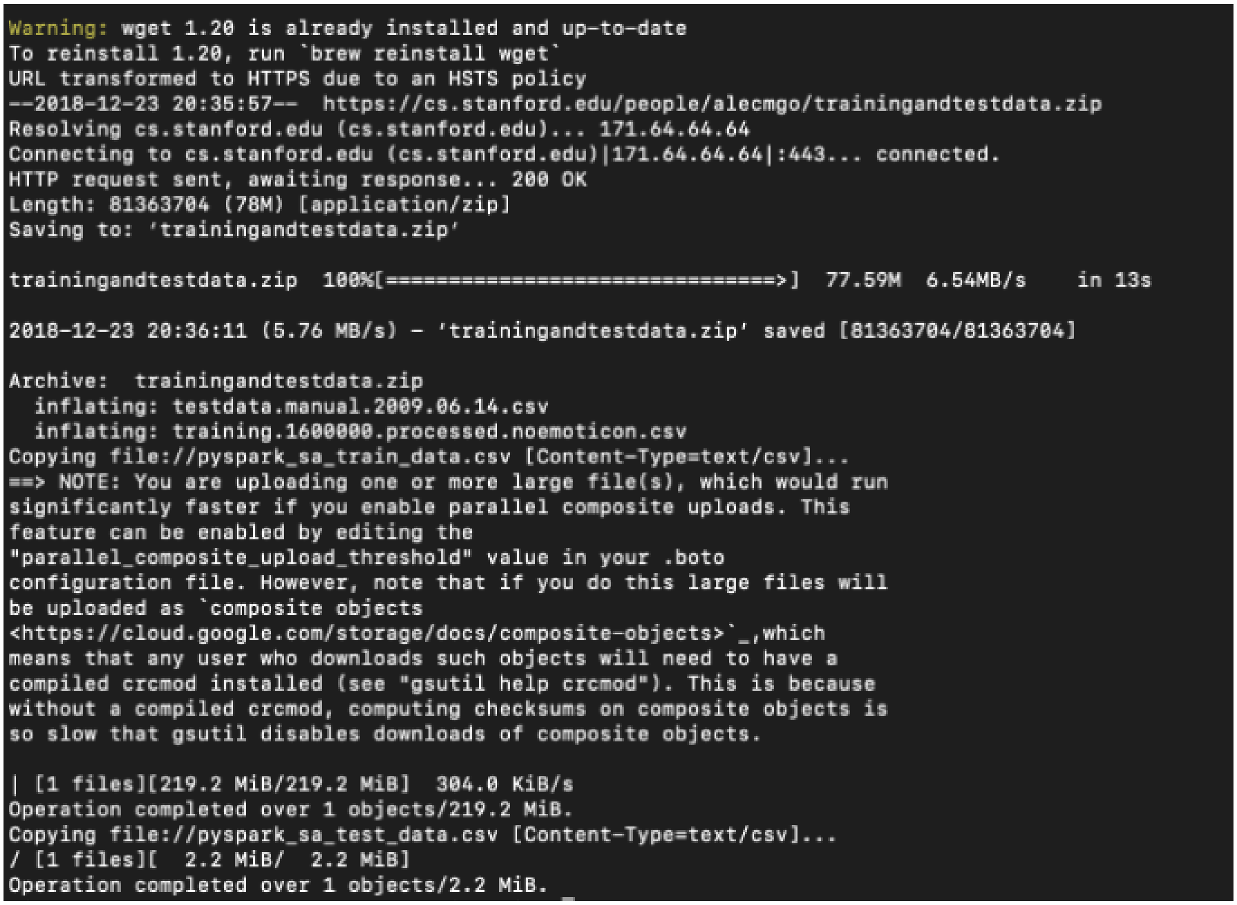
Nearline and Coldline data in multi-regional locations is now stored geo-redundantly. New locations nam4 and eur4 (available in beta) enable co-location of compute and storage for high performance with geo-redundancy. Learn more

