CIS 8392
Topics in Big Data Analytics

#Intro to Cloud and GCP

Yu-Kai Lin

Agenda

- What is Cloud?
- What is Google Cloud Platform (GCP)?
- Google Cloud Storage and BigQuery

[**Acknowledgements**] The materials in the following slides are based on the source(s) below:

- An Introduction to GCP for Students
- R interface to Keras

A little history

On Premise

Time-sharing

Cloud

What is Cloud?

Cloud is a model for enabling convenient, **on-demand** network access to **a shared pool of configurable computing resources** (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with **minimal management** effort or service provider interaction

Characteristics of cloud computing

1. Shared / pooled resources
    - Resources are drawn from a common pool
    - Common resources build economies of scale
2. Broadband network access
3. On-demand self-service
4. Scalable and elastic
    - Resources dynamically-allocated between users
    - Additional resources dynamically-released when needed
5. Metered by use

Conventional vs. cloud computing

Conventional

- Dedicated Hardware
- Fixed Capacity
- Pay for Capacity
- Capital & Operational Expenses
- Managed via system admins

Cloud

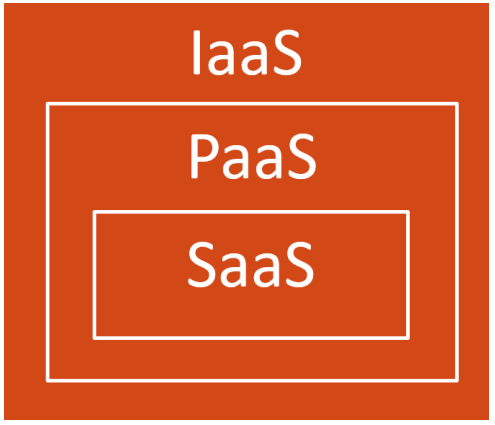
- Shared Hardware
- Elastic Capacity
- Pay for Use
- Operational Expenses
- Managed via APIs

Cloud service models

- **Infrastructure as a Service (IaaS)**
  - Consumers gets access to the infrastructure to deploy their stuff
  - Don’t manage or control the infrastructure
  - Time-sharing
  - E.g., Virtual machines

- **Platform as a Service (PaaS)**
  - Users deploy their applications on a cloud
  - Users control their apps
  - Vendor manages scaling
  - E.g., Remote web-hosting and databases

- **Software as a Service (SaaS)**
  - Use provider's applications over a network
  - E.g., RStudio Cloud



Why use cloud for big data analytics?

- Reduce cost
- Highly scalable resources
- Easy system administration and maintenance
  - OS, database, Hadoop, Spark, ...
  - Allow you to focus on your innovation/creativity
- Containerize other essential applications/software
  - https://console.cloud.google.com/marketplace
  - Rstudio, MongoDB, ...
- Collaborate with multiple team members on large projects
  - End-to-end continuous pipeline

AWS vs. Azure vs. Google

- **Amazon Web Services** – With a vast tool set that continues to grow exponentially, Amazon's capabilities are unmatched. Yet its cost structure can be confusing, and focus on public cloud rather than hybrid cloud or private cloud.

- **Microsoft Azure** – A close competitor to AWS with an exceptionally capable cloud infrastructure. Few companies have the enterprise background (and Windows support) as Microsoft. Hybrid cloud is a true strength.

- **Google Cloud Platform (GCP)** – A well-funded underdog in the competition, Google entered the cloud market later and doesn't have the enterprise focus that helps draw corporate customers. But its technical expertise is profound, and **its industry-leading tools in deep learning and artificial intelligence, machine learning and data analytics** are significant advantages.

Source and more detailed comparison: https://www.datamation.com/cloud-computing/aws-vs-azure-vs-google-cloud-comparison.html

GCP Pricing

It is important to keep in mind the pricing of various GCP services:

- Cloud Storage
- BigQuery
- Deep Learning VM
- Cloud AI and Machine Learning
    - Cloud Vision: First 1K units free (1 unit = 1 image detection feature)
    - Cloud Natural Language: First 5K free (1 unit = 1K characters)
    - Cloud Translation: $20 per 1 million characters
    - Cloud Text-to-Speech: $4.00 USD / 1 million characters
    - Cloud Speech-to-Text: First 60 minutes free

Set up your GCP project

1. Get and redeem the coupon from GCP Education Grants (Remember the Google account to which you received the credits)

2. Create a project in the GCP Console (the manage resources page).

   - Project Name: `CIS8392`
   - Location: `No organization`

3. Make sure that billing is enabled for the project.

   - Verify here: https://console.cloud.google.com/billing/projects

4. Install the Google Cloud SDK if you do not already have it. Choose the quickstart for you operating system.

   - The last screen of the installer should show an option to run `gcloud init`, which will guide you through the initialization.
   - Run `gcloud auth login`, which will open a page on your browser so that you can login your Google account.

