

Aims

This exercise aims to get you to practice:

- Create a Cloud Storage bucket in Dataproc
- Create a cluster in Dataproc
- Run Spark jobs in Dataproc

Background

Google Cloud:

Google Cloud consists of a set of physical assets, such as computers and hard disk drives, and virtual resources, such as virtual machines (VMs), that are contained in Google's data centers around the globe. Each data center location is in a region. Regions are available in Asia, Australia, Europe, North America, and South America. Each region is a collection of zones, which are isolated from each other within the region. Each zone is identified by a name that combines a letter identifier with the name of the region.

In cloud computing, what you might be used to thinking of as software and hardware products, become services. These services provide access to the underlying resources. The list of available Google Cloud services is long, and it keeps growing. When you develop your website or application on Google Cloud, you mix and match these services into combinations that provide the infrastructure you need, and then add your code to enable the scenarios you want to build. See more documentation at:

<https://cloud.google.com/docs/overview>

Dataproc:

Dataproc is a fully managed and highly scalable service for running Apache Spark, Apache Flink, Presto, and 30+ open source tools and frameworks. Use Dataproc for data lake modernization, ETL, and secure data science, at planet scale, fully integrated with Google Cloud, at a fraction of the cost. See more documentation at:

<http://docs.aws.amazon.com/AmazonS3/latest/gsg/GetStartedWithS3.html>

Caution: Before doing the lab, please make sure that you have a google account in Dataproc with **\$300 free credits**!!! We are NOT responsible for any charge of your credit cards if you do not follow the lab instructions.

Register Google Cloud

If you have an existing google account, you can use the same email and password for Google Cloud. Otherwise, please follow the below instructions:

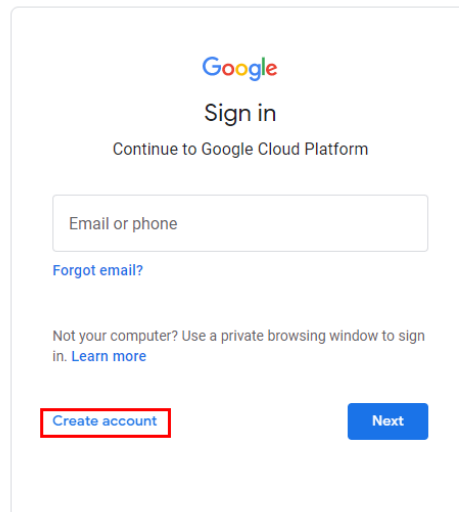
- Go to <https://cloud.google.com/free> and click “Get started for free”.

Solve real business challenges on Google Cloud

Get started for free

Contact sales

- Click “Create account”.

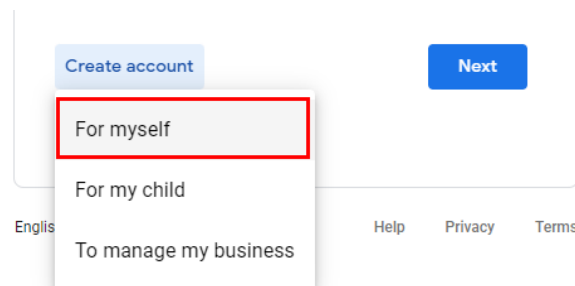


The image shows the Google sign-in page for the Google Cloud Platform. At the top is the Google logo, followed by the text "Sign in" and "Continue to Google Cloud Platform". Below this is a text input field labeled "Email or phone". A link "Forgot email?" is positioned below the input field. Further down, there is a note: "Not your computer? Use a private browsing window to sign in. [Learn more](#)". At the bottom of the sign-in area, there are two buttons: "Create account" (highlighted with a red box) and "Next".

English (United Kingdom) ▾

[Help](#) [Privacy](#) [Terms](#)

- Select “For myself”.



The image shows the Google account creation selection screen. It features a "Create account" button (highlighted with a blue box) and a "Next" button. A dropdown menu is open from the "Create account" button, showing three options: "For myself" (highlighted with a red box), "For my child", and "To manage my business". At the bottom of the screen, there is a language selector "English" and links for "Help", "Privacy", and "Terms".

- Enter your name and email, then verify your email address.
- Enter your personal information and, and you’ll need to agree to the Terms of Service to create a Google Account.