○ Multi-Regional
● Regional
○ Nearline
○ Coldline

Location

europe-west1

SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK



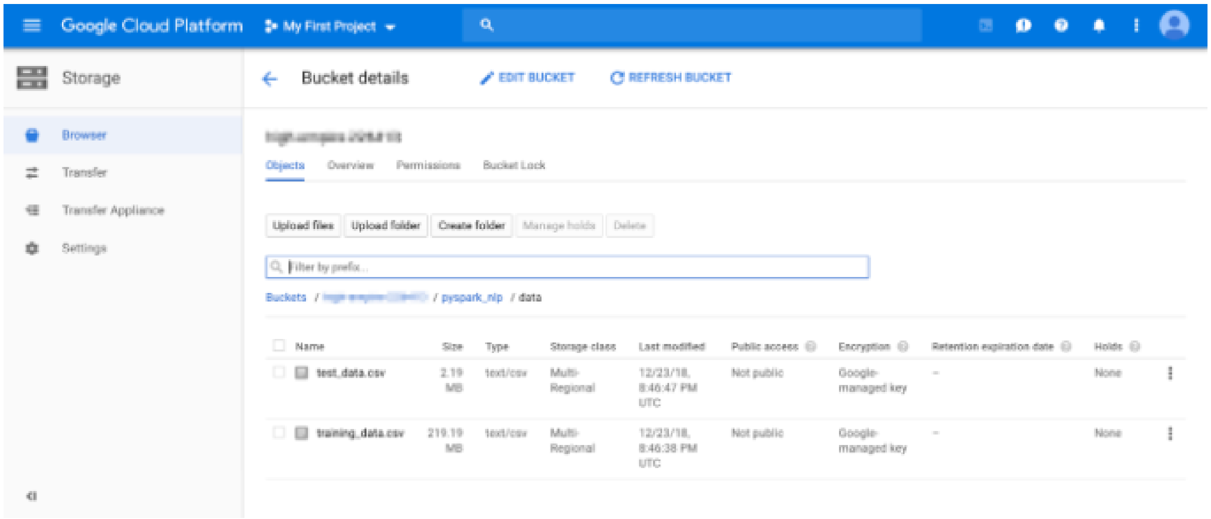
```
Warning: wget 1.20 is already installed and up-to-date
To reinstall 1.20, run `brew reinstall wget`
URL transformed to HTTPS due to an HSTS policy
--2018-12-23 20:35:57--  https://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip
Resolving cs.stanford.edu (cs.stanford.edu)... 171.64.64.64
Connecting to cs.stanford.edu (cs.stanford.edu)|171.64.64.64|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 81363704 (78M) [application/zip]
Saving to: 'trainingandtestdata.zip'

trainingandtestdata.zip  100%[===============================>]  77.59M  6.54MB/s    in 13s

2018-12-23 20:36:11 (5.76 MB/s) - 'trainingandtestdata.zip' saved [81363704/81363704]

Archive:  trainingandtestdata.zip
  inflating: testdata.manual.2009.06.14.csv
  inflating: training.1600000.processed.noemoticon.csv
Copying file://pyspark_sa_train_data.csv [Content-Type=text/csv]...
==> NOTE: You are uploading one or more large file(s), which would run
significantly faster if you enable parallel composite uploads. This
feature can be enabled by editing the
"parallel_composite_upload_threshold" value in your .boto
configuration file. However, note that if you do this large files will
be uploaded as `composite objects
<https://cloud.google.com/storage/docs/composite-objects>`_,which
means that any user who downloads such objects will need to have a
compiled crcmod installed (see "gsutil help crcmod"). This is because
without a compiled crcmod, computing checksums on composite objects is
so slow that gsutil disables downloads of composite objects.

| [1 files][219.2 MiB/219.2 MiB]  304.0 KiB/s
Operation completed over 1 objects/219.2 MiB.
Copying file://pyspark_sa_test_data.csv [Content-Type=text/csv]...
/ [1 files][  2.2 MiB/  2.2 MiB]
Operation completed over 1 objects/2.2 MiB.
```

