

# 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:

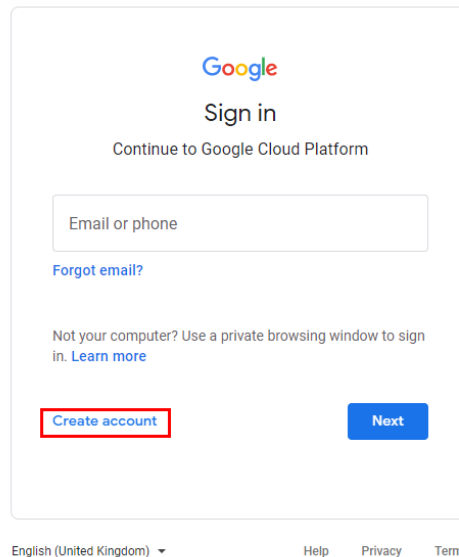
- Go to <https://cloud.google.com/free> and click “Get started for free”.

# Solve real business challenges on Google Cloud

Get started for free

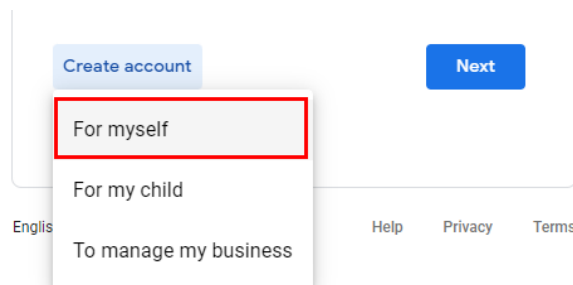
Contact sales

- 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". At the very bottom of the page, there is a language selector set to "English (United Kingdom)" and links for "Help", "Privacy", and "Terms".

- Select “For myself”.



The image shows the account creation selection screen. It features a "Create account" button (highlighted with a red 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 set to "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](https://myaccount.google.com)).

[More options](#) ▾

[Cancel](#)

I agree

- Enter your account information.

#### Step 1 of 3 Account Information



[SWITCH ACCOUNT](#)

Country

Australia

What best describes your organization or needs?

Please select  
Personal project

Terms of Service

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

Required to continue

Email updates

☒ I would like to receive periodic emails on news, product updates and special offers from Google Cloud and Google Cloud Partners.

[CONTINUE](#)

#### Access to all Cloud Platform Products

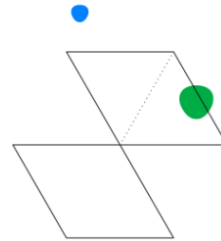
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

#### \$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.)

#### Payment method

☒ Add credit or debit card

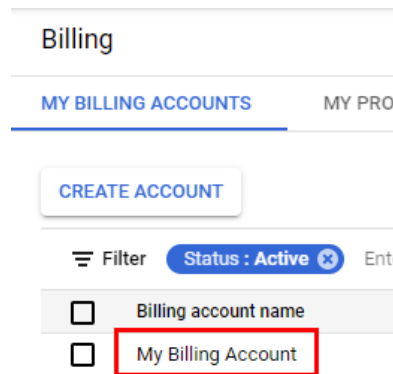
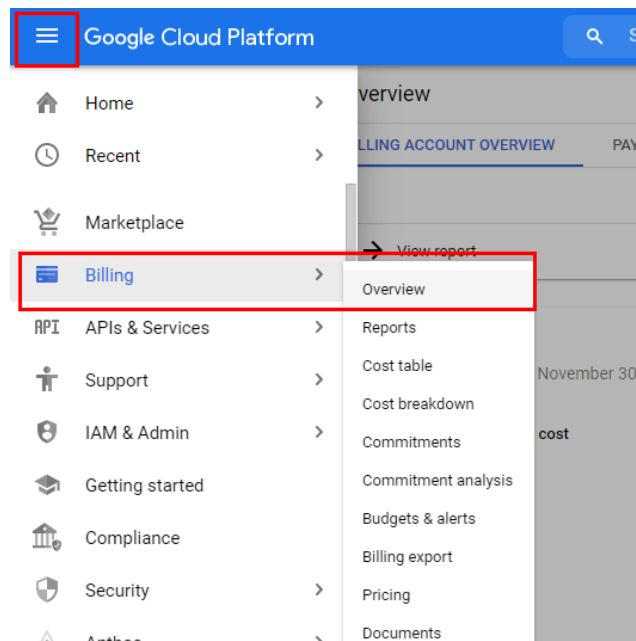
☐ Add PayPal

