California State University, Fresno

CSCI 191T

Assignment 4

GCP and Dataproc for Hadoop/Spark

Problem Introduction

Cloud Dataproc is a managed Spark and Hadoop service that lets you take advantage of open source data tools for batch processing, querying, streaming, and machine learning. Cloud Dataproc automation facilitates cluster creation, cluster management, and saves money by turning clusters off when not need.

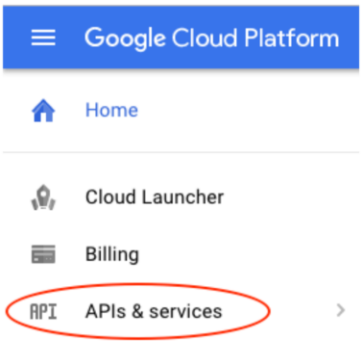
Objectives:

- How to create a managed Cloud Dataproc cluster (with Apache Spark pre-installed)
- How to ssh into the master node of a Dataproc cluster
- How to submit a Spark job
- How to resize a cluster
- How to use gcloud to examine clusters, jobs, and firewall rules
- How to shut down your cluster

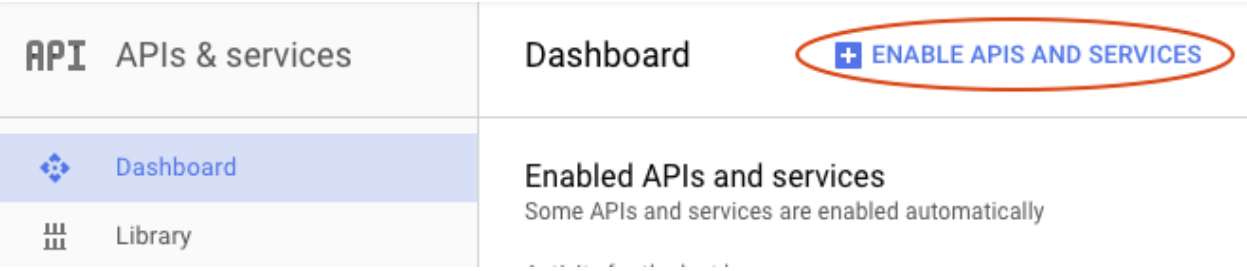
Part 1: Cloud Dataproc and Google Compute Engine APIs

Click on the menu icon in the top left of the screen in GCP.

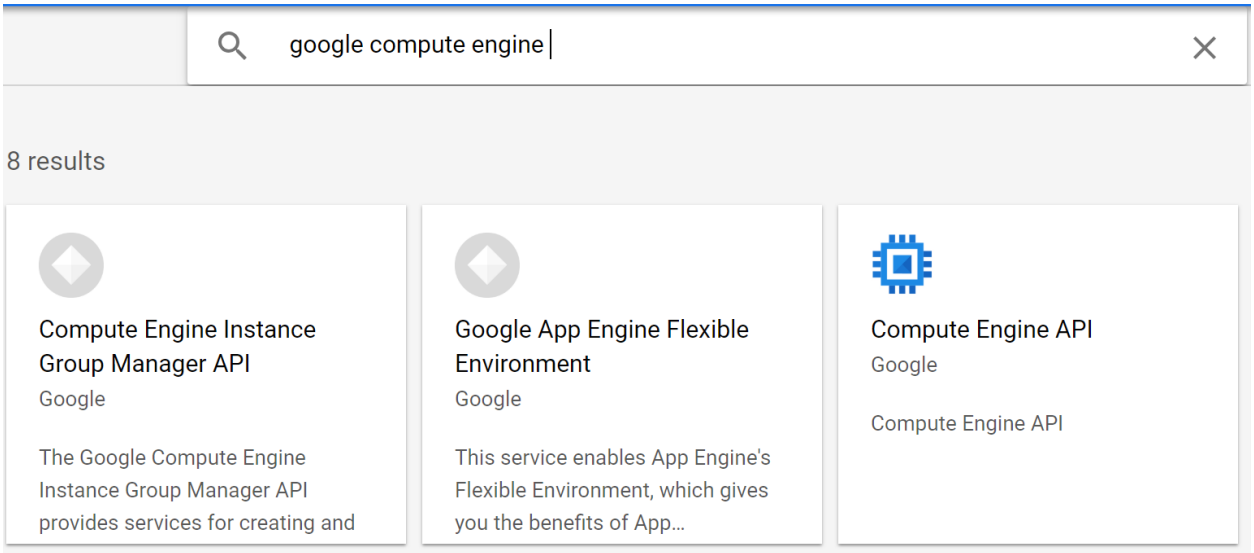
Select API Manager from the drop down.



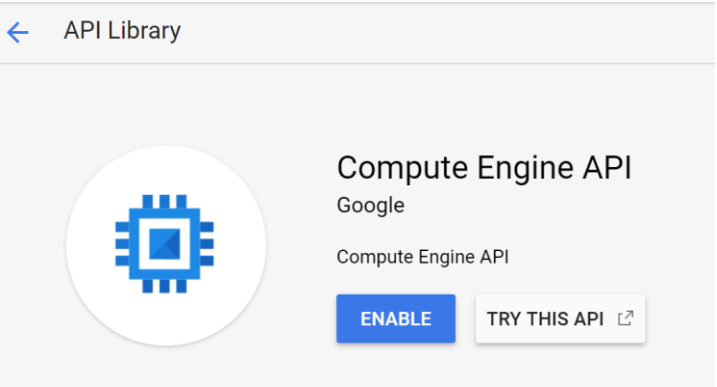
Click on Enable APIs and Services.



