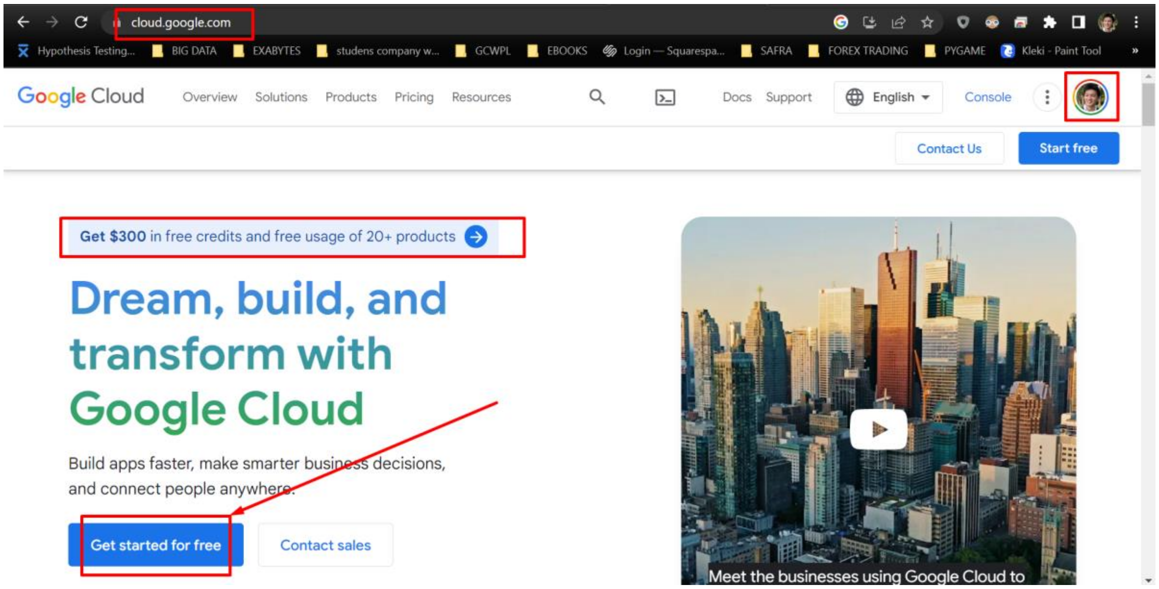
SETTING UP APACHE SPARK CLUSTER ON GOOGLE CLOUD

DR. ALVIN ANG

CONTENTS

I. SIGN UP FOR GOOGLE CLOUD


cloud.google.com

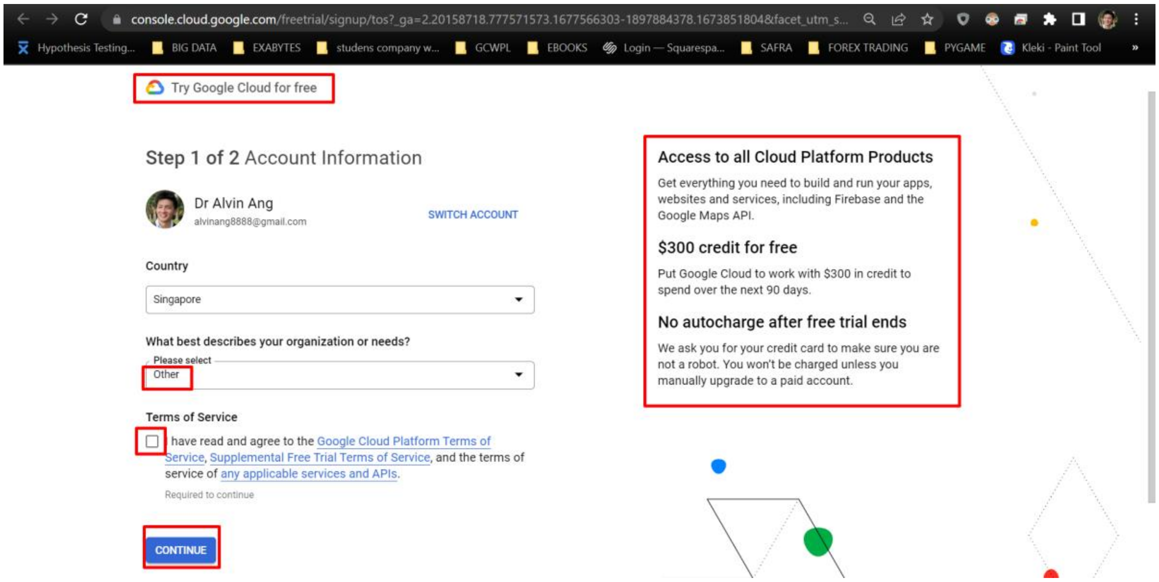
Get $300 in free credits and free usage of 20+ products

Dream, build, and transform with Google Cloud

Build apps faster, make smarter business decisions, and connect people anywhere.

Meet the businesses using Google Cloud to

https://cloud.google.com/


console.cloud.google.com/freetrial/signup/tos?_ga=2.20158718.777571573.1677566303-1897884378.1673851804&facet_utm_s...

Try Google Cloud for free

Step 1 of 2 Account Information

Dr Alvin Ang
alvinang8888@gmail.com

Country
Singapore

What best describes your organization or needs?
Please select
Other

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

Access to all Cloud Platform Products

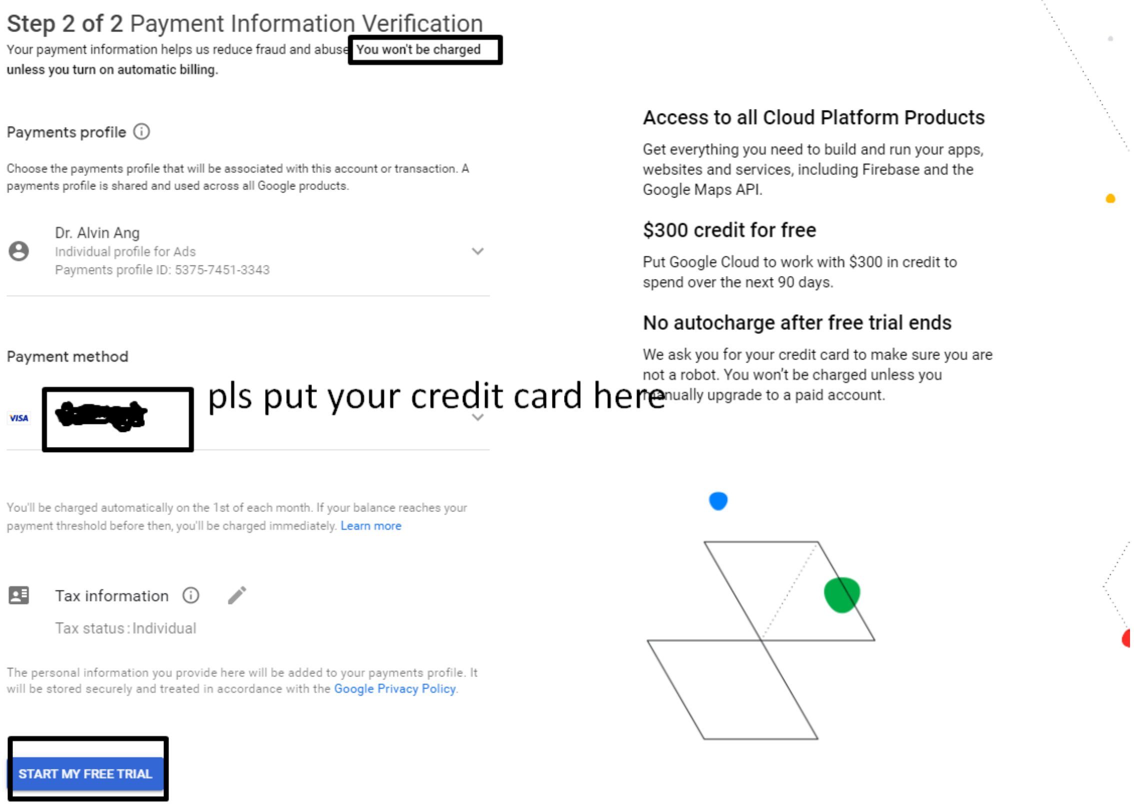
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

$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.



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.

Payments profile ⓘ

Choose the payments profile that will be associated with this account or transaction. A payments profile is shared and used across all Google products.

Dr. Alvin Ang
Individual profile for Ads
Payments profile ID: 5375-7451-3343

Payment method

VISA [redacted]

pls put your credit card here

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

Tax information ⓘ

Tax status : Individual

The personal information you provide here will be added to your payments profile. It will be stored securely and treated in accordance with the Google Privacy Policy.

START MY FREE TRIAL

Access to all Cloud Platform Products

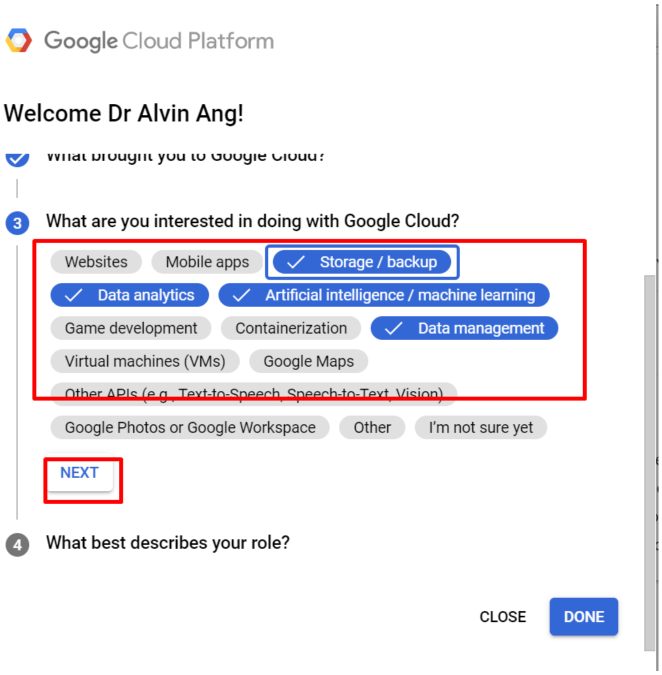
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

$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.



Google Cloud Platform

Welcome Dr Alvin Ang!

✓ What brought you to Google Cloud?

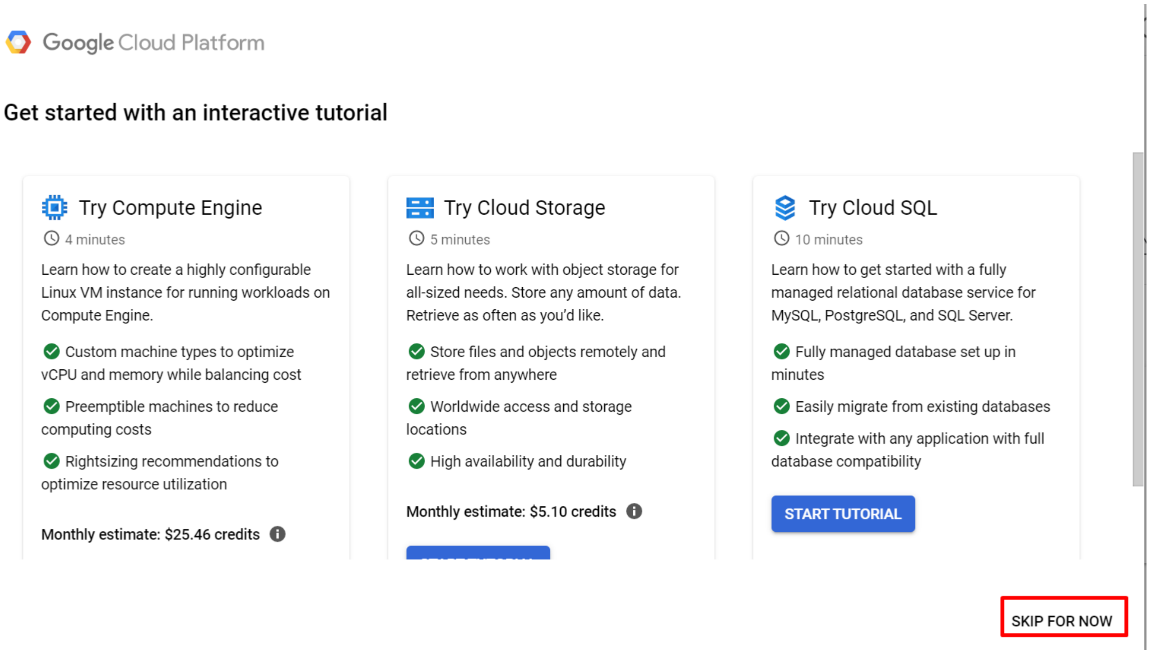
3 What are you interested in doing with Google Cloud?

Websites | Mobile apps | ✓ Storage / backup
✓ Data analytics | ✓ Artificial intelligence / machine learning
Game development | Containerization | ✓ Data management
Virtual machines (VMs) | Google Maps
Other APIs (e.g., Text-to-Speech, Speech-to-Text, Vision)
Google Photos or Google Workspace | Other | I'm not sure yet

NEXT

4 What best describes your role?

CLOSE DONE



Google Cloud Platform

Get started with an interactive tutorial

Try Compute Engine

4 minutes

Learn how to create a highly configurable Linux VM instance for running workloads on Compute Engine.

- ✓ Custom machine types to optimize vCPU and memory while balancing cost
- ✓ Preemptible machines to reduce computing costs
- ✓ Rightsizing recommendations to optimize resource utilization

Monthly estimate: $25.46 credits

Try Cloud Storage

5 minutes

Learn how to work with object storage for all-sized needs. Store any amount of data. Retrieve as often as you'd like.

- ✓ Store files and objects remotely and retrieve from anywhere
- ✓ Worldwide access and storage locations
- ✓ High availability and durability

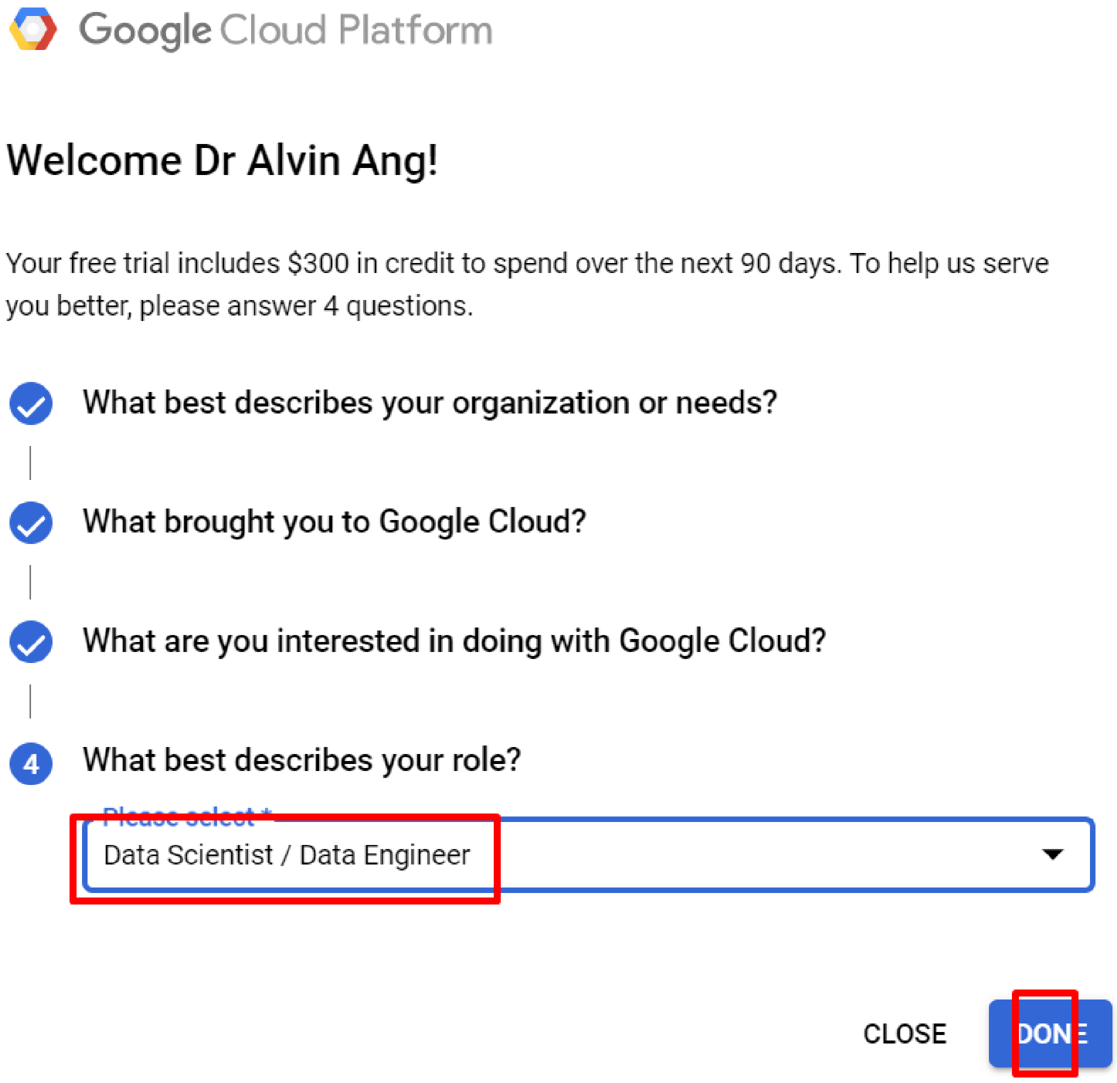
Monthly estimate: $5.10 credits

Try Cloud SQL

10 minutes

Learn how to get started with a fully managed relational database service for MySQL, PostgreSQL, and SQL Server.

- ✓ Fully managed database set up in minutes
- ✓ Easily migrate from existing databases
- ✓ Integrate with any application with full database compatibility



Google Cloud Platform

Welcome Dr Alvin Ang!

Your free trial includes $300 in credit to spend over the next 90 days. To help us serve you better, please answer 4 questions.

What best describes your organization or needs?

What brought you to Google Cloud?

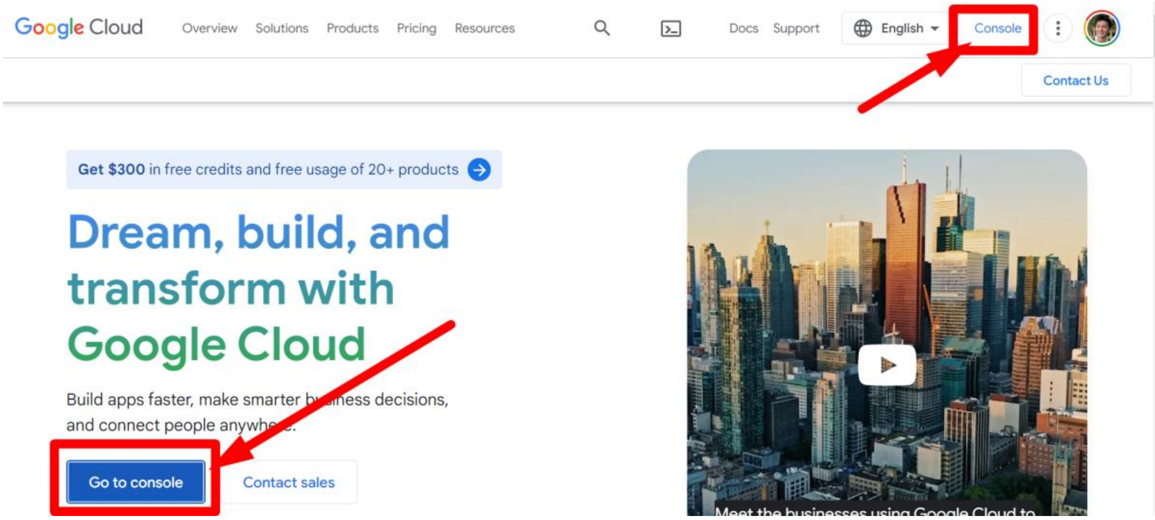
What are you interested in doing with Google Cloud?

4 What best describes your role?

Please select *

Data Scientist / Data Engineer

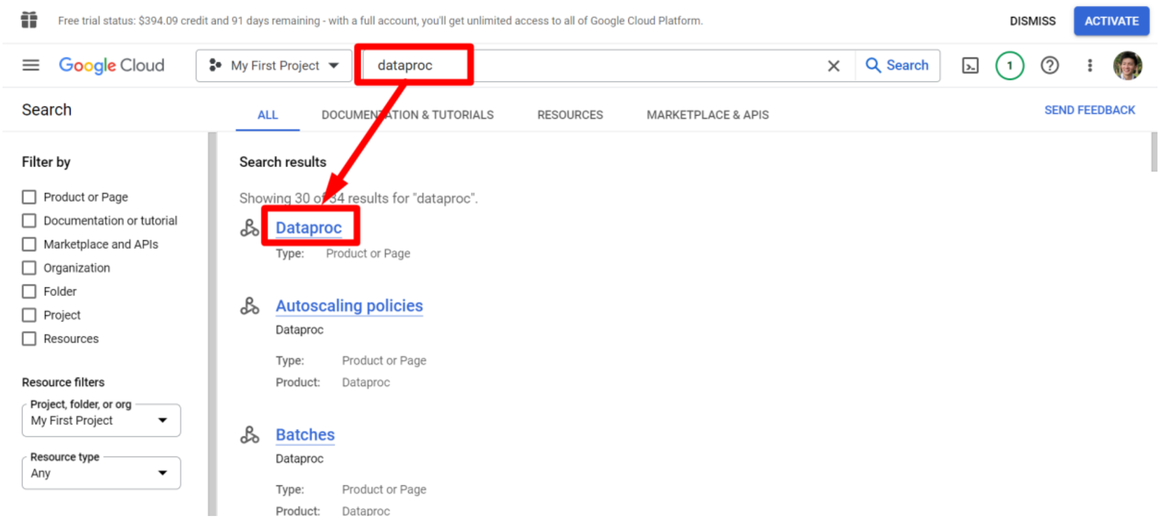
II. START DATAPROC



Get $300 in free credits and free usage of 20+ products

Dream, build, and transform with Google Cloud

Build apps faster, make smarter business decisions, and connect people anywhere.