Depending on your account settings, some of this data may be associated with your Google Account and we treat this data as personal information. You can control how we collect and use this data now by clicking 'More Options' below. You can always adjust your controls later or withdraw your consent for the future by visiting My Account (myaccount.google.com).


[More options](#) ▾

[Cancel](#)

I agree

- Enter your account information.

Step 1 of 2 Account Information

 Xin Cao
@gmail.com [SWITCH ACCOUNT](#)

Country
Australia ▼

What best describes your organization or needs?
Please select
Personal project ▼

Terms of Service
☒ I have read and agree to the [Google Cloud Platform Terms of Service](#), [Supplemental Free Trial Terms of Service](#), and the terms of service of [any applicable services and APIs](#).
Required to continue

[CONTINUE](#)

[Privacy policy](#) | [FAQs](#)

Access to all Cloud Platform Products

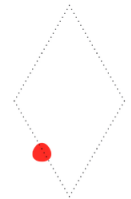
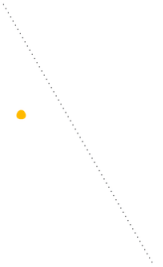
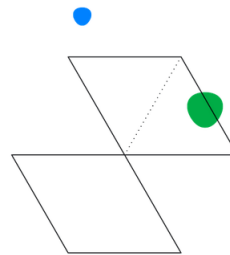
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

Put Google Cloud to work with \$300 in credit to spend over the next 90 days.

No autocharge after free trial ends



We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.







- Complete Identity Verification and Contact Information.
- Enter your payment information. (Google asks for your credit card or PayPal to make sure you are not a robot. **You won't be charged unless you manually upgrade to a paid account or the \$300 credits have been spent.**)

Step 2 of 2 Payment Information Verification

Your payment information helps us reduce fraud and abuse. **You won't be charged unless you turn on automatic billing.**

 Account type 
Individual
Only Business accounts can have multiple users. You cannot change the account type after signing up. In some countries, this selection affects your tax options. [Learn more](#)

Payment method
   MM / YY CVC
Card number is required
Cardholder name
Xin Cao

 Billing address

[START MY FREE TRIAL](#)

Access to all Cloud Platform Products

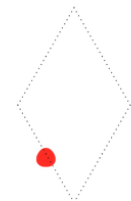
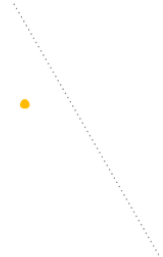
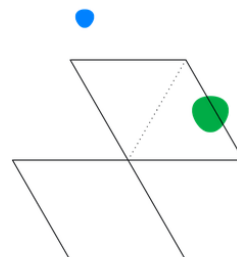
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

Put Google Cloud to work with \$300 in credit to spend over the next 90 days.

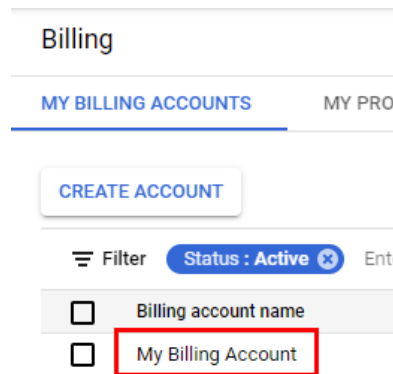
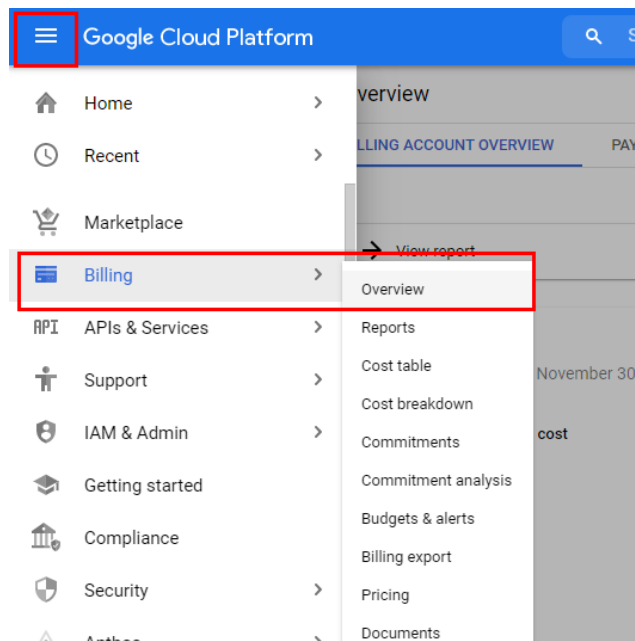
No autocharge after free trial ends

We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.



