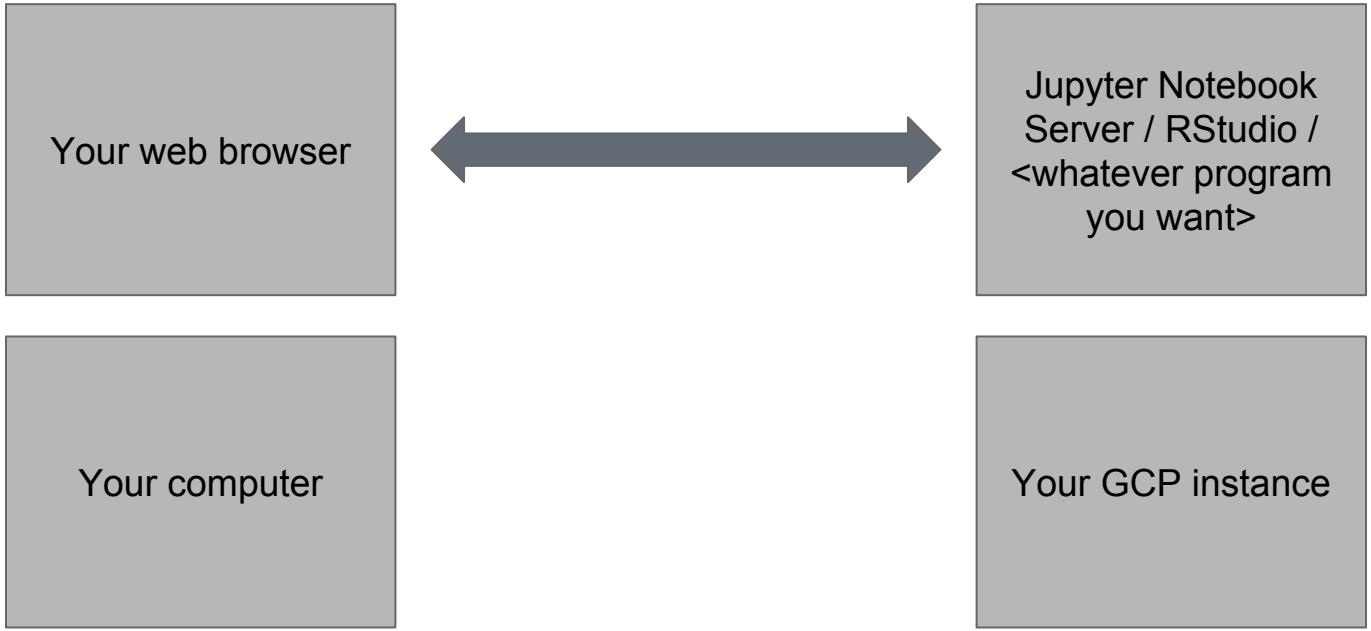
Using Google Cloud Platform for the Hackathon

Columbia Data Science Society

How does this actually work?



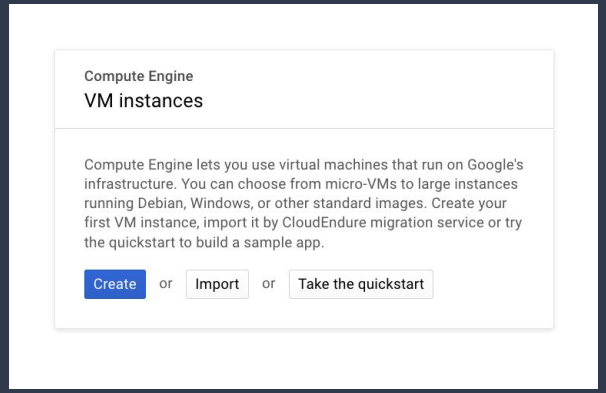
Why go through the hassle?

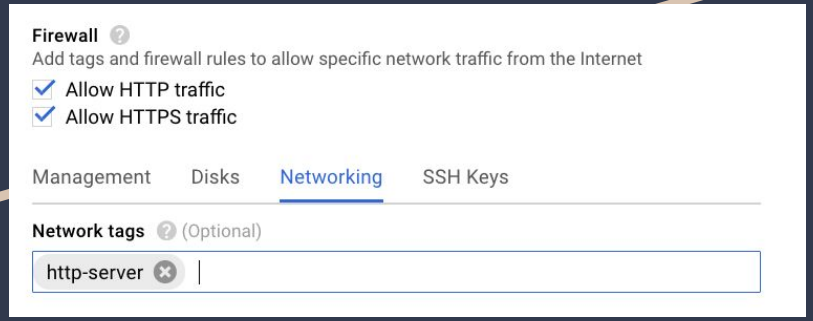
- Storing all data in memory
  - At least one of the datasets won't fit within your laptop's memory
- Downloading all data super fast!
  - Google -> Google transfers are super quick
- Getting results with a super fast computer
  - All the computing power!
- Only using web browser
  - All of us using GCP won't overload WiFi on downloading data locally