Free trial status: $394.09 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

□ Product or Page
□ Documentation or tutorial
□ Marketplace and APIs
□ Organization
□ Folder
□ Project
□ Resources

Project, folder, or org: My First Project

Resource type: Any

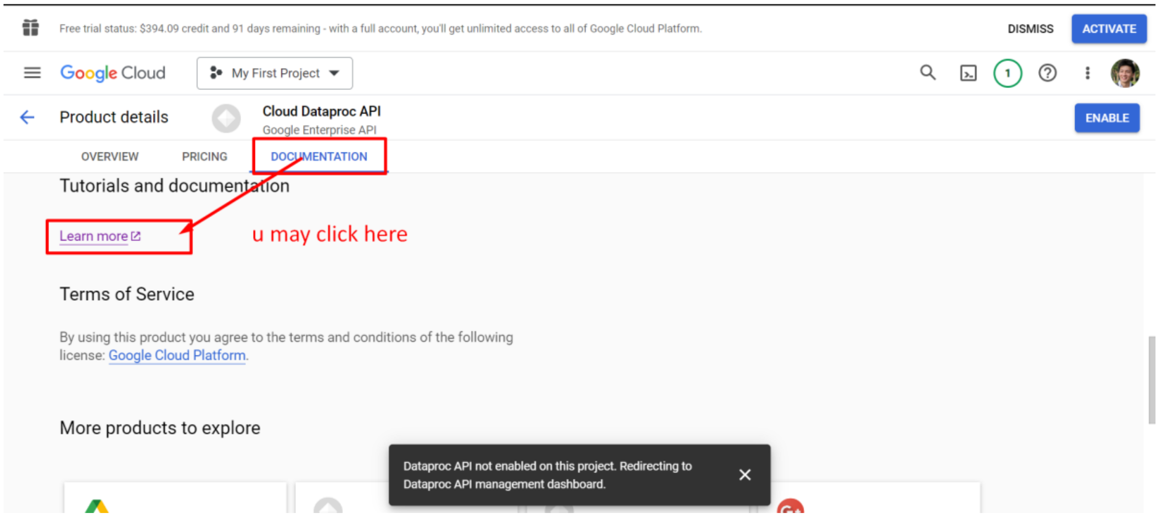
Search results

Showing 30 of 34 results for "dataproc".

Dataproc
Type: Product or Page

Autoscaling policies
Dataproc
Type: Product or Page
Product: Dataproc

Batches
Dataproc
Type: Product or Page
Product: Dataproc


Free trial status: $394.09 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
DISMISS
ACTIVATE
Google Cloud
My First Project
Product details
Cloud Dataproc API
Google Enterprise API
ENABLE
OVERVIEW
PRICING
DOCUMENTATION

Tutorials and documentation

Learn more

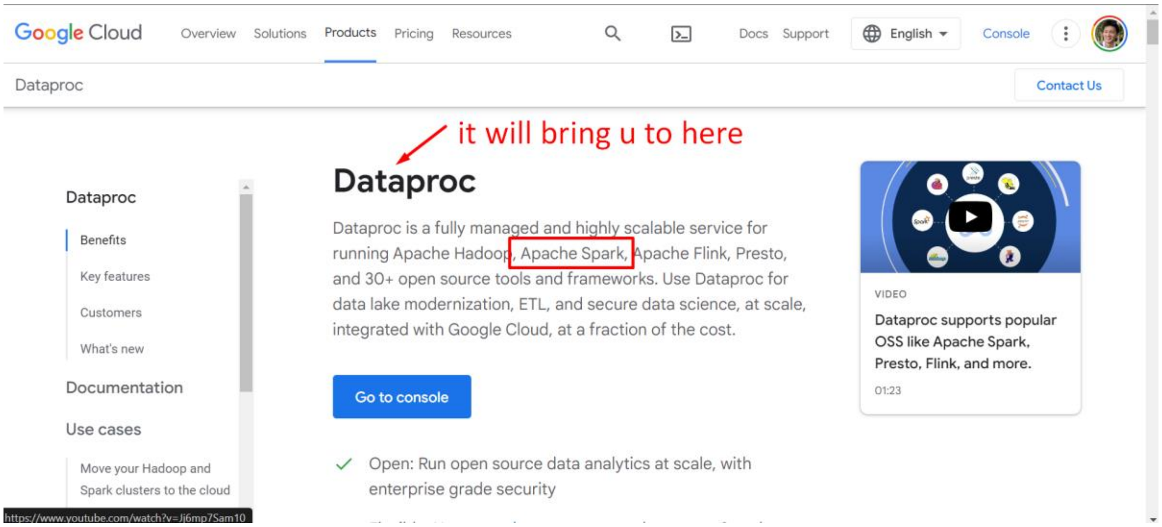
u may click here

Terms of Service

By using this product you agree to the terms and conditions of the following license: Google Cloud Platform.

More products to explore

Dataproc API not enabled on this project. Redirecting to Dataproc API management dashboard.



it will bring u to here

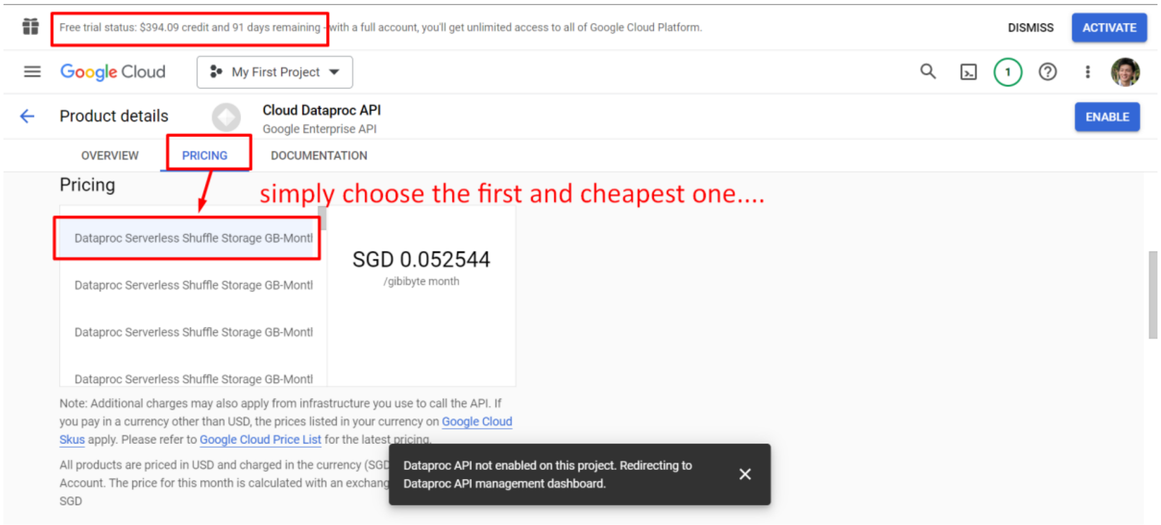
Dataproc

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

Dataproc supports popular OSS like Apache Spark, Presto, Flink, and more.

✓ Open: Run open source data analytics at scale, with enterprise grade security

IV. DATAPROC PRICING



Free trial status: $394.09 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Pricing

simply choose the first and cheapest one....

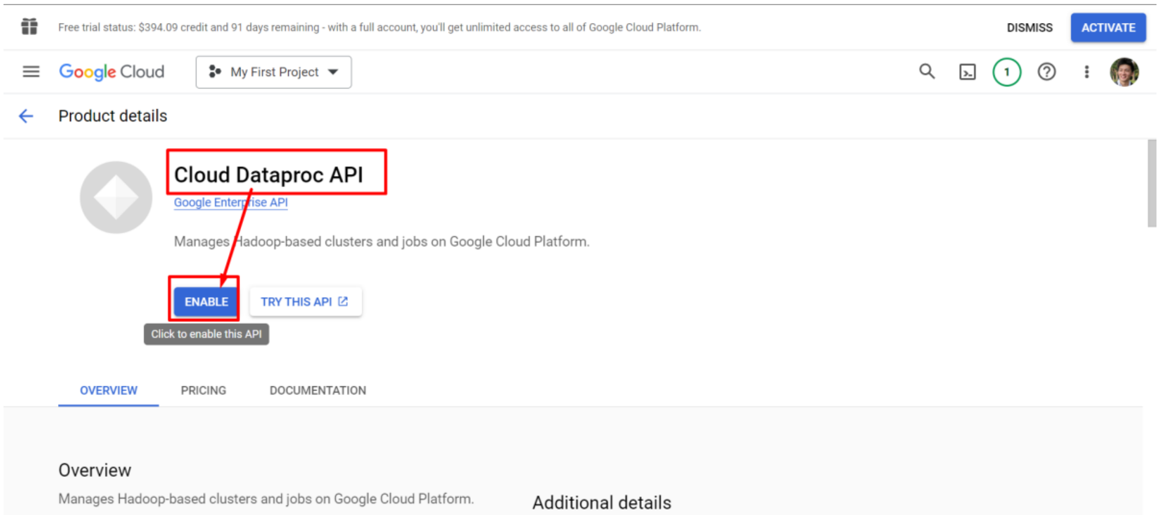
Dataproc Serverless Shuffle Storage GB-Montl	SGD 0.052544 /gibibyte month
Dataproc Serverless Shuffle Storage GB-Montl	
Dataproc Serverless Shuffle Storage GB-Montl	
Dataproc Serverless Shuffle Storage GB-Montl	

Note: Additional charges may also apply from infrastructure you use to call the API. If you pay in a currency other than USD, the prices listed in your currency on Google Cloud Skus apply. Please refer to Google Cloud Price List for the latest pricing.

All products are priced in USD and charged in the currency (SGD Account. The price for this month is calculated with an exchang SGD

Dataproc API not enabled on this project. Redirecting to Dataproc API management dashboard.

V. ENABLE DATAPROC


Free trial status: $394.09 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Product details

Cloud Dataproc API

Google Enterprise API

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

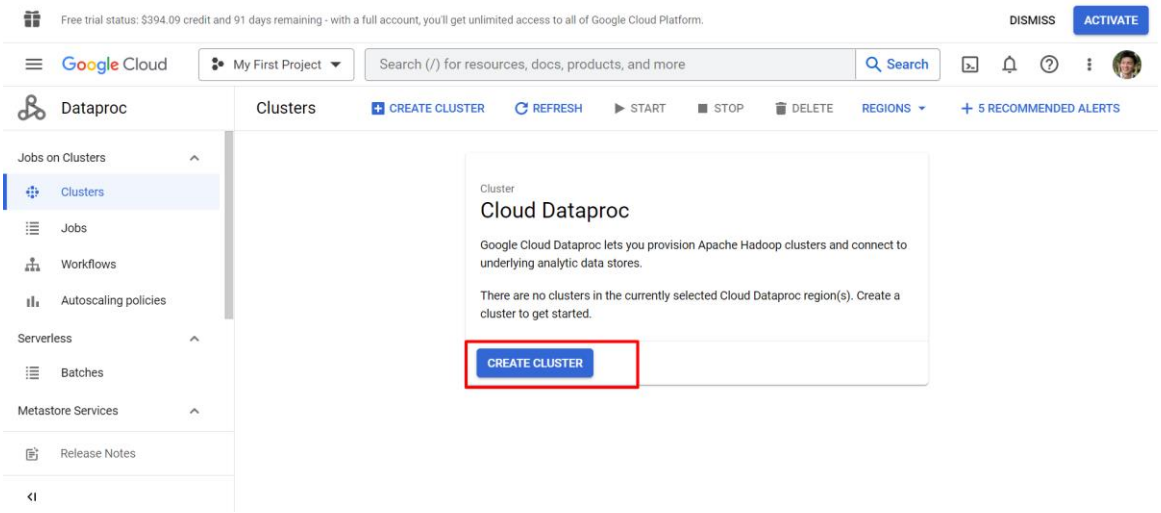
Click to enable this API

Overview

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

Additional details

VI. CREATE AND CONFIGURE CLUSTER



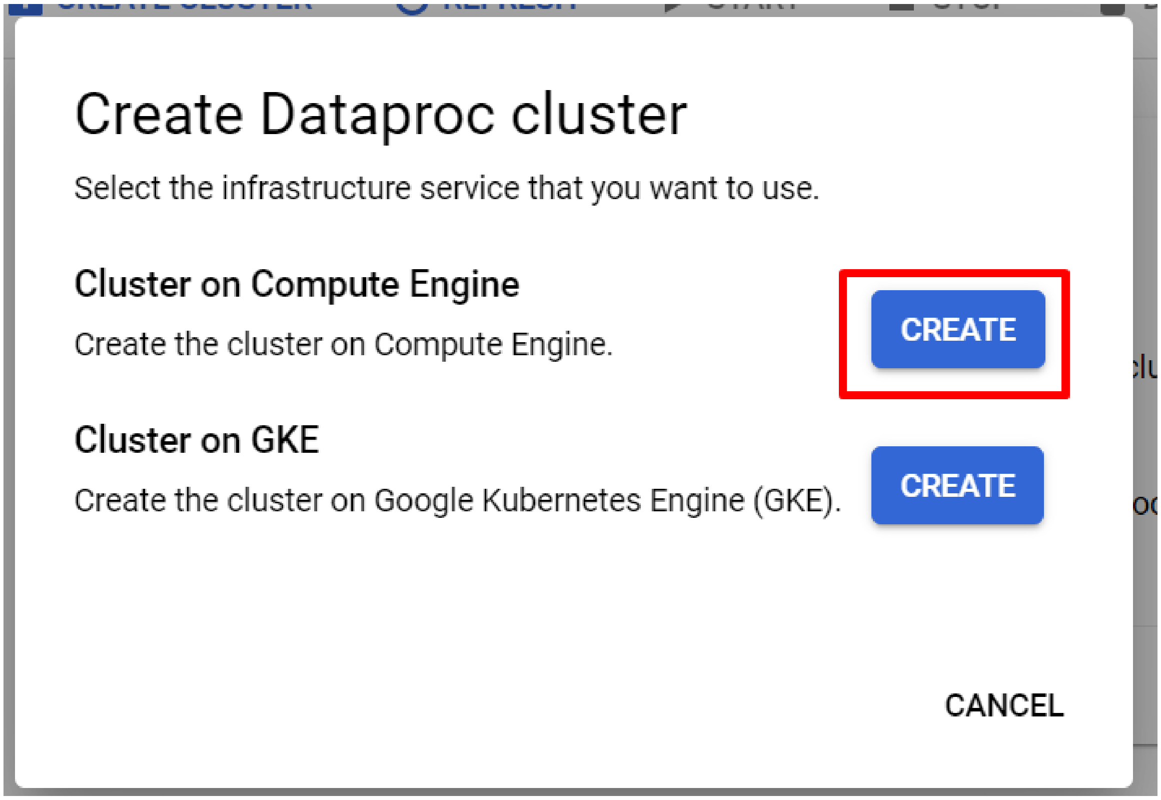
Free trial status: $394.09 credit and 91 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

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 Dataproc cluster

Select the infrastructure service that you want to use.

Cluster on Compute Engine

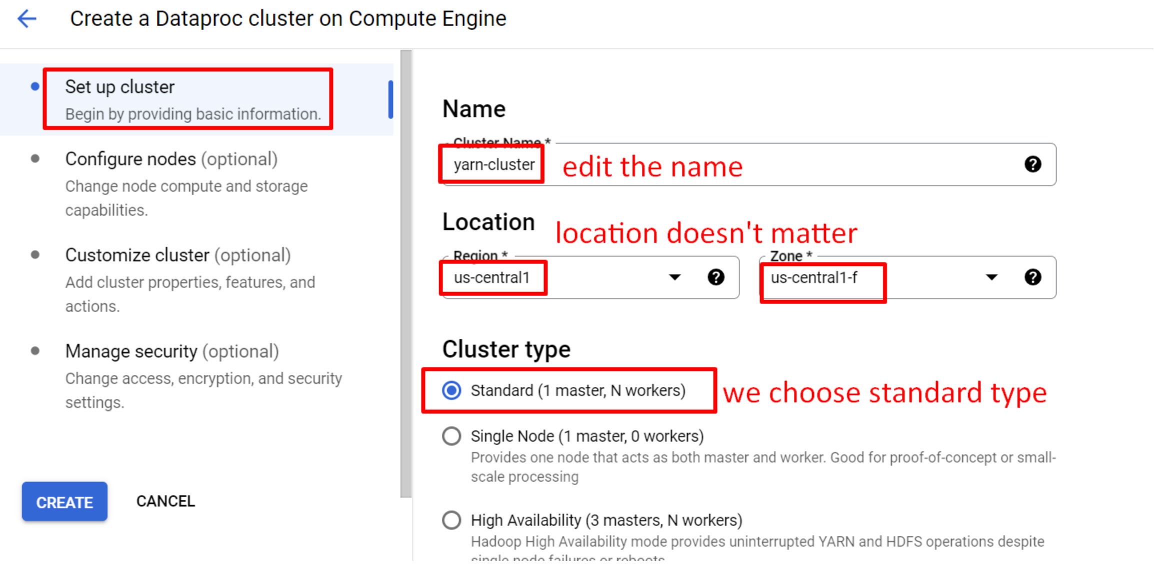
Create the cluster on Compute Engine.

Cluster on GKE

Create the cluster on Google Kubernetes Engine (GKE).

A. SETUP CLUSTER

1. SETUP NAME / LOCATION / CLUSTER TYPE



Create a Dataproc cluster on Compute Engine

- 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

Name

Cluster Name *: yarn-cluster

edit the name

Location

location doesn't matter

Region *: us-central1

Zone *: us-central1-f

Cluster type

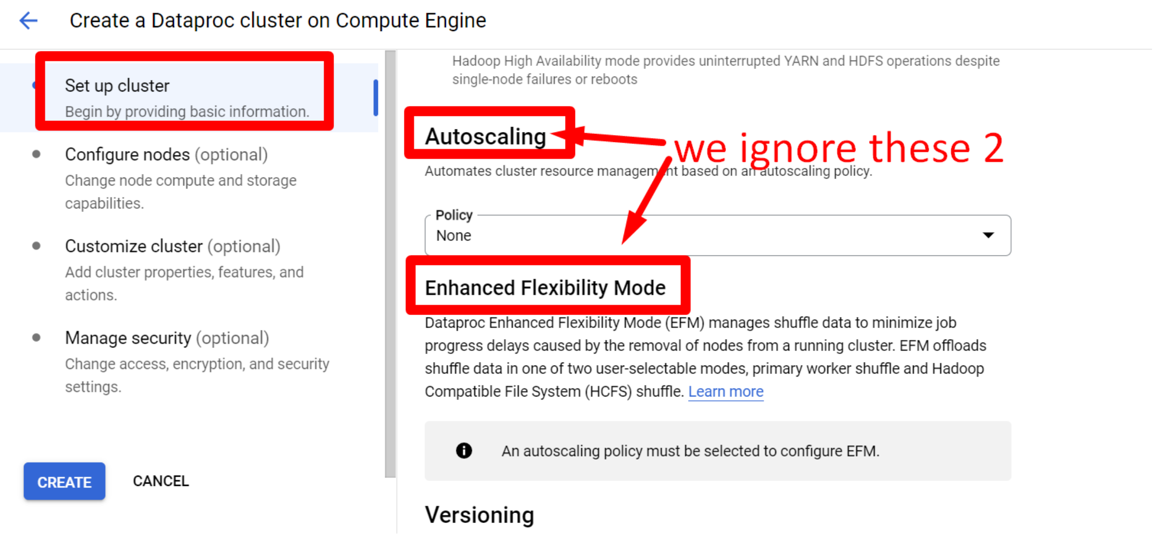
● Standard (1 master, N workers)

we choose standard type

○ 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

2. IGNORE AUTOSCALING AND FLEXIBILITY MODE



Create a Dataproc cluster on Compute Engine

- 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

Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Autoscaling

we ignore these 2

Automates cluster resource management based on an autoscaling policy.

Policy: None

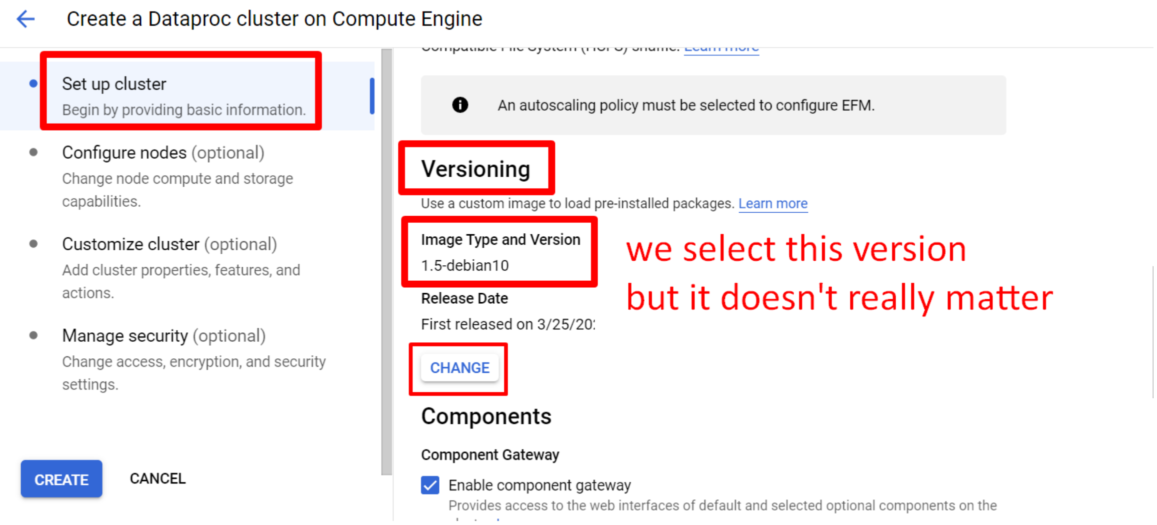
Enhanced Flexibility Mode

Dataproc Enhanced Flexibility Mode (EFM) manages shuffle data to minimize job progress delays caused by the removal of nodes from a running cluster. EFM offloads shuffle data in one of two user-selectable modes, primary worker shuffle and Hadoop Compatible File System (HCFS) shuffle. Learn more

An autoscaling policy must be selected to configure EFM.

Versioning

3. CHOOSE VERSIONING


Create a Dataproc cluster on Compute Engine
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
CHANGE

Compatible File System (HCFS) shuffle. Learn more

An autoscaling policy must be selected to configure EFM.

