

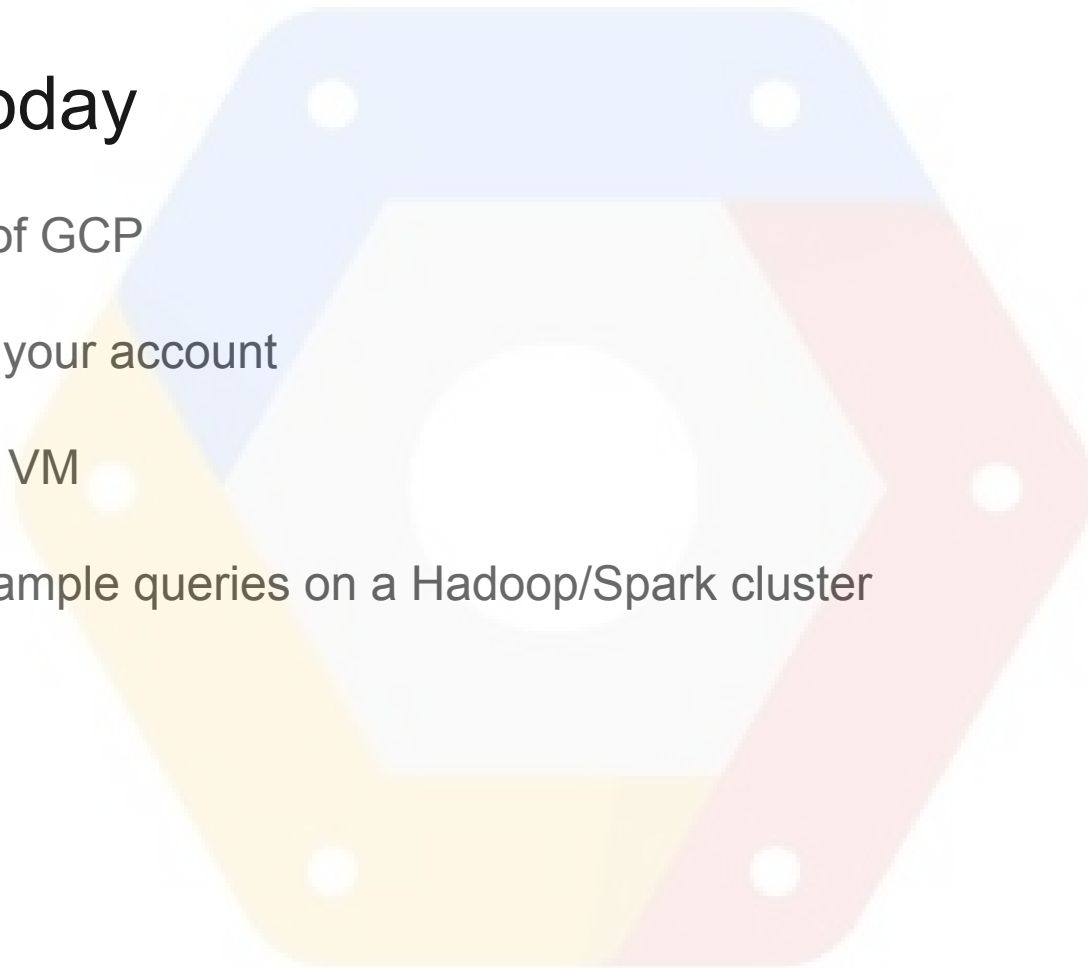
The background features the Google Cloud Platform logo, which is a stylized hexagon composed of three interlocking shapes in blue, yellow, and red. The hexagon has rounded corners and four small white circles, one on each side.

# Google Cloud Platform

CS341

# Plan for today

- Overview of GCP
- Setting up your account
- Creating a VM
- Running sample queries on a Hadoop/Spark cluster



# What is Google Cloud Platform?

Google's cloud computing service (using same infrastructure used by Google for products like search). Relevant for this class:

<b>Compute Engine</b>	Virtual Machines
<b>Storage Services</b>	Relational and NoSQL cloud storage
<b>Data Services</b>	Hadoop/Spark clusters, cloud ML service, APIs for natural language, vision, speech

Full list of products: <https://cloud.google.com/products/>

# Setup: Create account and set up billing

1. Login with your Google account (**NOT** stanford.edu account).
2. Visit <https://console.cloud.google.com/education> and enter the coupon code.
3. Click “Accept and Continue”

### Education grants

Please enter the coupon code provided to you via the Google Cloud Platform Education Grants program to receive credit for Google Cloud Platform. Get what you need to build and run your apps, websites and services.