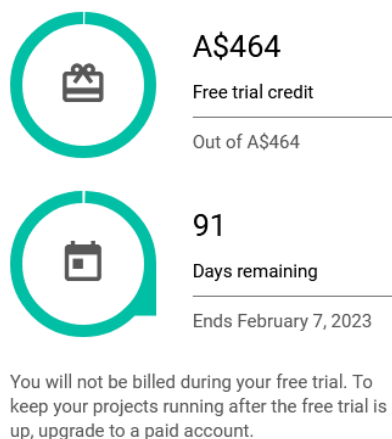
Check your free trial credit

- In the navigation menu of Google Cloud Platform, select “Billing -> overview”, or go to <https://console.cloud.google.com/billing/> and then select “My Billing Account”



- Make sure that you have the free trial credit.

Free trial credit

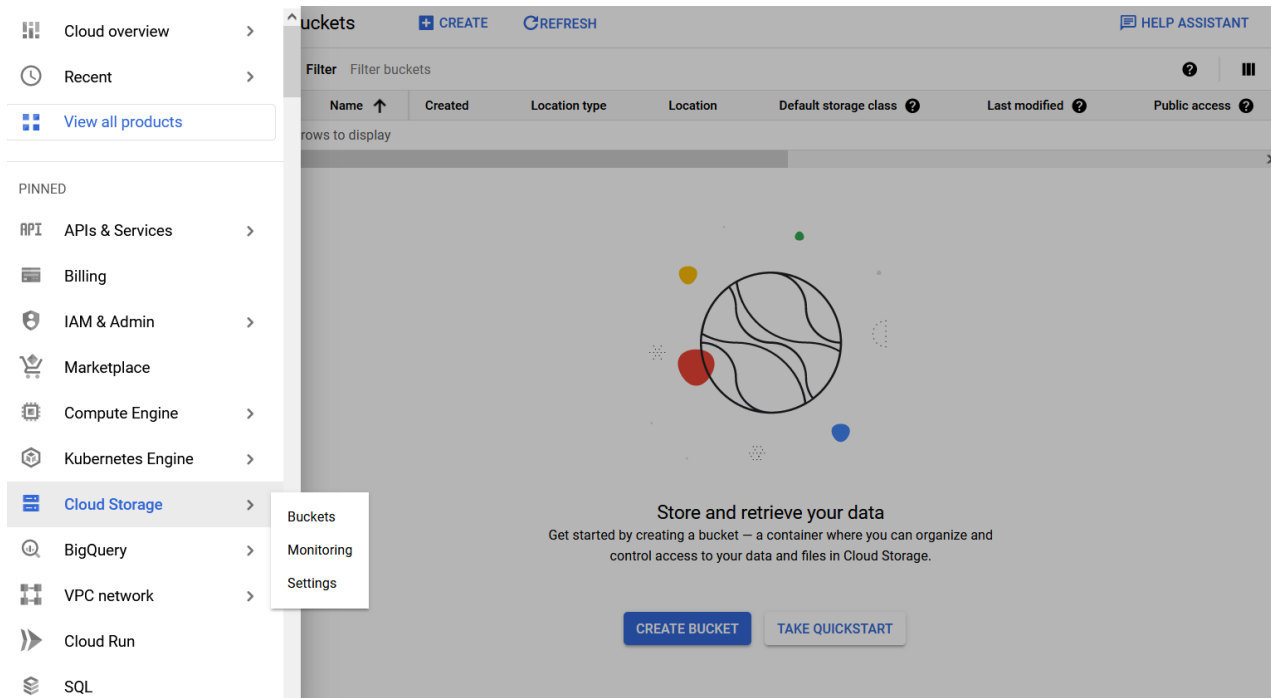


Create a Cloud Storage bucket

If you need to store some data in Google Cloud, you need to create a bucket for your data.