Versioning

Use a custom image to load pre-installed packages. Learn more

Image Type and Version
1.5-debian10

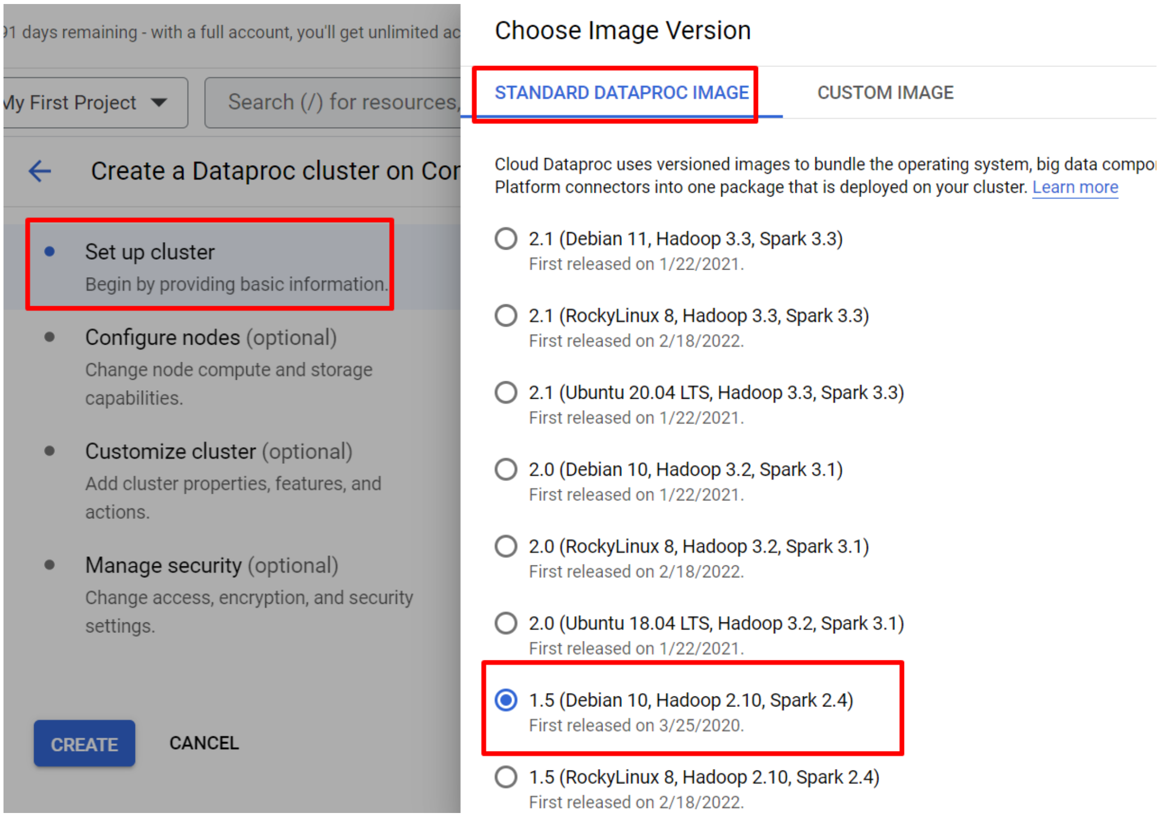
Release Date
First released on 3/25/20:

Components

Component Gateway

☑ Enable component gateway
Provides access to the web interfaces of default and selected optional components on the

we select this version
but it doesn't really matter


91 days remaining - with a full account, you'll get unlimited ac
My First Project
Search (/) for resources,
Create a Dataproc cluster on Cor
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
STANDARD DATAPROC IMAGE
CUSTOM IMAGE

Choose Image Version

Cloud Dataproc uses versioned images to bundle the operating system, big data compo
Platform connectors into one package that is deployed on your cluster. Learn more

○ 2.1 (Debian 11, Hadoop 3.3, Spark 3.3)
First released on 1/22/2021.

○ 2.1 (RockyLinux 8, Hadoop 3.3, Spark 3.3)
First released on 2/18/2022.

○ 2.1 (Ubuntu 20.04 LTS, Hadoop 3.3, Spark 3.3)
First released on 1/22/2021.

○ 2.0 (Debian 10, Hadoop 3.2, Spark 3.1)
First released on 1/22/2021.

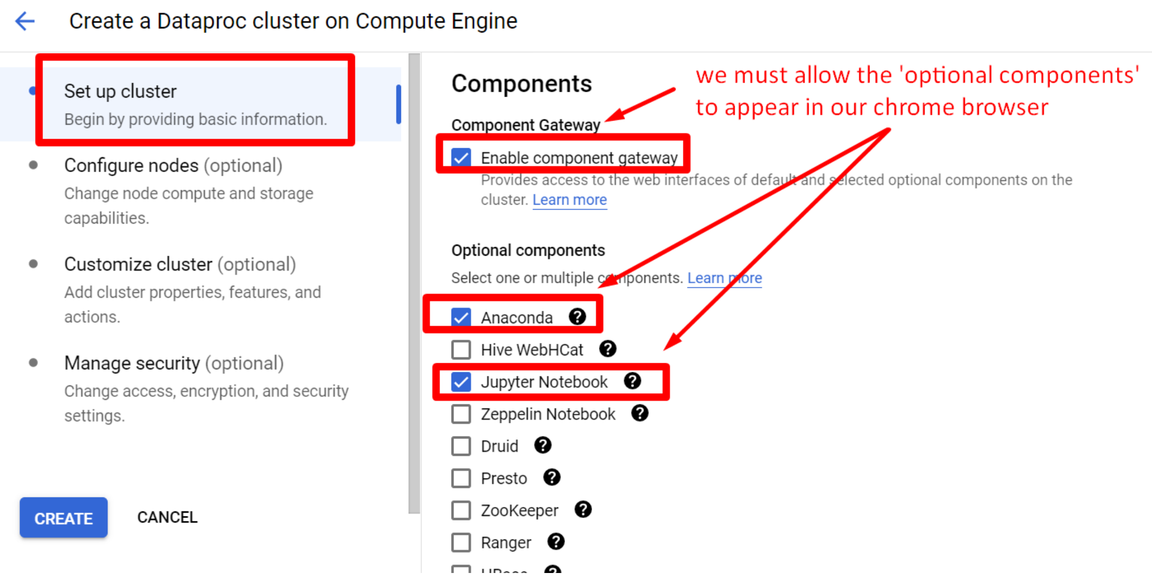
○ 2.0 (RockyLinux 8, Hadoop 3.2, Spark 3.1)
First released on 2/18/2022.

○ 2.0 (Ubuntu 18.04 LTS, Hadoop 3.2, Spark 3.1)
First released on 1/22/2021.

● 1.5 (Debian 10, Hadoop 2.10, Spark 2.4)
First released on 3/25/2020.

○ 1.5 (RockyLinux 8, Hadoop 2.10, Spark 2.4)
First released on 2/18/2022.

4. SELECT COMPONENTS



Create a Dataproc cluster on Compute Engine

- 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

Components

we must allow the 'optional components' to appear in our chrome browser

Component Gateway

☑ Enable component gateway
Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

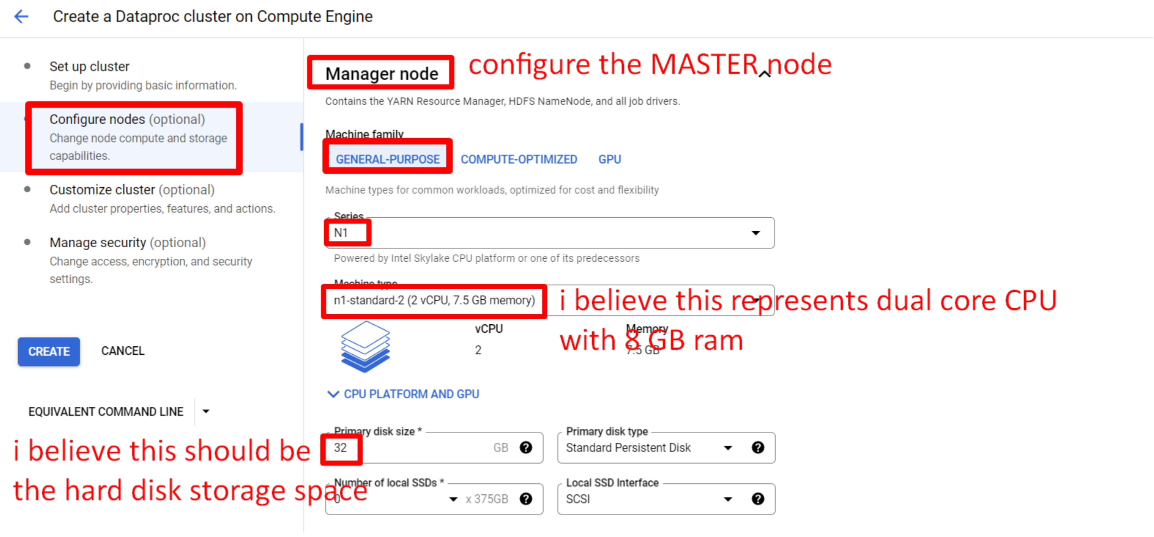
Optional components

Select one or multiple components. Learn more

☑ Anaconda
□ Hive WebHCat
☑ Jupyter Notebook
□ Zeppelin Notebook
□ Druid
□ Presto
□ ZooKeeper
□ Ranger

B. CONFIGURE NODES

1. CONFIGURE MASTER NODE



Create a Dataproc cluster on Compute Engine

- 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

Manager node

configure the MASTER node

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

Machine family: GENERAL-PURPOSE COMPUTE-OPTIMIZED GPU

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)

i believe this represents dual core CPU with 8 GB ram

vCPU: 2

Memory: 7.5 GB

CPU PLATFORM AND GPU

Primary disk size *: 32 GB

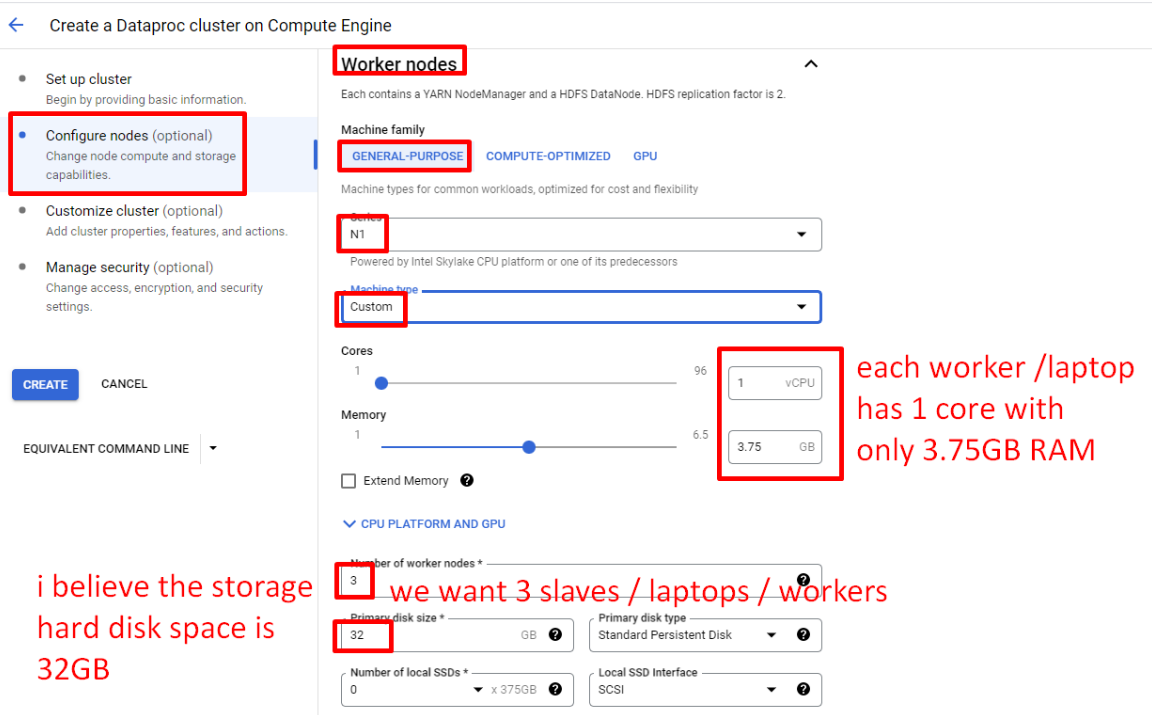
Primary disk type: Standard Persistent Disk

i believe this should be the hard disk storage space

Number of local SSDs *: 0 x 375GB

Local SSD Interface: SCSI

2. CONFIGURE SLAVE NODES



Create a Dataproc cluster on Compute Engine

- 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

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

Machine family: GENERAL-PURPOSE COMPUTE-OPTIMIZED GPU

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: Custom

Cores: 1 — 96: 1 vCPU

Memory: 1 — 6.5: 3.75 GB

☐ Extend Memory

each worker /laptop has 1 core with only 3.75GB RAM

CPU PLATFORM AND GPU

Number of worker nodes *: 3

we want 3 slaves / laptops / workers

Primary disk size *: 32 GB

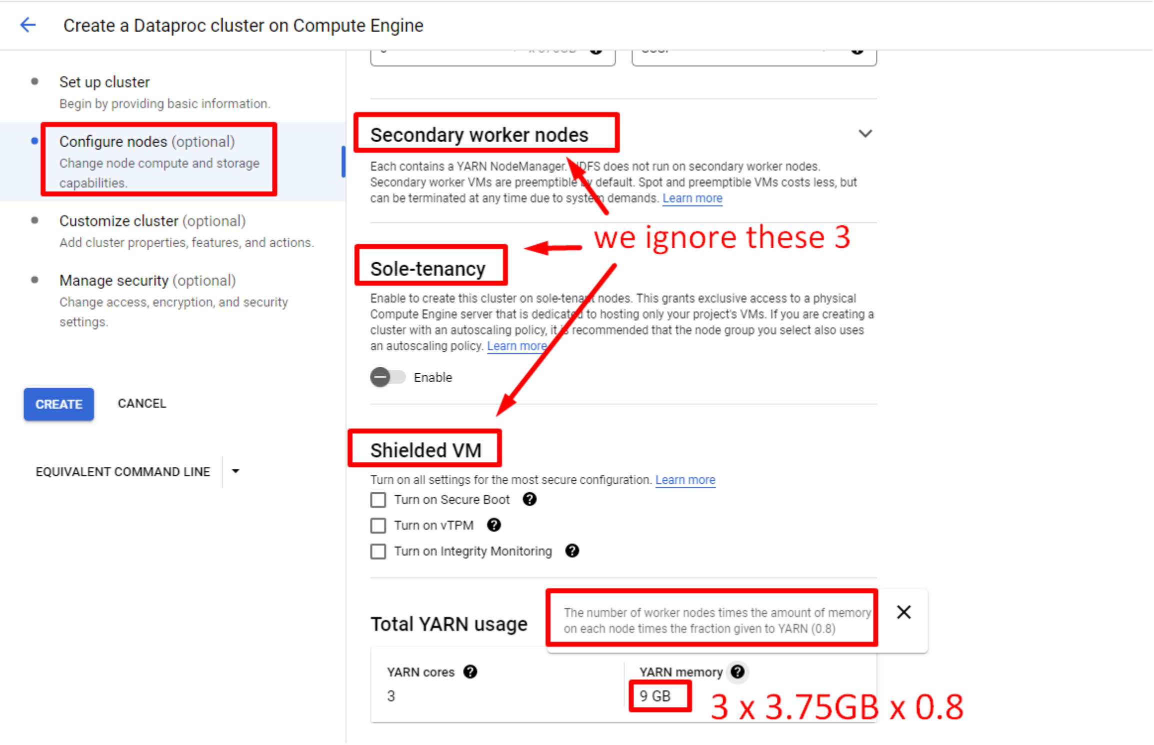
Primary disk type: Standard Persistent Disk

i believe the storage hard disk space is 32GB

Number of local SSDs *: 0 x 375GB

Local SSD Interface: SCSI

3. OBSERVE THE TOTAL YARN USAGE



Create a Dataproc cluster on Compute Engine

- 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.

Secondary worker nodes

Each contains a YARN NodeManager. HDFS does not run on secondary worker nodes. Secondary worker VMs are preemptible by default. Spot and preemptible VMs costs less, but can be terminated at any time due to system demands. Learn more

we ignore these 3

Sole-tenancy

Enable to create this cluster on sole-tenant nodes. This grants exclusive access to a physical Compute Engine server that is dedicated to hosting only your project's VMs. If you are creating a cluster with an autoscaling policy, it is recommended that the node group you select also uses an autoscaling policy. Learn more

Enable

Shielded VM

Turn on all settings for the most secure configuration. Learn more

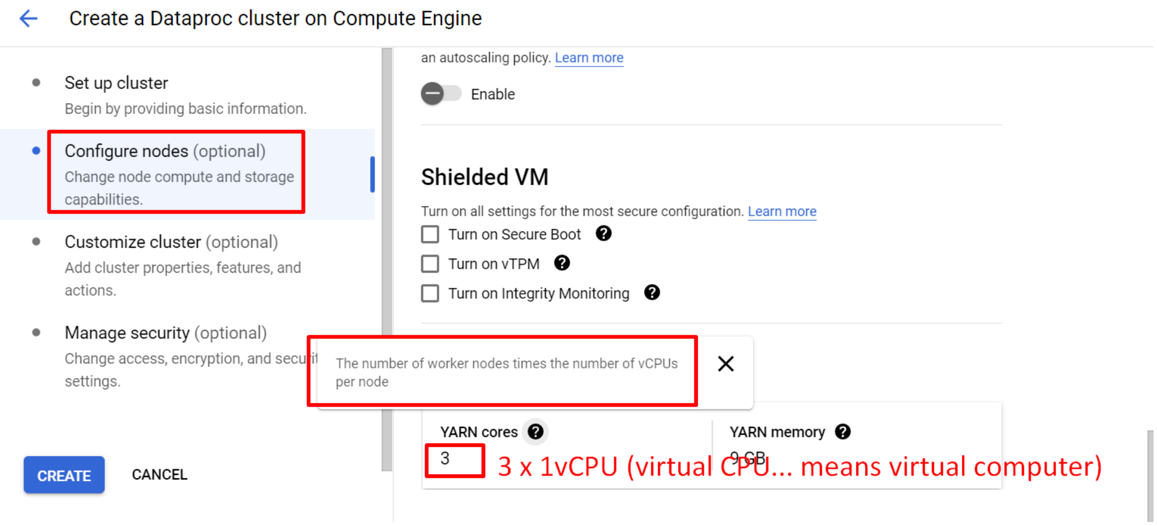
- □ Turn on Secure Boot
- □ Turn on vTPM
- □ Turn on Integrity Monitoring

Total YARN usage

The number of worker nodes times the amount of memory on each node times the fraction given to YARN (0.8)

YARN cores	YARN memory
3	9 GB

3 x 3.75GB x 0.8



Create a Dataproc cluster on Compute Engine

- 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.

an autoscaling policy. Learn more

Enable

Shielded VM

Turn on all settings for the most secure configuration. Learn more

- □ Turn on Secure Boot
- □ Turn on vTPM
- □ Turn on Integrity Monitoring

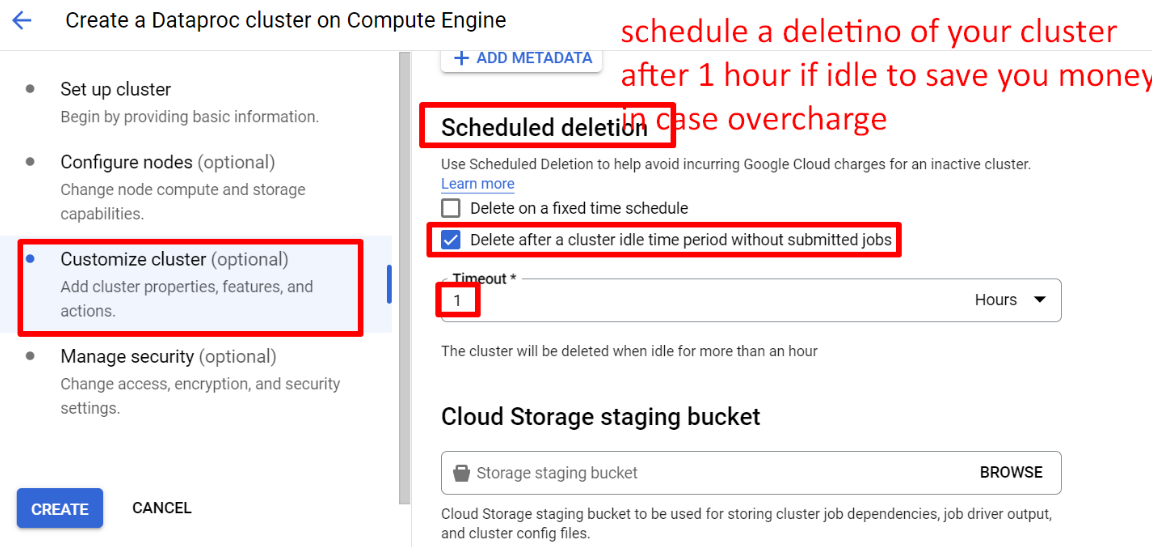
The number of worker nodes times the number of vCPUs per node

YARN cores	YARN memory
3	9 GB

3 x 1vCPU (virtual CPU... means virtual computer)

C. CUSTOMIZE CLUSTER

1. SCHEDULED DELETION OF CLUSTER



schedule a deletino of your cluster after 1 hour if idle to save you money in case overcharge

Create a Dataproc cluster on Compute Engine

- 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

+ ADD METADATA

Scheduled deletion

Use Scheduled Deletion to help avoid incurring Google Cloud charges for an inactive cluster. Learn more

□ Delete on a fixed time schedule

☑ Delete after a cluster idle time period without submitted jobs

Timeout * 1 Hours

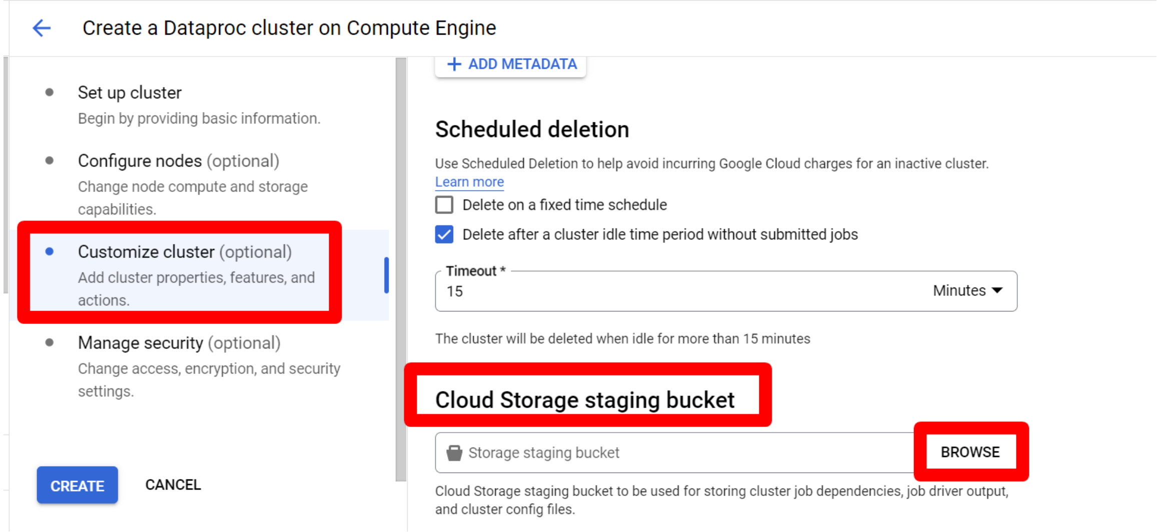
The cluster will be deleted when idle for more than an hour

Cloud Storage staging bucket

Storage staging bucket BROWSE

Cloud Storage staging bucket to be used for storing cluster job dependencies, job driver output, and cluster config files.

