



EECS E6893 Big Data Analytics

Intro to Big Data Analytics on GCP

Frank Ou Yang

Agenda

- GCP
- Cloud Storage
- BigQuery
- Dataproc



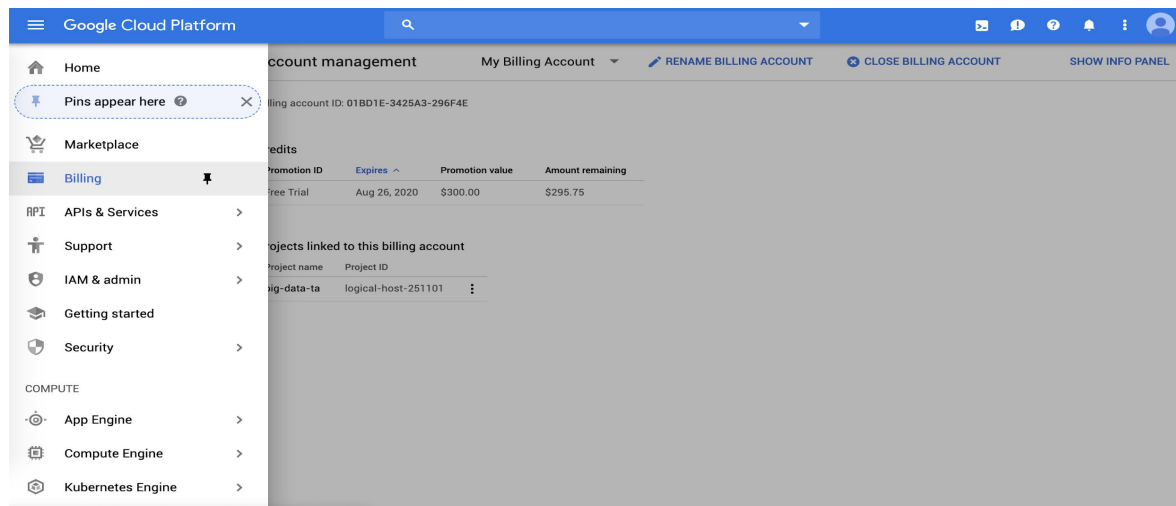
Google Cloud Platform (GCP)

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
 - Big data products
 - BigQuery: Data warehouse for analytics
 - Dataproc: Hadoop and Spark
 - Storage & DBs
 - Cloud Storage: Object storage system
 - More at <https://cloud.google.com/products/>

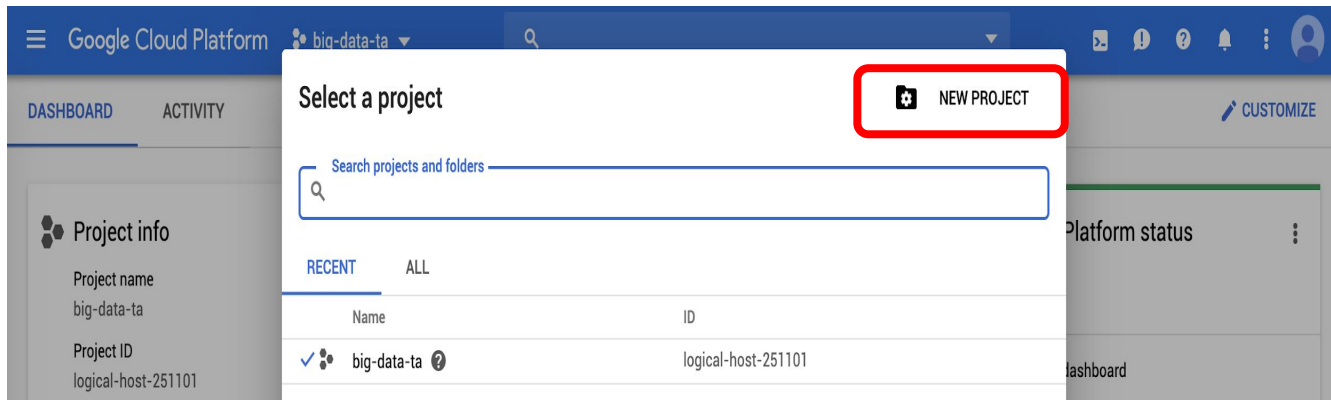
GCP Setup

- 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 is blue and contains the Google Cloud Platform logo, the project name 'big-data-ta', a search bar, and several utility icons. A red rectangle highlights the search bar. Below the navigation bar, the dashboard is divided into several sections. On the left, there is a 'Project info' section with details like Project name, Project ID, and Project number, and a 'Resources' section listing Compute Engine instances, Storage buckets, and BigQuery datasets. In the center, there is a 'Compute Engine' section showing a line graph of CPU usage over time, with a red rectangle highlighting the 'API APIs' section below it. On the right, there is a 'Google Cloud Platform status' section indicating 'All services normal', a 'Billing' section showing estimated charges, and an 'Error Reporting' section. A red rectangle highlights the 'Billing' section.