You'll be charged automatically on the 1st of each month. If your balance reaches your payment threshold before then, you'll be charged immediately. [Learn more](#)

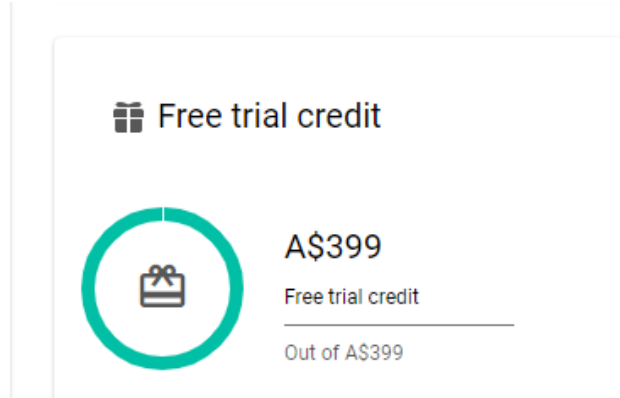
[START MY FREE TRIAL](#)

## 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.



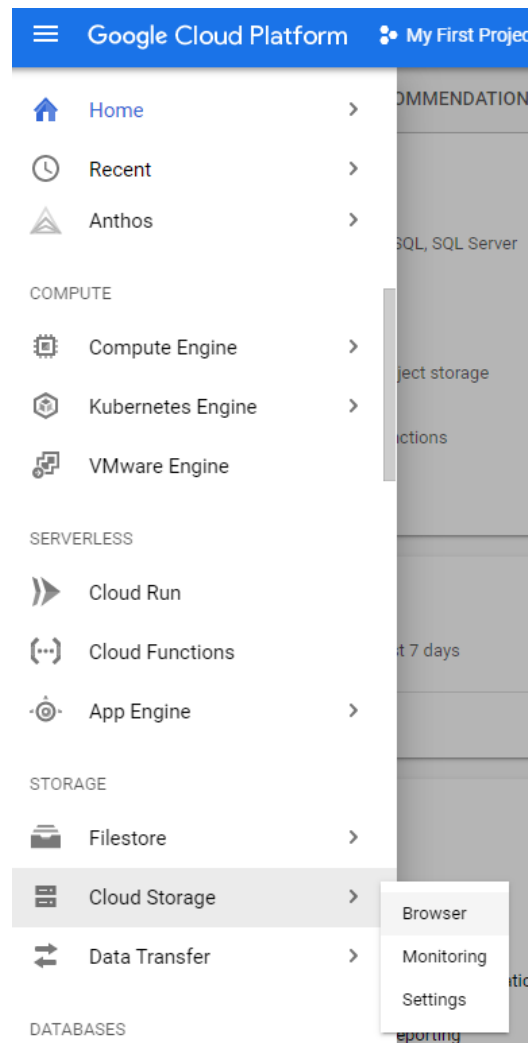
## Create a Cloud Storage bucket

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

### Navigate to Cloud Storage

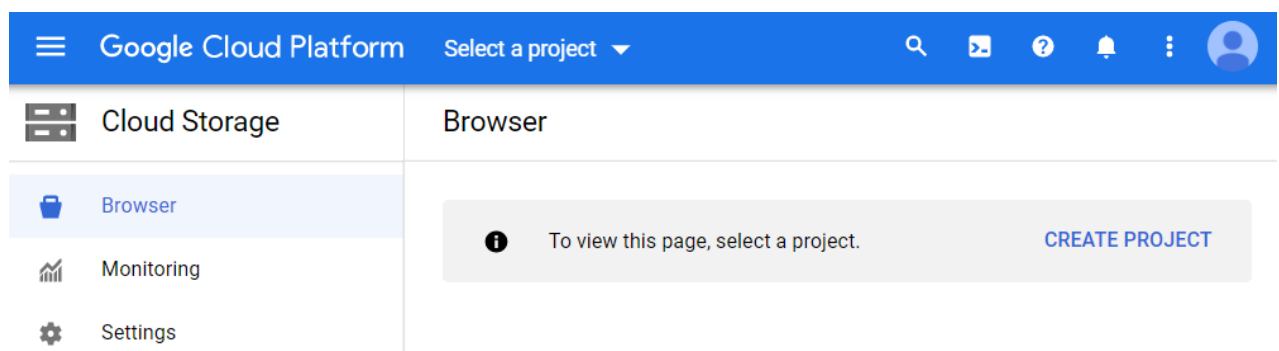
- Open the menu on the left side of the console.

- In the **Storage** section, click **Cloud Storage->Browser**.

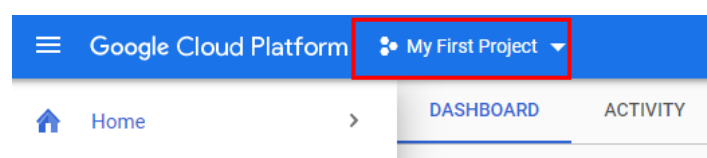


## Set up a Project

- In order to create a bucket, you need to first create a project if it does not exist.

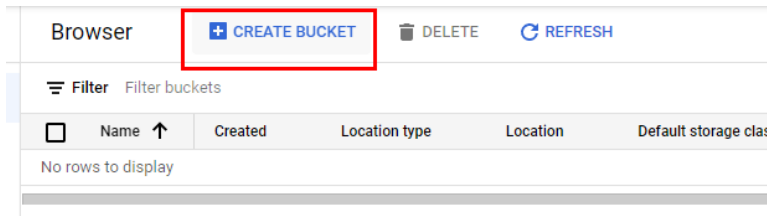


- For example, you can name your project as “My First Project”.



## Name your bucket

- Begin by clicking **Create Bucket**.
- Enter a name for your bucket. (You can use “comp9313-zID” by replacing “zID” with your own zID). *Note:* Bucket names must be **globally unique** (among all buckets ever created by any user).