2. EDIT BUCKET NOW…



Create a Dataproc cluster on Compute Engine

- 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

+ ADD METADATA

Scheduled deletion

Use Scheduled Deletion to help avoid incurring Google Cloud charges for an inactive cluster. Learn more

□ Delete on a fixed time schedule

☑ Delete after a cluster idle time period without submitted jobs

Timeout * 15 Minutes

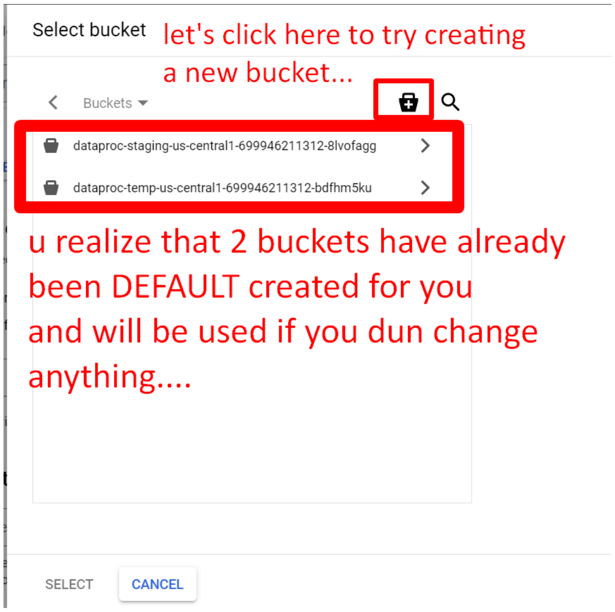
The cluster will be deleted when idle for more than 15 minutes

Cloud Storage staging bucket

Storage staging bucket BROWSE

Cloud Storage staging bucket to be used for storing cluster job dependencies, job driver output, and cluster config files.

a) Create a New Bucket



Select bucket

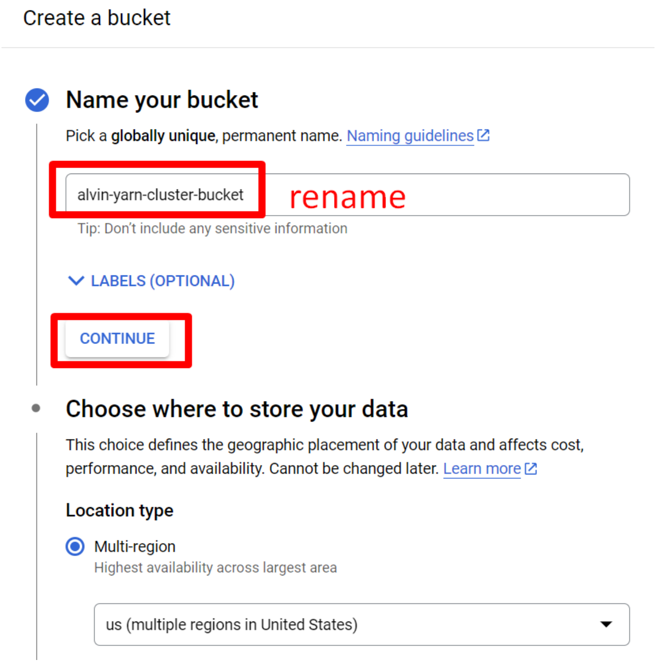
let's click here to try creating a new bucket...

dataproc-staging-us-central1-699946211312-8lvofagg

dataproc-temp-us-central1-699946211312-bdfhm5ku

u realize that 2 buckets have already been DEFAULT created for you and will be used if you dun change anything....

b) Name Your New Bucket



Create a bucket

Name your bucket

Pick a **globally unique**, permanent name. Naming guidelines

alvin-yarn-cluster-bucket

rename

Tip: Don't include any sensitive information

Choose where to store your data

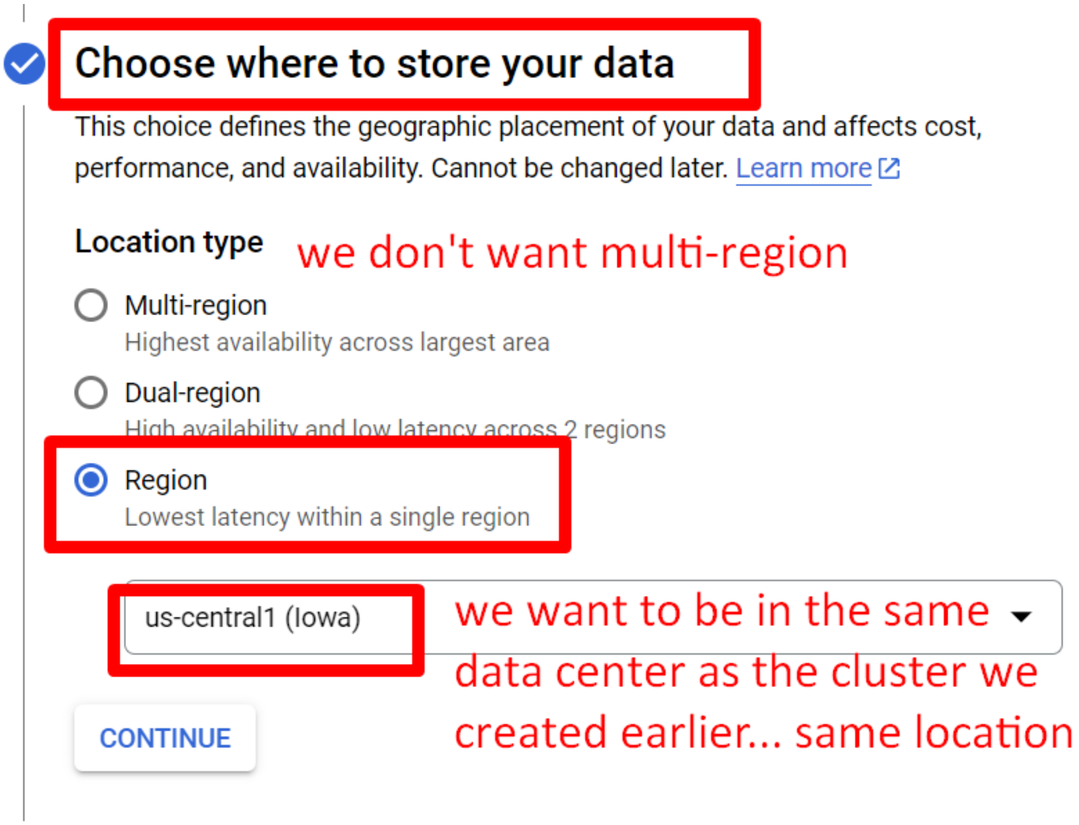
This choice defines the geographic placement of your data and affects cost, performance, and availability. Cannot be changed later. Learn more

Location type

● Multi-region
Highest availability across largest area

us (multiple regions in United States)

c) Where to Store Your Data?



Choose where to store your data

This choice defines the geographic placement of your data and affects cost, performance, and availability. Cannot be changed later. Learn more

Location type

we don't want multi-region

○ Multi-region
Highest availability across largest area

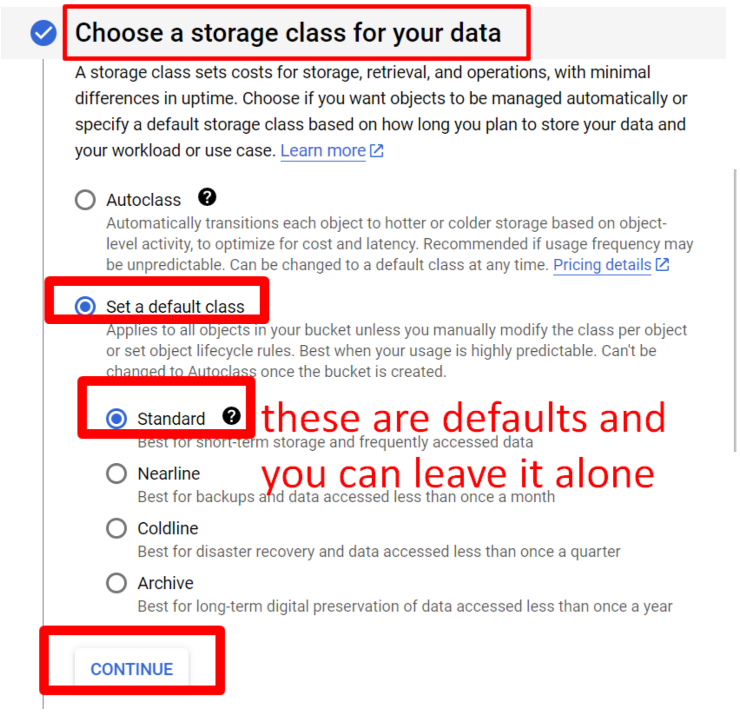
○ Dual-region
High availability and low latency across 2 regions

● Region
Lowest latency within a single region

us-central1 (Iowa)

we want to be in the same data center as the cluster we created earlier... same location

d) Leave the Rest of the Default Settings



Choose a storage class for your data

A storage class sets costs for storage, retrieval, and operations, with minimal differences in uptime. Choose if you want objects to be managed automatically or specify a default storage class based on how long you plan to store your data and your workload or use case. Learn more

○ Autoclass

Automatically transitions each object to hotter or colder storage based on object-level activity, to optimize for cost and latency. Recommended if usage frequency may be unpredictable. Can be changed to a default class at any time. Pricing details

● Set a default class

Applies to all objects in your bucket unless you manually modify the class per object or set object lifecycle rules. Best when your usage is highly predictable. Can't be changed to Autoclass once the bucket is created.

● Standard

Best for short-term storage and frequently accessed data

○ Nearline

Best for backups and data accessed less than once a month

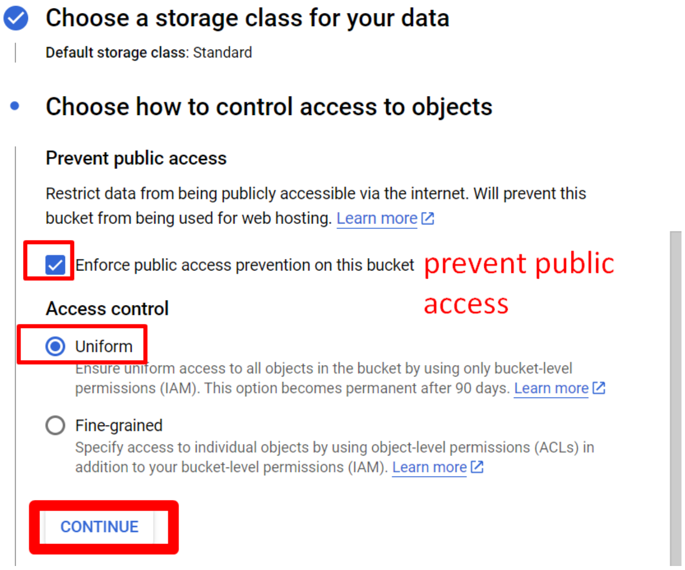
○ Coldline

Best for disaster recovery and data accessed less than once a quarter

○ Archive

Best for long-term digital preservation of data accessed less than once a year

these are defaults and you can leave it alone



Choose a storage class for your data

Default storage class: Standard

Choose how to control access to objects

Prevent public access

Restrict data from being publicly accessible via the internet. Will prevent this bucket from being used for web hosting. Learn more

☑ Enforce public access prevention on this bucket

prevent public access

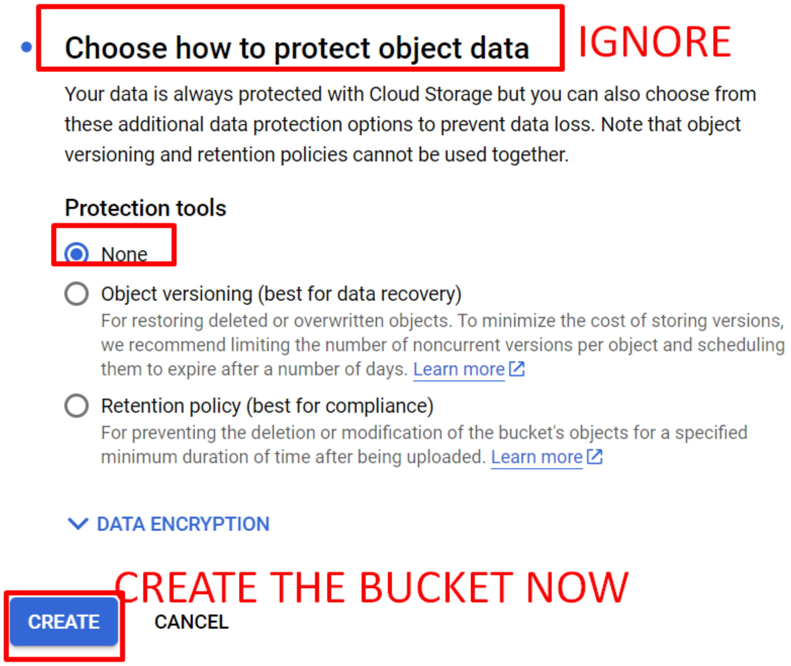
Access control

● Uniform

Ensure uniform access to all objects in the bucket by using only bucket-level permissions (IAM). This option becomes permanent after 90 days. Learn more

○ Fine-grained

Specify access to individual objects by using object-level permissions (ACLs) in addition to your bucket-level permissions (IAM). Learn more



Choose how to protect object data

IGNORE

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.

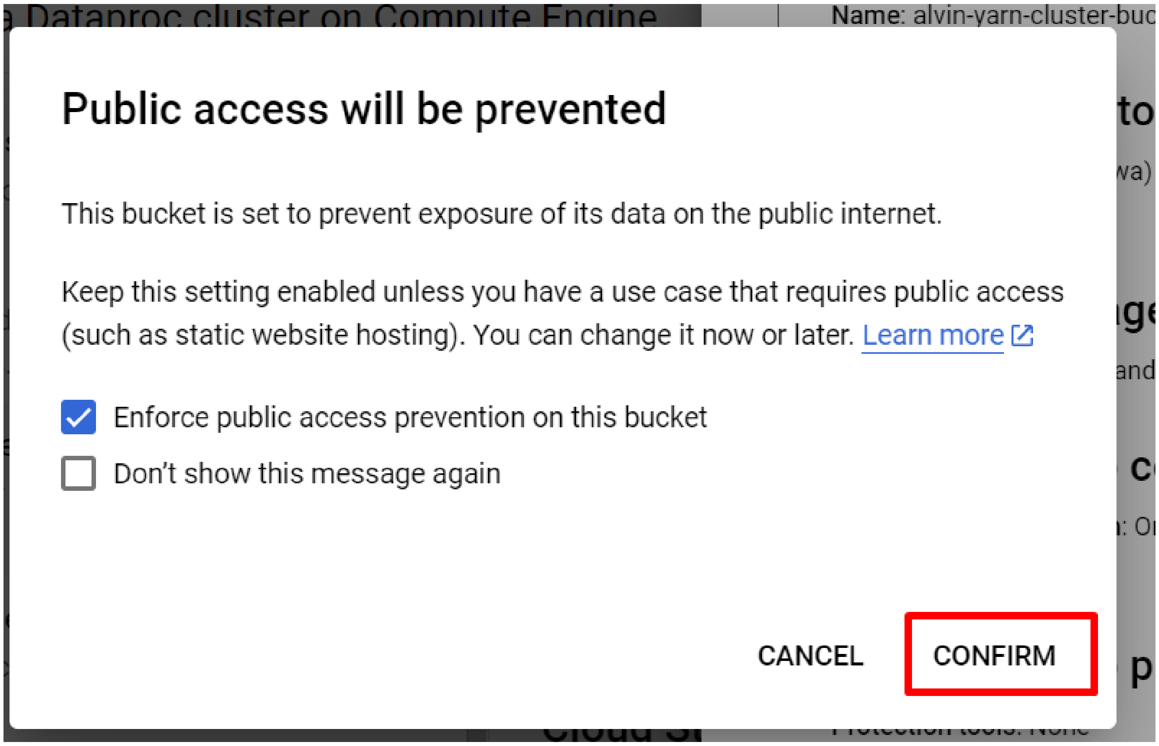
Protection tools

● None

○ Object versioning (best for data recovery)
For restoring deleted or overwritten objects. To minimize the cost of storing versions, we recommend limiting the number of noncurrent versions per object and scheduling them to expire after a number of days. Learn more

○ Retention policy (best for compliance)
For preventing the deletion or modification of the bucket's objects for a specified minimum duration of time after being uploaded. Learn more

CREATE THE BUCKET NOW



a Dataproc cluster on Compute Engine

Name: alvin-yarn-cluster-buc

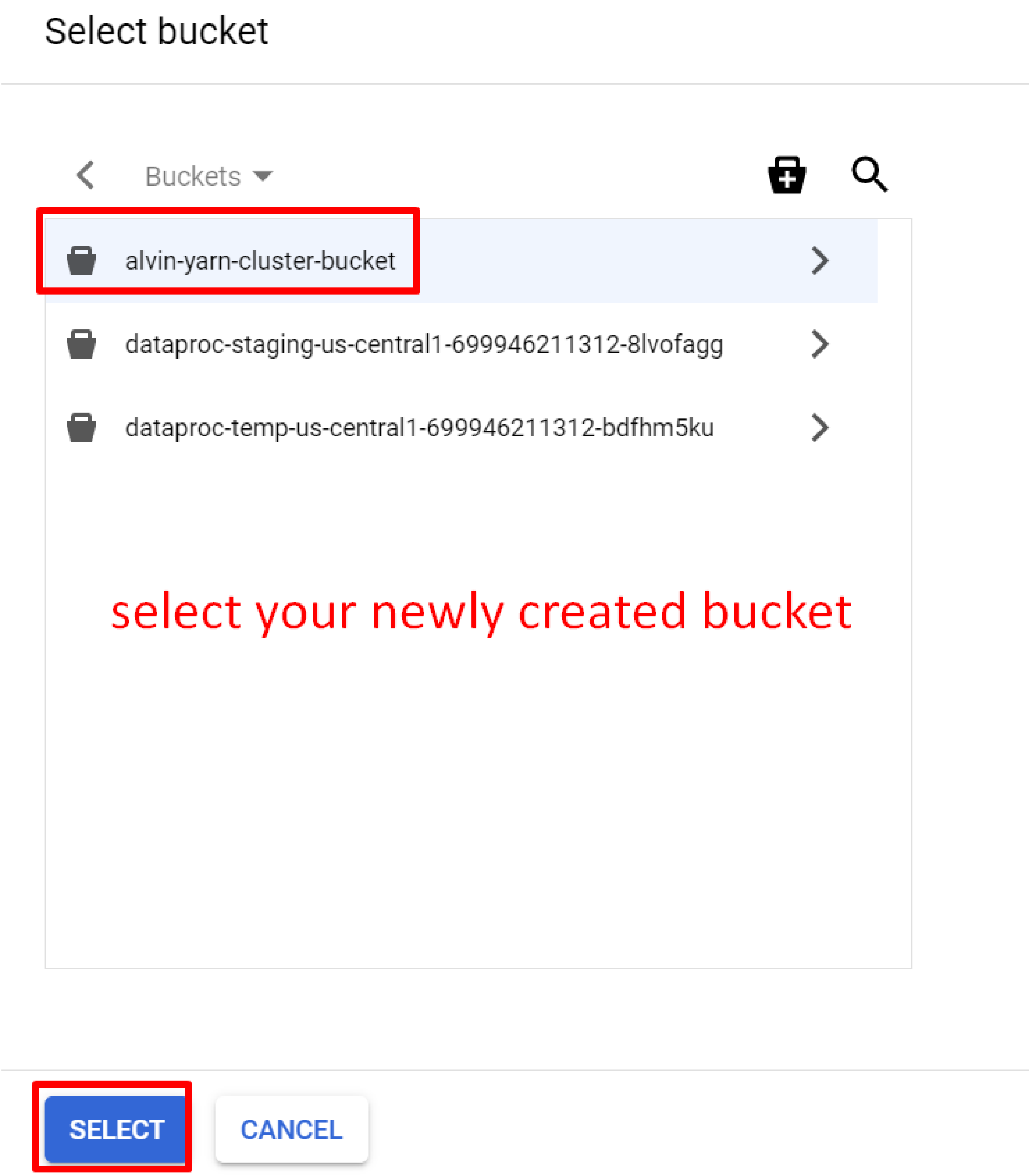
Public access will be prevented

This bucket is set to prevent exposure of its data on the public internet.