Navigate to Cloud Storage

- Open the menu on the left side of the console.
- In the **Storage** section, click **Cloud Storage**->**Buckets**.
- Begin by clicking **Create Bucket**.



Name your bucket

- Enter a name for your bucket. (E.g., you can use your own zID). *Note:* Bucket names must be **globally unique** (among all buckets ever created by any user).

• Name your bucket

Pick a **globally unique**, permanent name. [Naming guidelines](#)

Tip: Don't include any sensitive information

▼ LABELS (OPTIONAL)

CONTINUE

• Choose where to store your data

Location: us (multiple regions in United States)

Location type: Multi-region

• Choose a storage class for your data

Default storage class: Standard

• Choose how to control access to objects

Public access prevention: On

Access control: Uniform

• Choose how to protect object data

Protection tools: None

Data encryption: Google-managed key

CREATE

CANCEL

Good to know

Location pricing

Storage rates vary depending on the storage class of your data and location of your bucket. [Pricing details](#)

Current configuration: Multi-region / Standard

Item	Cost
us (multiple regions in United States)	\$0.026 per GB-month
With default replication	\$0.020 per GB written

ESTIMATE YOUR MONTHLY COST

Choose storage location

- **Choose where to store your data**

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. [Learn more](#)

Location type

- ☐ **Multi-region**
Highest availability across largest area
- ☐ **Dual-region**
High availability and low latency across 2 regions
- ☒ **Region**
Lowest latency within a single region

australia-southeast1 (Sydney) ▼

CONTINUE

- Select the Location Type for your data.
 - The default, **Multi-region**, delivers the highest availability.
 - For lower latency, you may wish to choose **Region**.
 - Choosing **Dual-region** strikes a balance between them.
- Select “Australia-southeast1(Sydney)” as the location of your storage.
- Click Continue (you can also skip the following and click “**Create**” directly).

(optional) Select Storage Class (use the default in this lab)

- Select a default storage class for data in this bucket. The default is **Standard**, but you may wish to choose a different option based on your needs.
 - This decision should be based on how long you plan to store your data and how often it will be accessed. [Learn more about storage classes](#).
- Click Continue.

(optional) Access Control (use the default in this lab)

- Specify how to control access to objects, whether you want to control access at the bucket level only (Uniform), or to also enable individual stored objects to have additional permission settings (Fine-grained). [Learn more about the differences here](#).
- Click Continue.

(optional) Choose how to protect object data (use the default in this lab)

- Your data is always protected with Cloud Storage but you can also choose from these additional data protection options to prevent data loss. Note that object versioning and retention policies cannot be used together.

After configuring your bucket setting, you can click the “**CREATE**” button.

✓

Name your bucket

Name: comp9313-zid

✓

Choose where to store your data

Location: asia (multiple regions in Asia)

Location type: Multi-region

✓

Choose a default storage class for your data

Default storage class: Standard

✓

Choose how to control access to objects

Public access prevention: Off

Access control: Uniform

•

Choose how to protect object data

Protection tools: None

Data encryption: Google-managed key

CREATE

CANCEL

Create a cluster

In the navigation menu of Google Cloud Platform, click Dataproc->Clusters, and then in the new page click CREATE CLUSTER. You can also access Dataproc by searching it at the head of the webpage. In the creating cluster panel, most fields are filled with default values already. You can change these default values to customize your own cluster.

ANALYTICS

Composer

Dataproc

Pub/Sub

Dataflow

Datastream

IoT Core

BigQuery

Dataplex

Looker

Data Catalog

Data Fusion

Dataprep

Financial Services

Jobs ON CLUSTERS

Clusters

Jobs

Workflows

Autoscaling policies

SERVERLESS

Batches

METASTORE SERVICES

Metastore

Federation

UTILITIES

Component exchange

Workbench

comp9313-22t3

Location	Storage class	Public access	Protection
asia (multiple regions in Asia)	Standard	Not public	None

CONFIGURATION

PERMISSIONS

PROTECTION

LOAD FOLDER

CREATE FOLDER

TRANSFER DATA

Filter

Filter objects and folders

Type	Created	Storage class	Last modified
------	---------	---------------	---------------

Click “ENABLE” to use the Dataproc API.