- Click Continue and you will finish creating the bucket.



## Choose storage location

- Select the Location Type for your data.
  - The default, **Multi-region**, delivers the highest availability.
  - For lower latency, you may wish to choose **Regional**.
  - Choosing **Dual-region** strikes a balance between them.
- Select “asian1” 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

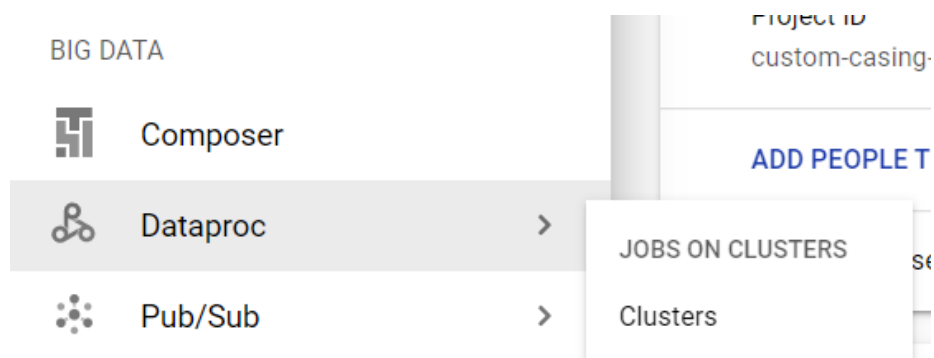
Data encryption: Google-managed key

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. In the creating cluster panel, most fields are filled with default values already. You can change these default values to customize your own cluster.



## Set up cluster

You need to at least give a name, select a location, and select a cluster type for your cluster, like below:

← Create a cluster

• 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.

Name

Cluster Name \* lab8

Location

Region \* australia-southeast1

Zone \* australia-southeast1-c

Cluster type

☒ Standard (1 master, N workers)

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.

