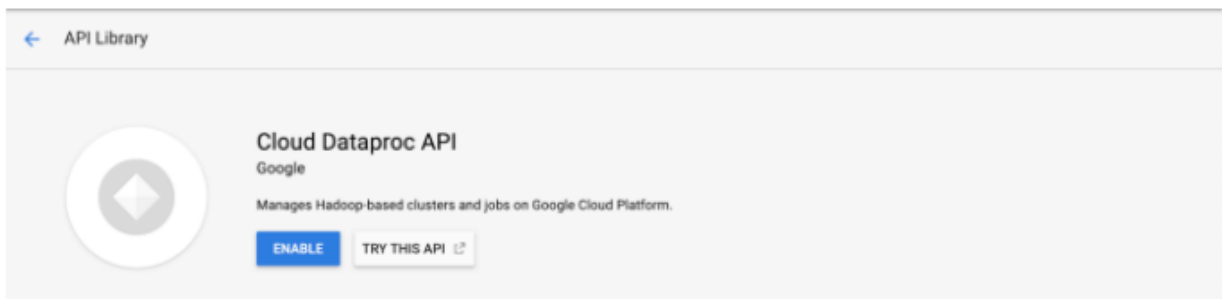
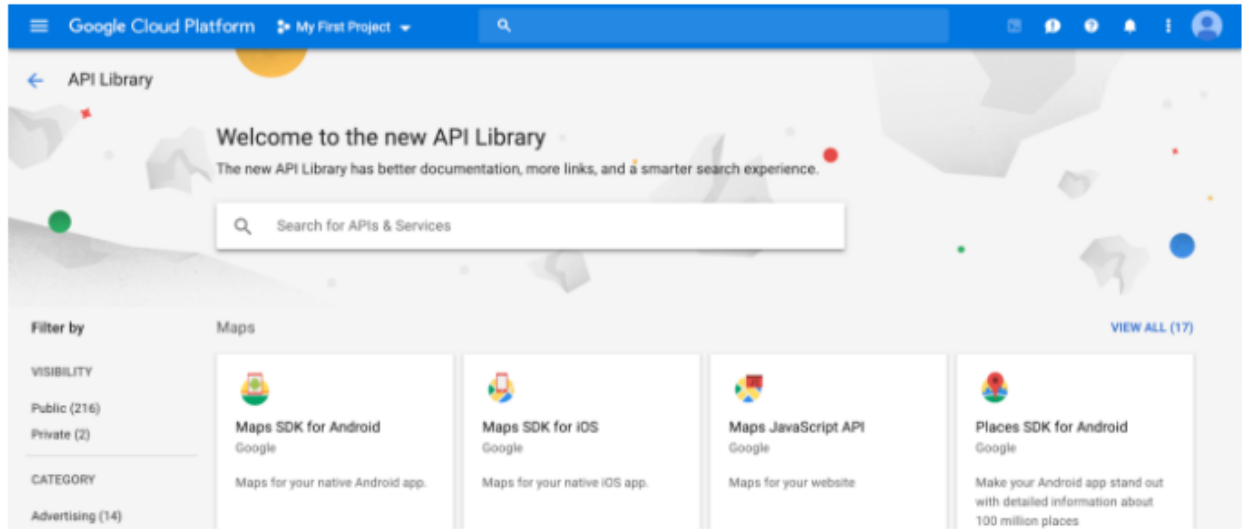