Cloud Dataproc API

[Google Enterprise API](#)

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

ENABLE

TRY THIS API [↗](#)

Click “CREATE CLUSTERS”.

Cluster

Cloud Dataproc

Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

There are no clusters in the currently selected Cloud Dataproc region(s). Create a cluster to get started.

CREATE CLUSTER

Select “Cluster on Compute Engine” or “Cluster on GKE”.

Create Dataproc cluster

Select the infrastructure service that you want to use.

Cluster on Compute Engine

Create the cluster on Compute Engine.

CREATE

Cluster on GKE

Create the cluster on Google Kubernetes Engine (GKE).

CREATE

CANCEL

Set up cluster

You need to at least give a name, select a location, like below:

- Set up cluster**
Begin by providing basic information.
- Configure nodes (optional)**
Change node compute and storage capabilities.
- Customize cluster (optional)**
Add cluster properties, features, and actions.
- Manage security (optional)**
Change access, encryption, and security settings.

CREATE **CANCEL**

EQUIVALENT COMMAND LINE ▾

Name

Cluster Name *
lab9 ?

Location

Region *
asia-southeast1 ▾ ?

Zone *
asia-southeast1-c ▾ ?

Cluster type

☒ **Standard (1 master, N workers)**

☐ **Single Node (1 master, 0 workers)**
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

☐ **High Availability (3 masters, N workers)**
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

The cluster name appears on the Clusters page, and its status is updated to Running after the cluster is provisioned. Click the cluster name to open the cluster details page where you can examine jobs, instances, and configuration settings for your cluster and connect to web interfaces running on your cluster.

You can optionally configure the nodes you are going to use for both master and worker nodes. For example, you can set the machine type as “n1-standard-2”, the disk sizes of master and worker nodes to 30GB as below:

For the panels of “Customize cluster” and “Manage security”, you just need to use the default values in this lab.

Error

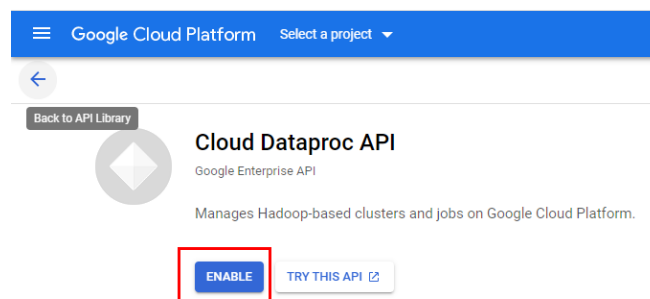
Cloud Dataproc API has not been used in project 973340955223 before or it is disabled. Enable it by visiting <https://console.developers.google.com/apis/api/dataproc.googleapis.com/c?project=973340955223> then retry. If you enabled this API recently, wait a few minutes for the action to propagate to our systems and retry.

Request ID: 1651644594890892054

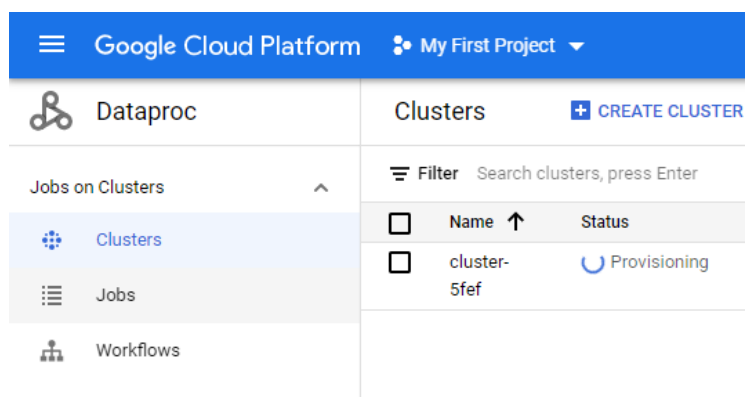
SEND FEEDBACK

CLOSE

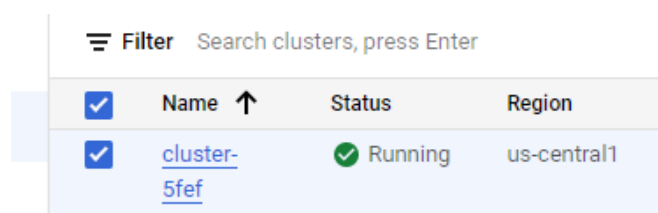
You should visit the link shown in the message, and enable the Cloud Dataproc API. Then, try to create the cluster again.