Making an account

- Only one person from each team needs to create an account!
  - Everyone on the team will be using the same GCP instance
- Use the GCP credit code you received here: https://console.cloud.google.com/billing/redeem
- Then head over to: https://console.cloud.google.com
  - Click “Create Project”
  - Name it anything you like

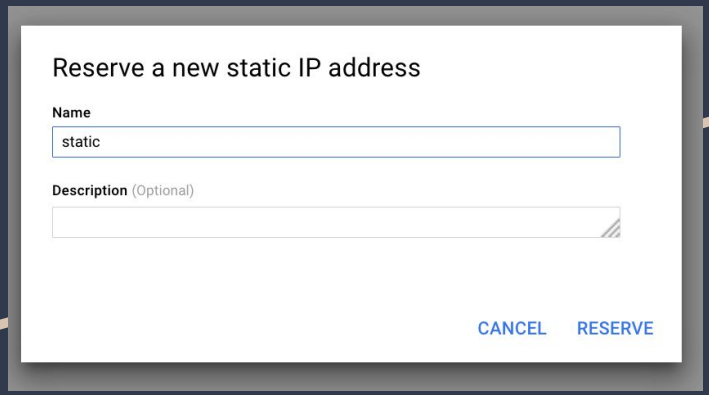
Create an instance





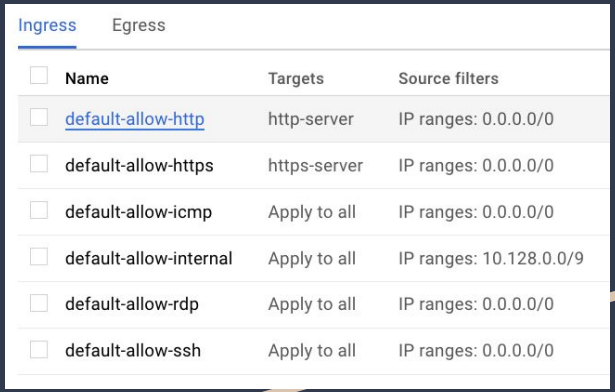
- Go to this link: https://console.cloud.google.com/compute/instances
- Hit “Create”
- Settings
  - Name: any name you want
  - Zone: us-east1-b
  - Machine Type: 8 vCPUs + 52 GB memory
  - Boot Disk: Ubuntu 16.04 LTS + 64 GB SSD persistent disk
  - Firewall: Make sure “Allow HTTP traffic” and “Allow HTTPS traffic” are checked
  - Networking (expand “Management, disks, networking, SSH keys”)
    - Network tags: “http-server” (should become a text bubble once you hit space)
- Hit “Create” at the bottom of the page

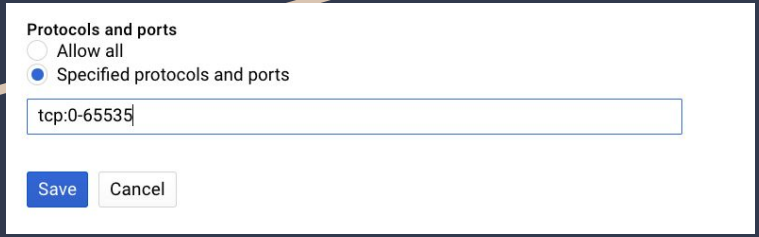
External IP Address for GCP instance



- Go to this link: https://console.cloud.google.com/networking/addresses/list
- Change from “Ephemeral” to “Static”
- Call it “static”
- Click on “Reserve”

Allowing Network Traffic on GCP Instance





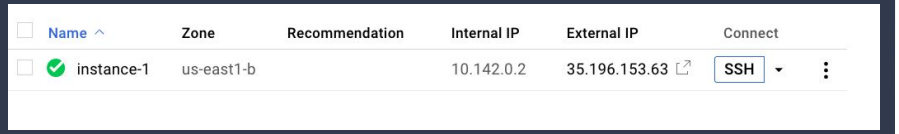
- Go to this link: https://console.cloud.google.com/networking/firewalls/list
- Click on “default-allow-http”
- Click on “Edit”
- Protocols and ports (last item): change to “tcp:0-65535”
- Click on “Save”

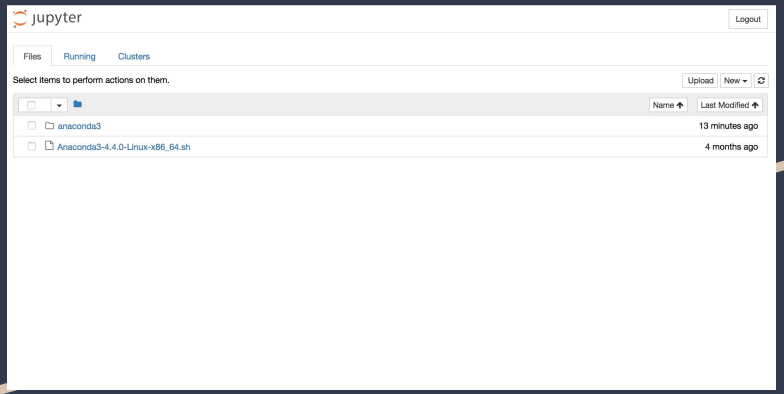
General Setup for Multiple Users Using SSH


```
Connected, host fingerprint: ssh-rsa 2048 FC:F8:B9:4C:C0:04:5C:35:92:3
:03:55:EE:9B:74:FA:B1:08:DF:66:5F:B6:05:DD:20
Welcome to Ubuntu 16.04.3 LTS (GNU/Linux 4.10.0-35-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

0 packages can be updated.
0 updates are security updates.



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

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

ashutosh_nanda@instance-1:~$
```