Other optional setups based on your needs

Running R at Scale on Compute Engine

1. Enable the Compute Engine API
2. Install Python, if it isn't already installed on the host
3. Install virtualenv
4. Installing and configuring ElastiCluster ElastiCluster

Running RStudio Server on a Cloud Dataproc Cluster

1. Enable the Cloud Dataproc and Cloud Storage APIs
2. Creating a Cloud Dataproc cluster
3. Installing RStudio Server and its dependencies on the master node

Other optional setups based on your needs

Using a Windows VM

1. install a third-party RDP client such as Chrome RDP by FusionLabs.
2. In the GCP Console, go to the VM Instances page and create a virtual machine instance

Connecting to BigQuery from Microsoft Excel using ODBC

1. Downloading the driver
   - Check whether your version of Excel is 32-bit or 64-bit.
   - Download the latest version of the ODBC driver from the Simba Drivers for Google BigQuery page which matches your version of Excel.
   - Run the ODBC driver installer.
2. Configuring the driver
3. Running a query

Google Cloud SDK

Command-line interface for Google Cloud Platform products and services

It contains `gcloud`, `gsutil`, and `bq` command-line tools, which you can use to access Google Compute Engine, Google Cloud Storage, Google BigQuery, and other products and services from the command-line.

- **gcloud command-line tool**: The `gcloud` CLI manages authentication, local configuration, developer workflow, and interactions with the Cloud Platform APIs.
- **gsutil Tool**: `gsutil` provides command line access to manage Cloud Storage buckets and objects.
- **bq Tool**: `bq` allows you to run queries, manipulate datasets, tables, and entities in BigQuery through the command line.

There are web interfaces for these tools. Why learn the command-line?

The gcloud command-line tool

If you follow the Google Cloud SDK installer with the default setting, you should see the following screen pop-up once you finish (which is equivalent to you running `gcloud init` by yourself):

C:\WINDOWS\SYSTEM32\cmd.exe - gcloud init

```
---
Welcome! This command will take you through the configuration of gcloud.

Settings from your current configuration [default] are:
compute:
  region: us-east1
  zone: us-east1-b
core:
  account: yklin82@gmail.com
  disable_usage_reporting: 'True'
  project: cloudml-test-210413

Pick configuration to use:
 [1] Re-initialize this configuration [default] with new settings
 [2] Create a new configuration
Please enter your numeric choice:  2

Enter configuration name. Names start with a lower case letter and
contain only lower case letters a-z, digits 0-9, and hyphens '-':  cis8392
```

Before you can run the `bq` and `gsutil` commands, you also need to run

```
gcloud auth login
```

C:\WINDOWS\SYSTEM32\cmd.exe

```
C:\Users\yukai\AppData\Local\Google\Cloud SDK>bq query "select * from [data-science-getting-
started:preprocessing.meteors] limit 10"
ERROR: (bq) There was a problem refreshing your current auth tokens: invalid_grant: Bad Requ
est
Please run:

  $ gcloud auth login

to obtain new credentials, or if you have already logged in with a
different account:

  $ gcloud config set account ACCOUNT

to select an already authenticated account to use.

C:\Users\yukai\AppData\Local\Google\Cloud SDK> gcloud auth login
Your browser has been opened to visit:
```