If it is successful, you can find a cluster in your Clusters panel.



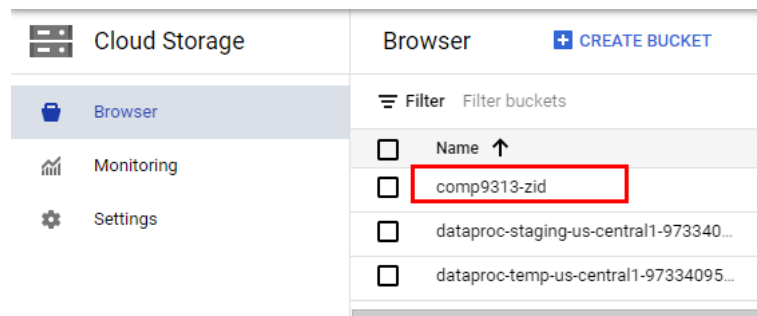
The status will change from “Provisioning” to “Running” when it is ready.



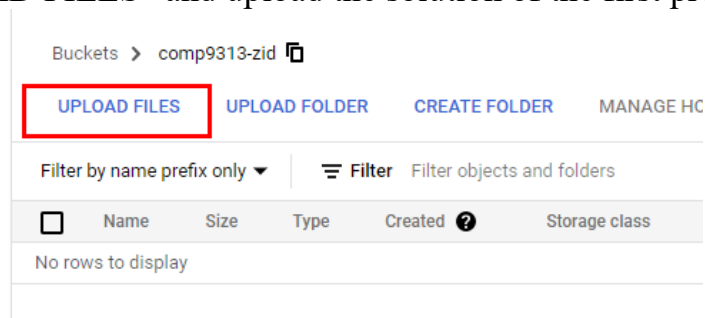
Run Spark Jobs in Google Dataproc

Upload Python file to Google Cloud Storage

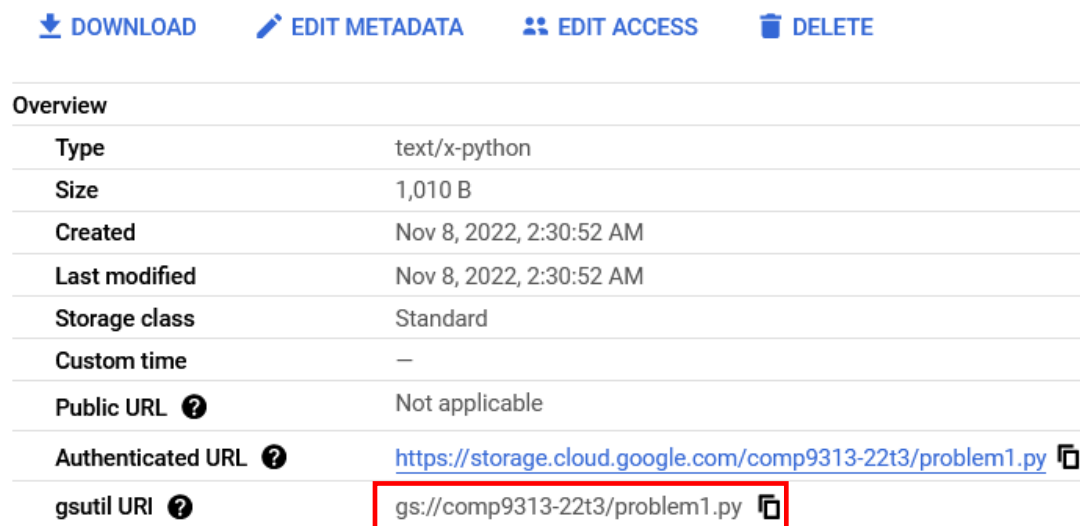
- Click the bucket you just created with name comp9313-<ZID>



- Select “UPLOAD FILES” and upload the solution of the first problem in Lab 6:








- Click the file, then in the new page find its gsutil URI.