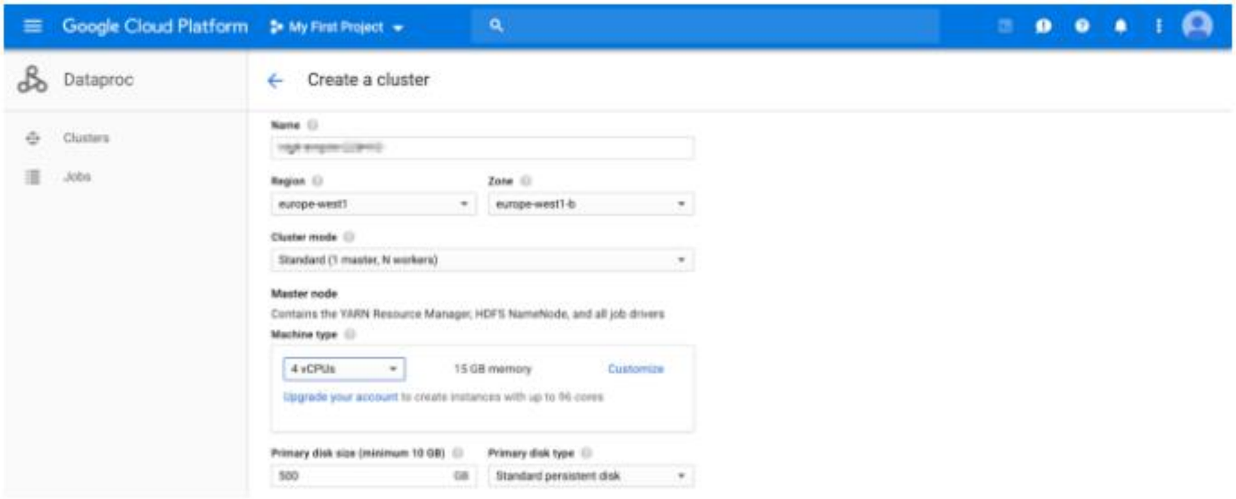


[redacted]

Buckets / [redacted] / pyspark_nlp / data

Name	Size	Type	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
test_data.csv	2.19 MB	text/csv	Multi-Regional	12/23/18, 8:46:47 PM UTC	Not public	Google-managed key	–	None
training_data.csv	219.19 MB	text/csv	Multi-Regional	12/23/18, 8:46:38 PM UTC	Not public	Google-managed key	–	None

SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK



Name: [redacted]

Region: europe-west1

Zone: europe-west1-b

Cluster mode: Standard (1 master, N workers)

Master node

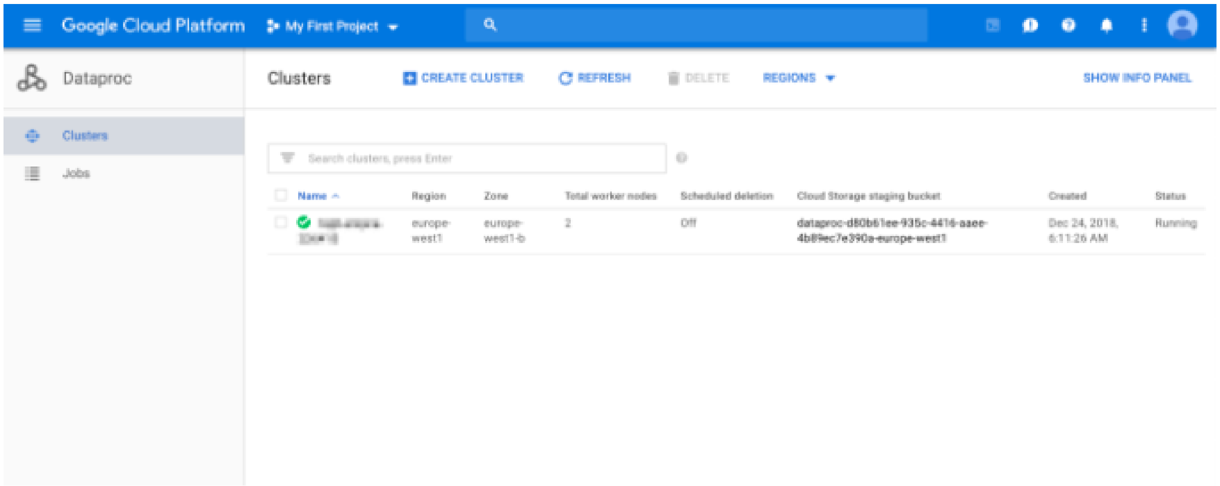
Contains the YARN Resource Manager, HDFS NameNode, and all job drivers

