



EECS E6893 Big Data Analytics

Intro to Big Data Analytics on GCP

Hritik Jain, hj2533@columbia.edu

Agenda

- GCP
 - Setup
 - Interaction
- Services
 - Cloud Storage
 - BigQuery
 - Dataproc (Spark)
- HW0



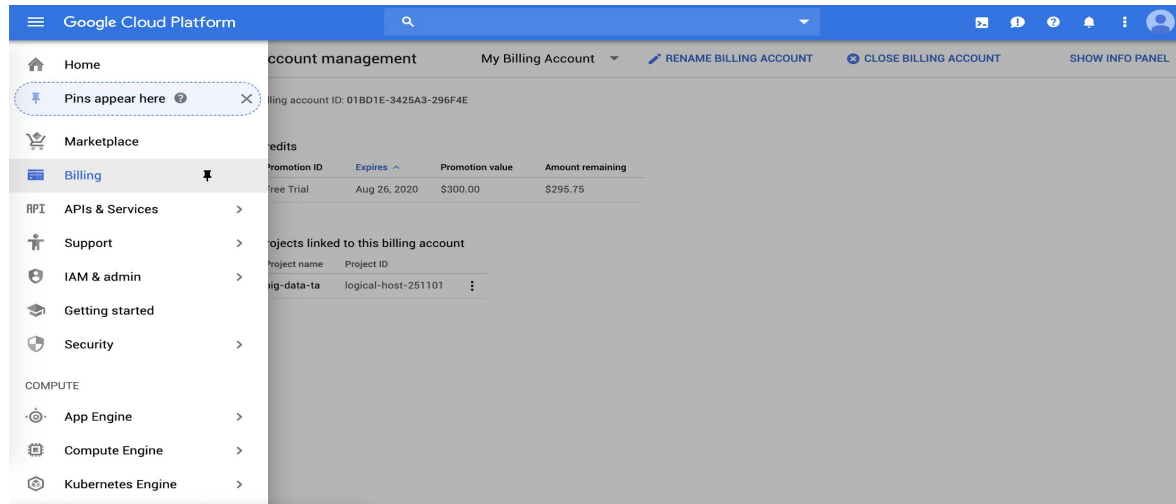
Google Cloud Platform (GCP)

GCP

- Cloud computing platform
 - Flexibility: on-demand and scale as you want
 - Efficiency: no need to maintain infra
- Services (relevant to this assignment)
 - Compute
 - Compute Engines: VMs / Servers (automatically created by Dataproc)
 - Big data products
 - BigQuery: Data warehouse for analytics
 - Dataproc: Hadoop and Spark
 - Storage
 - Cloud Storage: Object storage system
 - Much much more at <https://cloud.google.com/products/>