### (Optional) Configure nodes

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.:

Machine types for common workloads, optimized for cost and flexibility

Series  
N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type  
n1-standard-2 (2 vCPU, 7.5 GB memory)

	vCPU	Memory
	2	7.5 GB

✓ CPU PLATFORM AND GPU

Primary disk size \*  
30 GB ?

Primary disk type  
Standard Persistent Disk ?

Number of local SSDs \*  
0 x 375GB ?

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

After clicking the “CREATE” button, if you get an error message like this:

**Error**

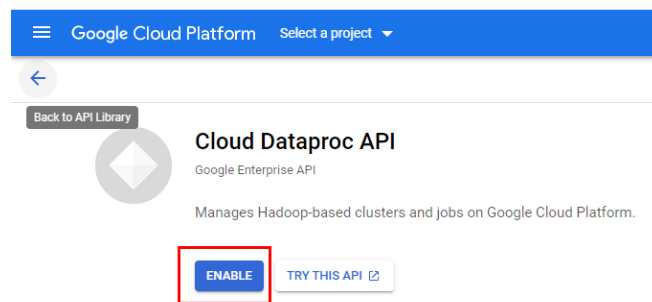
Cloud Dataproc API has not been used in project 973340955223 before or it is disabled. Enable it by visiting [https://console.developers.google.com/apis/api/dataproc.googleapis.com/authorize?project=\[redacted\]](https://console.developers.google.com/apis/api/dataproc.googleapis.com/authorize?project=[redacted]) 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

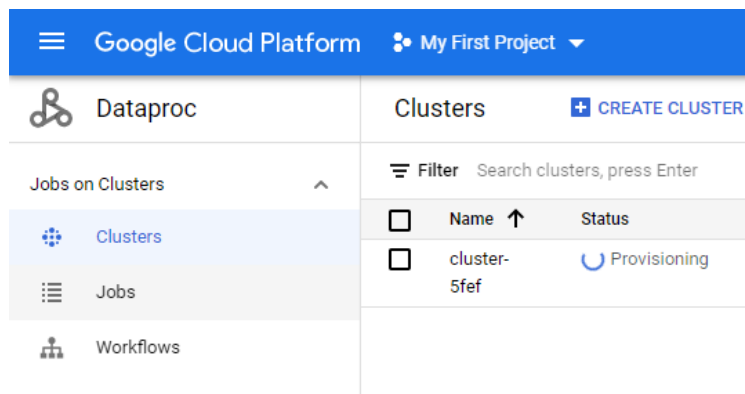
SEND FEEDBACK CLOSE

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

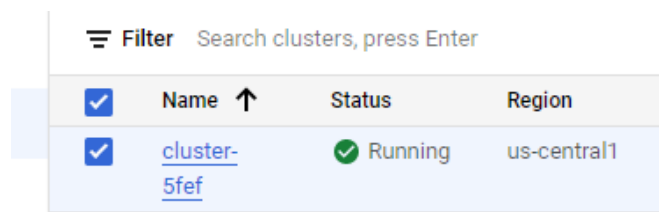




If it is successful, you can find a cluster in your Clusters panel.