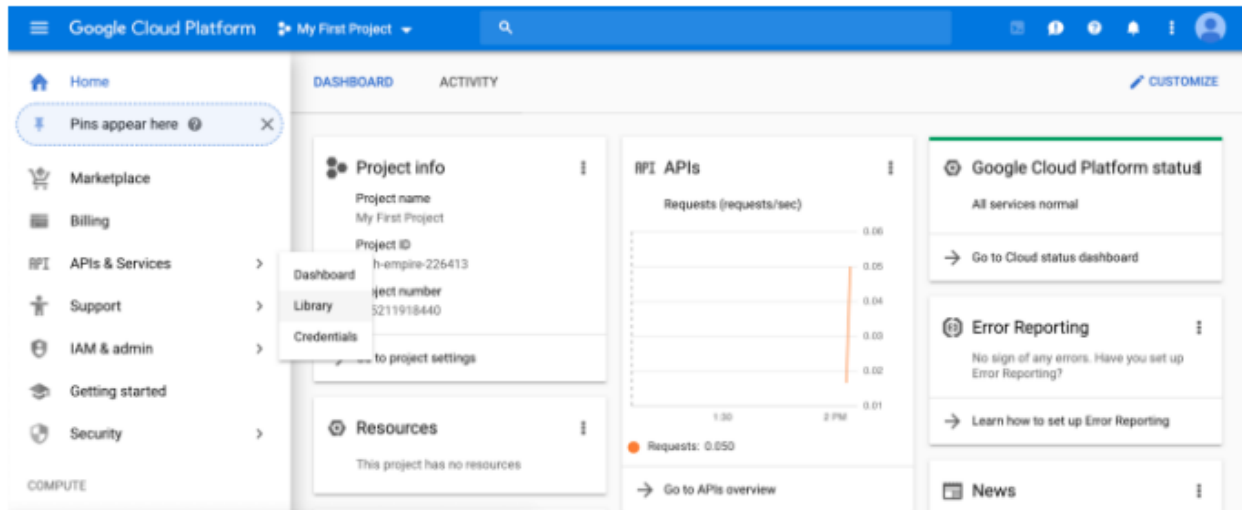
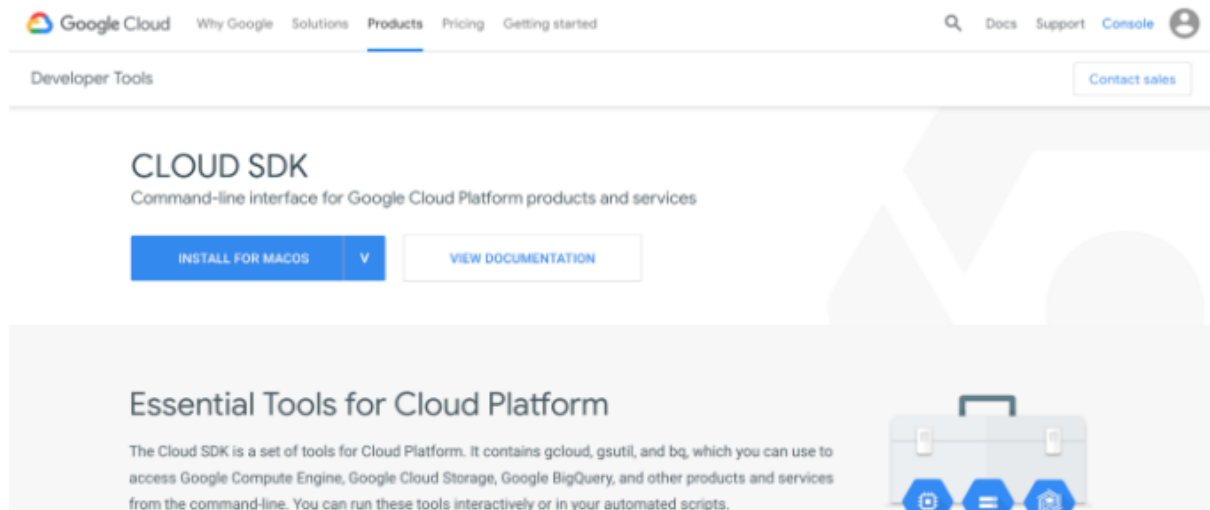


## SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK

### SERIES OF STEPS PERFORMED IN GOOGLE CLOUD PLATFORM USING CLOUD DATAPRO



# SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK



The image shows the Google Cloud SDK landing page. At the top, there's a navigation bar with links for Google Cloud, Why Google, Solutions, Products, Pricing, and Getting started. Below this, a 'Developer Tools' section highlights the 'CLOUD SDK' as a 'Command-line interface for Google Cloud Platform products and services'. It features two buttons: 'INSTALL FOR MACOS' and 'VIEW DOCUMENTATION'. A large graphic of a person is in the background. Below the SDK section, there's a section titled 'Essential Tools for Cloud Platform' which describes the Cloud SDK as a set of tools (gcloud, gsutil, bq) for accessing various Google Cloud services. It includes an illustration of a toolbox.

Google Cloud

Why Google Solutions Products Pricing Getting started

Developer Tools

Contact sales

## CLOUD SDK

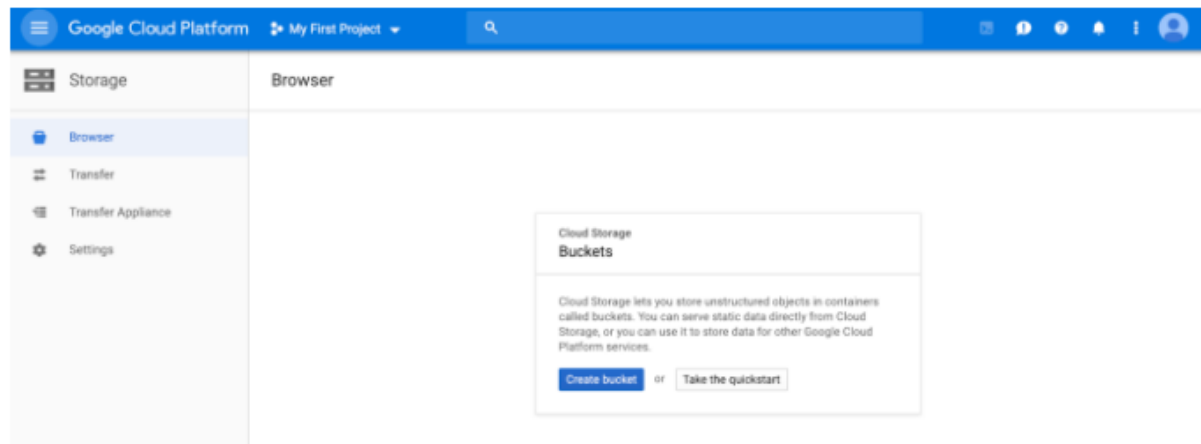
Command-line interface for Google Cloud Platform products and services

INSTALL FOR MACOS VIEW DOCUMENTATION

## 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/>



The image shows the Google Cloud Platform 'Storage' section, specifically the 'Browser' view. The left sidebar contains links for Storage, Browser, Transfer, Transfer Appliance, and Settings. The main content area is titled 'Cloud Storage Buckets' and provides a brief explanation of what buckets are. It includes two buttons: 'Create bucket' and 'Take the quickstart'.

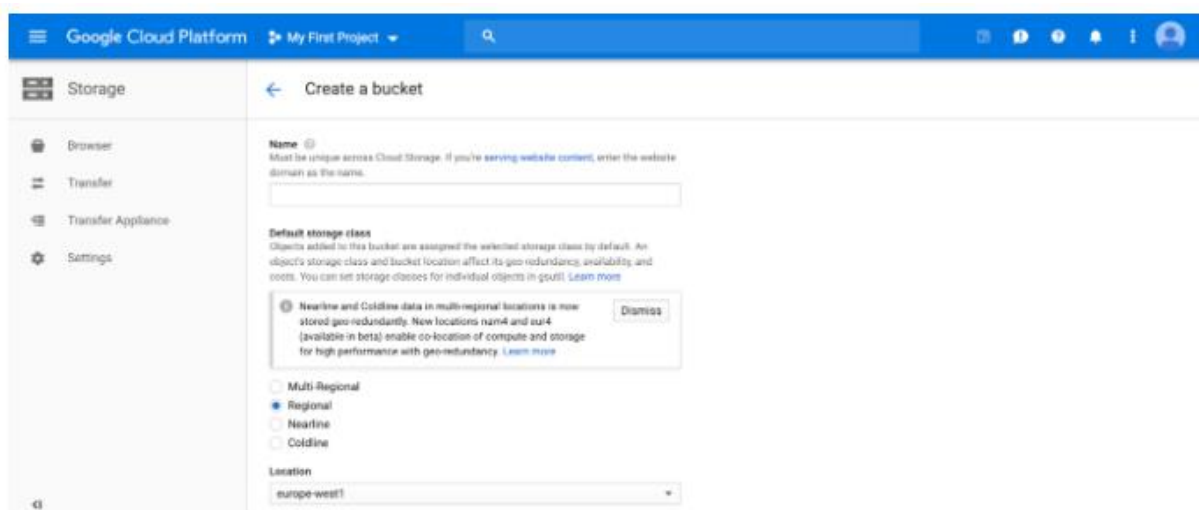
Google Cloud Platform My First Project

Storage Browser

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 bucket or Take the quickstart



The image shows the 'Create a bucket' form in the Google Cloud Platform. The left sidebar is the same as the previous image. The main content area is titled 'Create a bucket' and contains several fields and options. The 'Name' field is required and must be unique. The 'Default storage class' section has a warning about Nearline and Coldline data in multi-regional locations. Below this, there are radio buttons for 'Multi-Regional', 'Regional' (selected), 'Nearline', and 'Coldline'. The 'Location' dropdown is set to 'europe-west1'.