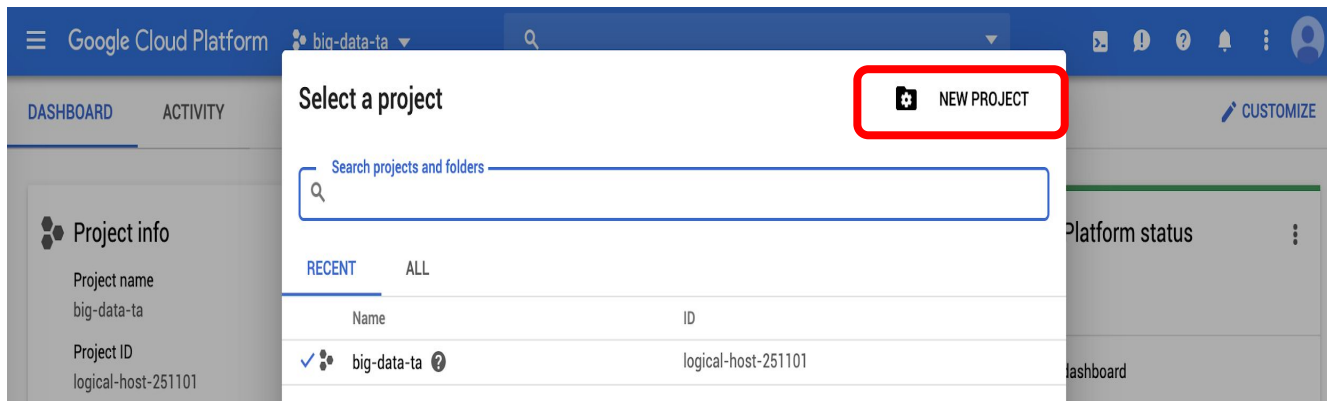
GCP Setup

- Create a google account, you could use your Columbia account
- Apply for \$300 credit for the first year: <https://cloud.google.com/free/>
- Go to [Console dashboard](#) -> Billing to check credit is there



GCP: Create project

- Project: basic unit for creating, enabling, and using all GCP services
 - managing APIs, billing, permissions
 - adding and removing collaborators
- Visit console dashboard or [cloud resource manager](#)
- Click on “create project / new project” and complete the flow
- Ensure billing is pointing to the \$300 credit



GCP: Interaction

- [Graphical UI / console](#): Useful to create VMs, set up clusters, provision resources, manage teams, etc
- [Command line tools / Cloud SDK](#): Useful for interacting from local host and using the resources once provisioned. E.x. ssh into instances, submit jobs, copy files, etc
- [Cloud Shell](#): Same as command line, but web-based and pre-installed with SDK and tools

GCP: console

Search for services here

The screenshot shows the Google Cloud Platform console dashboard for the project 'big-data-ta'. The top navigation bar includes the Google Cloud Platform logo, the project name, a search bar (highlighted with a red box), and user profile icons. Below the navigation bar, the dashboard is organized into several widgets:

- Project info:** Displays project details such as Project name (big-data-ta), Project ID (logical-host-251101), and Project number (312759131343). It includes a link to 'ADD PEOPLE TO THIS PROJECT' and 'Go to project settings'.
- Resources:** Lists active resources: 1 Compute Engine instance, 2 Storage buckets, and 1 BigQuery dataset.
- Compute Engine:** Shows a CPU usage graph for 'instance/cpu/utilization' with a current value of 0.016. The graph shows a peak around 1:30. A link 'Go to Compute Engine' is provided.
- API APIs:** Shows 'Requests (requests/sec)' with a current value of 0.26. This section is highlighted with a red box.
- Google Cloud Platform status:** Indicates 'All services normal' and provides a link to 'Go to Cloud status dashboard'.
- Billing:** Shows 'Estimated charges' for the billing period Sep 1 - 6, 2019, amounting to USD \$0.00. A link to 'View detailed charges' is provided. This section is highlighted with a red box.
- Error Reporting:** States 'No sign of any errors. Have you set up Error Reporting?' and provides a link to 'Learn how to set up Error Reporting'.

Manage / Enable APIs

GCP: Cloud SDK

- Install the SDK that is suitable for your local environment:

<https://cloud.google.com/sdk/docs/quickstarts>

- Some testing after installation:

- `gcloud info`
- `gcloud auth list`
- `gcloud components list`

- Change default config:

- `gcloud init`

```
[dyn-129-236-216-148:~ frank$ gcloud components list
```

```
Your current Cloud SDK version is: 259.0.0  
The latest available version is: 261.0.0
```