Google Cloud Platform big-data-ta

Search for services here

DASHBOARD ACTIVITY CUSTOMIZE

Project info

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

[ADD PEOPLE TO THIS PROJECT](#)

[Go to project settings](#)

Resources

- Compute Engine: 1 instance
- Storage: 2 buckets
- BigQuery: 1 dataset

Compute Engine

CPU (%)

0.0300

0.0225

0.0150

1:30 1:45 2 PM 2:15

instance/cpu/utilization: 0.016

[Go to Compute Engine](#)

API APIs

Requests (requests/sec)

0.26

Google Cloud Platform status

All services normal

[Go to Cloud status dashboard](#)

Billing

Estimated charges: USD \$0.00

For the billing period Sep 1 – 6, 2019

[View detailed charges](#)

Error Reporting

No sign of any errors. Have you set up Error Reporting?

[Learn how to set up Error Reporting](#)

Manage / Enable APIs

GCP: Cloud SDK setup

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

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

- Some testing after installation:

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

```
[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	kubect1	kubect1	< 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
```



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 bar, and various utility icons. The left sidebar contains navigation links for Storage, 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 titled 'Buckets' displays the following data:

<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 console. The left sidebar shows the navigation menu with options: Storage, Browser, Transfer, Transfer Appliance, and Settings. The main content area is titled 'Bucket details' for the bucket 'big_data_ta'. It includes tabs for 'Objects', 'Overview', 'Permissions', and 'Bucket Lock'. A red box highlights the action buttons: 'Upload files', 'Upload folder', 'Create folder', 'Manage holds', and 'Delete'. Below these buttons is a search bar labeled 'Filter by prefix...'. The breadcrumb trail shows 'Buckets / big_data_ta / data'. A table lists the objects in the bucket, with one object 'citibike_stations.csv' visible.

Google Cloud Platform big-data-ta

Storage

Browser

Transfer

Transfer Appliance

Settings

Bucket details EDIT BUCKET REFRESH BUCKET

big_data_ta

Objects Overview Permissions Bucket Lock

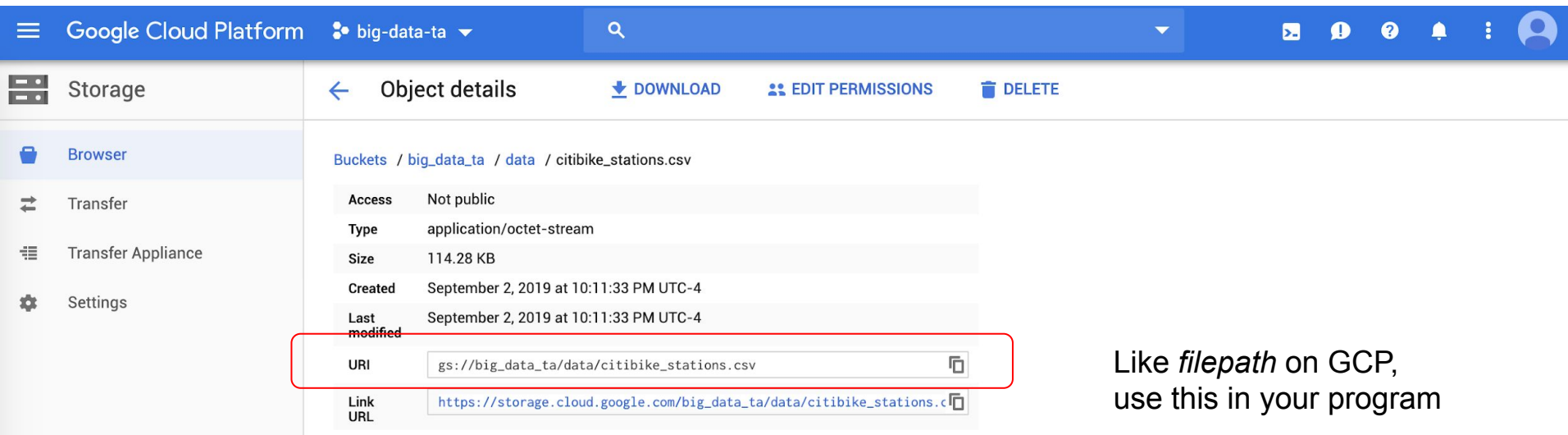
Upload files Upload folder Create folder Manage holds Delete

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. The left sidebar shows the 'Storage' section with options like 'Browser', 'Transfer', 'Transfer Appliance', and 'Settings'. The main area shows the 'Object details' for a file named 'citibike_stations.csv' in the 'big_data_ta' bucket. 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), and Last modified (September 2, 2019 at 10:11:33 PM UTC-4). A red box highlights the 'URI' field, which contains the text 'gs://big_data_ta/data/citibike_stations.csv'. Below the URI field is the 'Link URL' field, which contains the text 'https://storage.cloud.google.com/big_data_ta/data/citibike_stations.c'.

