

What you need to do

- You need to enable “Dataflow API” on GCP & other required APIs
- You need to grant me (hlee84@usfca.edu) “admin access” to your entire GCP Project
- You need to create a “bucket” in Google Cloud Storage (which is needed by Dataflow pipelines to “stage” necessary resource files).
- Finally, you can run your first Dataflow job on GCP afterwards.

Step-by-step guide.

TASK A: Grant me admin (Owner) access to your GCP project.

1. Go to your GCP “IAM” Dashboard page (replace the last URL parameter below to your GCP Project ID; mine is **beer-spear**, for instance).

<https://console.cloud.google.com/iam-admin/iam?project=beer-spear>

You’ll see something like this:

The screenshot shows the Google Cloud Platform console interface for the IAM & Admin section of the 'beer-spear' project. The left sidebar contains navigation links for IAM, Identity & Organization, Policy Troubleshooter, Organization Policies, Quotas, Service Accounts, Labels, Settings, Privacy & Security, Cryptographic Keys, Identity-Aware Proxy, Roles, and Audit Logs. The main content area is titled 'Permissions for project "beer-spear"' and includes a note that these permissions affect the project and all its resources. Below this, there are tabs for 'MEMBERS' and 'ROLES', with 'MEMBERS' currently selected. A 'Filter table' is visible above a table listing the project's members. The table has columns for 'Type', 'Member', 'Name', 'Role', 'Over granted permissions', and 'Inheritance'. The members listed include several service accounts (e.g., 748087422216-compute@developer.gserviceaccount.com, 748087422216@cloudservices.gserviceaccount.com) and two human users (haden.lee@moloco.com and hlee84@usfca.edu). The roles assigned are primarily 'Editor' and 'Owner'.

Type	Member	Name	Role	Over granted permissions	Inheritance
<input type="checkbox"/>	748087422216-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor	1501/1514	
<input type="checkbox"/>	748087422216@cloudservices.gserviceaccount.com	Google APIs Service Agent	Editor		
<input type="checkbox"/>	haden.lee@moloco.com		Owner	1513/1679	
<input type="checkbox"/>	hlee84@usfca.edu	Hooyeon Lee	Owner	1464/1679	
<input type="checkbox"/>	service-748087422216@compute-system.iam.gserviceaccount.com	Compute Engine Service Agent	Compute Engine Service Agent		
<input type="checkbox"/>	service-748087422216@container-engine-robot.iam.gserviceaccount.com	Kubernetes Engine Service Agent	Kubernetes Engine Service Agent		
<input type="checkbox"/>	service-748087422216@containerregistry.iam.gserviceaccount.com	Google Container Registry Service Agent	Editor		
<input type="checkbox"/>	service-748087422216@dataflow-service-producer-prod.iam.gserviceaccount.com	Cloud Dataflow Service Account	Cloud Dataflow Service Agent		

2. Click on “+ Add” Icon at the top. It'll open up a new panel on the right.

Add my email address (hlee84@usfca.edu), and select “Project-Owner” role.

This means I can access everything in your GCP project. I need this permission to let me grading system access your Dataflow/BigQuery/GCS stuff. Otherwise, the grading system will not be able to grade your future labs/projects.

Make sure you “send notification email” so that I can follow up and verify on my end (see the next image).

The screenshot shows the Google Cloud Platform IAM Admin console for the project 'beer-spear'. The left sidebar contains navigation links for IAM & Admin, IAM, Identity & Organization, Policy Troubleshooter, Organization Policies, Quotas, Service Accounts, Labels, Settings, Privacy & Security, Cryptographic Keys, Identity-Aware Proxy, and Roles. The main panel displays 'Permissions for project "beer-spear"' with a table of members. The 'ADD' button is visible at the top. On the right, the 'Add members to "beer-spear"' dialog is open, showing the 'New members' field with 'hlee84@usfca.edu' entered. The 'Role' dropdown is set to 'Owner', and the 'Send notification email' checkbox is checked. The 'SAVE' button is highlighted.