Keep this setting enabled unless you have a use case that requires public access (such as static website hosting). You can change it now or later. Learn more

☑ Enforce public access prevention on this bucket

□ Don't show this message again



Select bucket

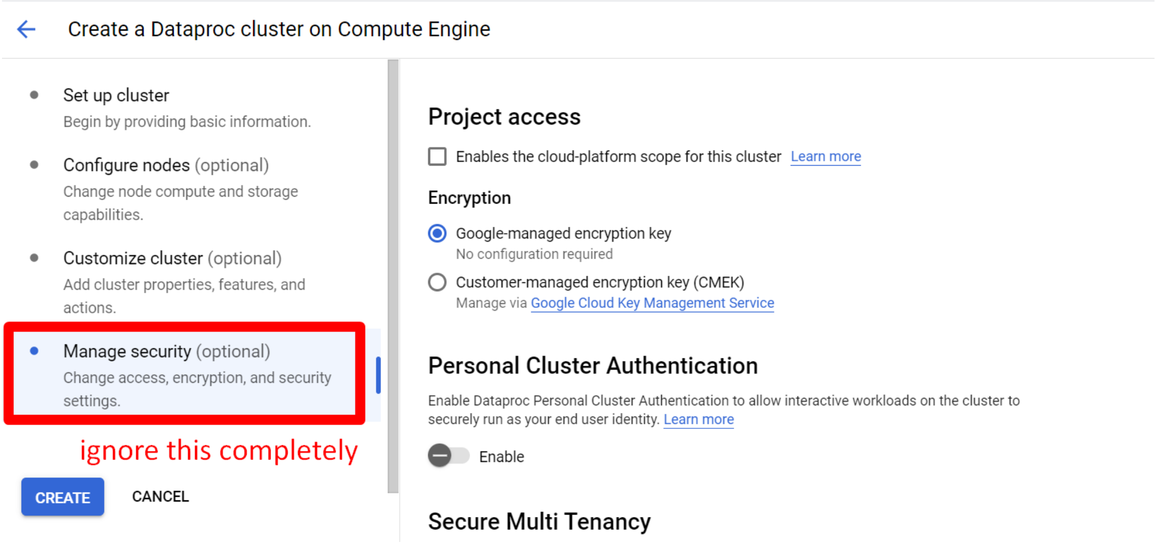
alvin-yarn-cluster-bucket

dataproc-staging-us-central1-699946211312-8lvofagg

dataproc-temp-us-central1-699946211312-bdfhm5ku

select your newly created bucket

D. IGNORE MANAGING SECURITY



Create a Dataproc cluster on Compute Engine

- 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.

ignore this completely

Project access

□ Enables the cloud-platform scope for this cluster Learn more

Encryption

● Google-managed encryption key
No configuration required

○ Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

Personal Cluster Authentication

Enable Dataproc Personal Cluster Authentication to allow interactive workloads on the cluster to securely run as your end user identity. Learn more

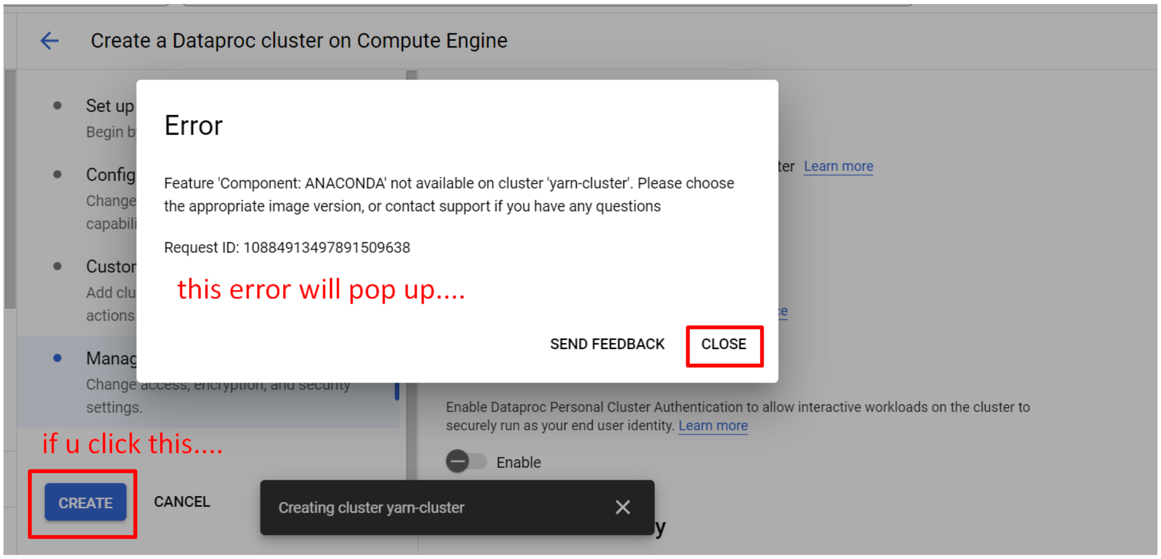
Enable

Secure Multi Tenancy

VII. CREATE THE CLUSTER

A. NO ANACONDA ISSUE….

When u try to create a new cluster….



Create a Dataproc cluster on Compute Engine

Error

Feature 'Component: ANACONDA' not available on cluster 'yarn-cluster'. Please choose the appropriate image version, or contact support if you have any questions

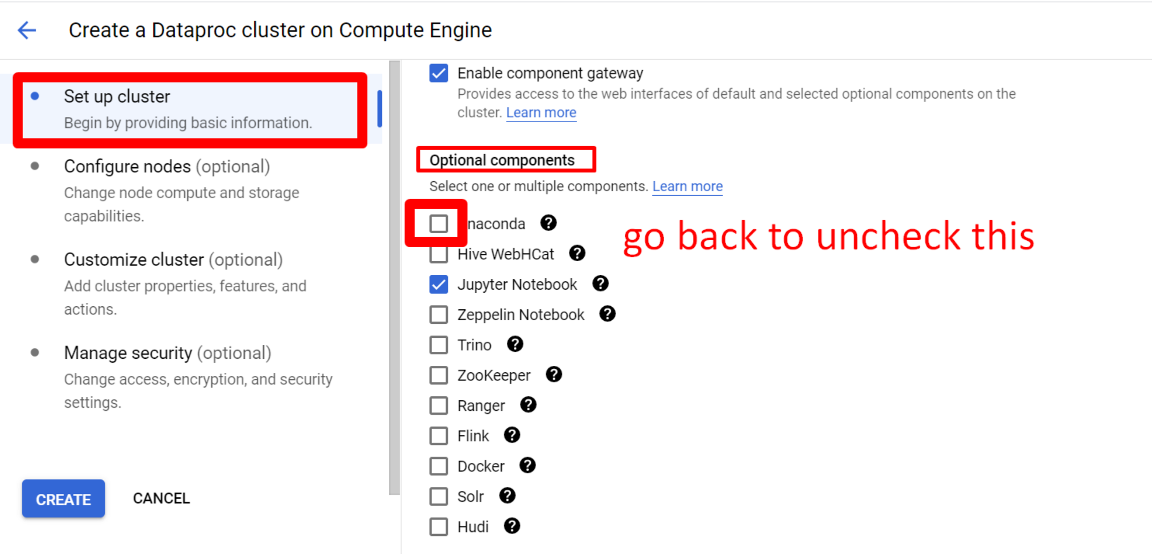
Request ID: 10884913497891509638

this error will pop up....

if u click this....

Enable Dataproc Personal Cluster Authentication to allow interactive workloads on the cluster to securely run as your end user identity. Learn more

Creating cluster yarn-cluster



Create a Dataproc cluster on Compute Engine

☑ Enable component gateway
Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

Optional components

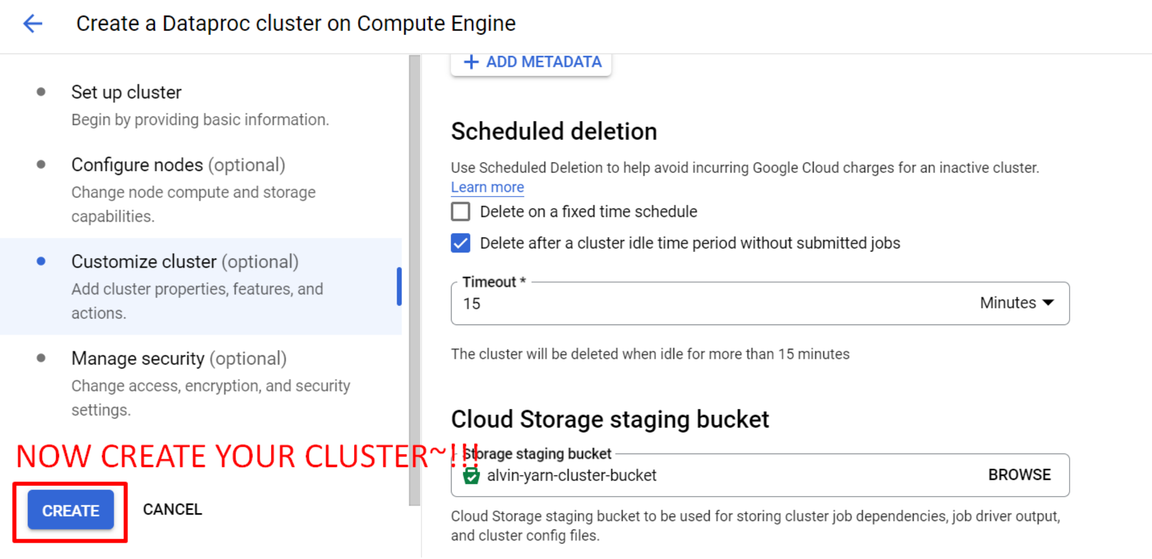
Select one or multiple components. Learn more

- □ naconda
- □ Hive WebHCat
- ☑ Jupyter Notebook
- □ Zeppelin Notebook
- □ Trino
- □ ZooKeeper
- □ Ranger
- □ Flink
- □ Docker
- □ Solr
- □ Hudi

go back to uncheck this

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

B. CREATE YOUR CLUSTER!



Create a Dataproc cluster on Compose Engine

- 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.

NOW CREATE YOUR CLUSTER~!!!

Scheduled deletion

Use Scheduled Deletion to help avoid incurring Google Cloud charges for an inactive cluster.
Learn more

☐ Delete on a fixed time schedule

☑ Delete after a cluster idle time period without submitted jobs

Timeout *: 15 Minutes

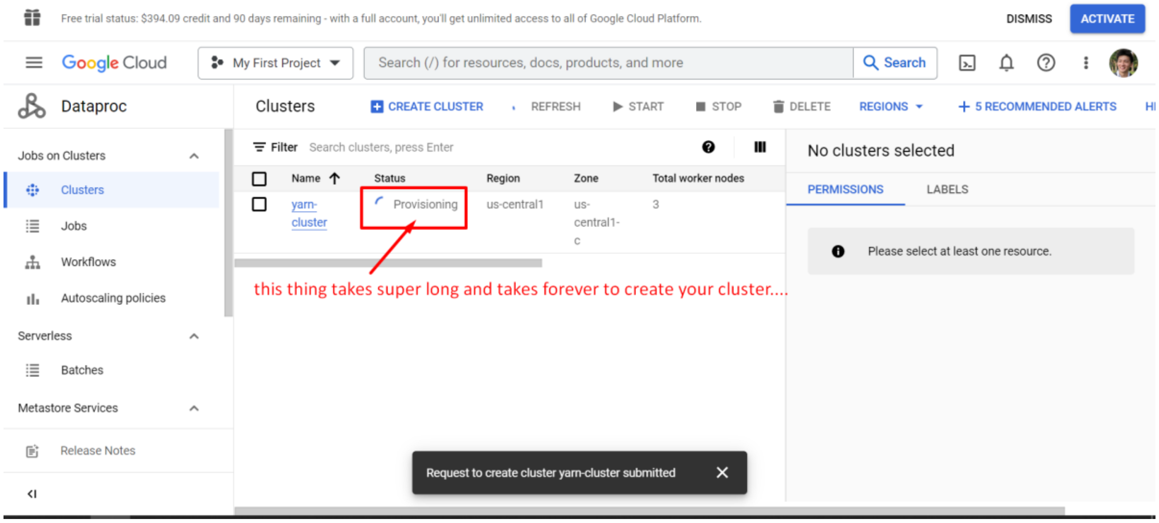
The cluster will be deleted when idle for more than 15 minutes

Cloud Storage staging bucket

Storage staging bucket: alvin-yarn-cluster-bucket

Cloud Storage staging bucket to be used for storing cluster job dependencies, job driver output, and cluster config files.

C. IT TAKES SUPER LONG....



Free trial status: $394.09 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

☐	Name ↑	Status	Region	Zone	Total worker nodes
☐	yarn-cluster	Provisioning	us-central1	us-central1-c	3

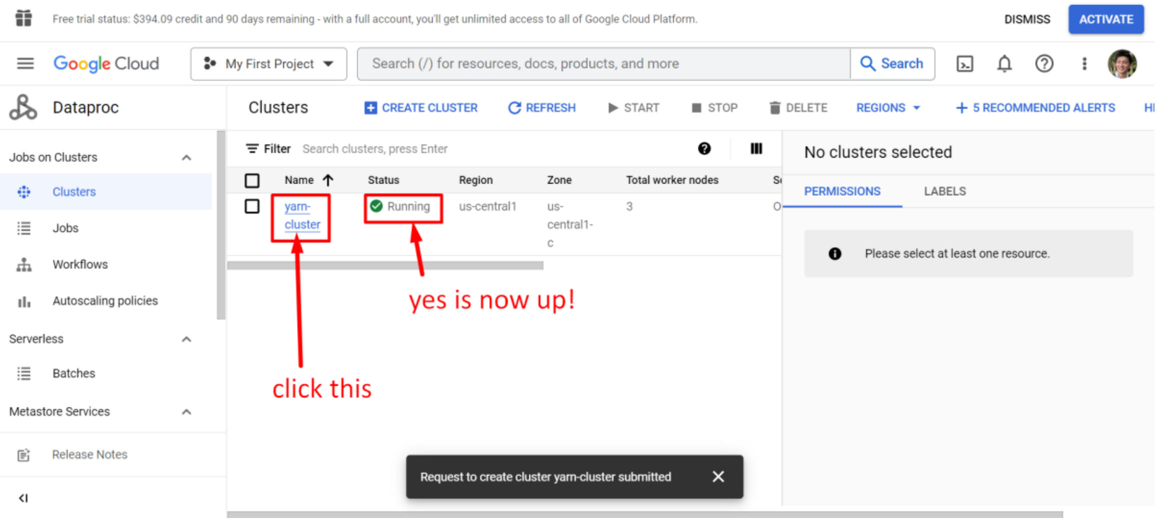
this thing takes super long and takes forever to create your cluster....

No clusters selected

Please select at least one resource.

Request to create cluster yarn-cluster submitted

D. MEANWHILE YOU MAY CHECK YOUR DASHBOARD…



Free trial status: $394.09 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Search (/) for resources, docs, products, and more

Search clusters, press Enter

Name	Status	Region	Zone	Total worker nodes
yarn-cluster	Running	us-central1	us-central1-c	3

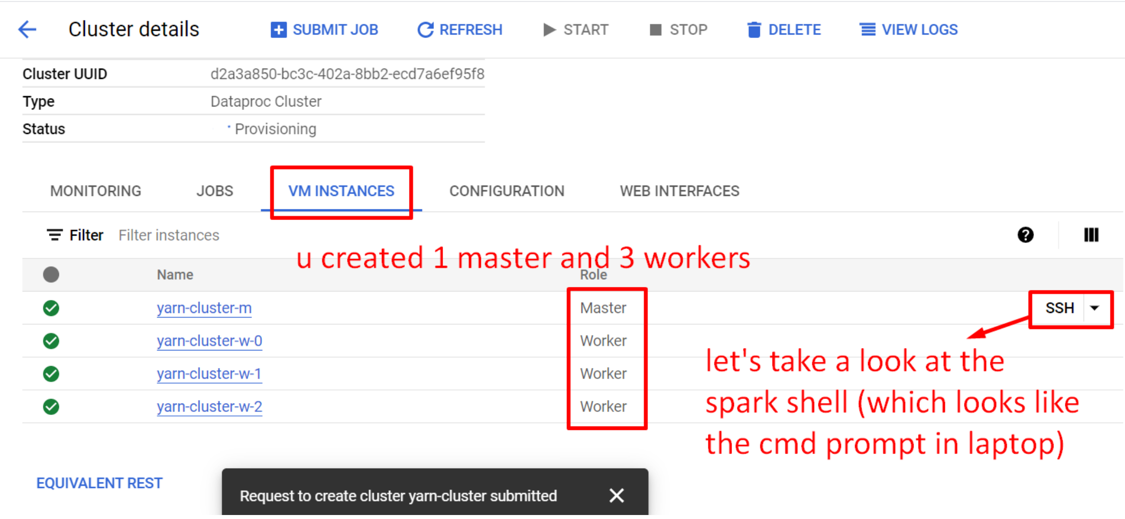
No clusters selected

Please select at least one resource.

Request to create cluster yarn-cluster submitted

yes is now up!

click this



Cluster UUID	d2a3a850-bc3c-402a-8bb2-ecd7a6ef95f8
Type	Dataproc Cluster
Status	Provisioning

Filter instances

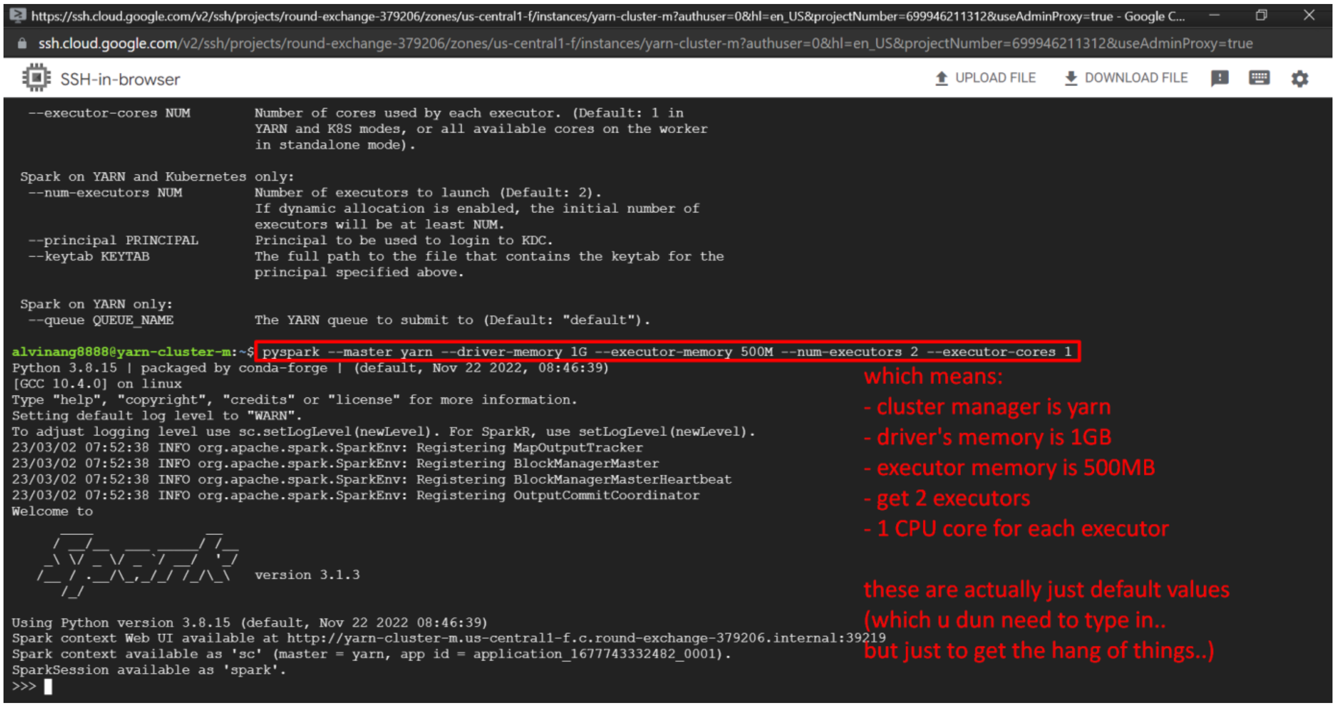
Name	Role
yarn-cluster-m	Master
yarn-cluster-w-0	Worker
yarn-cluster-w-1	Worker
yarn-cluster-w-2	Worker

Request to create cluster yarn-cluster submitted

u created 1 master and 3 workers

let's take a look at the spark shell (which looks like the cmd prompt in laptop)

E. SPARK SHELL

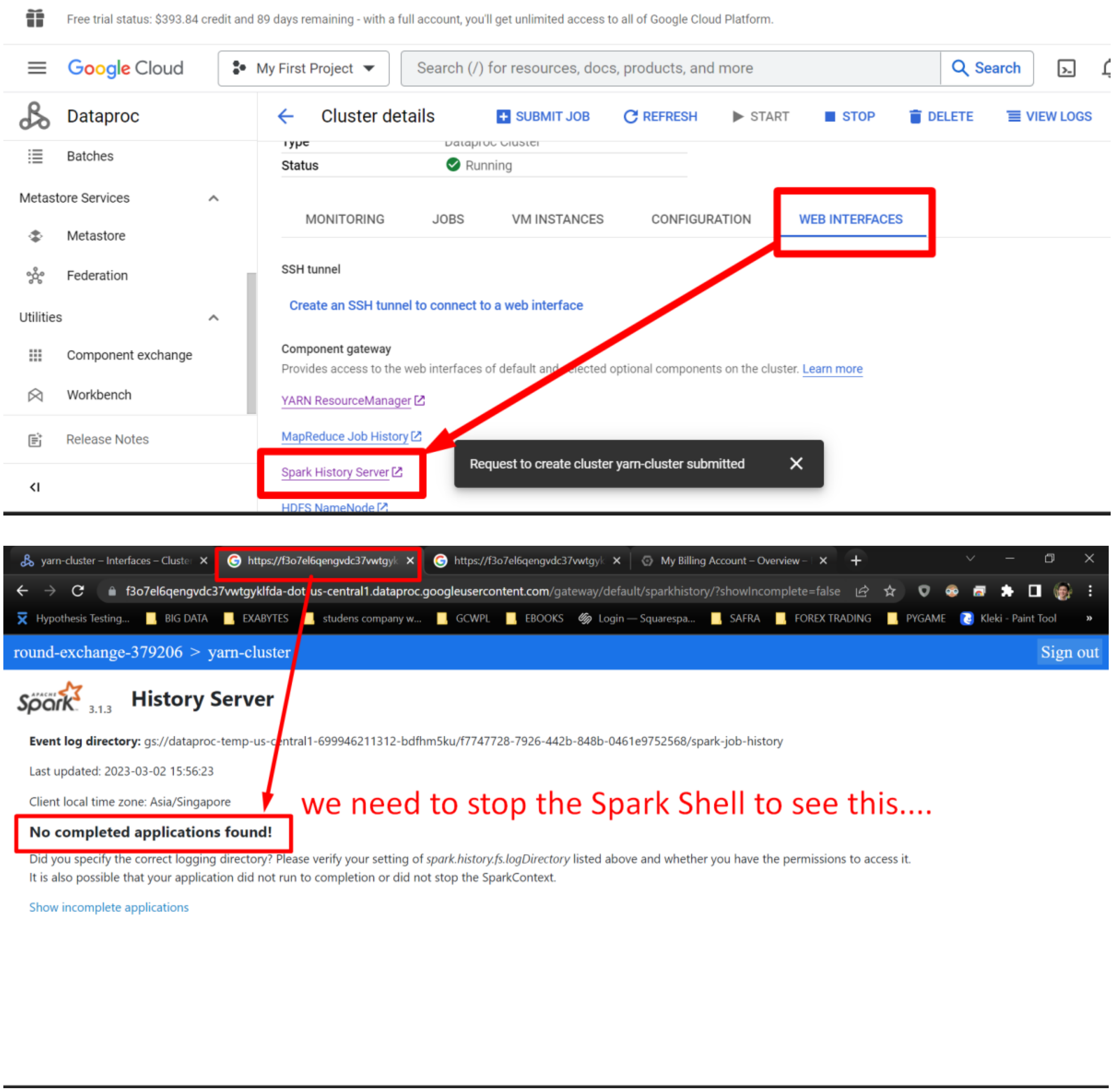

https://ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true - Google C...
ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
  --executor-cores NUM        Number of cores used by each executor. (Default: 1 in
                              YARN and K8S modes, or all available cores on the worker
                              in standalone mode).