Upload Input File to Google Cloud Storage


Download the testing input file pg100.txt, and upload it to your bucket as well. After the file is uploaded, check its gsutil URI, which will be used later.

Overview	
Type	text/plain
Size	5.3 MB
Created	Nov 8, 2022, 2:32:56 AM
Last modified	Nov 8, 2022, 2:32:56 AM
Storage class	Standard
Custom time	—
Public URL 	Not applicable
Authenticated URL 	https://storage.cloud.google.com/comp9313-22t3/pg100.txt 
gsutil URI 	gs://comp9313-22t3/pg100.txt 


Run Your Spark Job in Dataproc

- In the navigation menu of Google Cloud Platform, click Dataproc->Jobs. In the new page, click “SUBMIT JOB”.
- Configure your PySpark job in the new page. First, select the region as “Australia-southeast1”, the one you used when creating the cluster. Then, the created cluster would be visible to you:

Job ID *
 job-836317e5

Region *
 asia-southeast1 

Specifies the Cloud Dataproc regional service, which determines what clusters are available.

Cluster *
 lab9 

- Next, select the job type, configure the class, the jar file, and the arguments. Please make sure that there is no unexpected char (e.g. unexpected space) following your arguments. If you paste all these arguments into the webpage, you must be careful about this issue.

Cluster *
lab9

Job type *
PySpark

Main python file *
gs://comp9313-22t3/problem1.py

Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix

Additional python files

Jar files

Jar files are included in the CLASSPATH. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix.

Archive files

Archive files are extracted in the Spark working directory. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix. Supported file types: .jar, .tar, .tar.gz, .tgz, .zip.

Arguments

gs://comp9313-22t3/pg100.txt × gs://comp9313-22t3/output ×

Press <Return> to add more arguments

Additional arguments to pass to the main class. Press Return after each argument.

- *Job type:* PySpark
- *Main python file:* problem1.py in your bucket
- *Arguments:* gs://your-bucket-name/pg100.txt gs://your-bucket-name/output
- Click **Submit** to start the job. You will see the details of the job running.
- Once the job starts, it is added to the Jobs list. The elapsed time of the job is also displayed to you after the job completes successfully.

Filter Filter jobs

<input type="checkbox"/>	Job ID	Status	Region	Type	Cluster	Start time	Elapsed time	Labels
<input type="checkbox"/>	job-836317e5	✓ Succeeded	asia-southeast1	PySpark	lab9	Nov 8, 2022, 2:39:46 AM	44 sec	None

- Click the Job ID to open the **Jobs** page, where you can view the job's driver output
- You can see your output in your bucket now:

Buckets > comp9313-22t3 > output

[UPLOAD FILES](#)
[UPLOAD FOLDER](#)
[CREATE FOLDER](#)
[TRANSFER DATA](#)
[MANAGE HOLDS](#)
[DOWNLOAD](#)
[DELETE](#)

Filter by name prefix only Filter objects and folders ☐ Show deleted data

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access	Version histo	
<input type="checkbox"/>	_SUCCESS	0 B	application/octet-stream	Nov 8, 20...	Standard	Nov 8, 202...	Not public	—	
<input type="checkbox"/>	part-00000	180 B	application/octet-stream	Nov 8, 20...	Standard	Nov 8, 202...	Not public	—	
<input type="checkbox"/>	part-00001	148 B	application/octet-stream	Nov 8, 20...	Standard	Nov 8, 202...	Not public	—	

Caution: Do not forget to stop the cluster after you finish all labs (Click “STOP”) and delete all the data in your bucket!!!

Google Cloud Platform My First Project

Dataproc Clusters [CREATE CLUSTER](#) [REFRESH](#) [START](#) [STOP](#) [DELETE](#) [REGIO](#)

Filter Search clusters, press Enter

<input checked="" type="checkbox"/>	Name	Status	Region	Zone	Total worker nodes	Scheduled deletion
<input checked="" type="checkbox"/>	cluster-5fef	Running	us-central1	us-central1-a	2	Off

Jobs on Clusters

- Clusters
- Jobs
- Workflows

You can try submitting your solutions to problems in Labs 6 and 7 to Dataproc and check the running time.

Before submitting a Spark job to Dataproc, you always need to start a cluster first, and remember to stop the cluster when your job completes.