**Coupon code**

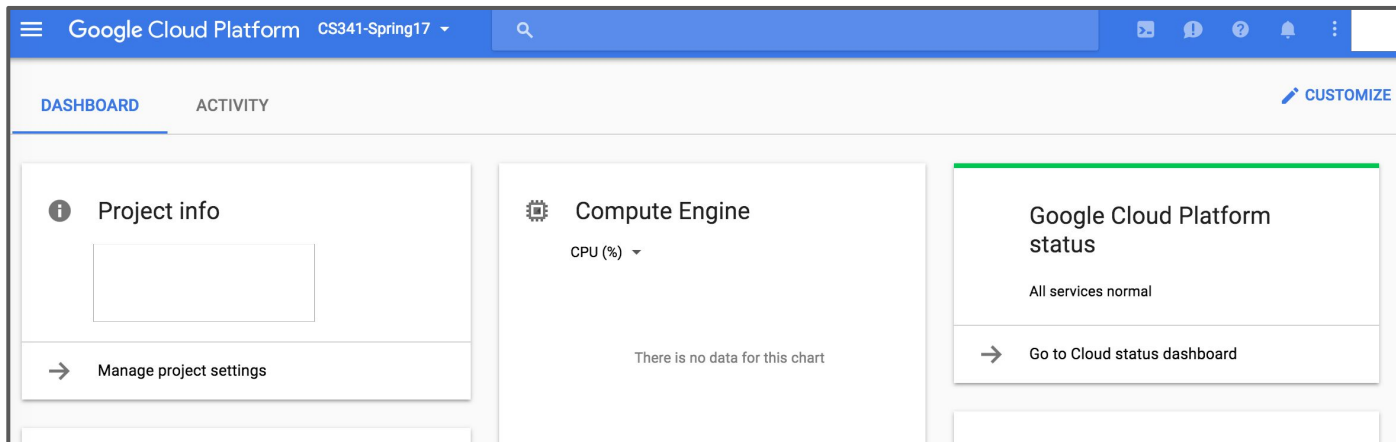
**Please email me updates regarding feature announcements, performance suggestions, feedback surveys and special offers.**

Yes  No

**Google Cloud Platform education grants credits terms and conditions**

# Setup: Create a project

1. Visit <https://console.cloud.google.com>
2. Click on “Create a Project” and complete the flow. Billing should be set up automatically to use the EDU credits
3. Go to “IAM” from main menu, add rest of team members (using Google accounts, **NOT** stanford.edu account)



# Interacting with Google Cloud Platform

Broadly you can interact with GCP in two ways:

1. Graphical UI (<https://console.cloud.google.com/>): Useful to create VMs, set up clusters, provision resources, manage teams etc
2. Command line (gcloud sdk tools): Useful for using the resources once provisioned. E.g. ssh into instances, submit jobs, copy files etc

# Setup: Command line tools

1. Make sure you have Python 2.7.9 or higher
2. Download SDK: <https://cloud.google.com/sdk/docs/>
3. Install: run `./install.sh` and follow the installation steps
4. Authorize using your credentials: Run `./bin/gcloud init`
5. Test: `gcloud components list`, `gcloud auth list`

# Setup: Command line tools

```
nihit@nihit-lp1:~/Documents$ gcloud components list

Your current Cloud SDK version is: 149.0.0
The latest available version is: 149.0.0
```

Status	Name	ID	Size
Not Installed	App Engine Go Extensions	app-engine-go	47.7 MiB
Not Installed	Cloud Datastore Emulator	cloud-datastore-emulator	15.4 MiB
Not Installed	Cloud Datastore Emulator (Legacy)	gcd-emulator	38.1 MiB
Not Installed	Cloud Pub/Sub Emulator	pubsub-emulator	21.0 MiB
Not Installed	Emulator Reverse Proxy	emulator-reverse-proxy	56.8 MiB
Not Installed	Google Container Registry's Docker credential helper	docker-credential-gcr	3.4 MiB
Not Installed	gcloud app Java Extensions	app-engine-java	128.6 MiB
Not Installed	gcloud app PHP Extensions (Mac OS X)	app-engine-php-darwin	21.9 MiB
Not Installed	gcloud app Python Extensions	app-engine-python	6.1 MiB
Not Installed	kubectl	kubectl	11.4 MiB
Installed	BigQuery Command Line Tool	bq	< 1 MiB
Installed	Bigtable Command Line Tool	cbt	3.9 MiB
Installed	Cloud Datalab Command Line Tool	datalab	< 1 MiB
Installed	Cloud SDK Core Libraries	core	5.8 MiB
Installed	Cloud Storage Command Line Tool	gsutil	2.8 MiB
Installed	Default set of gcloud commands	gcloud	
Installed	gcloud Alpha Commands	alpha	< 1 MiB
Installed	gcloud Beta Commands	beta	< 1 MiB



# Configure and use a VM

1. Visit <https://console.cloud.google.com/compute/instances>.
2. Click on the “Create Instance” button.
3. Configure instance name, zone, machine type, network traffic, etc.
4. Congrats, your VM has been created! Use “View gcloud command” and copy the message in the pop-up dialog to your bash shell.