Components			
Status	Name	ID	Size
Update Available	BigQuery Command Line Tool	bq	< 1 MiB
Update Available	Cloud SDK Core Libraries	core	11.5 MiB
Not Installed	App Engine Go Extensions	app-engine-go	56.4 MiB
Not Installed	Cloud Bigtable Command Line Tool	cbt	7.3 MiB
Not Installed	Cloud Bigtable Emulator	bigtable	6.6 MiB
Not Installed	Cloud DataLab Command Line Tool	datalab	< 1 MiB
Not Installed	Cloud Datastore Emulator	cloud-datastore-emulator	18.4 MiB
Not Installed	Cloud Datastore Emulator (Legacy)	gcd-emulator	38.1 MiB
Not Installed	Cloud Firestore Emulator	cloud-firestore-emulator	36.8 MiB
Not Installed	Cloud Pub/Sub Emulator	pubsub-emulator	34.8 MiB
Not Installed	Cloud SQL Proxy	cloud_sql_proxy	3.7 MiB
Not Installed	Emulator Reverse Proxy	emulator-reverse-proxy	14.5 MiB
Not Installed	Google Cloud Build Local Builder	cloud-build-local	5.9 MiB
Not Installed	Google Container Registry's Docker credential helper	docker-credential-gcr	1.8 MiB
Not Installed	gcloud Alpha Commands	alpha	< 1 MiB
Not Installed	gcloud app Java Extensions	app-engine-java	85.9 MiB
Not Installed	gcloud app PHP Extensions	app-engine-php	21.9 MiB
Not Installed	gcloud app Python Extensions	app-engine-python	6.0 MiB
Not Installed	gcloud app Python Extensions (Extra Libraries)	app-engine-python-extras	28.5 MiB
Not Installed	kubect	kubectl	< 1 MiB
Installed	Cloud Storage Command Line Tool	gsutil	3.6 MiB
Installed	gcloud Beta Commands	beta	< 1 MiB

```
To install or remove components at your current SDK version [259.0.0], run:
```

```
$ gcloud components install COMPONENT_ID  
$ gcloud components remove COMPONENT_ID
```

```
To update your SDK installation to the latest version [261.0.0], run:
```

```
$ gcloud components update
```

GCP: Cloud Shell

The screenshot shows the Google Cloud Platform dashboard for the project 'big-data-ta'. The top navigation bar includes the GCP logo, the project name, a search bar, and utility icons for help, notifications, and user profile. A red box highlights the 'Activate Cloud Shell' button, which is located next to the help and notification icons. Below the navigation bar, the dashboard is divided into several sections: 'Project info' (showing project name, ID, and number), 'RPI APIs' (a line graph showing requests per second), 'Google Cloud Platform status' (indicating all services are normal), and 'Billing' (showing estimated charges of USD \$0.00 for the period Sep 1 - 12, 2019).

Google Cloud Platform big-data-ta

DASHBOARD ACTIVITY

Activate Cloud Shell CUSTOMIZE

Project info

- Project name: big-data-ta
- Project ID: logical-host-251101
- Project number: 312759131343

→ Go to project settings

RPI APIs

Requests (requests/sec)

2.5
2.0
1.5
1.0
0.5

Google Cloud Platform status

All services normal

→ Go to Cloud status dashboard

Billing

Estimated charges: USD \$0.00
For the billing period Sep 1 - 12, 2019

persistent home directory :)

GCP: Cloud Shell (Cont')

The screenshot displays the Google Cloud Platform (GCP) dashboard interface. At the top, the navigation bar includes the GCP logo, the project name 'big-data-ta', a search bar, and utility icons for help, notifications, and user profile. Below the navigation bar, the 'DASHBOARD' and 'ACTIVITY' tabs are visible, along with a 'CUSTOMIZE' button.

The dashboard content is organized into three main panels:

