EECS E6893 Big Data Analytics
Intro to Big Data Analytics on GCP

Frank Ou Yang

Agenda

- GCP
- Cloud Storage
- BigQuery
- Dataproc

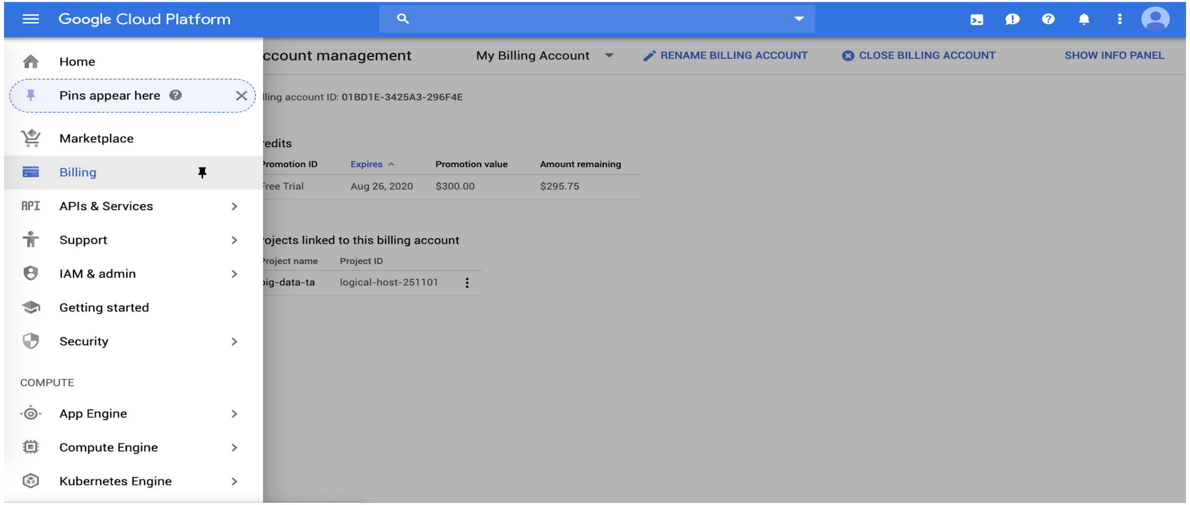
Google Cloud Platform (GCP)

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



ccount management My Billing Account

lling account ID: 01BD1E-3425A3-296F4E

redits

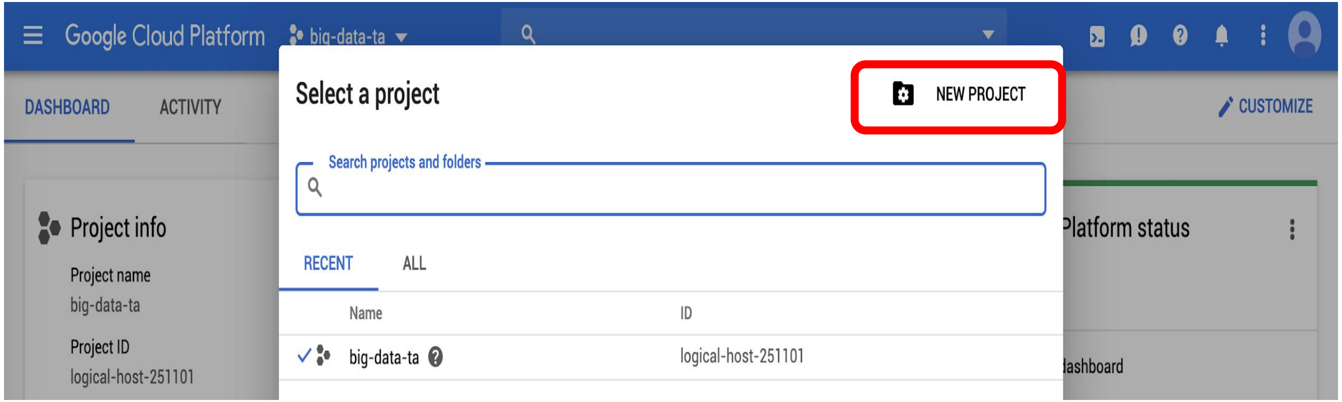
romotion ID	Expires	Promotion value	Amount remaining
ree Trial	Aug 26, 2020	$300.00	$295.75

