08. Google Cloud Storage

Telung Pan Ph.D.
telung@mac.com

Unified Object Storage

- Google Cloud Storage is unified object storage.
- Live data serving, data analytics/ML to data archiving.
- Coldline: cold storage - data to touch less than once a year.
- Nearline: data expect to access less than once a month

Cloud Storage Features

- Durable
- Available
- Scalable
- Consistent

Creating Storage Buckets

- gcloud auth login

- gsutil mb gs://my-awesome-bucket/

Upload an object into your bucket

- gsutil cp Downloads/cloud-storage-logo.png gs://my-awesome-bucket

Download an object from your bucket

- gsutil cp gs://my-awesome-bucket/cloud-storage-logo.png Desktop

Copy an object to a folder in the bucket

- Use the gsutil cp command to create a folder and copy the icon into it:
gsutil cp gs://my-awesome-bucket/cloud-storage-logo.png gs://my-awesome-bucket/just-a-folder/

List contents of a bucket or folder

- Use the gsutil ls command to list the contents of the bucket:
  gsutil ls gs://my-awesome-bucket

- List details for an object
  Use the gsutil ls command, with the -l flag to get some details about an object:
  gsutil ls -l gs://my-awesome-bucket/cloud-storage-logo.png

Make your object publicly accessible

- Use the gsutil acl ch command to grant all users read permission for the object stored in your bucket:
  gsutil acl ch -u AllUsers:R gs://my-awesome-bucket/cloud-storage-logo.png

- To remove this permission, use the command:
  gsutil acl ch -d AllUsers gs://my-awesome-bucket/cloud-storage-logo.png

Give someone access to your bucket

- Use the gsutil acl ch command to give a specific email address read and write permission for your bucket:

gsutil acl ch -u user@gmail.com:W gs://my-awesome-bucket

gsutil acl ch -d user@gmail.com gs://my-awesome-bucket

Delete objects

- Use the gsutil rm command to delete an object:
  gsutil rm gs://my-awesome-bucket/cloud-storage-logo.png

- Use the gsutil rm command with the -r flag to delete the bucket and anything inside of it:
  gsutil rm -r gs://my-awesome-bucket

Using Google Cloud Storage with Big Data

"Spotify chose Google in part because its services for analyzing large amounts of data ... are more advanced than data services from other cloud providers."

— Wired, February 2016

Using Google Cloud Storage with Big Data

- Google Cloud Storage is a key part of storing and working with Big Data on Google Cloud Platform. Examples include:

  - Loading data into Google BigQuery.

  - Configuring Cloud Storage as the default file system for Hadoop on Google Cloud Platform.

  - Using a bucket to hold staging files and temporary data for Google Cloud Dataflow pipelines.

  - Using a bucket to import your genomic data to a Google Genomics dataset.

Copying many files to a bucket

- If you have a large number of files to upload you can use the gsutil -m option, to perform a parallel (multi-threaded/multi-processing) copy. To recursively copy subdirectories, use the -R flag of the cp command. For example, to copy files including subdirectories from a local directory named top-level-dir to a bucket, you can use:
  gsutil -m cp -R top-level-dir gs://example-bucket

- You can use wildcards to match a specific set of names for an operation. For example, to copy only files that start with image:
  gsutil -m cp -R top-level-dir/subdir/image* gs://example-bucket

- You can remove files using the same wildcard:
  gsutil -m rm gs://example-bucket/top-level-dir/subdir/image*

- In addition to copying local files to the cloud and vice versa, you can also copy in the cloud, for example:

  gsutil -m cp gs://example-bucket/top-level-dir/subdir/ ** gs://example-bucket/top-level-dir/subdir/subdir2

- gsutil automatically detects that you're moving multiple files and creates them in a new directory named subdir2.

Analyzing Entities

- Entity Analysis inspects the given text for known entities (proper nouns such as public figures, landmarks, etc.), and returns information about those entities.

- To perform entity analysis, use the gcloud command line tool and use the --content flag to identify the content to analyze:

  gcloud ml language analyze-entities --content="President Obama is speaking at the White House."

Analyzing Entities in a Remote File

- For your convenience, the Natural Language API can perform entity analysis directly on a file located in Google Cloud Storage, without the need to send the contents of the file in the body of your request.

  gcloud ml language analyze-entities --content-file=gs://YOUR_BUCKET_NAME/YOUR_FILE_NAME