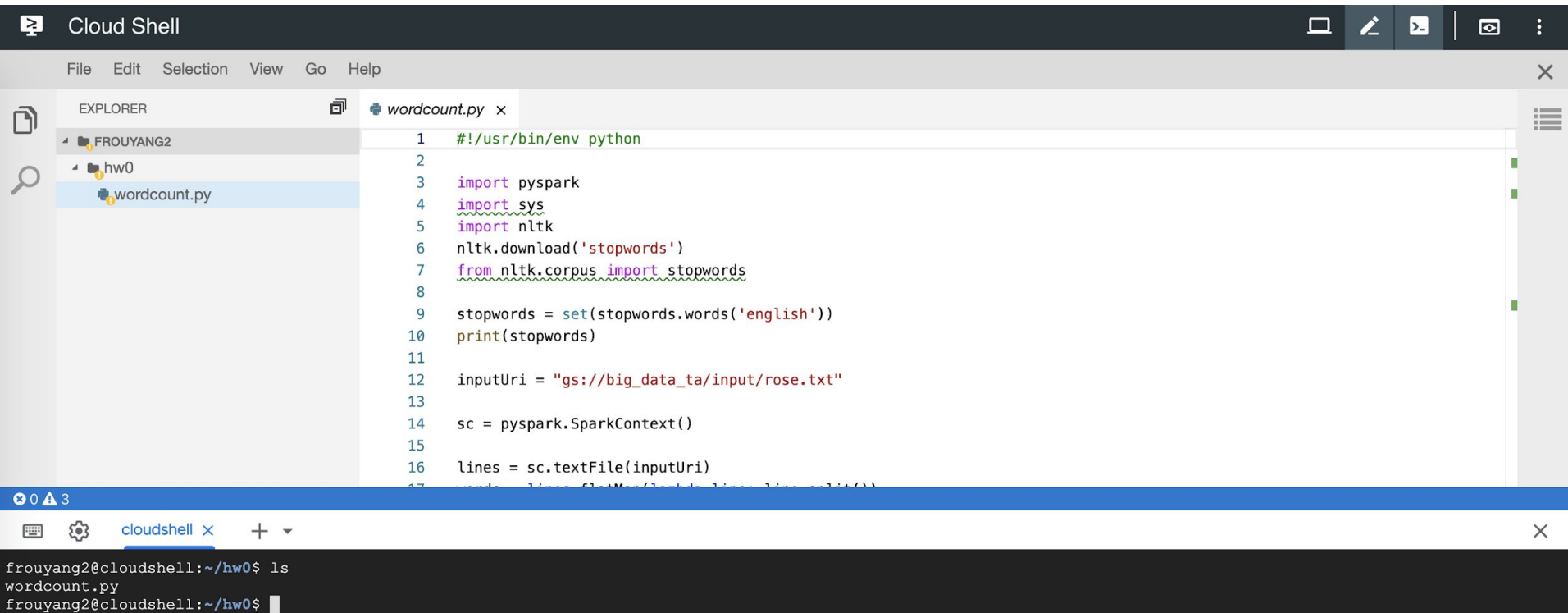
- Project info:** Displays key project details:
 - Project name: big-data-ta
 - Project ID: logical-host-251101
 - Project number: [obscured]
- API APIs:** Shows a line chart for 'Requests (requests/sec)'. The y-axis ranges from 2.0 to 2.5. A single data point is visible at approximately 2.1 requests/sec.
- Google Cloud Platform status:** Indicates 'All services normal' and provides a link to 'Go to Cloud status dashboard'.

At the bottom, a terminal window titled 'cloudshell' shows the following commands and output:

```
froyang2@cloudshell:~$ ls
hw0 README-cloudshell.txt
froyang2@cloudshell:~$
```

A 'Launch code editor BETA' button is located in the bottom right corner of the terminal area.

GCP: Cloud Shell Code Editor



Cloud Shell

File Edit Selection View Go Help

EXPLORER

wordcount.py x

```
1 #!/usr/bin/env python
2
3 import pyspark
4 import sys
5 import nltk
6 nltk.download('stopwords')
7 from nltk.corpus import stopwords
8
9 stopwords = set(stopwords.words('english'))
10 print(stopwords)
11
12 inputUri = "gs://big_data_ta/input/rose.txt"
13
14 sc = pyspark.SparkContext()
15
16 lines = sc.textFile(inputUri)
17 words = lines.flatMap(lambda line: line.split())
```

cloudshell x

```
froyang2@cloudshell:~/hw0$ ls
wordcount.py
froyang2@cloudshell:~/hw0$
```



Cloud Storage

Cloud Storage

- Online file storage system
- Graphical UI through console

The screenshot shows the Google Cloud Platform Storage console. The top navigation bar includes the Google Cloud Platform logo, the project name 'big-data-ta', a search icon, and various utility icons. The left sidebar shows the 'Storage' menu with options for 'Browser', 'Transfer', 'Transfer Appliance', and 'Settings'. The main content area is titled 'Browser' and features a 'CREATE BUCKET' button (highlighted with a red box), a 'REFRESH' button, and a 'DELETE' button. Below these buttons is a search bar labeled 'Filter by prefix...' and a 'Columns' dropdown menu. A table of buckets is displayed below, with columns for Name, Default storage class, Location, Location Type, Public access, Lifecycle, Access control model, and Labels. The table contains one entry for the bucket 'big_data_ta'.