ojects linked to this billing account

roject name	Project ID
ig-data-ta	logical-host-251101

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



Select a project

NEW PROJECT

Search projects and folders

RECENT ALL

Name	ID
✓ big-data-ta	logical-host-251101

Project name
big-data-ta

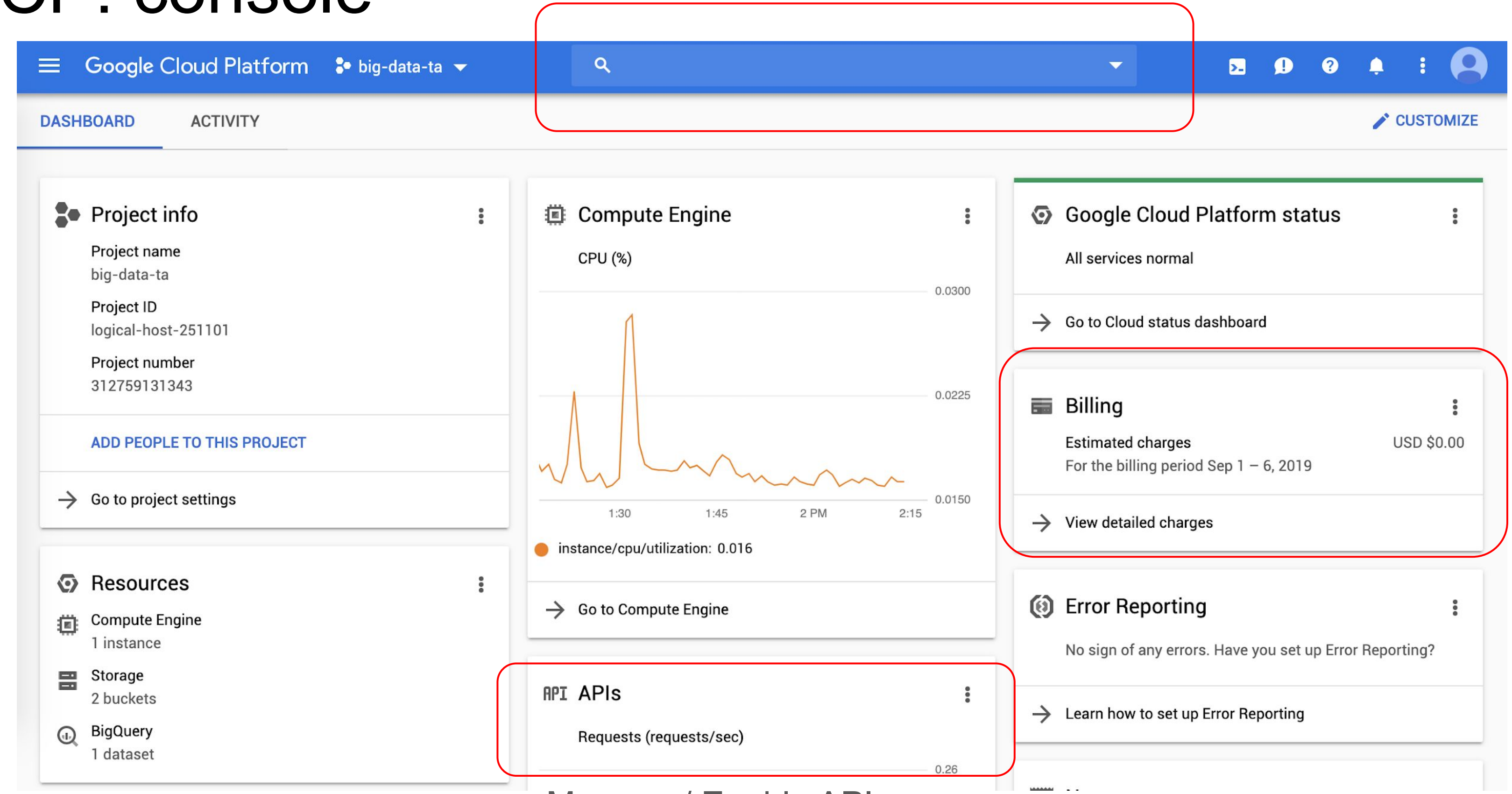
Project ID
logical-host-251101

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



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 (%)

instance/cpu/utilization: 0.016

Go to Compute Engine

APIs

Requests (requests/sec)