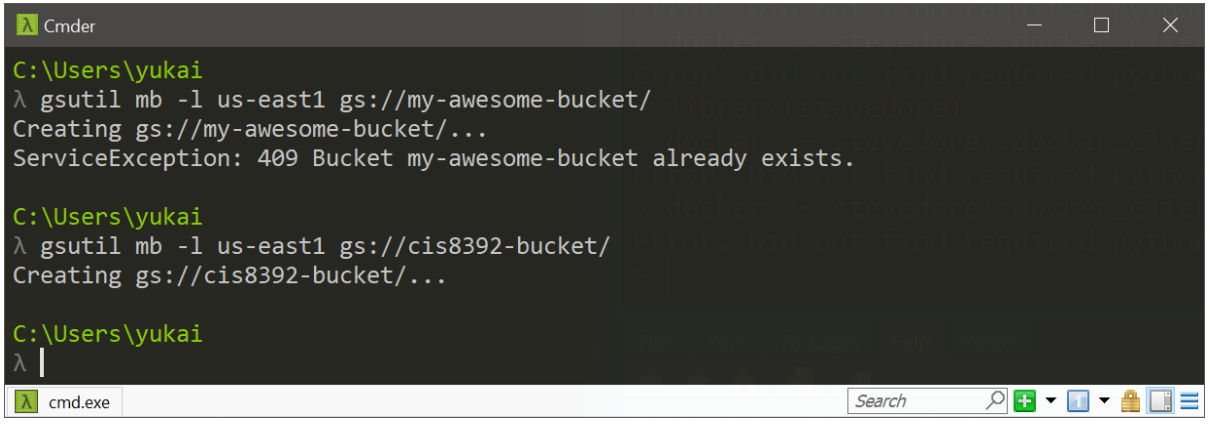
The gsutil command-line tool

Create a bucket

```
gsutil mb -l us-east1 gs://cis8392-bucket/
```

This uses a bucket named "cis8392-bucket." You must choose your own, globally-unique, bucket name. Please use the following name rule:

- `cis8392-[TODAY_in_yyyymmdd]-[you_first_name_lowercase]-bucket`
- In the following slides, be sure to replace the bucket name to yours



Upload an object into your bucket

Suppose you have an image file (you can download it from here)

```
gsutil cp C:/CIS8392/kitten.png gs://cis8392-bucket/
```

Download an object from your bucket

```
gsutil cp gs://cis8392-bucket/kitten.png C:/CIS8392/kitten2.png
```

Note: If the file path has a space, the path needs to be surrounded by quotes:

```
gsutil cp gs://cis8392-bucket/kitten.png "C:/CIS 8392/kitten2.png"
```

Copy an object to a folder in the bucket

```
gsutil cp gs://cis8392-bucket/kitten.png gs://cis8392-bucket/just-a-folder/kitten3.png
```

List contents of a bucket or folder

```
gsutil ls gs://cis8392-bucket
```

List details for an object

```
gsutil ls -l gs://cis8392-bucket/kitten.png
```

Make your object publicly accessible

```
acl = Access Control Lists
```

```
gsutil acl ch -u AllUsers:R gs://cis8392-bucket/kitten.png
```

Note: If an object is publicly accessible, you can download it from your browser (replace `cis8392-bucket` with your bucket name):
`https://storage.googleapis.com/cis8392-bucket/kitten.png`

To remove this permission, use the command:

```
gsutil acl ch -d AllUsers gs://cis8392-bucket/kitten.png
```

Give someone access to your bucket

```
iam = Identity and Access Management
```

```
gsutil iam ch user:jane@gmail.com:objectCreator,objectViewer gs://cis8392-bucket
```

To remove this permission, use the command:

```
gsutil iam ch -d user:sam@gmail.com:objectCreator,objectViewer gs://cis8392-bucket
```

Note: `gsutil` will return an error if the email address does not exist.

Delete objects

```
gsutil rm gs://cis8392-bucket/kitten.png
```

Delete a bucket

```
gsutil rm -r gs://cis8392-bucket
```

Your turn

1. If you haven't already, create a bucket with the following name: `cis8392-[TODAY_in_yyyymmdd]-[your_first_name_lowercase]-bucket`
2. Download the baby names zip file from here and unzip it
3. Upload the `yob2010.txt` file into the bucket
4. List the objects in your bucket
5. Make the `yob2010.txt` file publicly accessible
6. Share the file URL with another student and ask him/her to download the `yob2010.txt` file to his/her local machine

BigQuery

BigQuery is not just a command-line SQL querying tool

- In addition to support for standard SQL, BigQuery also provides some useful additional aggregate functions.
- There's also has an extensive REST API with client libraries for programmatic management and querying of data within BigQuery.
- Other Google Cloud Platform tools integrate with BigQuery, such as
  - loading data from Cloud Storage to BigQuery
  - using BigQuery as a data source and/or sink in a Cloud Dataflow pipeline.
  - performing BigQuery queries directly from a Cloud Datalab notebook
  - creating graphs in Data Studio.