(something like: `gcloud compute --project "yourProjectID" ssh --zone "yourInstanceZone" "yourInstanceName"`)

<input type="checkbox"/>	Name ^	Zone	Recommendation	Internal IP	External IP	Connect
<input type="checkbox"/>	<input checked="" type="checkbox"/> instance-1	us-west1-a		10.138.0.2	<a href="#">35.185.216.114</a> ↗	SSH ▾ ⋮ Open in browser window Open in browser window on custom port <b>View gcloud command</b> Use another SSH client

# Configure and use a VM (Cont'd)

5. Stop your machine when not in use to avoid unexpected charges.
6. For more details, see <https://cloud.google.com/compute/docs/quickstart-linux>.

*FAQ: My bash shell is complaining gcloud command not found. :( Reload your bash\_profile using the "source" command, OR simply restart your bash shell.*

# Create a Cluster

## 1. Two ways to create a cluster:

Use command line (easier): `gcloud dataproc clusters create <cluster-name>`

**OR** Use GUI: visit <https://console.cloud.google.com/dataproc/clusters>.

## 2. View your clusters: <https://console.cloud.google.com/dataproc/clusters>.

*Clusters:*

<input type="checkbox"/> Name ^	Zone	Total worker nodes	Cloud Storage staging bucket	Created	Status
<input checked="" type="checkbox"/> cluster-1	us-central1-a	2	dataproc-a60e0265-f815-44e1-83e2-8b7284431f9e-us	Apr 4, 2017, 11:34:03 PM	Running

*Instances: 1 master node and 2 worker nodes have been created*

<input type="checkbox"/> Name ^	Zone	Recommendation	Internal IP	External IP	Connect
<input checked="" type="checkbox"/> cluster-1-m	us-central1-a		10.128.0.2	104.198.52.60	SSH ▾ ⋮
<input checked="" type="checkbox"/> cluster-1-w-0	us-central1-a		10.128.0.4	35.184.84.218	SSH ▾ ⋮
<input checked="" type="checkbox"/> cluster-1-w-1	us-central1-a		10.128.0.3	104.154.182.220	SSH ▾ ⋮

# Submit a Job

## 1. Create your job.

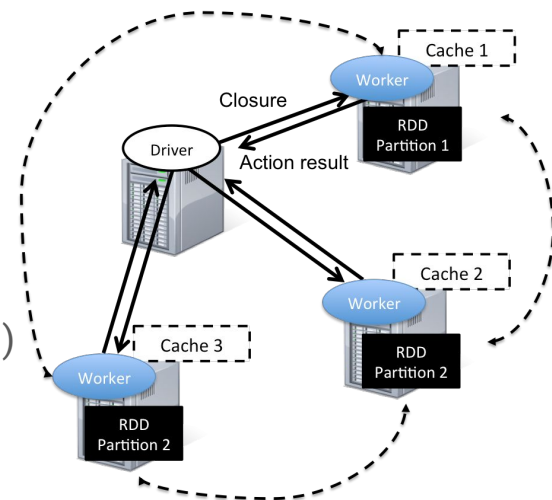
*Simple example: add one to every element in an array.*

```
import pyspark
sc = pyspark.SparkContext()
original_array_rdd = sc.parallelize([3,2,5,1,4])
new_array_rdd = original_array_rdd.map(lambda x: x+1)
new_array = sorted(new_array_rdd.collect())
print new_array
```

## 2. Submit your job:

```
gcloud dataproc jobs submit pyspark --cluster
<my-dataproc-cluster> my-first-job.py
```

## 3. View your jobs: <https://console.cloud.google.com/dataproc/jobs>.



Data shuffling across machines  
(wide dependencies)

# Attach a Disk to Your VM

## 1. **Create your blank disk.**

(1) VM instances -> click on your instance -> “Edit” button at the top -> additional disks -> “Add item” button.

(2) Select “Name” dropdown -> Create disk -> Source type: select “blank disk” -> configure whatever nickname and size to your disk.

## 2. **Format and mount your disk.** [live demo]

## 3. **Every time you reboot, you need to mount your disk again:**

```
sudo mount -o discard,defaults /dev/[DEVICE_ID] /mnt/disks/[MNT_DIR]
```

## 4. For more details, see

<https://cloud.google.com/compute/docs/disks/add-persistent-disk>

# Storage Solutions for Clusters

1. You can choose to use

(1) cloud storage

(2) share a persistent disk among your cluster

(3) Other solutions depending on your needs

This page offers detailed explanation

<https://cloud.google.com/solutions/filers-on-compute-engine#cloud-storage>.

2. To set up **cloud storage**, see tutorial on

<https://cloud.google.com/compute/docs/disks/gcs-buckets>.

3. To **share a persistent disk** among all machines in your cluster, see tutorial on

[https://cloud.google.com/compute/docs/disks/add-persistent-disk#use\\_multi\\_instances](https://cloud.google.com/compute/docs/disks/add-persistent-disk#use_multi_instances).

# Other services that might be useful

- Natural Language: <https://cloud.google.com/natural-language/>
- BigQuery: <https://cloud.google.com/bigquery>
- DataPrep: <https://cloud.google.com/dataprep/>
- DataProc: <https://cloud.google.com/dataproc/>
- Cloud ML Engine: <https://cloud.google.com/ml-engine/>