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

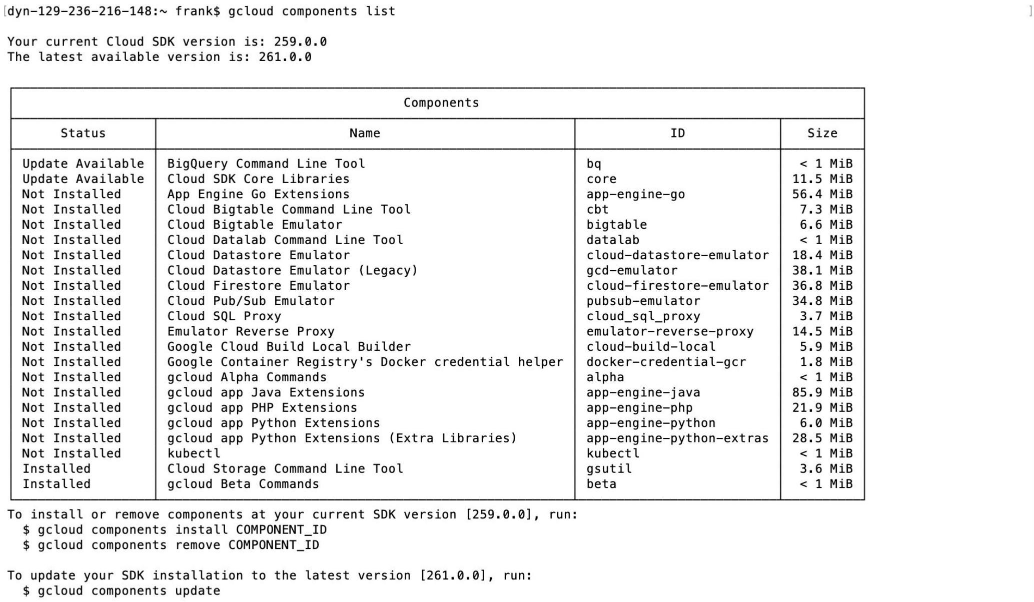
No sign of any errors. Have you set up Error Reporting?

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`

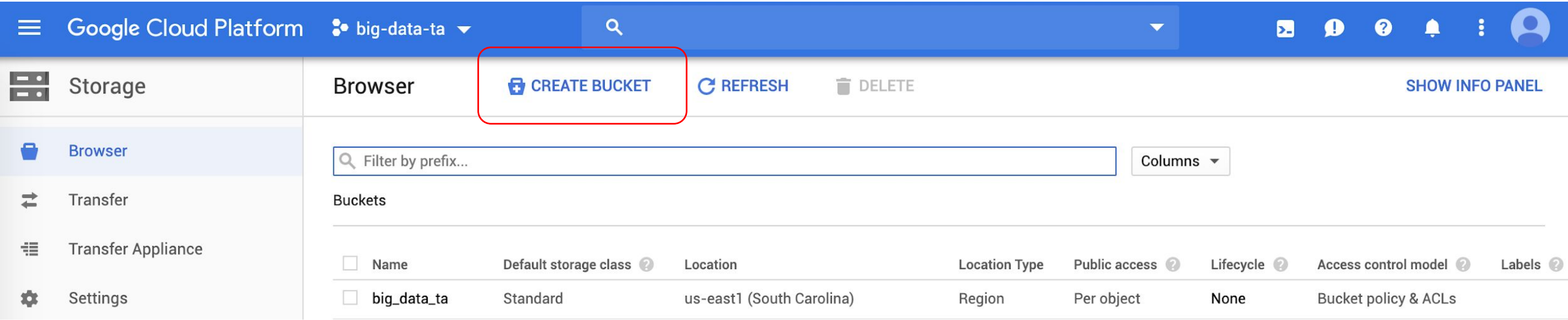


```
dyn-129-236-216-148:~ frank$ gcloud components list

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    │ kubectl                                              │ 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
```