The bq command-line tool

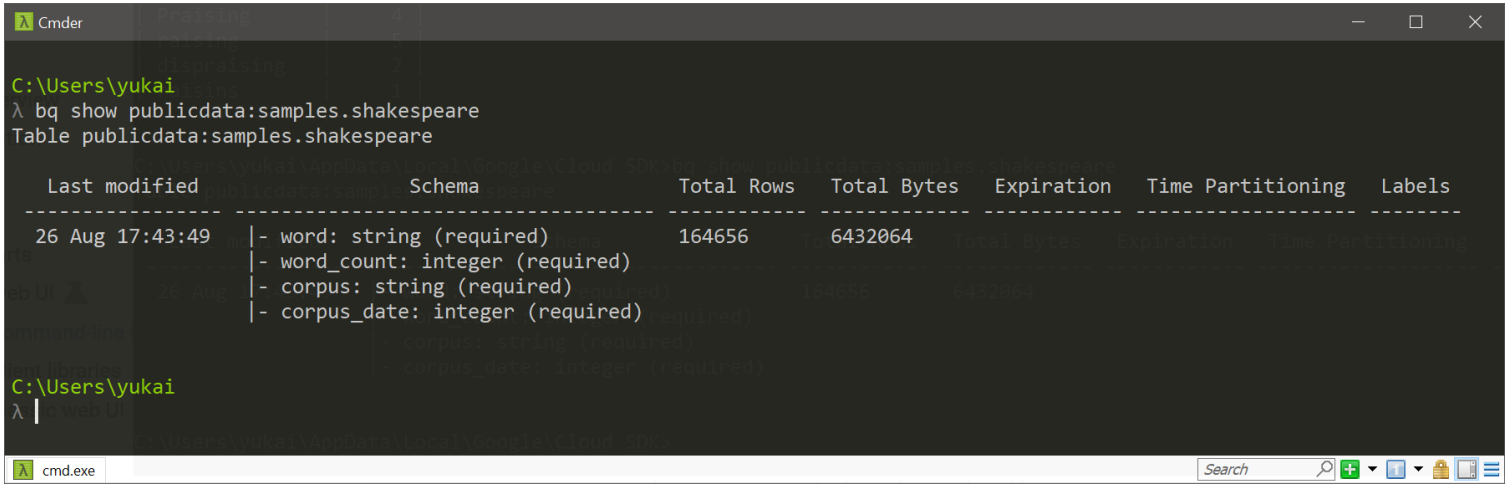
Examine a table

Template:

```
bq show projectId:datasetId.tableId
```

Example:

```
bq show publicdata:samples.shakespeare
```



Run a query

Template:

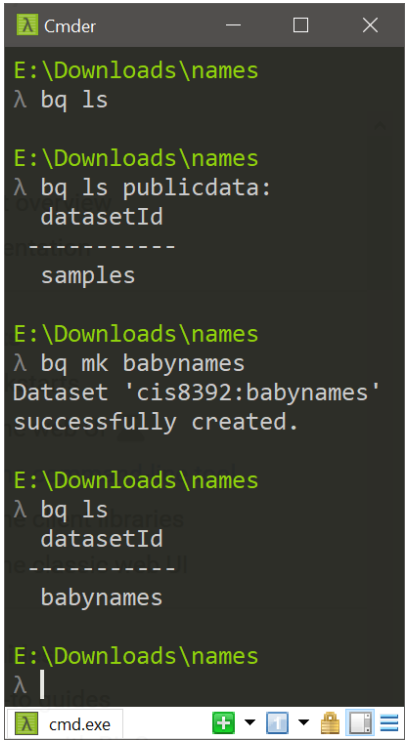
```
bq query "[SQL_STATEMENT]"
```

Example (IMPORTANT: Although I put the command in multiple lines on the slide, you should put it in one single line when you run it; otherwise, you will get an error):

```
bq query "SELECT word, SUM(word_count) as count
FROM publicdata:samples.shakespeare
WHERE word CONTAINS 'raisin' GROUP BY word"
```

Create and load a new table

1. Use the `bq ls` command to see whether your default project has any existing datasets
2. Run `bq ls` again to list the datasets in a specific project by including the project ID followed by a colon (:).
3. Use the `bq mk` command to create a new dataset
4. Run the `bq load` command to load your source file into a new table
   - Download the baby names zip file from here and unzip it
   - Copy or move the `yob2010.txt` file into the directory you are using to run bq commands.