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

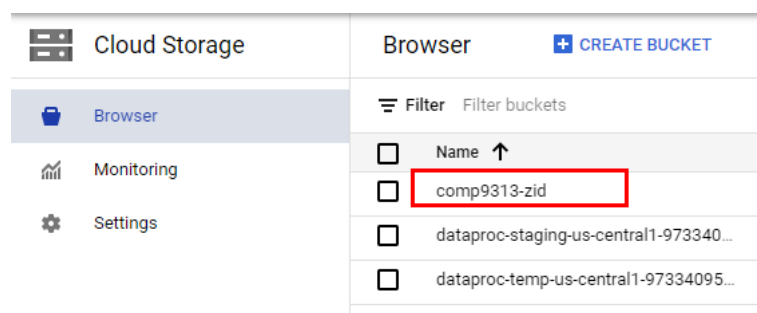


## Run Spark Jobs in Google Dataproc

### Upload jar file to Google Cloud Storage

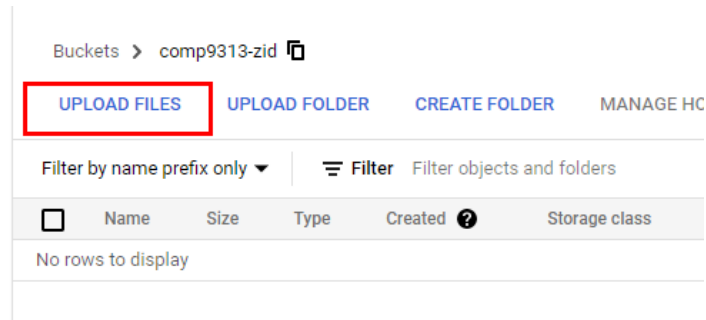
In Lab 6, you have learned how to create a jar file with SBT for your Spark project. Now you should first upload the jar file to Google Cloud Storage.

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




- Select “UPLOAD FILES” and upload the word-count\_2.12-1.0.jar file on

<https://webcms3.cse.unsw.edu.au/COMP9313/21T3/resources/69124>



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

 word-count\_2.12-1.0.jar

Overview

Type	application/octet-stream
Size	5.4 KB
Created	Nov 9, 2021, 12:08:31 AM
Last modified	Nov 9, 2021, 12:08:31 AM
Storage class	Standard
Custom time	—
Public URL ?	Not applicable
Authenticated URL ?	<a href="https://storage.cloud.google.com/comp9313-z3515164/word-count_2.12-1.0.jar?authuser=1">https://storage.cloud.google.com/comp9313-z3515164/word-count_2.12-1.0.jar?authuser=1</a>
gsutil URI ?	<a href="gs://comp9313-z3515164/word-count_2.12-1.0.jar">gs://comp9313-z3515164/word-count_2.12-1.0.jar</a>

## Upload Input File to Google Cloud Storage

Download the testing input file from:

<https://webcms3.cse.unsw.edu.au/COMP9313/21T3/resources/69126>, 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	40 B
Created	Nov 9, 2021, 12:16:36 AM
Last modified	Nov 9, 2021, 12:16:36 AM
Storage class	Standard
Custom time	—
Public URL ?	Not applicable
Authenticated URL ?	<a href="https://storage.cloud.google.com/comp9313-z3515164/input.txt?authuser=1">https://storage.cloud.google.com/comp9313-z3515164/input.txt?authuser=1</a>
gsutil URI ?	<a href="gs://comp9313-z3515164/input.txt">gs://comp9313-z3515164/input.txt</a>

## 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 Spark 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 \*

Region \*

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

Cluster \*

- Next, select the job type, configure the class, the jar file, and the arguments.

Job type \*

Main class or jar \*

The fully qualified name of a class in a provided or standard jar file, for example, com.example.wordcount, or a provided jar file to use the main class of that jar file

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

Press <Return> to add more arguments

- Job type: Spark
- Main class: comp9313.lab8.WordCount
- Jar file: Specify the Cloud Storage URI path to your WordCount jar (gs://**your-bucket-name**/word-count\_2.12-1.0.jar).
- Archive files: gs://**your-bucket-name**/input.txt
- Arguments: gs://**your-bucket-name**/input.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
<input type="checkbox"/>	job-e466ebb9	✓ Succeeded	australia-southeast1	Spark	lab8	Nov 9, 2021, 2:24:58 AM	42 sec

- 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-z3515164 > output

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

Filter by name prefix only Filter objects and folders

<input type="checkbox"/>	Name	Size	Type	Created	Storage class
<input type="checkbox"/>	_SUCCESS	0 B	application/octet-stream	Nov 9, 20...	Standard
<input type="checkbox"/>	part-00000	20 B	application/octet-stream	Nov 9, 20...	Standard
<input type="checkbox"/>	part-00001	21 B	application/octet-stream	Nov 9, 20...	Standard

**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 Search products and resources

Dataproc Clusters CREATE CLUSTER REFRESH START STOP DELETE REGIO

Jobs on Clusters Clusters Jobs Workflows

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

**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.**