Cloud Storage

Cloud Storage

- Online file storage system
- Graphical UI through console

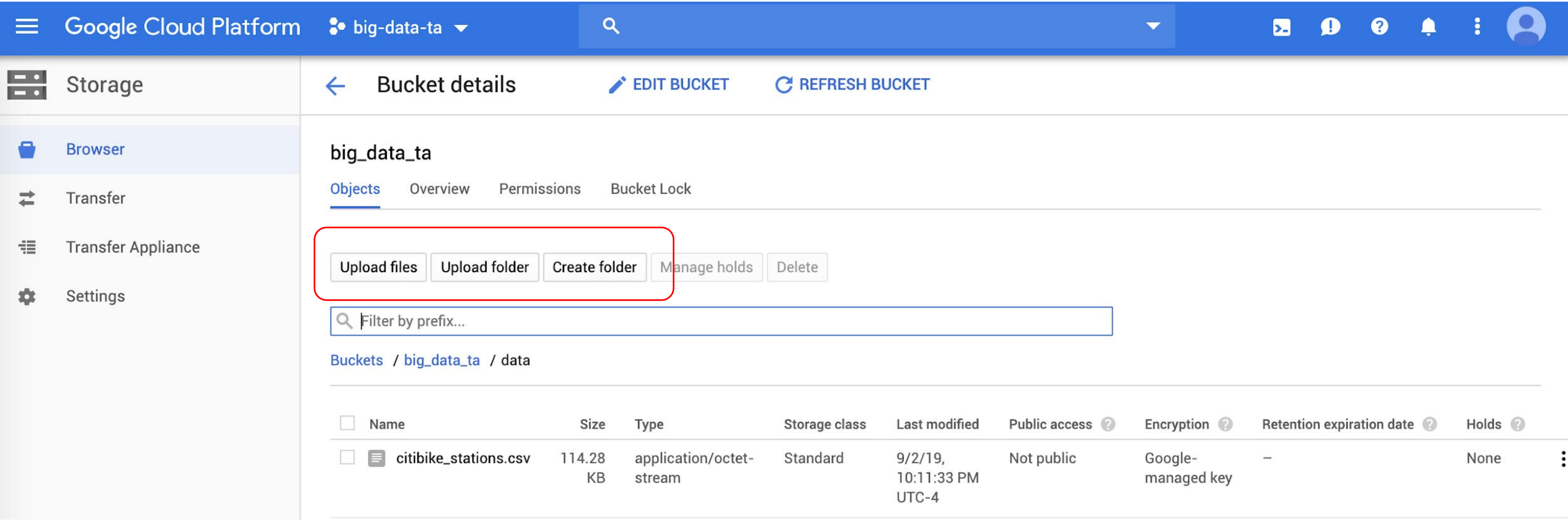


Buckets

□	Name	Default storage class	Location	Location Type	Public access	Lifecycle	Access control model	Labels
□	big_data_ta	Standard	us-east1 (South Carolina)	Region	Per object	None	Bucket policy & ACLs	