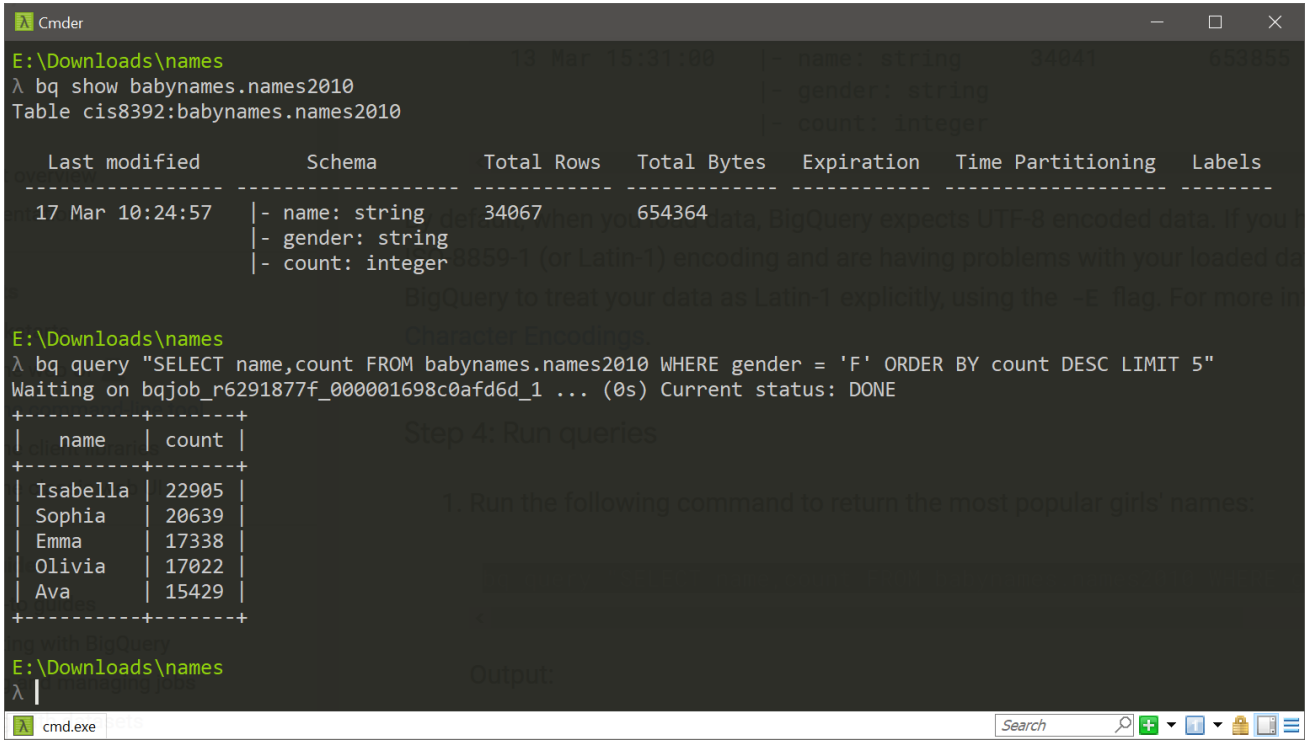

```
bq mk babynames

# THE FOLLOWING COMMAND IS IN A SINGLE LINE
bq load babynames.names2010 yob2010.txt
 name:string,gender:string,count:integer
```

Run `bq show` to see the schema:

```
bq show babynames.names2010
```

Run queries (again, remember to put the command in a single line):

```
bq query "SELECT name,count FROM babynames.names2010
WHERE gender = 'F' ORDER BY count DESC LIMIT 5"
```