Machine type: 4 vCPUs 15 GB memory

Upgrade your account to create instances with up to 96 cores

Primary disk size (minimum 10 GB): 500 GB

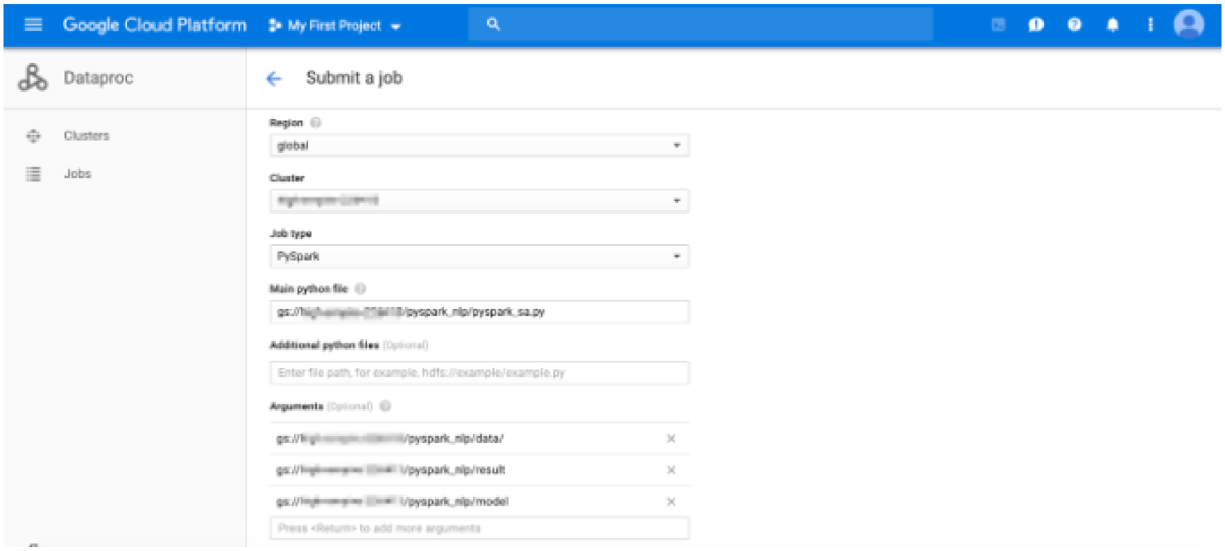
Primary disk type: Standard persistent disk



Clusters

Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
[redacted]	europe-west1	europe-west1-b	2	Off	dataproc-d80b61ee-935c-4416-aaee-4b89ec7e390a-europe-west1	Dec 24, 2018, 6:11:26 AM	Running

SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK



Submit a job

Region: global

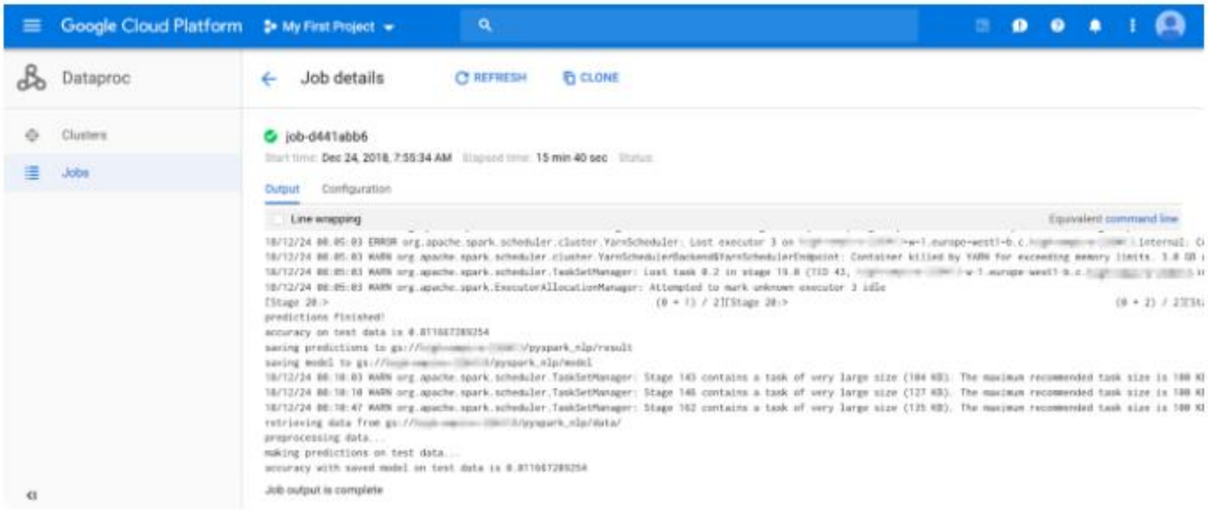
Cluster: [redacted]

Job type: PySpark

Main python file: gs://[redacted]/pyspark_nlp/pyspark_sa.py

Additional python files (Optional): Enter file path, for example, hdfs://example/example.py