- Command line tool: `gsutil`

Cloud Storage - graphical UI

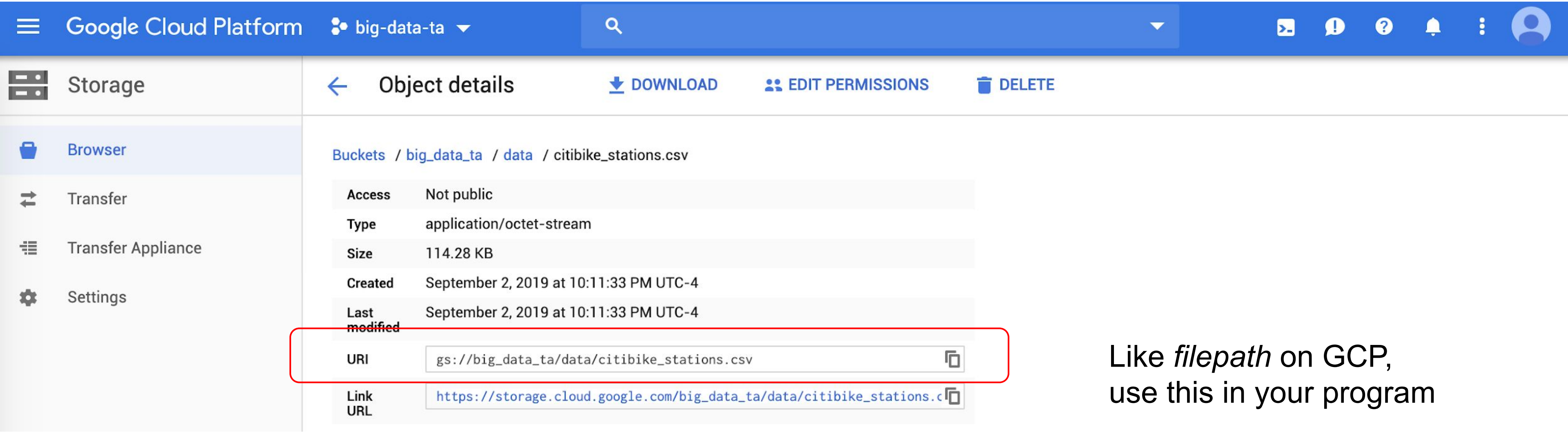


big_data_ta

Buckets / big_data_ta / data

□	Name	Size	Type	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
□	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’)



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

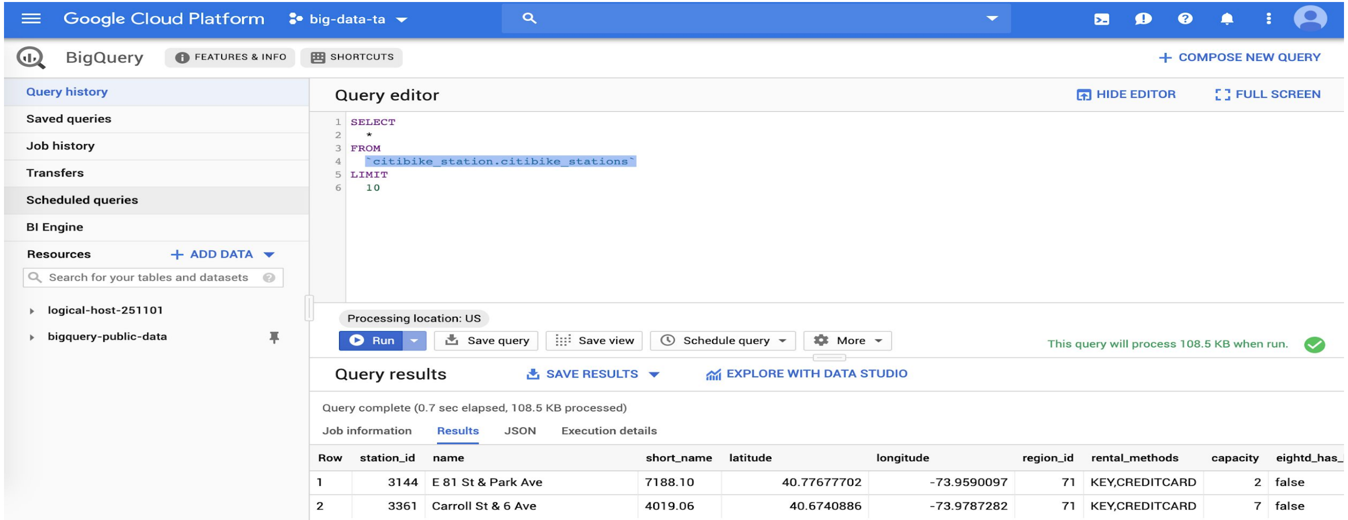
Cloud Storage - `gsutil`

- 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



```
SELECT
  *
FROM
  `citibike_station.citibike_stations`
LIMIT
  10
```

This query will process 108.5 KB when run.

Query complete (0.7 sec elapsed, 108.5 KB processed)