Remove dataset

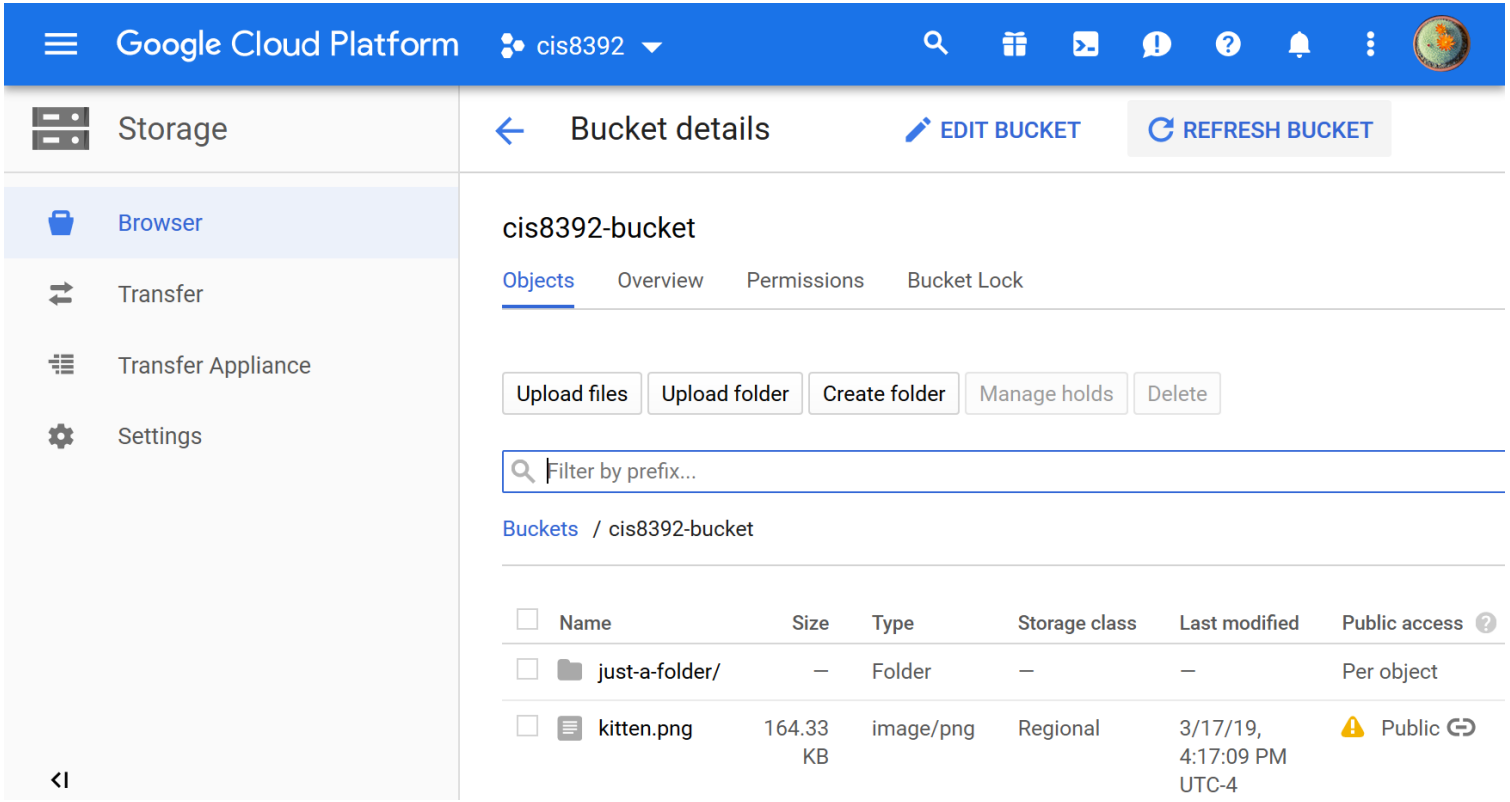
1. Run the `bq rm` command to remove the babynames dataset. Use the `-r` flag to delete all tables in the dataset, include the `names2010` table.
2. Confirm the delete command by typing `y`.

Your turn

1. Create a dataset `babynames2`
2. Load the following data into the dataset `babynames2`:
   - `yob1880.txt`, `yob1881.txt`, ..., `yob1889.txt`
3. Write a query to find out the most popular baby names in 1880s
4. Run the `bq query` command and compare your results with another student

The Cloud Storage web UI

https://console.cloud.google.com/storage



The BigQuery web UI

https://console.cloud.google.com/bigquery