<input type="checkbox"/>	Name	Default storage class [?]	Location	Location Type	Public access [?]	Lifecycle [?]	Access control model [?]	Labels [?]
<input type="checkbox"/>	big_data_ta	Standard	us-east1 (South Carolina)	Region	Per object	None	Bucket policy & ACLs	

- Command line tool: `gsutil`

Cloud Storage - graphical UI

The screenshot displays the Google Cloud Platform Storage interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'big-data-ta', a search bar, and various utility icons. The left sidebar shows navigation options: Storage, Browser, Transfer, Transfer Appliance, and Settings. The main content area is titled 'Bucket details' for the bucket 'big_data_ta'. It features tabs for 'Objects', 'Overview', 'Permissions', and 'Bucket Lock'. Below the tabs, a row of action buttons is visible: 'Upload files', 'Upload folder', 'Create folder', 'Manage holds', and 'Delete'. The 'Upload files' button is highlighted with a red rectangular box. Below the buttons is a search input field labeled 'Filter by prefix...'. The breadcrumb path is 'Buckets / big_data_ta / data'. A table lists the objects in the bucket, with one object 'citibike_stations.csv' shown.

Storage

Browser

Transfer

Transfer Appliance

Settings

Bucket details

EDIT BUCKET

REFRESH BUCKET

big_data_ta

Objects Overview Permissions Bucket Lock

Upload files Upload folder Create folder Manage holds Delete

Filter by prefix...

Buckets / big_data_ta / data

<input type="checkbox"/>	Name	Size	Type	Storage class	Last modified	Public access ?	Encryption ?	Retention expiration date ?	Holds ?
<input type="checkbox"/>	citibike_stations.csv	114.28 KB	application/octet-stream	Standard	9/2/19, 10:11:33 PM UTC-4	Not public	Google-managed key	-	None

Cloud Storage - graphical UI (cont')

The screenshot displays the Google Cloud Platform Storage console interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'big-data-ta', a search bar, and various utility icons. The left sidebar contains navigation options: Storage, Browser, Transfer, Transfer Appliance, and Settings. The main content area shows the 'Object details' for a file named 'citibike_stations.csv' located in the bucket 'big_data_ta / data'. The details include:

- Access: Not public
- Type: application/octet-stream
- Size: 114.28 KB
- Created: September 2, 2019 at 10:11:33 PM UTC-4
- Last modified: September 2, 2019 at 10:11:33 PM UTC-4
- URI: `gs://big_data_ta/data/citibike_stations.csv` (highlighted with a red box)
- Link URL: `https://storage.cloud.google.com/big_data_ta/data/citibike_stations.c`