Google Cloud Platform big-data-ta

Storage

Object details

DOWNLOAD EDIT PERMISSIONS DELETE

Buckets / big_data_ta / data / citibike_stations.csv

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

Link URL https://storage.cloud.google.com/big_data_ta/data/citibike_stations.c

Like *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 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. The left sidebar contains a 'Query history' section with links to 'Saved queries', 'Job history', 'Transfers', 'Scheduled queries', 'BI Engine', and 'Resources'. The main area is the 'Query editor', which contains a SQL query:

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

 Below the query editor, there are buttons for 'Run', 'Save query', 'Save view', 'Schedule query', and 'More'. A status message indicates 'This query will process 108.5 KB when run.' with a green checkmark. The 'Query results' section shows 'Query complete (0.7 sec elapsed, 108.5 KB processed)' and a table of results. The table has columns: Row, station_id, name, short_name, latitude, longitude, region_id, rental_methods, capacity, and eighthd_has_.

Row	station_id	name	short_name	latitude	longitude	region_id	rental_methods	capacity	eighthd_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
```

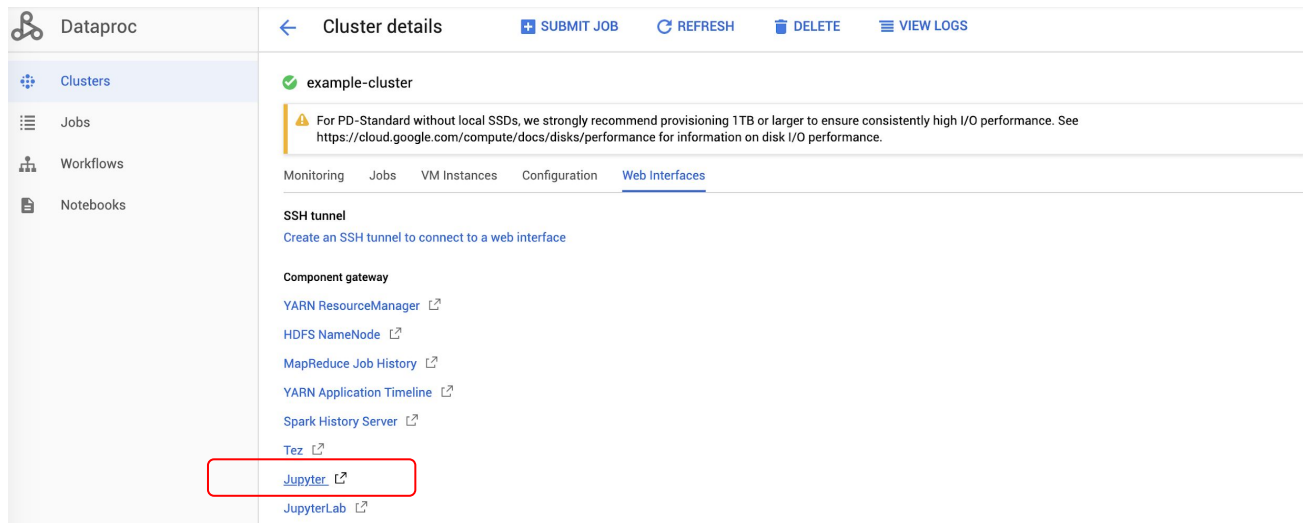
Works like `pip install <your package>`

Cloud Storage
bucket: where
your jupyter
notebooks are
saved

Install
Jupyter
Notebook

Dataproc - Spark execution / submit jobs

- Jupyter notebook:



- Cloud SDK:

- `gcloud dataproc jobs submit pyspark <your_program.py> --cluster=<cluster-name>`

- [View your jobs in console](#)

- program could be Cloud Storage URI or local path
- Data should be on Cloud storage

Dataproc - Spark execution / submit jobs (cont')

- Spark shell
 - ssh into master node

- 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

  ____      __
 / ___ |__ /  /  ___
/  _ \|_ \|  /  / _ \
/  ___/ __ \|_/  / ___/
/___/____/____/____/
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.