- Will be using SSH - like command line but for the instance
  - Go to https://console.cloud.google.com/compute/instances
  - Click on SSH
- Everyone will work out of a “/data” folder
  - `sudo mkdir /data`
  - `sudo chmod 777 /data`
- Need to create user accounts for all users
  - If you don’t have a team yet, just add yourself; you can add others later
  - `sudo adduser team_member_1`
    - eg. `sudo adduser john`
    - Follow the prompts to setup password, then just keep hitting enter, and finally `Y` to confirm everything
  - `sudo adduser team_member_2`
  - ...
  - Repeat this step until all team members have accounts

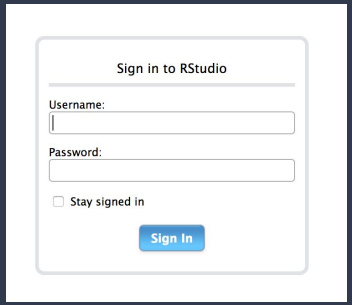
Setting Up Jupyter

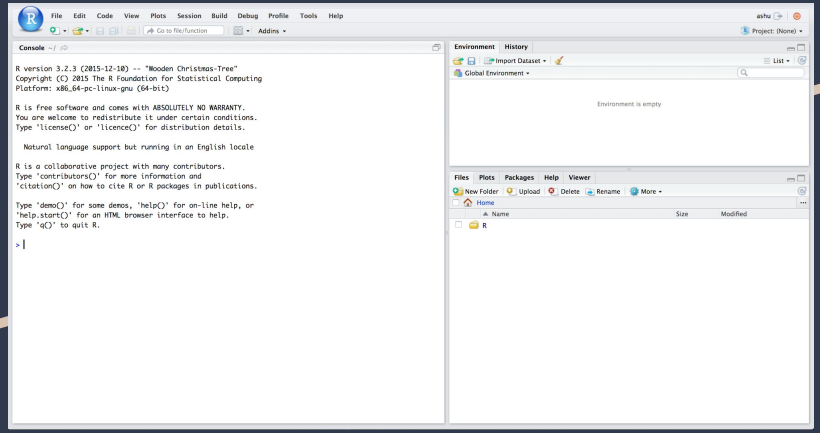




- `wget https://repo.continuum.io/archive/Anaconda3-4.4.0-Linux-x86_64.sh`
- `sudo bash Anaconda3-4.4.0-Linux-x86_64.sh`
  - Press enter, keep hitting enter to get through agreement, `yes`, press enter, wait for a minute or two, `yes`
- `exit`
  - This will quit the SSH console window
- Open a new SSH console window
- `cd /data`
- `jupyter notebook --NotebookApp.token=InsertAnyTokenYourTeamWantsHere --ip=0.0.0.0 --port=8888 &`
- Navigate to “1.1.1.1:8888” where 1.1.1.1 is whatever external IP your instance has
  - It’s listed on your instances page: https://console.cloud.google.com/compute/instances
- Have fun hacking in Python!
  - You’ll all be using the same folder for seamless collaboration!

Setting Up RStudio





- `sudo apt-get install r-base gdebi-core`
  - `Y` to confirm
- `wget https://download2.rstudio.org/rstudio-server-1.0.153-amd64.deb`
- `sudo gdebi rstudio-server-1.0.153-amd64.deb`
  - `y` to confirm
- Navigate to “1.1.1.1:8787” where 1.1.1.1 is whatever external IP your instance has
  - It’s listed on your instances page: https://console.cloud.google.com/compute/instances
- Log in using the usernames and passwords you created in the “General Setup” stage
- `setwd('/data')`
  - Run this command first so that you are all working in the same folder for seamless collaboration
- Have fun hacking in R!

Getting Data

- Bloomberg
  - `curl -O https://storage.googleapis.com/2017cdsdatasets/bbg_cdss_hackathon_2017.tar.gz`
- Digital Reasoning
  - `curl -O https://storage.googleapis.com/2017cdsdatasets/enron.zip`
- Enigma
  - `curl -O https://storage.googleapis.com/2017cdsdatasets/govt2000_2012.zip; curl -O https://storage.googleapis.com/2017cdsdatasets/govt2013_2015.zip; curl -O https://storage.googleapis.com/2017cdsdatasets/govt2016_2017.zip`

Questions?

- Feel free to ask volunteers (green shirts) for help
- Double-check the instructions
- No rush -- everyone will be set up soon!