Search for "Compute Engine" in the search box. Click on "Compute Engine API" in the results list that appears.



Open the Google Compute Engine page, next click Enable.

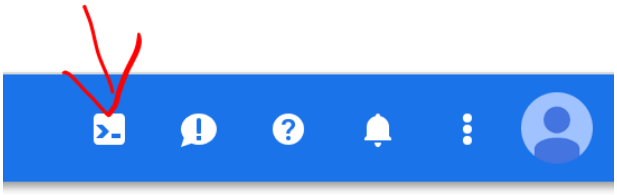


Once it has enabled, click the arrow pointing left to go back. Now search for "Google Cloud Dataproc API" and enable it as well.

Part 2: Create a cluster

Create a cluster(use class PPTs for guid) Your default virtual machine is a Debian-based virtual machine loaded with various development tools. It offers a persistent home directory, and runs on the Google Cloud, greatly enhancing network performance and authentication.

You can either use command line(we already installed cloud SDK) on your machine or activate Google Cloud Shell, from the developer console, click the button on the top right-hand side (it should only take a few moments to provision and connect to the environment):



Then accept the terms of service and click the "Start Cloud Shell" link: Once connected to the cloud shell, you should see that you are already authenticated and that the project is already set to your PROJECT_ID :

$gcloud auth list

Command output Credentialed accounts:

- <myaccount>@<mydomain>.com (active)

Note: gcloud is the powerful and unified command-line tool for Google Cloud Platform. Full documentation is available from https://cloud.google.com/sdk/gcloud. It comes pre-installed on CloudShell and has support for tab-completion. gcloud config list project Command output [core] project = <PROJECT_ID> If for some reason the project is not set, simply issue the following command :

gcloud config set project <PROJECT_ID>

Looking for your PROJECT_ID? Check out what ID you used in the setup steps or look it up in the console dashboard: IMPORTANT: Finally, set the default zone and project configuration:

gcloud config set compute/zone us-west1-a

You can choose a variety of different zones. Learn more in the Regions & Zones documentation. Note: When you run gcloud on your own machine, the config settings would've been persisted across sessions. But in Cloud Shell, you will need to set this for every new session or reconnection.Create a Cloud Dataproc cluster

Part 3: Upload files and run Spark

As explained in class slides, upload your names.csv and names.json files. run spark and show the lines from the files you uploaded.

Create internal and external Hive tables using spark. Show the created tables and the elements inside them.

Part 4: Resize Cluster

For running larger computations, you might want to add more nodes to your cluster to speed it up. Dataproc lets you add nodes to and remove nodes from your cluster at any time. Examine the cluster configuration:

$ gcloud dataproc clusters describe ${CLUSTERNAME}

Make the cluster larger by adding some preemptible nodes:

$ gcloud dataproc clusters update ${CLUSTERNAME} --num-preemptible-workers=2

Examine the cluster again:

$ gcloud dataproc clusters describe ${CLUSTERNAME}

Note that in addition to the workerConfig from the original cluster description, there is now also a secondaryWorkerConfig that includes two instanceNames for the preemptible workers. Dataproc shows the cluster status as being ready while the new nodes are booting. Since you started with two nodes and now have four, your Spark jobs should run about twice as fast.

Part 5: SSH into Cluster

Connect via ssh to the master node, whose instance name is always the cluster name with -m appended:

$ gcloud compute ssh ${CLUSTERNAME}-m --zone=us-west1-a

The first time you run an ssh command on Cloud Shell it will generate ssh keys for your account there. You can choose a passphrase, or use a blank passphrase for now and change it later using ssh-keygen if you want. On the instance, check the hostname:

$ hostname Because you specified --scopes=cloud-platform

when you created the cluster, you can run gcloud commands on your cluster. List the clusters in your project:

$ gcloud dataproc clusters list

Log out of the ssh connection when you are done:

$ logout

Part 6: Examine tags

When you created your cluster you included a --tags option to add a tag to each node in the cluster. Tags are used to attach firewall rules to each node. You did not create any matching firewall rules in this lab, but you can still examine the tags on a node and the firewall rules on the network. Print the description of the master node:

$ gcloud compute instances describe ${CLUSTERNAME}-m --zone us-west1-a

Print the firewall rules:

$ gcloud compute firewall-rules list

Note the SRC_TAGS and TARGET_TAGS columns. By attaching a tag to a firewall rule, you can specify that it should be used on all nodes that have that tag.

Your report should include the followings:

1. Statement of objectives
2. The screenshots of completed task
3. A summary of what you learnt
4. Any difficulties you have faced and how did you overcome the issues

Assignment Rubric

Criteria	Pts
Organization Formatting of the sections/subsections (fonts, font sizes, indentation, paragraphization) Order/flow of the report Proper details of each section	1.0 pts
Statement of Objectives (a minimum of 100 words) Introduction, background, and justification	1.0 pts
Tasks Performing all tasks and including the screenshots	5.0 pts
Conclusion (a minimum of 100 words) Comprehensiveness and completeness of summary	1.0 pts
Encountered Problems Detailed reflection on issues encountered	1.0 pts
Procedure Detailed steps of the procedure	1.0 pts