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:

<table>
<thead>
<tr>
<th>Compute Engine</th>
<th>Virtual Machines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage Services</td>
<td>Relational and NoSQL cloud storage</td>
</tr>
<tr>
<td>Data Services</td>
<td>Hadoop/Spark clusters, cloud ML service, APIs</td>
</tr>
<tr>
<td></td>
<td>for natural language, vision, speech</td>
</tr>
</tbody>
</table>

Full list of products: [https://cloud.google.com/products/](https://cloud.google.com/products/)
Setup: Create account and set up billing

1. Login with your Google account (NOT stanford.edu account).
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

1. Make sure you have Python 2.7.9 or higher installed.
2. Download SDK using `gcloud.google.com`
3. Install: run `gcloud components install` and follow the on-screen instructions.
4. Authorize using `gcloud auth application-default login`.
5. Test: gcloud components list, gcloud auth list.
Configure and use a VM

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")
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>`

2. **View your clusters:** [https://console.cloud.google.com/dataproc/clusters](https://console.cloud.google.com/dataproc/clusters).

   *Clusters:*

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

1. **Create your job.**
   
   *Simple example: add one to every element in an array.*
   
   ```python
   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:**

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

3. **View your jobs:** [https://console.cloud.google.com/dataproc/jobs](https://console.cloud.google.com/dataproc/jobs)
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

2. To set up cloud storage, see tutorial on

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