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
    ```

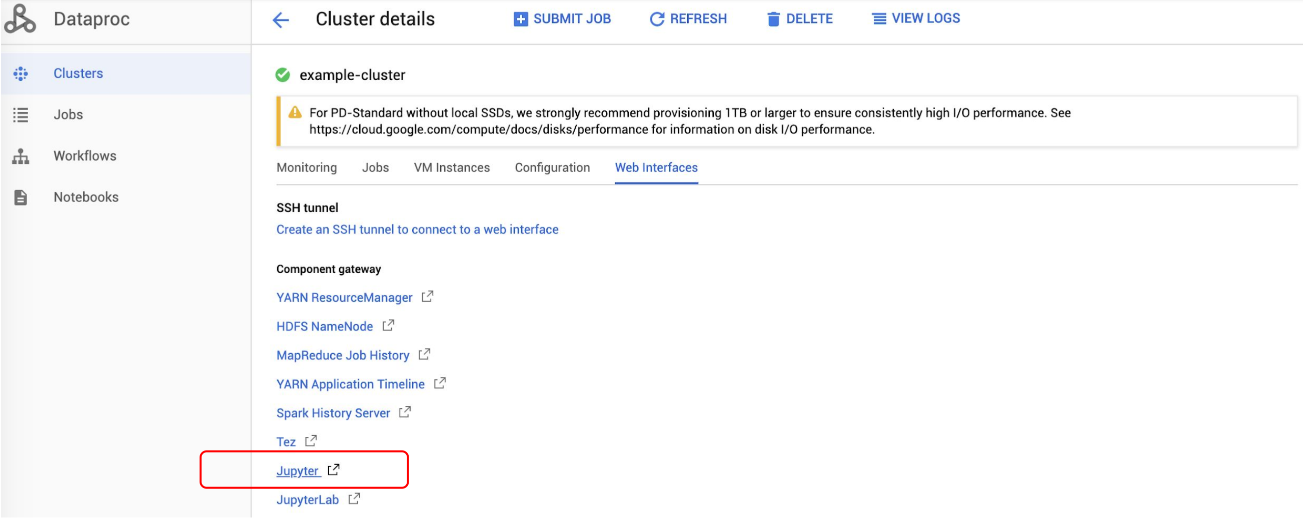
Install Jupyter Notebook

Cloud Storage bucket: where your jupyter notebooks are saved

Works like pip install <your package>

Dataproc - Spark execution / submit jobs

- Jupyter notebook:



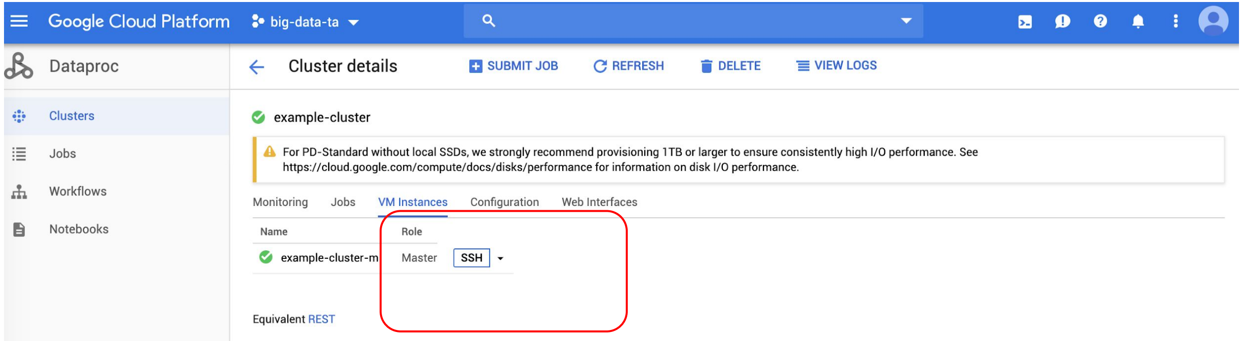
For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.

- 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

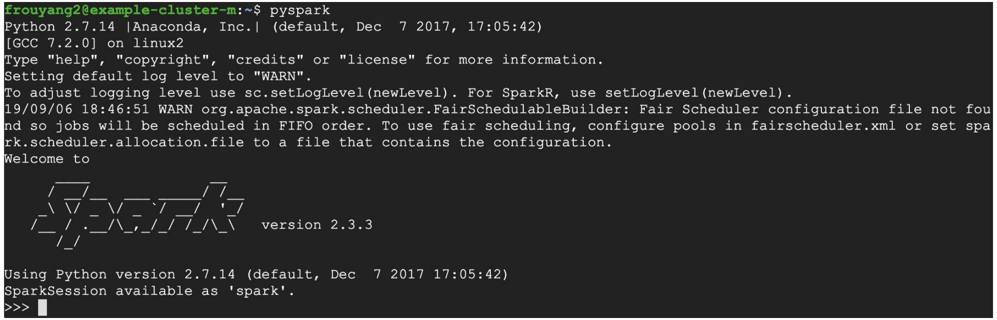


example-cluster

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.

Name	Role	
example-cluster-m	Master	SSH ▾

  - 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 fou
nd so jobs will be scheduled in FIFO order. To use fair scheduling, configure pools in fairscheduler.xml or set spa
rk.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.