Uniform Resource Identifier, like *a filepath* on GCP, use this in your program

Cloud Storage - gsutil

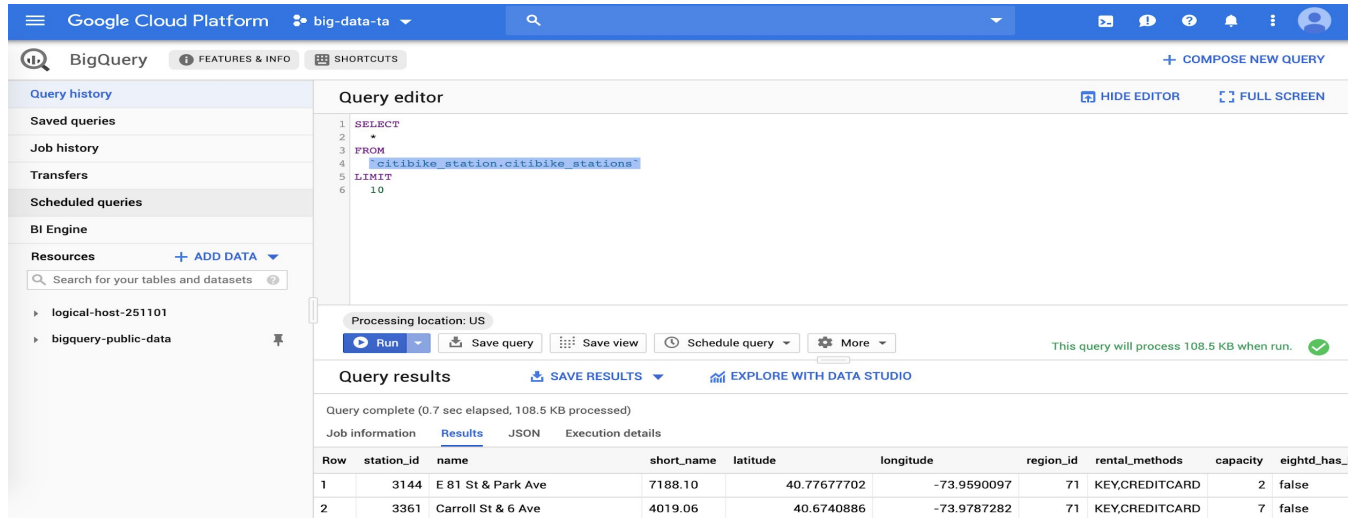
- Interact with Cloud Storage through command line
- Works similar to unix command line
- Useful commands:
 - Concatenate object content to stdout:
`gsutil cat [-h] url...`
 - Copy file:
`gsutil cp [OPTION]... src_url dst_url`
 - List files:
`gsutil ls [OPTION]... url...`
- Explore more at <https://cloud.google.com/storage/docs/gsutil>



BigQuery

BigQuery

- Data warehouse for analytics
- SQL-like languages to interact with DB
- RESTful APIs / client libraries for programmatic access
- Graphical UI



The screenshot displays the Google Cloud Platform BigQuery interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'big-data-ta', and a search bar. Below the navigation bar, the 'Query editor' is active, showing a SQL query:

```
1 SELECT
2 *
3 FROM
4 citibike_station.citibike_stations
5 LIMIT
6 10
```

 The query is highlighted in blue. Below the query editor, the 'Query results' section shows a table with 10 columns: Row, station_id, name, short_name, latitude, longitude, region_id, rental_methods, capacity, and eightd_has. The table contains two rows of data. The 'Query complete' message indicates that the query was executed successfully in 0.7 seconds, processing 108.5 KB of data.

Row	station_id	name	short_name	latitude	longitude	region_id	rental_methods	capacity	eightd_has
1	3144	E 81 St & Park Ave	7188.10	40.77677702	-73.9590097	71	KEY,CREDITCARD	2	false
2	3361	Carroll St & 6 Ave	4019.06	40.6740886	-73.9787282	71	KEY,CREDITCARD	7	false



Dataproc

Dataproc

- On-demand, fully managed cloud service for running Apache Hadoop and Spark on GCP
- Cluster creation (using Cloud SDK):

- Automatically creates VMs with Spark pre-installed
- `gcloud dataproc clusters create <cluster-name>`
- `gcloud beta dataproc clusters create <cluster-name>`

```
--optional-components=ANACONDA,JUPYTER --image-version=1.3  
--enable-component-gateway --bucket <bucket-name> --project  
<project-id> --single-node --metadata  
'PIP_PACKAGES=graphframes==0.6' --initialization-actions  
gs://dataproc-initialization-actions/python/pip-install.sh
```