 Spark on YARN and Kubernetes only:
  --num-executors NUM         Number of executors to launch (Default: 2).
                              If dynamic allocation is enabled, the initial number of
                              executors will be at least NUM.
  --principal PRINCIPAL       Principal to be used to login to KDC.
  --keytab KEYTAB             The full path to the file that contains the keytab for the
                              principal specified above.

 Spark on YARN only:
  --queue QUEUE_NAME          The YARN queue to submit to (Default: "default").

alvinang8888@yarn-cluster-m:~$ pyspark --master yarn --driver-memory 1G --executor-memory 500M --num-executors 2 --executor-cores 1
Python 3.8.15 | packaged by conda-forge | (default, Nov 22 2022, 08:46:39)
[GCC 10.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/03/02 07:52:38 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/03/02 07:52:38 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/03/02 07:52:38 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
23/03/02 07:52:38 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.3
      /_/

Using Python version 3.8.15 (default, Nov 22 2022 08:46:39)
Spark context Web UI available at http://yarn-cluster-m.us-central1-f.c.round-exchange-379206.internal:39219
Spark context available as 'sc' (master = yarn, app id = application_1677743332482_0001).
SparkSession available as 'spark'.
>>> 
```

which means:
- cluster manager is yarn
- driver's memory is 1GB
- executor memory is 500MB
- get 2 executors
- 1 CPU core for each executor

these are actually just default values
(which u dun need to type in..
but just to get the hang of things..)

1. SPARK HISTORY SERVER



Type: Dataproc Cluster

Status: Running

SSH tunnel

Create an SSH tunnel to connect to a web interface

Component gateway

Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

YARN ResourceManager

MapReduce Job History

Spark History Server

HDFS NameNode

Request to create cluster yarn-cluster submitted

f3o7el6qengvdc37vwtgyklfda-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/?showIncomplete=false

Event log directory: gs://dataproc-temp-us-central1-699946211312-bdfhm5ku/f7747728-7926-442b-848b-0461e9752568/spark-job-history

Last updated: 2023-03-02 15:56:23

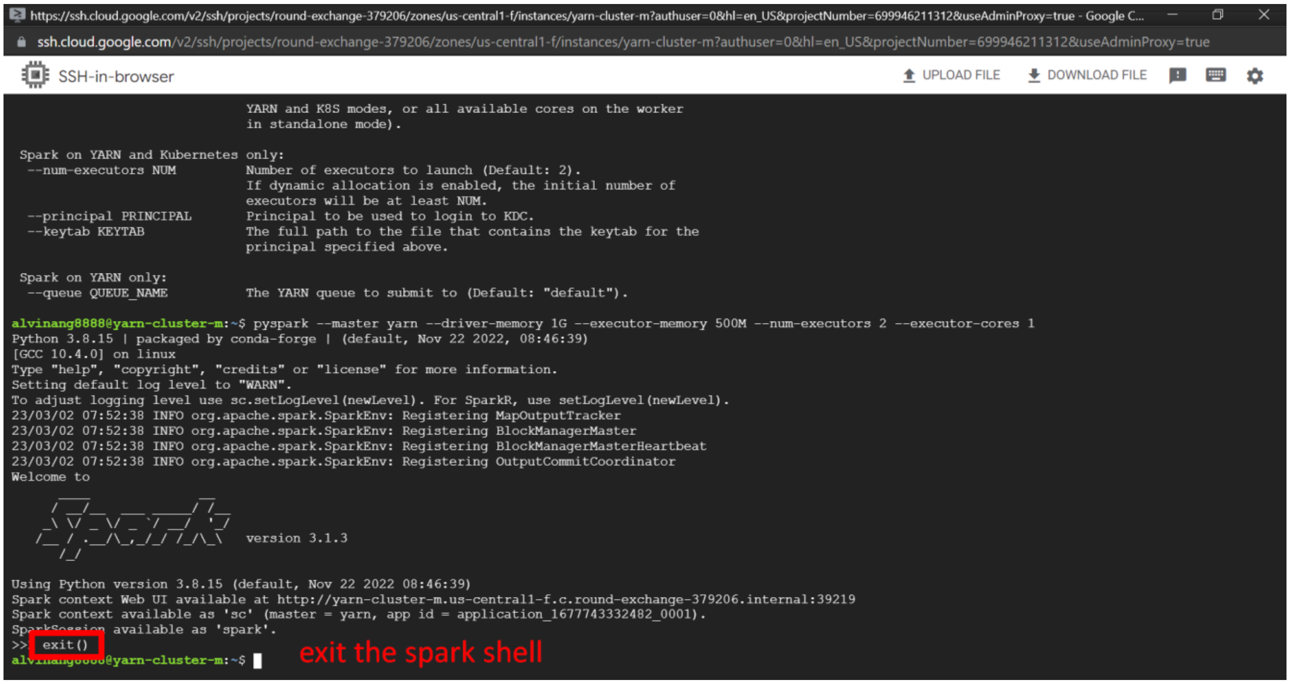
Client local time zone: Asia/Singapore

No completed applications found!

we need to stop the Spark Shell to see this....

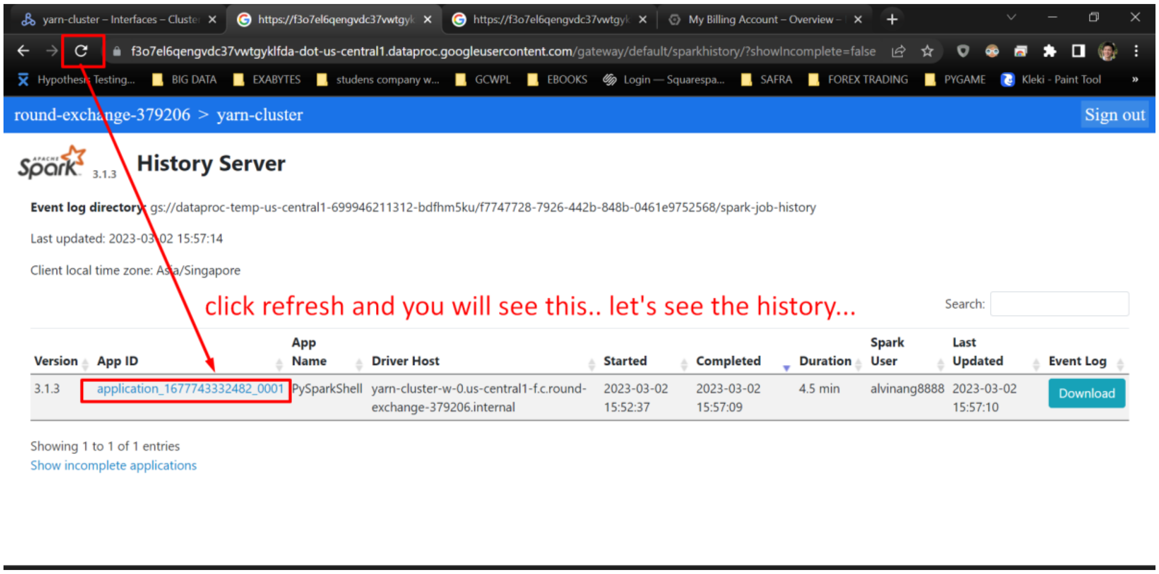
Did you specify the correct logging directory? Please verify your setting of *spark.history.fs.logDirectory* listed above and whether you have the permissions to access it.
It is also possible that your application did not run to completion or did not stop the SparkContext.

Show incomplete applications


https://ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true - Google C...
ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
                             YARN and K8S modes, or all available cores on the worker
                             in standalone mode).

 Spark on YARN and Kubernetes only:
  --num-executors NUM        Number of executors to launch (Default: 2).
                             If dynamic allocation is enabled, the initial number of
                             executors will be at least NUM.
  --principal PRINCIPAL      Principal to be used to login to KDC.
  --keytab KEYTAB            The full path to the file that contains the keytab for the
                             principal specified above.

 Spark on YARN only:
  --queue QUEUE_NAME         The YARN queue to submit to (Default: "default").

alvinang8888@yarn-cluster-m:~$ pyspark --master yarn --driver-memory 1G --executor-memory 500M --num-executors 2 --executor-cores 1
Python 3.8.15 | packaged by conda-forge | (default, Nov 22 2022, 08:46:39)
[GCC 10.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/03/02 07:52:38 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
23/03/02 07:52:38 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
23/03/02 07:52:38 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
23/03/02 07:52:38 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.1.3
      /_/

Using Python version 3.8.15 (default, Nov 22 2022 08:46:39)
Spark context Web UI available at http://yarn-cluster-m.us-central1-f.c.round-exchange-379206.internal:39219
Spark context available as 'sc' (master = yarn, app id = application_1677743332482_0001).
SparkSession available as 'spark'.
>>> exit()
alvinang8888@yarn-cluster-m:~$
```

exit the spark shell


yarn-cluster – Interfaces – Cluste
https://f3o7el6qengvdc37vwtgyk
https://f3o7el6qengvdc37vwtgyk
My Billing Account – Overview –
f3o7el6qengvdc37vwtgyklfda-dot-us-central1.dataproc.googleusercontent.com/gateway/default/sparkhistory/?showIncomplete=false
Hypothesis Testing...
BIG DATA
EXABYTES
studens company w...
GCWPL
EBOOKS
Login — Squarespa...
SAFRA
FOREX TRADING
PYGAME
Kleki - Paint Tool
round-exchange-379206 > yarn-cluster
Sign out
Spark 3.1.3

History Server

Event log directory: gs://dataproc-temp-us-central1-699946211312-bdfhm5ku/f7747728-7926-442b-848b-0461e9752568/spark-job-history

Last updated: 2023-03-02 15:57:14

Client local time zone: Asia/Singapore

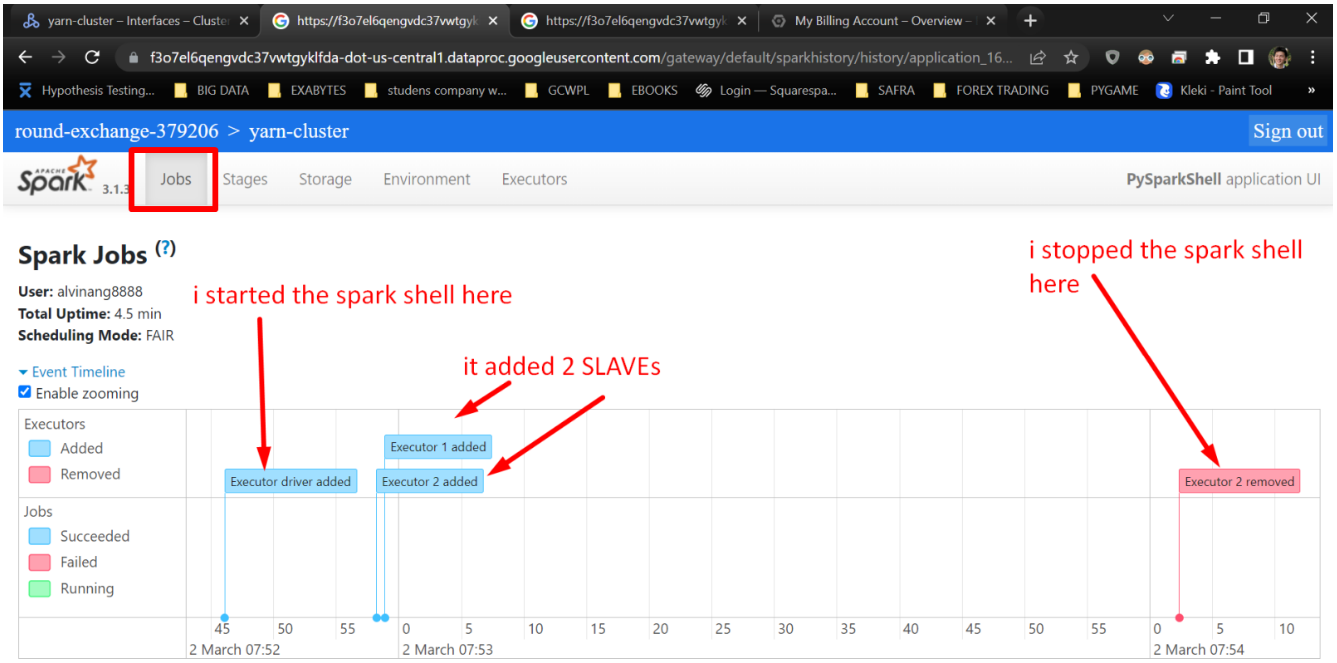
click refresh and you will see this.. let's see the history...

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.1.3	application_1677743332482_0001	PySparkShell	yarn-cluster-w-0.us-central1-f.c.round-exchange-379206.internal	2023-03-02 15:52:37	2023-03-02 15:57:09	4.5 min	alvinang8888	2023-03-02 15:57:10	Download

Showing 1 to 1 of 1 entries

Show incomplete applications



round-exchange-379206 > yarn-cluster

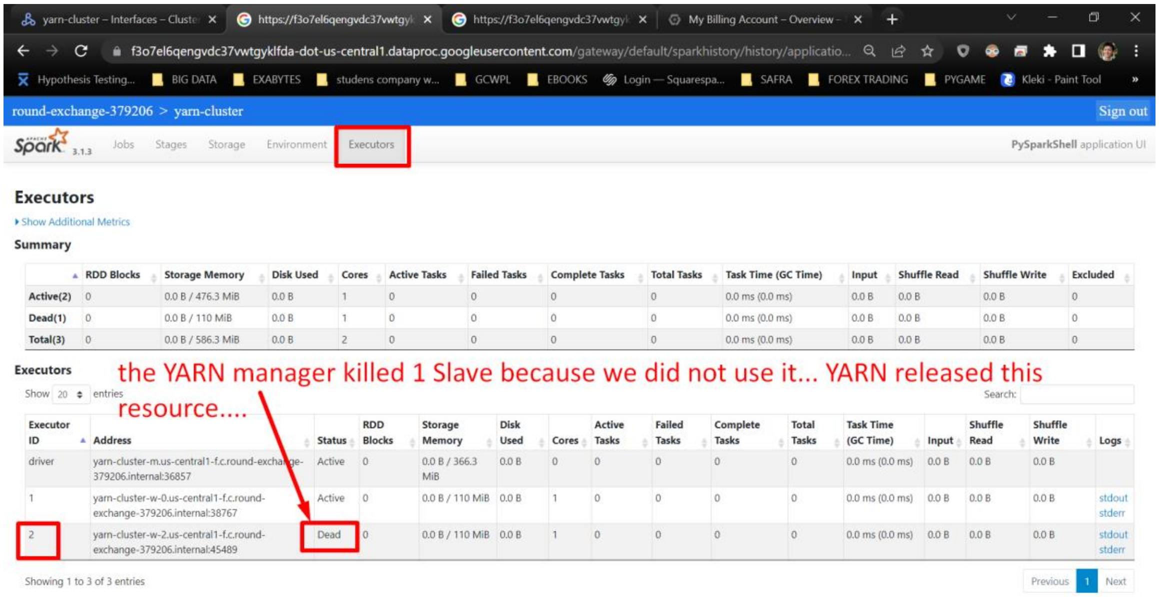
Spark Jobs (?)

User: alvinang8888
Total Uptime: 4.5 min
Scheduling Mode: FAIR

i started the spark shell here

it added 2 SLAVEs

i stopped the spark shell here



round-exchange-379206 > yarn-cluster

Executors

▸ Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(2)	0	0.0 B / 476.3 MiB	0.0 B	1	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Dead(1)	0	0.0 B / 110 MiB	0.0 B	1	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(3)	0	0.0 B / 586.3 MiB	0.0 B	2	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0

Executors

Show 20 entries

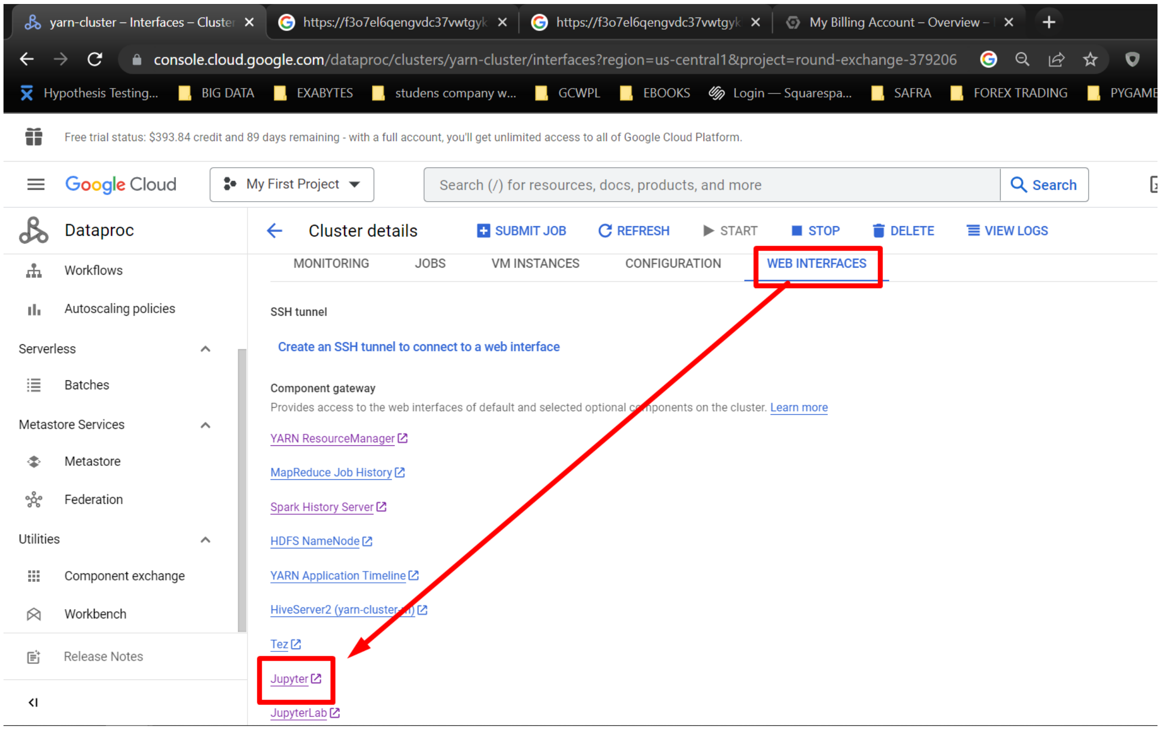
Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs
driver	yarn-cluster-m.us-central1-f.c.round-exchange-379206.internal:36857	Active	0	0.0 B / 366.3 MiB	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	
1	yarn-cluster-w-0.us-central1-f.c.round-exchange-379206.internal:38767	Active	0	0.0 B / 110 MiB	0.0 B	1	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr
2	yarn-cluster-w-2.us-central1-f.c.round-exchange-379206.internal:45489	Dead	0	0.0 B / 110 MiB	0.0 B	1	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr

Showing 1 to 3 of 3 entries

the YARN manager killed 1 Slave because we did not use it... YARN released this resource....