Arguments (Optional):
gs://[redacted]/pyspark_nlp/data/
gs://[redacted]/pyspark_nlp/result
gs://[redacted]/pyspark_nlp/model
Press <Return> to add more arguments



Job details

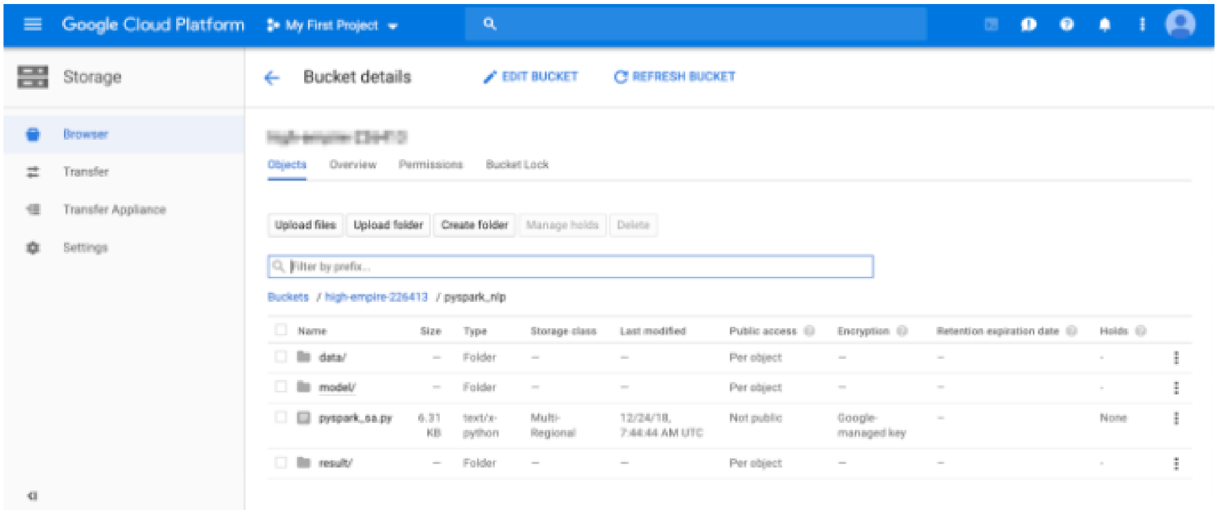
job-d441abb6

Start time: Dec 24, 2018, 7:55:34 AM Elapsed time: 15 min 40 sec Status:

```
18/12/24 08:05:03 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 3 on [redacted]-w-1.europe-west1-b.c.[redacted].internal: C
18/12/24 08:05:03 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Container killed by YARN for exceeding memory limits. 3.0 GB o
18/12/24 08:05:03 WARN org.apache.spark.scheduler.TaskSetManager: Lost task 0.2 in stage 19.0 (TID 43, [redacted]-w-1.europe-west1-b.c.[redacted] e
18/12/24 08:05:03 WARN org.apache.spark.ExecutorAllocationManager: Attempted to mark unknown executor 3 idle
[Stage 20:>                                                   (0 + 1) / 2][Stage 20:>                                                   (0 + 2) / 2][St
predictions finished!
accuracy on test data is 0.811667289254
saving predictions to gs://[redacted]/pyspark_nlp/result
saving model to gs://[redacted]/pyspark_nlp/model
18/12/24 08:10:03 WARN org.apache.spark.scheduler.TaskSetManager: Stage 143 contains a task of very large size (104 KB). The maximum recommended task size is 100 KB
18/12/24 08:10:10 WARN org.apache.spark.scheduler.TaskSetManager: Stage 146 contains a task of very large size (127 KB). The maximum recommended task size is 100 KB
18/12/24 08:10:47 WARN org.apache.spark.scheduler.TaskSetManager: Stage 162 contains a task of very large size (125 KB). The maximum recommended task size is 100 KB
retrieving data from gs://[redacted]/pyspark_nlp/data/
preprocessing data...
making predictions on test data...
accuracy with saved model on test data is 0.811667289254
```

Job output is complete

SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK



[redacted]

Buckets / high-empire-226413 / pyspark_nlp

	Name	Size	Type	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
□	data/	–	Folder	–	–	Per object	–	–	-
□	model/	–	Folder	–	–	Per object	–	–	-
□	pyspark_sa.py	6.31 KB	text/x-python	Multi-Regional	12/24/18, 7:44:44 AM UTC	Not public	Google-managed key	–	None
□	result/	–	Folder	–	–	Per object	–	–	-