GCP Project

Account Setup

1. Go to cloud.google.com. Create a Google Account/Sign-In with a Google Account.
2. Towards the upper-left side, near the Google Cloud logo, Click the Box.
3. A dialog box will open, click the box in the upper-right corner, “New Project.”

Storage Bucket Setup

1. Select the navigation menu (hamburger icon) in the upper right corner; Hover over Cloud Storage; Select Buckets
2. Enter the following information: Storage Bucket name, storage location, storage class (default), storage control access (e.g., public access prevention, and uniform access control), storage object data (e.g., soft delete policy, object versioning, bucket retention policy, object retention, encryption type). Click create.

Compute Engine API Setup

1. Select the navigation menu (hamburger icon) in the upper right corner: hover over APIs and Services; Select “Enable APIs & Services”
2. Select “+ ENABLE APIS AND SERVICES”
3. Scroll down until you find “Compute Engine API” under Google Enterprise APIs; Select “Compute Engine API”.
4. Select “Enable” on the Compute Engine API product details page.

DataProc Setup

1. Select the hamburger icon in the upper right corner; Hover over Dataproc; Click/Select Dataproc
2. You will be directed to the Cloud Dataproc API product details page. Select “Enable” to enable Dataproc API – This will enable Hadoop Clusters for big data.

DataProc Cluster Creation

1. Directed to Dataproc Cluster Page. Select CREATE CLUSTER
2. Select CREATE to Create a Cluster on the Compute Engine
3. Enter the following information: Name, Select Region, Zone, Cluster type: Standard (1 master, N workers)
4. Select Change to change the release date. This allows you to change the version images to a specific OS bundle of your choice. For this project, we will be sticking with Debian 12; however, Ubuntu 12 or Rocky Linux are good alternatives.
5. Select Configure nodes in the left-hand menu; Create a Dataproc cluster and its manager and worker nodes on the compute engine using the following information:
   1. Crete a name
   2. Under the Machine family GENERAL –PURPOSE tab
   3. Select Series: E2 (for basic).
   4. Select Machine type: e2-standard-4 (4 vCPUs, 16 GB memory) as standard or basic computation.
   5. Enter Primary disk size 128 GB
   6. Select Primary disk type: Standard Persistent Disk
   7. Scroll down to the worker nodes
6. Select Configure nodes in the left-hand menu; Create a Dataproc cluster and its manager and worker nodes on the compute engine using the following information:
   1. Crete a name
   2. Under the Machine family GENERAL –PURPOSE tab
   3. Select Series: E2 (for basic).
   4. Select Machine type: e2-standard-4 (4 vCPUs, 16 GB memory) as standard or basic computation.
   5. Enter Primary disk size 128 GB
   6. Select Primary disk type: Standard Persistent Disk
   7. Scroll down to the worker nodes
7. Select Customize cluster
8. Scroll down to “Cloud Storage staging bucket” and Select Browse.
9. Find the storage bucket you created [storage bucket name]
10. Select your bucket and then Click SELECT
11. After you’ve verified that you can see your storage bucket, select CREATE. Wait for the provisioning of your cluster.
12. Your new Hadoop cluster HAS BEEN CREATED SUCCESSFULLY and IT IS RUNNING.

NOTE: If you intend to use your cluster immediately after creation, then allow it to run, but if not, stop the cluster by ticking the box next to the cluster and selecting “STOP.”

Running VM Instances

1. Select the Navigation Menu (hamburger icon menu), scroll down to Compute Engine and Select VM instances
2. Select the arrow next to SSH of the manager node to expand a drop-down menu
3. Select “Open in a browser window”
4. Select Authorize
   1. Note: It takes several seconds for the authentication to be done so wait patiently
5. After authentication, an SSH terminal will pop up in the browser.
   1. Note: Utilizing GCP documentation will allow you to become knowledgeable about how to use OpenSSH, PuTTY app, or Secure Shell Chrome App to SSH GCP VM into your local network.

BigQuery

1. Enable BigQuery API Verify BigQuery API is enabled; if not, you need to enable the API now. Type in BigQuery API in the search bar and click on BigQuery API under Marketplace results (NOT Documentation results). Then click Enable if it is not enabled.
2. Step 2: Find the public dataset 2.1 Go to the Navigation Menu and Select BigQuery and BigQuery Studio. If a Welcome screen opens, read it, and select DONE. Next, in the Type to search field, enter a desired public data and ENTER or click SEARCH ALL PROJECTS if nothing displays or “ + ADD” next to “Explorer to add your dataset. For this project, I will connect the storage bucket data.
3. Out of Local File, Google Cloud Storage, and Connections to external data sources, Select “Google Cloud Storage.” If the dataset you want to use is somewhere else, scroll down and find the appropriate selection.
4. In the Create table menu under Source:
   1. Select Google Cloud Storage from Create Table Form (or wherever your dataset is based)
   2. In the “Select file from GCS bucket,” select Browse.
      1. Select the desired storage bucket from the storage bucket file selection (i.e., Elden-ring-ultimate-bucket)
      2. Choose the desired individual file from the storage bucket (e.g., bosses.csv)
      3. Click Select
   3. For file format, select the format according to your dataset. I selected CSV.
5. In the Create Table menu under Destination:
   1. Leave the name in Project the same to create consistency.
   2. Select dataset, open Create Dataset menu.
      1. Enter dataset ID/name
      2. Select region (typically multi-region)
      3. Leave Default table expiration and tags alone for the moment.
      4. Make sure Google Encryption is selected under advanced options
      5. Click Create Dataset
      6. Enter Table name (like dataset to keep consistency) (e.g., elden\_ring\_bosses.
      7. Since we are using a dataset from Google Cloud Storage, choose Native Table under Table type. If we were using an external dataset, we would’ve chosen External Table.
6. In the Create Table menu under Schema, Select Auto-detect, unless you want to fine-tune the dataset’s schema.
7. Leave Partition and cluster settings, Tags, and Advanced Options alone (advanced topics for a later date).
8. Select Create Table.
9. The table is now viewable and can be edited or queried
   1. NOTE: TO add another dataset to the project, repeat the same steps.

Exploring and Querying BigQuery Datasets (Coming Soon with SQL Tutorial)