Google Cloud Platform My First Project

Storage 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.
```

Google Cloud Platform My First Project

Storage

Bucket details EDIT BUCKET REFRESH BUCKET

test\_data.csv

Objects Overview Permissions Bucket Lock

Upload file Upload folder Create folder Manage holds Delete

Filter by prefix...

Buckets / pyspark\_nlp / data

Name	Size	Type	Storage class	Last modified	Public access	Encryption	Retention expiration date	Hold
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

The screenshot shows the 'Create a cluster' page in the Google Cloud Platform console. The left sidebar has 'Dataproc' selected, with 'Clusters' and 'Jobs' sub-options. The main area is titled 'Create a cluster' and contains the following configuration fields:

- Name:** A text input field with the value 'high-usage-cluster'.
- Region:** A dropdown menu set to 'europe-west1'.
- Zone:** A dropdown menu set to 'europe-west1-b'.
- Cluster mode:** A dropdown menu set to 'Standard (1 master, N workers)'.
- Master node:** A section containing the text 'Contains the YARN Resource Manager, HDFS NameNode, and all job drivers'.
- Machine type:** A dropdown menu set to '4 vCPUs', with '15 GB memory' and a 'Customize' link.
- Primary disk size (minimum 10 GB):** A text input field with the value '500'.
- Primary disk type:** A dropdown menu set to 'Standard persistent disk'.

The screenshot shows the 'Clusters' page in the Google Cloud Platform console. The left sidebar has 'Dataproc' selected, with 'Clusters' and 'Jobs' sub-options. The main area is titled 'Clusters' and contains a table of clusters. The table has the following columns: Name, Region, Zone, Total worker nodes, Scheduled deletion, Cloud Storage staging bucket, Created, and Status.

Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
<input type="checkbox"/> <a href="#">high-usage-cluster</a>	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 PYSARK

Google Cloud Platform My First Project

Dataproc Submit a job

Clusters Jobs

Region: global

Cluster: high-memory-224613

Job type: PySpark

Main python file: gs://high-memory-224613/pyspark\_nlp/pyspark\_sa.py

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

Arguments (Optional):

- gs://high-memory-224613/pyspark\_nlp/data/
- gs://high-memory-224613/pyspark\_nlp/result
- gs://high-memory-224613/pyspark\_nlp/model

Press <Return> to add more arguments

Google Cloud Platform My First Project

Dataproc Job details REFRESH CLONE

Clusters Jobs

job-d441abb6

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

Output Configuration

Line mapping Equivalent command line

```
18/12/24 08:05:03 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 3 on high-memory-224613-w-1-europe-west1-b.c.high-memory-224613-internal: 0
18/12/24 08:05:03 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Container killed by YARN for exceeding memory limits. 3.8 GB
18/12/24 08:05:03 WARN org.apache.spark.scheduler.TaskSetManager: Lost task 0.2 in stage 10.0 (TID 43), high-memory-224613-w-1-europe-west1-b.c.high-memory-224613-internal: 0
18/12/24 08:05:03 WARN org.apache.spark.ExecutorAllocationManager: Attempted to mark unknown executor 3 idle
(Stage 28-> (0 + 1) / 2)Stage 28-> (0 + 2) / 2(34)
predictions Finished!
accuracy on test data is 0.811067289254
saving predictions to gs://high-memory-224613/pyspark_nlp/result
saving model to gs://high-memory-224613/pyspark_nlp/model
18/12/24 08:18:03 WARN org.apache.spark.scheduler.TaskSetManager: Stage 143 contains a task of very large size (184 KB). The maximum recommended task size is 100 KB
18/12/24 08:18:10 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 08:18:47 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://high-memory-224613/pyspark_nlp/data/
preprocessing data...
making predictions on test data...
accuracy with saved model on test data is 0.811067289254
Job output is complete
```

## SENTIMENT ANALYSIS ON GOOGLE CLOUD DATAPROC USING PYSPARK

The screenshot displays the Google Cloud Platform Storage interface. The left sidebar shows the 'Storage' section with options like 'Browser', 'Transfer', 'Transfer Appliance', and 'Settings'. The main area is titled 'Bucket details' for the bucket 'high-empire-226413'. It includes tabs for 'Objects', 'Overview', 'Permissions', and 'Bucket Lock'. Below these are buttons for 'Upload files', 'Upload folder', 'Create folder', 'Manage holds', and 'Delete'. A search bar labeled 'Filter by prefix...' is present. The 'Objects' tab shows a list of items in the bucket 'pyspark\_nlp'.

<input type="checkbox"/>	Name	Size	Type	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
<input type="checkbox"/>	data/	—	Folder	—	—	Per object	—	—	—
<input type="checkbox"/>	model/	—	Folder	—	—	Per object	—	—	—
<input type="checkbox"/>	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
<input type="checkbox"/>	result/	—	Folder	—	—	Per object	—	—	—