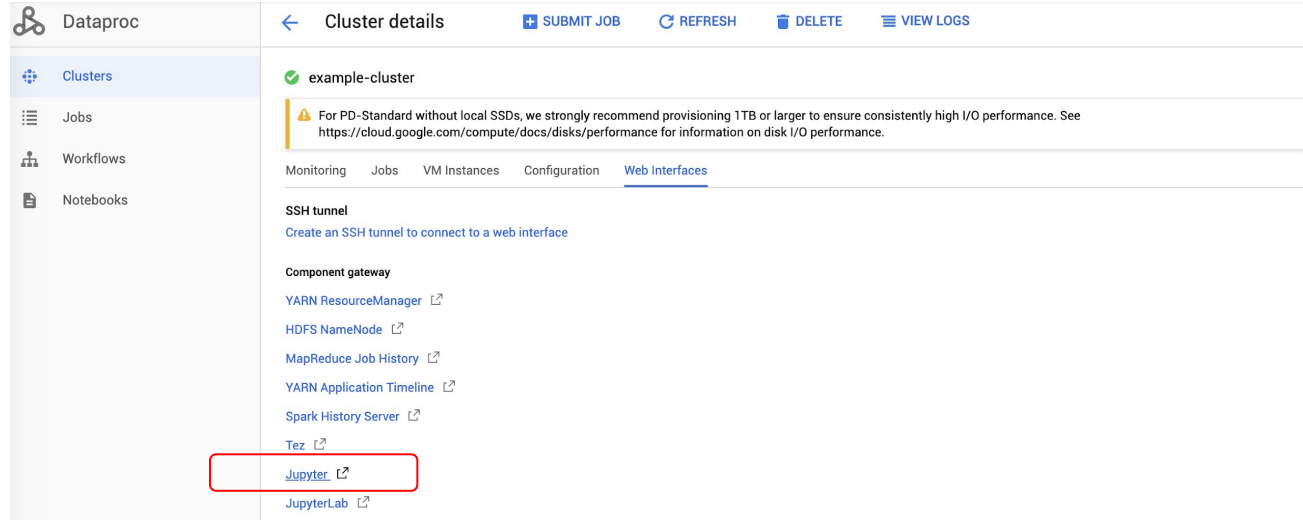
Works like `pip install <your package>`

Install
Jupyter
Notebook

Cloud Storage
bucket: where
your jupyter
notebooks are
saved

Dataproc - Spark execution / submit jobs

- Jupyter notebook:



The screenshot shows the Google Cloud Dataproc console interface. On the left is a navigation sidebar with 'Clusters' selected. The main area displays 'Cluster details' for 'example-cluster'. A warning message is visible at the top. Below that, there are tabs for 'Monitoring', 'Jobs', 'VM Instances', 'Configuration', and 'Web Interfaces'. Under the 'Web Interfaces' tab, a list of services is shown: 'SSH tunnel', 'Component gateway', 'YARN ResourceManager', 'HDFS NameNode', 'MapReduce Job History', 'YARN Application Timeline', 'Spark History Server', 'Tez', 'Jupyter', and 'JupyterLab'. The 'Jupyter' link is highlighted with a red rectangular box.

- Cloud SDK:

- `gcloud dataproc jobs submit pyspark <your_program.py>`
`--cluster=<cluster-name>`
- [View your jobs in console](#)

- Program could be Cloud Storage URI / local path / Cloud Shell path
- Data should be on Cloud storage

Dataprocc - Spark execution / submit jobs (cont')

- Spark shell
 - ssh into master node

Google Cloud Platform big-data-ta

Dataprocc

Cluster details

+ SUBMIT JOB REFRESH DELETE VIEW LOGS

example-cluster

For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See <https://cloud.google.com/compute/docs/disks/performance> for information on disk I/O performance.

Monitoring Jobs VM Instances Configuration Web Interfaces

Name	Role
example-cluster-m	Master SSH

Equivalent REST

- pyspark

```
frouyang2@example-cluster-m:~$ pyspark
Python 2.7.14 [Anaconda, Inc.] (default, Dec 7 2017, 17:05:42)
[GCC 7.2.0] on linux2
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).
19/09/06 18:46:51 WARN org.apache.spark.scheduler.FairSchedulableBuilder: Fair Scheduler configuration file not found so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spark.scheduler.allocation.file to a file that contains the configuration.
Welcome to

  SPARK  version 2.3.3

Using Python version 2.7.14 (default, Dec 7 2017 17:05:42)
SparkSession available as 'spark'.
>>>
```

HW0

1. Read documentations and tutorials
 - a. Setup GCP and Cloud SDK
 - b. Run Spark examples on Dataproc - Pi calculation and word count
 - c. Familiar yourself with BigQuery
2. Two light programming questions
 - a. BigQuery
 - b. Spark program - Find top k most frequent words

Remember to delete your dataproc clusters when you finish executions to save money.