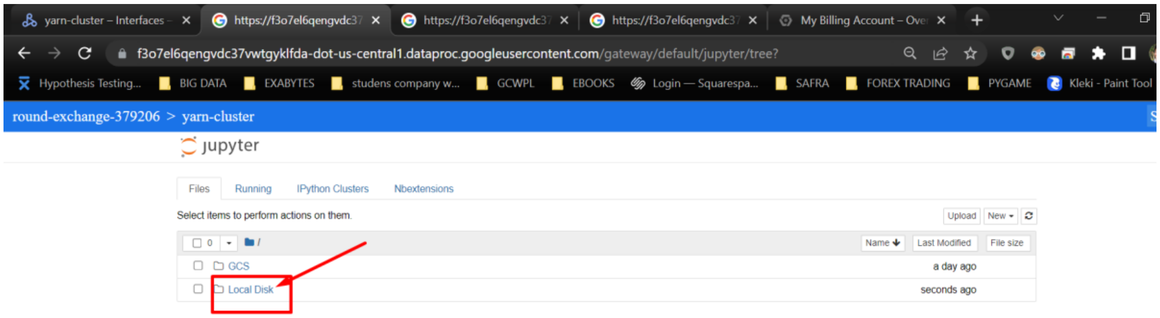
However, we are not likely to use the Spark Shell… we will most probably use the Jupyter Notebook…

F. JUPYTER NOTEBOOK



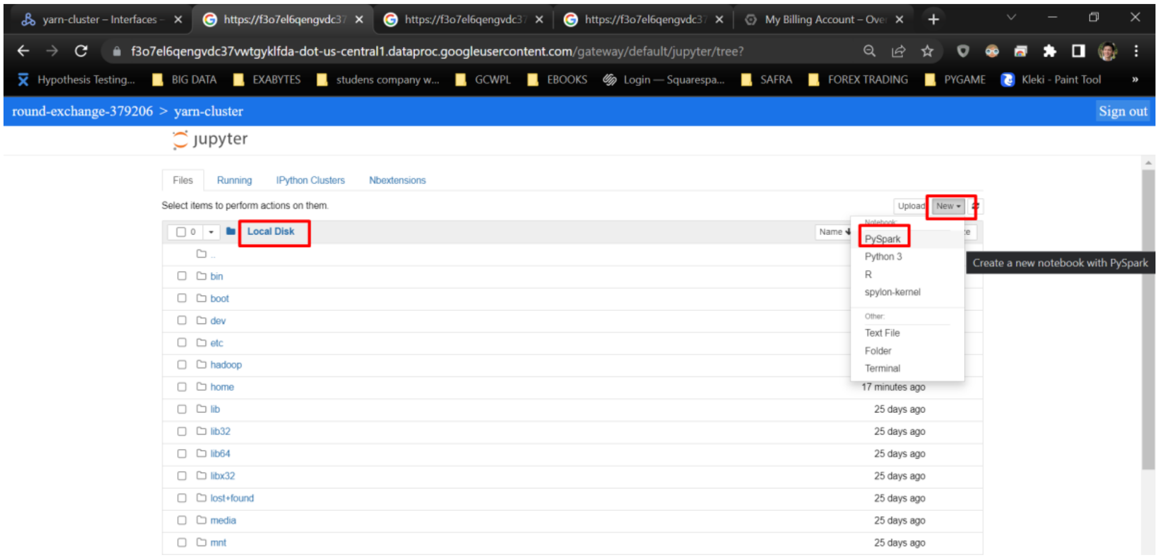
Free trial status: $393.84 credit and 89 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

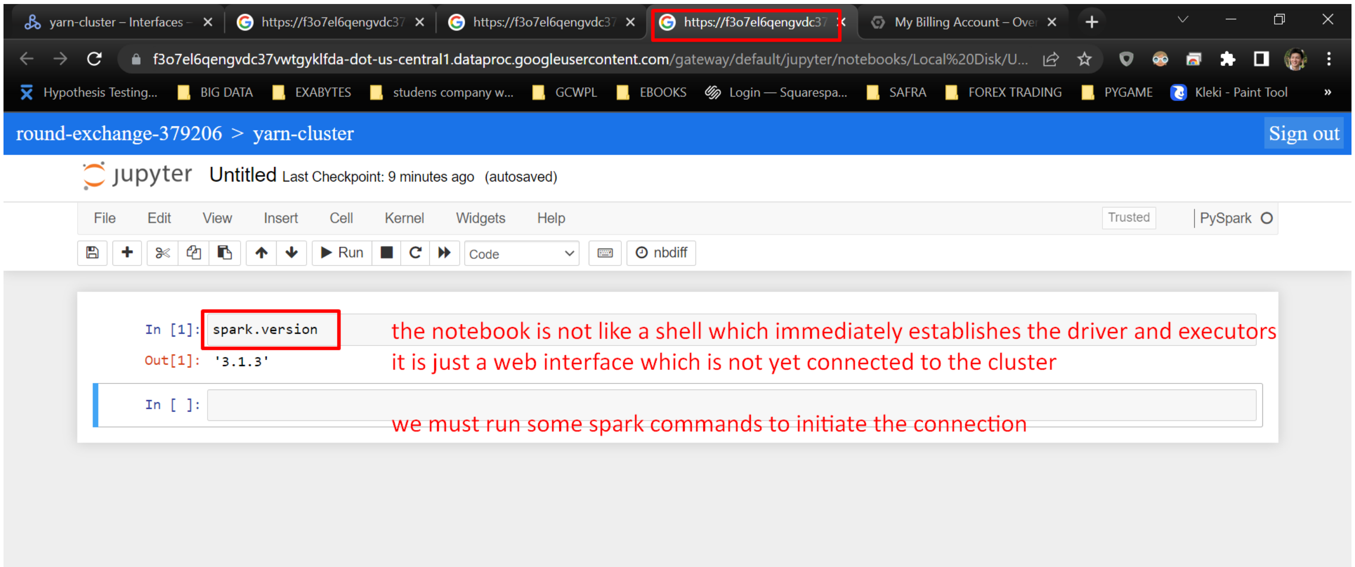


Select items to perform actions on them.

Name	Last Modified
GCS	a day ago
Local Disk	seconds ago



Name	Last Modified
bin	
boot	
dev	
etc	
hadoop	
home	17 minutes ago
lib	25 days ago
lib32	25 days ago
lib64	25 days ago
libx32	25 days ago
lost+found	25 days ago
media	25 days ago
mnt	25 days ago

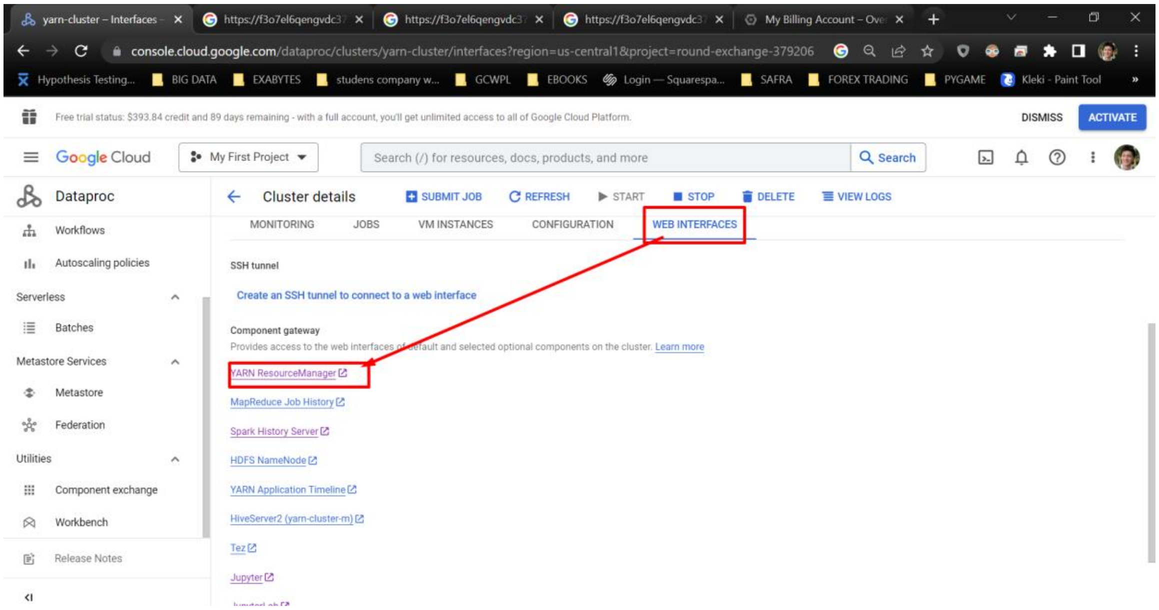

round-exchange-379206 > yarn-cluster
Sign out
jupyter Untitled Last Checkpoint: 9 minutes ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
PySpark
Run
Code
nbdiff

```
In [1]: spark.version
Out[1]: '3.1.3'

In [ ]:
```

the notebook is not like a shell which immediately establishes the driver and executors
it is just a web interface which is not yet connected to the cluster

we must run some spark commands to initiate the connection

1. YARN RESOURCE MANAGER



Free trial status: $393.84 credit and 89 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

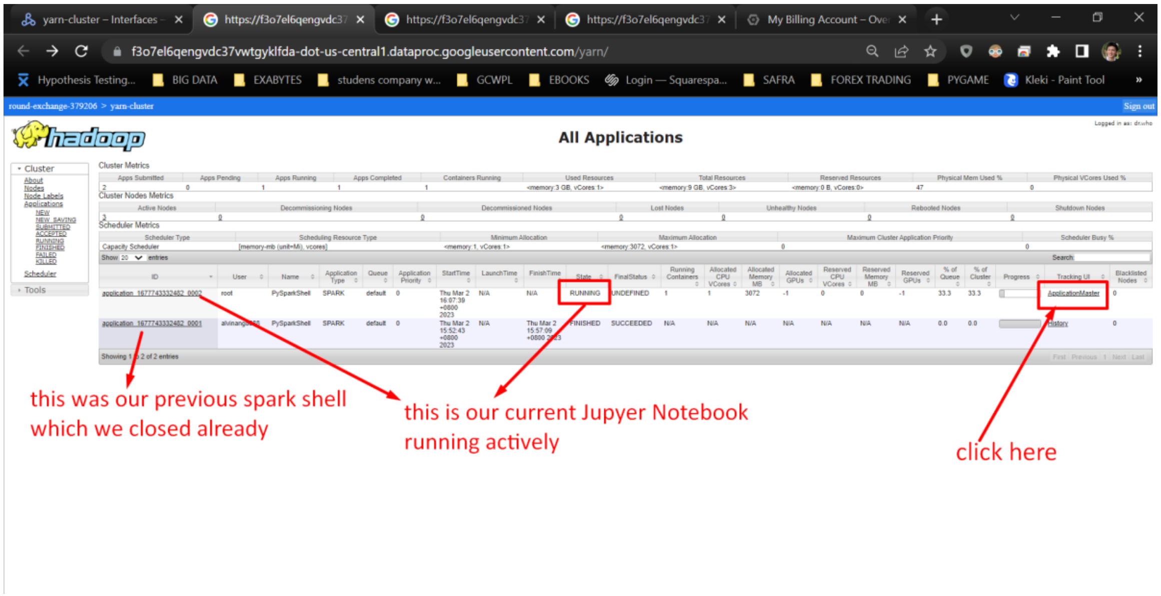
SSH tunnel

Create an SSH tunnel to connect to a web interface

Component gateway

Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

- YARN ResourceManager
- MapReduce Job History
- Spark History Server
- HDFS NameNode
- YARN Application Timeline
- HiveServer2 (yarn-cluster-m)
- Tez
- Jupyter



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
2	0	1	1	1	<memory:3 GB, vCores:1>	<memory:9 GB, vCores:3>	<memory:0 B, vCores:0>	47	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
3	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority	Scheduler Busy %
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1, vCores:1>	<memory:3072, vCores:1>	0	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1677743332482_0002	root	PySparkShell	SPARK	default	0	Thu Mar 2 16:07:39 +0800 2023	N/A	N/A	RUNNING	UNDEFINED	1	1	3072	-1	0	0	-1	33.3	33.3		ApplicationMaster	0
application_1677743332482_0001	alvinang[illegible]	PySparkShell	SPARK	default	0	Thu Mar 2 15:52:43 +0800 2023	N/A	Thu Mar 2 15:57:09 +0800 2023	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.0	0.0		History	0

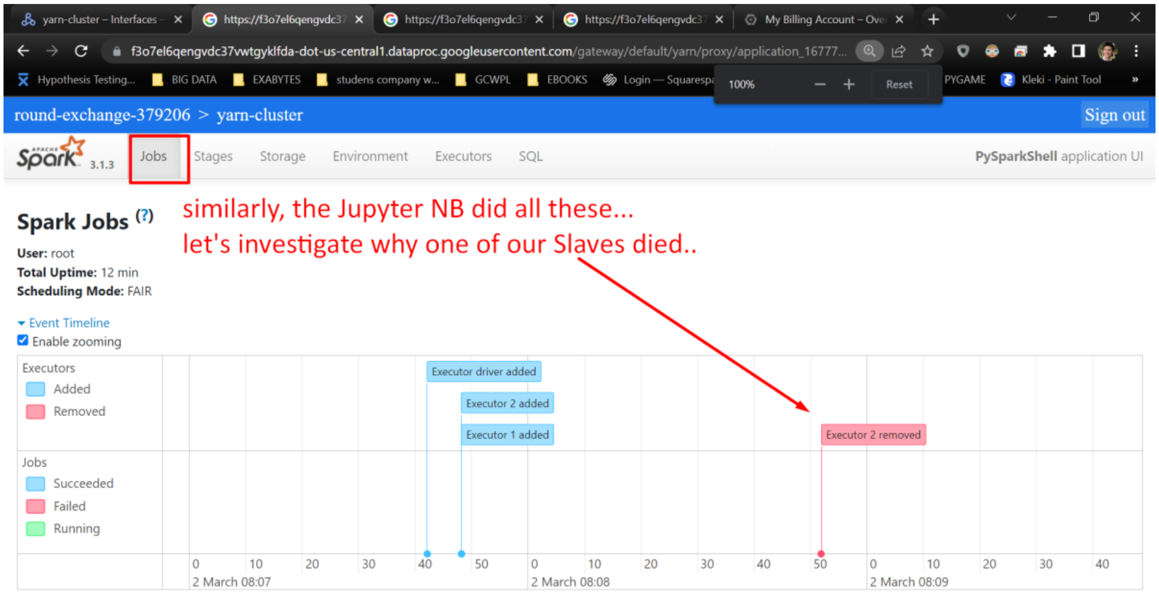
Showing 1 to 2 of 2 entries First Previous 1 Next Last

this was our previous spark shell which we closed already

this is our current Jupyer Notebook running actively

click here

2. CHECK PREVIOUS JOBS

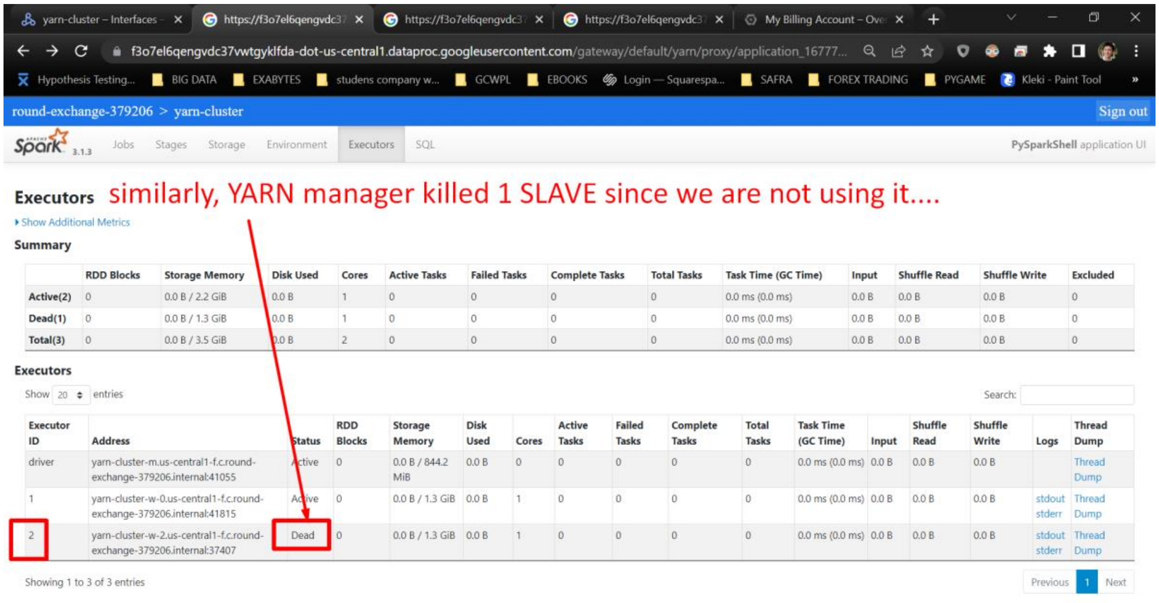


similarly, the Jupyter NB did all these...
let's investigate why one of our Slaves died..

Spark Jobs (?)

User: root
Total Uptime: 12 min
Scheduling Mode: FAIR

3. CHECK EXECUTORS



Executors similarly, YARN manager killed 1 SLAVE since we are not using it....

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(2)	0	0.0 B / 2.2 GiB	0.0 B	1	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Dead(1)	0	0.0 B / 1.3 GiB	0.0 B	1	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(3)	0	0.0 B / 3.5 GiB	0.0 B	2	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0

Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Thread Dump
driver	yarn-cluster-m.us-central1-f.c.round-exchange-379206.internal:41055	Active	0	0.0 B / 844.2 MiB	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B		Thread Dump
1	yarn-cluster-w-0.us-central1-f.c.round-exchange-379206.internal:41815	Active	0	0.0 B / 1.3 GiB	0.0 B	1	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr	Thread Dump
2	yarn-cluster-w-2.us-central1-f.c.round-exchange-379206.internal:37407	Dead	0	0.0 B / 1.3 GiB	0.0 B	1	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr	Thread Dump

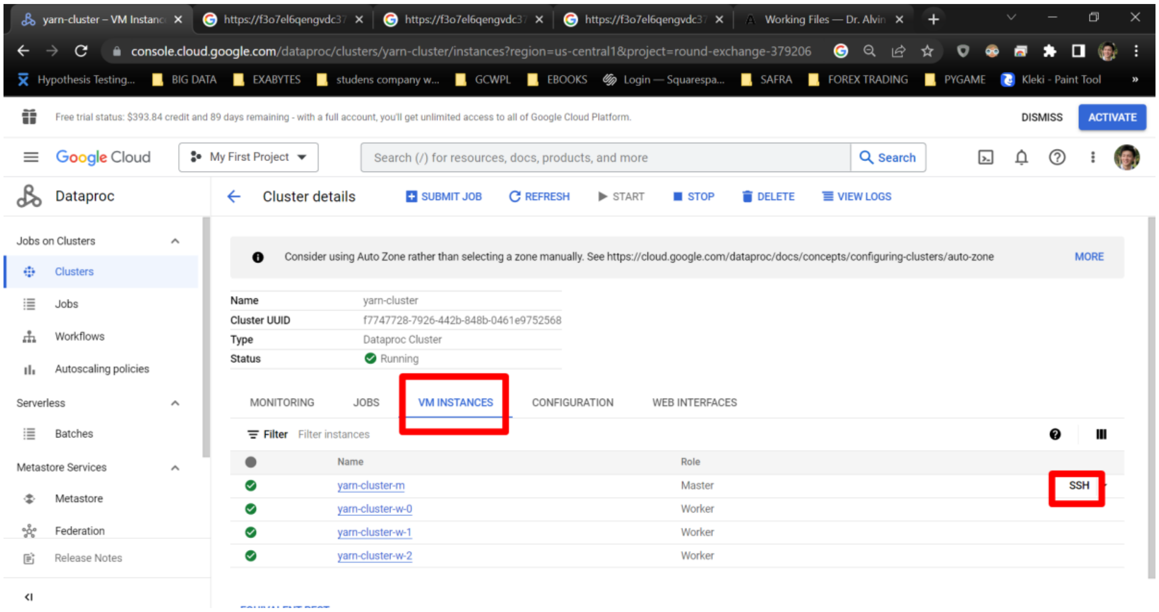
Showing 1 to 3 of 3 entries

VIII. SUBMITTING A JOB TO THE CLUSTER

https://www.alvinang.sg/s/wordcount.py

https://www.alvinang.sg/s/pi.py

A. WE WILL TRY SUBMITTING AN APPLICATION TO THE CLUSTER NOW



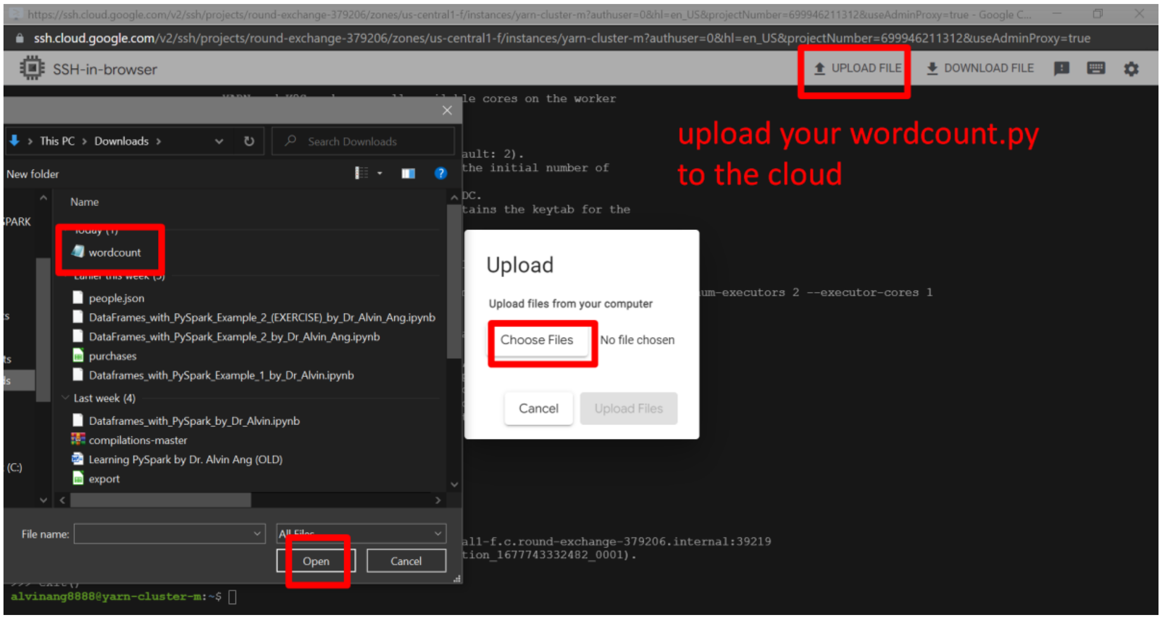
Free trial status: $393.84 credit and 89 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Consider using Auto Zone rather than selecting a zone manually. See https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone

Name	yarn-cluster
Cluster UUID	f7747728-7926-442b-848b-0461e9752568
Type	Dataproc Cluster
Status	✓ Running

●	Name	Role
✓	yarn-cluster-m	Master
✓	yarn-cluster-w-0	Worker
✓	yarn-cluster-w-1	Worker
✓	yarn-cluster-w-2	Worker

B. UPLOAD YOUR .PY FILE



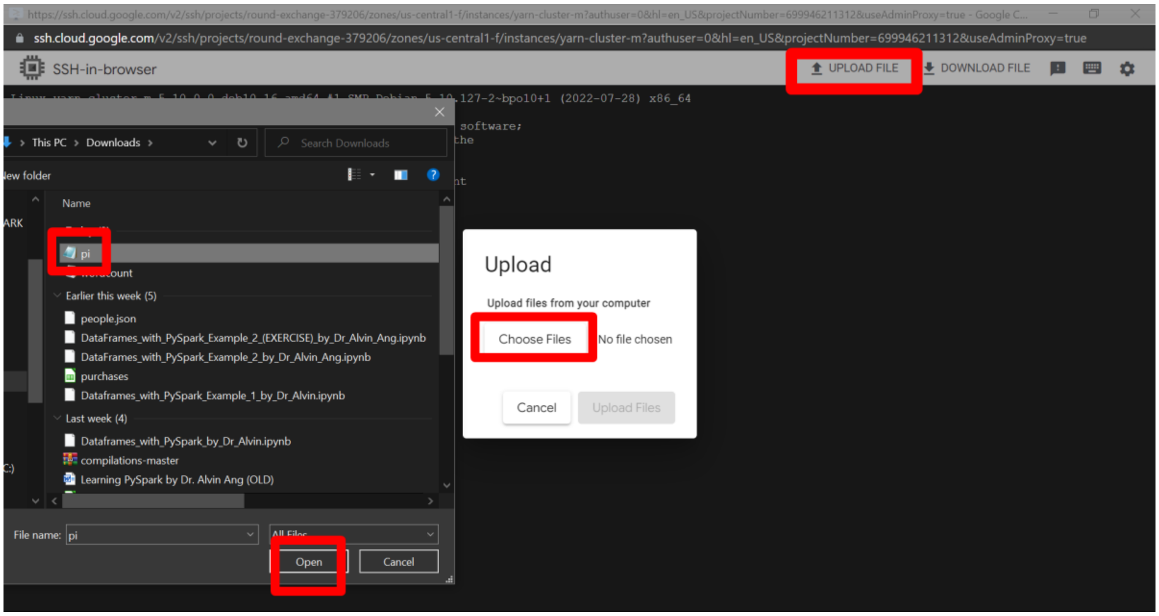
upload your wordcount.py to the cloud

```
https://ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true
```

```
le cores on the worker
ault: 2).
the initial number of
DC.
tains the keytab for the
um-executors 2 --executor-cores 1
all-f.c.round-exchange-379206.internal:39219
tion_1677743332482_0001).
alvinang8888@yarn-cluster-m:~$
```

SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
This PC › Downloads ›
Search Downloads
New folder
Name
wordcount
people.json
DataFrames_with_PySpark_Example_2_(EXERCISE)_by_Dr_Alvin_Ang.ipynb
DataFrames_with_PySpark_Example_2_by_Dr_Alvin_Ang.ipynb
purchases
Dataframes_with_PySpark_Example_1_by_Dr_Alvin.ipynb
Last week (4)
Dataframes_with_PySpark_by_Dr_Alvin.ipynb
compilations-master
Learning PySpark by Dr. Alvin Ang (OLD)
export
File name:
All Files
Open
Cancel
Upload
Upload files from your computer
Choose Files
No file chosen
Cancel
Upload Files

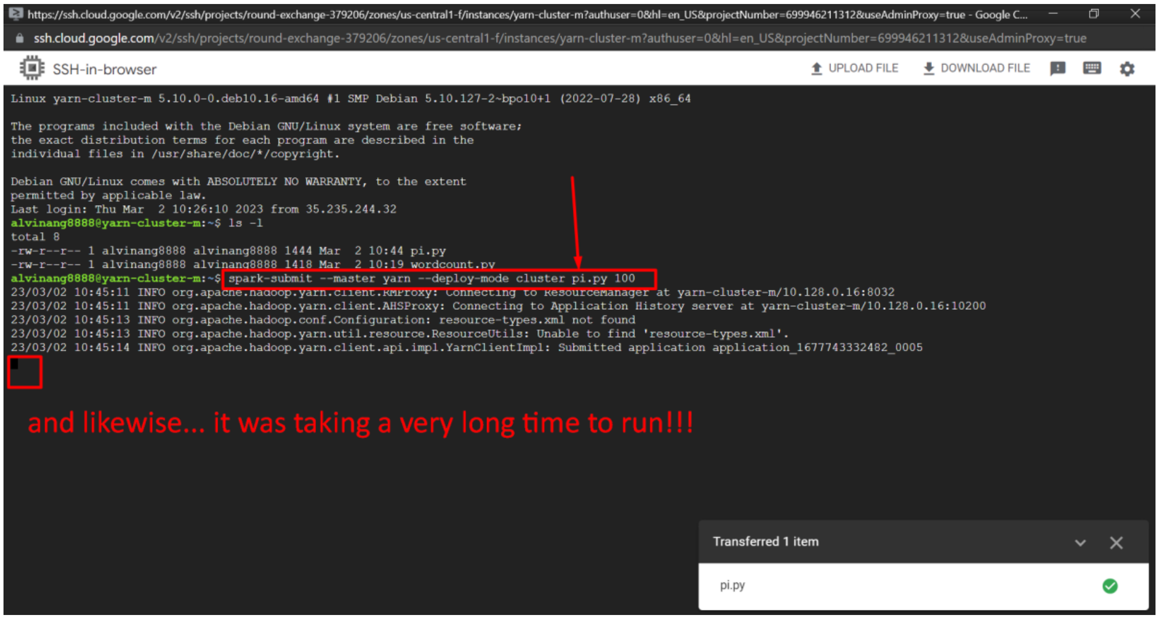
I was trying out wordcount.py… but it seems to get stuck forever… so I decided to try using pi.py…



```
.127-2~bpo10+1 (2022-07-28) x86_64
software;
the
```

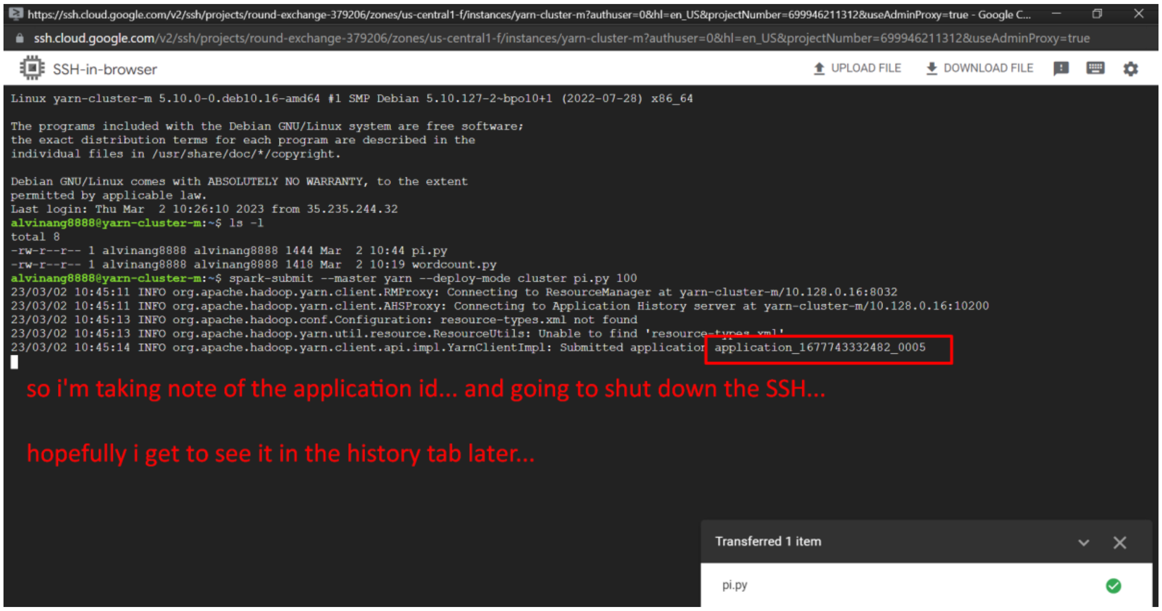
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
This PC › Downloads ›
Search Downloads
New folder
Name
pi
wordcount
Earlier this week (5)
people.json
DataFrames_with_PySpark_Example_2_(EXERCISE)_by_Dr_Alvin_Ang.ipynb
DataFrames_with_PySpark_Example_2_by_Dr_Alvin_Ang.ipynb
purchases
Dataframes_with_PySpark_Example_1_by_Dr_Alvin.ipynb
Last week (4)
Dataframes_with_PySpark_by_Dr_Alvin.ipynb
compilations-master
Learning PySpark by Dr. Alvin Ang (OLD)
File name: pi
All Files
Open
Cancel
Upload
Upload files from your computer
Choose Files
No file chosen
Cancel
Upload Files

C. RUN THE CODE (SPARK-SUBMIT)


https://ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true - Google C...
ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Transferred 1 item
pi.py

```
Linux yarn-cluster-m 5.10.0-0.deb10.16-amd64 #1 SMP Debian 5.10.127-2~bpo10+1 (2022-07-28) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Thu Mar  2 10:26:10 2023 from 35.235.244.32
alvinang8888@yarn-cluster-m:~$ ls -l
total 8
-rw-r--r-- 1 alvinang8888 alvinang8888 1444 Mar  2 10:44 pi.py
-rw-r--r-- 1 alvinang8888 alvinang8888 1418 Mar  2 10:19 wordcount.py
alvinang8888@yarn-cluster-m:~$ spark-submit --master yarn --deploy-mode cluster pi.py 100
23/03/02 10:45:11 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at yarn-cluster-m/10.128.0.16:8032
23/03/02 10:45:11 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at yarn-cluster-m/10.128.0.16:10200
23/03/02 10:45:13 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
23/03/02 10:45:13 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/03/02 10:45:14 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1677743332482_0005
```

and likewise... it was taking a very long time to run!!!

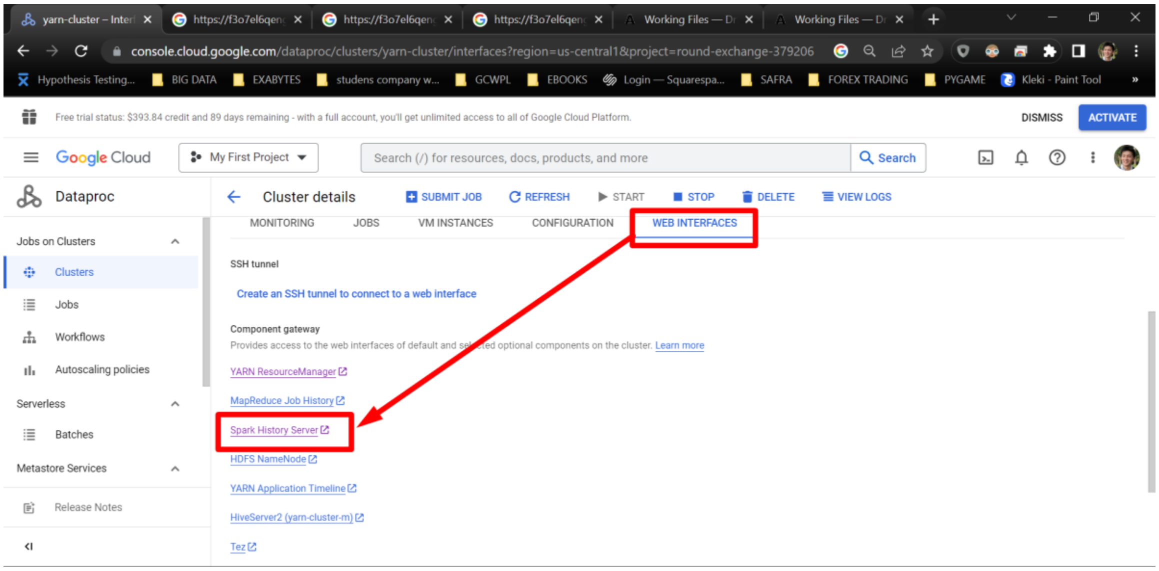

https://ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true - Google C...
ssh.cloud.google.com/v2/ssh/projects/round-exchange-379206/zones/us-central1-f/instances/yarn-cluster-m?authuser=0&hl=en_US&projectNumber=699946211312&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
Transferred 1 item
pi.py

```
Linux yarn-cluster-m 5.10.0-0.deb10.16-amd64 #1 SMP Debian 5.10.127-2~bpo10+1 (2022-07-28) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Thu Mar  2 10:26:10 2023 from 35.235.244.32
alvinang8888@yarn-cluster-m:~$ ls -l
total 8
-rw-r--r-- 1 alvinang8888 alvinang8888 1444 Mar  2 10:44 pi.py
-rw-r--r-- 1 alvinang8888 alvinang8888 1418 Mar  2 10:19 wordcount.py
alvinang8888@yarn-cluster-m:~$ spark-submit --master yarn --deploy-mode cluster pi.py 100
23/03/02 10:45:11 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at yarn-cluster-m/10.128.0.16:8032
23/03/02 10:45:11 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at yarn-cluster-m/10.128.0.16:10200
23/03/02 10:45:13 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
23/03/02 10:45:13 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/03/02 10:45:14 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1677743332482_0005
```

so i'm taking note of the application id... and going to shut down the SSH...

hopefully i get to see it in the history tab later...

D. CHECK OUT THE SPARK HISTORY…



Free trial status: $393.84 credit and 89 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

SSH tunnel

Create an SSH tunnel to connect to a web interface

Component gateway

Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

YARN ResourceManager

MapReduce Job History

Spark History Server

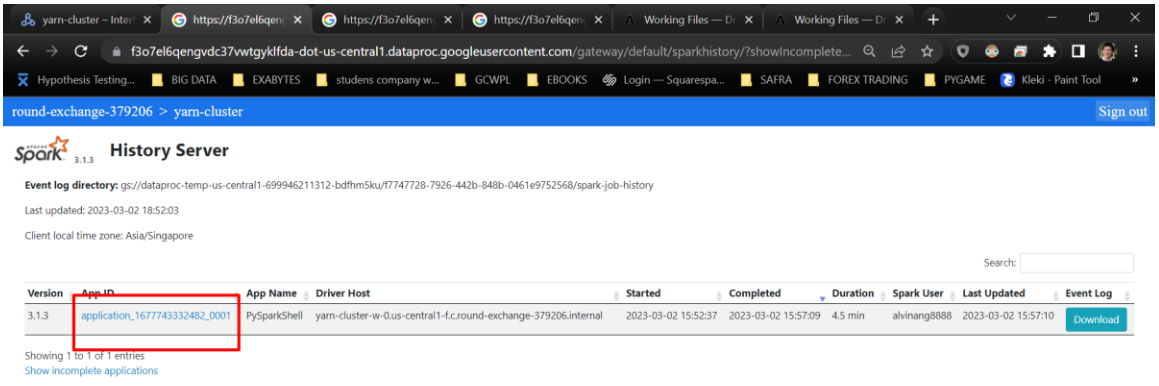
HDFS NameNode

YARN Application Timeline

HiveServer2 (yarn-cluster-m)

Tez

E. FAILED…..SIGH….



History Server

Event log directory: gs://dataproc-temp-us-central1-699946211312-bdfhm5ku/f7747728-7926-442b-848b-0461e9752568/spark-job-history

Last updated: 2023-03-02 18:52:03

Client local time zone: Asia/Singapore

Search:

Version	App ID	App Name	Driver Host	Started	Completed	Duration	Spark User	Last Updated	Event Log
3.1.3	application_1677743332482_0001	PySparkShell	yarn-cluster-w-0.us-central1-f.c.round-exchange-379206.internal	2023-03-02 15:52:37	2023-03-02 15:57:09	4.5 min	alvinang8888	2023-03-02 15:57:10	Download

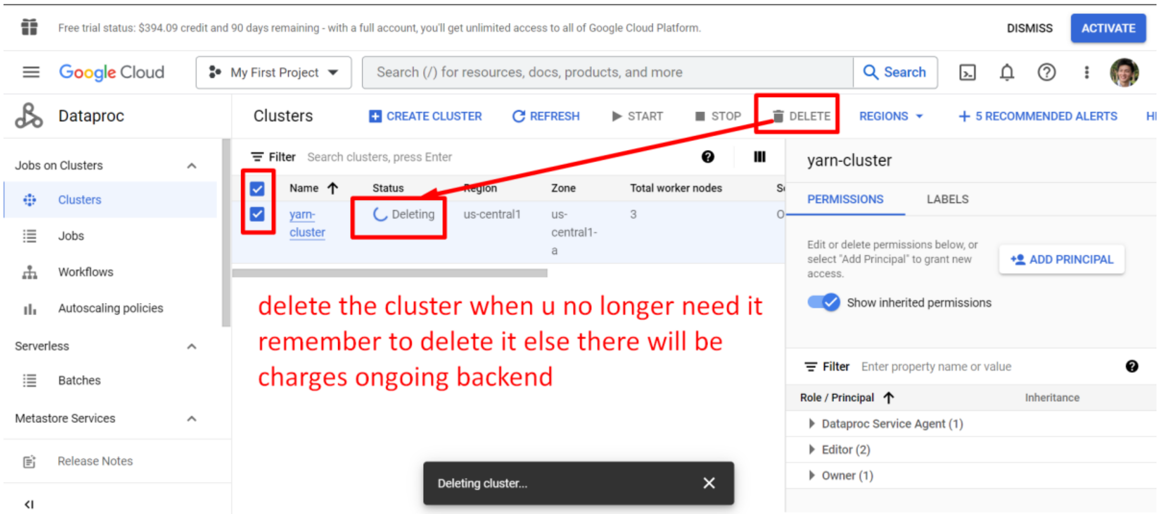
Showing 1 to 1 of 1 entries
Show incomplete applications

it just shows me back the old 'app ID' which i did earlier on.. its supposed to show me all the new 'app IDs' like 0005 etc...

perhaps i need to delete and restart my entire cluster and run everything again and hopefully it will work smoothly...

IX. DELETING YOUR CLUSTER

A. TO PREVENT OVERCHARGING



Free trial status: $394.09 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

☑	Name	Status	Region	Zone	Total worker nodes
☑	yarn-cluster	Deleting	us-central1	us-central1-a	3

Edit or delete permissions below, or select "Add Principal" to grant new access.

☑ Show inherited permissions

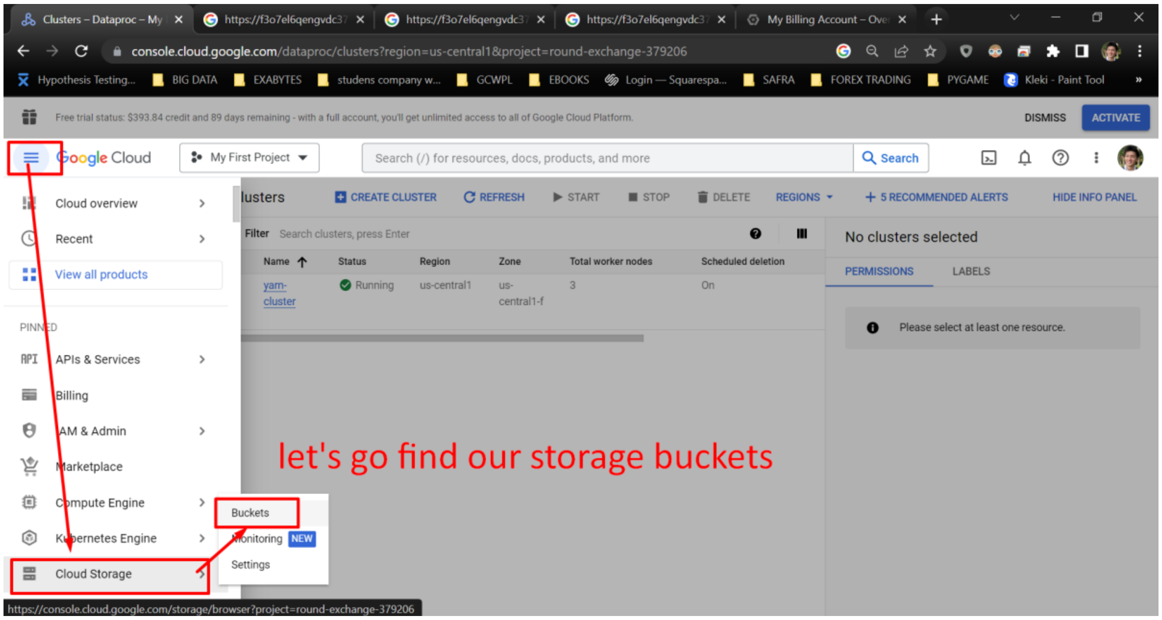
Role / Principal	Inheritance
Dataproc Service Agent (1)	
Editor (2)	
Owner (1)	

delete the cluster when u no longer need it remember to delete it else there will be charges ongoing backend

Deleting cluster...

This process takes super super long…..

B. DELETE YOUR STORAGE BUCKET TOO



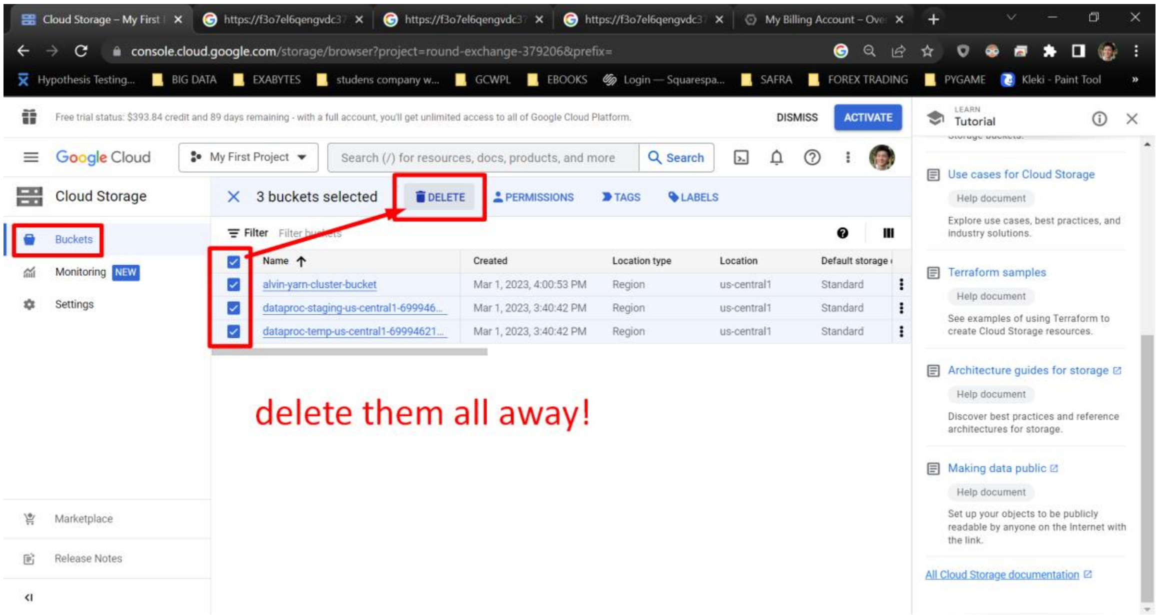
Free trial status: $393.84 credit and 89 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.

Name	Status	Region	Zone	Total worker nodes	Scheduled deletion
yarn-cluster	Running	us-central1	us-central1-f	3	On

No clusters selected

Please select at least one resource.

let's go find our storage buckets


console.cloud.google.com/storage/browser?project=round-exchange-379206&prefix=
Google Cloud
My First Project
Cloud Storage
Buckets
Monitoring NEW
Settings
Marketplace
Release Notes
DELETE
PERMISSIONS
TAGS
LABELS
Tutorial
Use cases for Cloud Storage
Terraform samples
Architecture guides for storage
Making data public
All Cloud Storage documentation

3 buckets selected

	Name	Created	Location type	Location	Default storage c
☑	alvin-yarn-cluster-bucket	Mar 1, 2023, 4:00:53 PM	Region	us-central1	Standard
☑	dataproc-staging-us-central1-699946...	Mar 1, 2023, 3:40:42 PM	Region	us-central1	Standard
☑	dataproc-temp-us-central1-69994621...	Mar 1, 2023, 3:40:42 PM	Region	us-central1	Standard

delete them all away!

ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.