Type	Member
<input type="checkbox"/>	748087422216-compute@developer.gserv
<input type="checkbox"/>	748087422216@cloudservices.gserviceacc
<input type="checkbox"/>	haden.lee@molocoads.com
<input type="checkbox"/>	hlee84@usfca.edu
<input type="checkbox"/>	service-748087422216@compute-system.iam.gserviceaccount.com

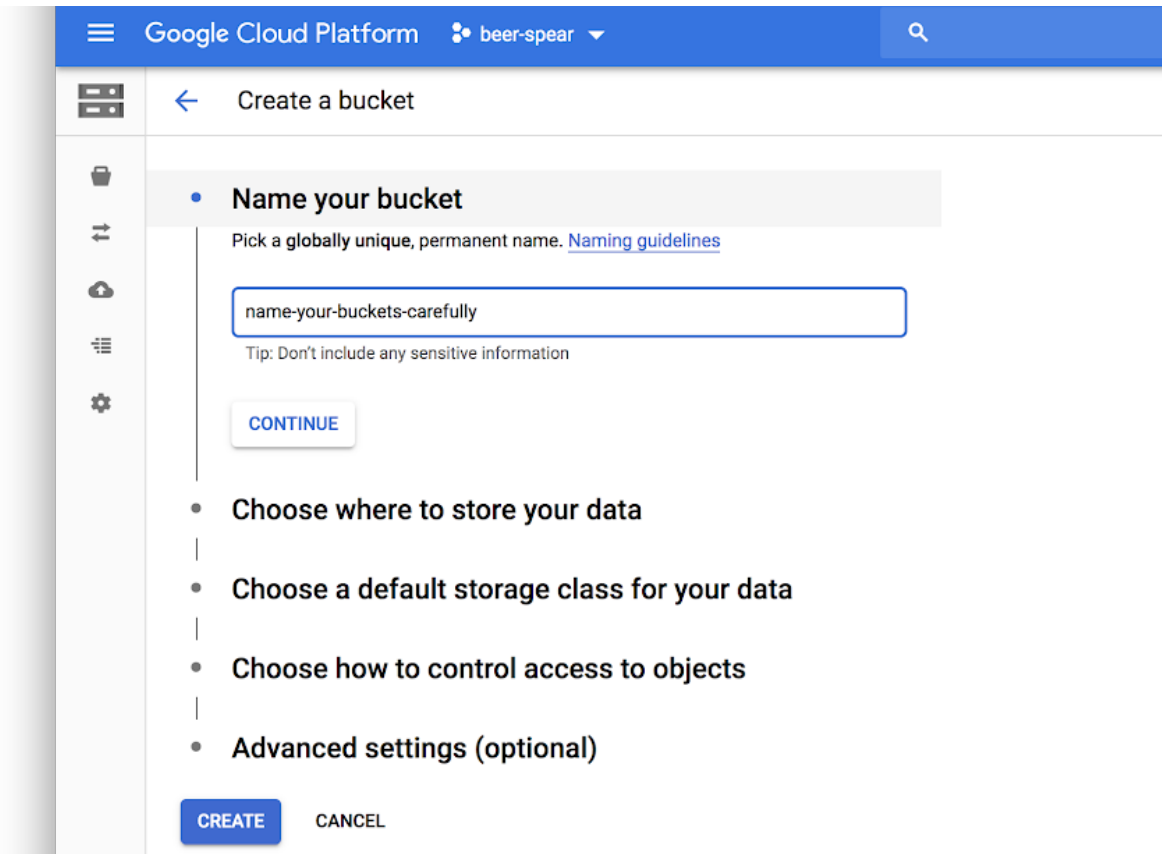
This screenshot shows the same Google Cloud IAM Admin console interface, but with the 'Select a role' dropdown menu open. The dropdown lists various roles, including Project, Access Approval, Access Context Manag..., Actions, AI Notebooks, Android Management, Apigee, and App Engine. The 'Owner' role is highlighted, showing its description: 'Full access to all resources.' The 'SAVE' button remains highlighted.

Type	Member
<input type="checkbox"/>	748087422216-compute@developer.gserv
<input type="checkbox"/>	748087422216@cloudservices.gserviceacc
<input type="checkbox"/>	haden.lee@molocoads.com
<input type="checkbox"/>	hlee84@usfca.edu
<input type="checkbox"/>	service-748087422216@compute-system.iam.gserviceaccount.com
<input type="checkbox"/>	service-748087422216@container-engine-robot.iam.gserviceaccount.com
<input type="checkbox"/>	service-748087422216@containerregistry.iam.gse

TASK B: Create your GCS Bucket.

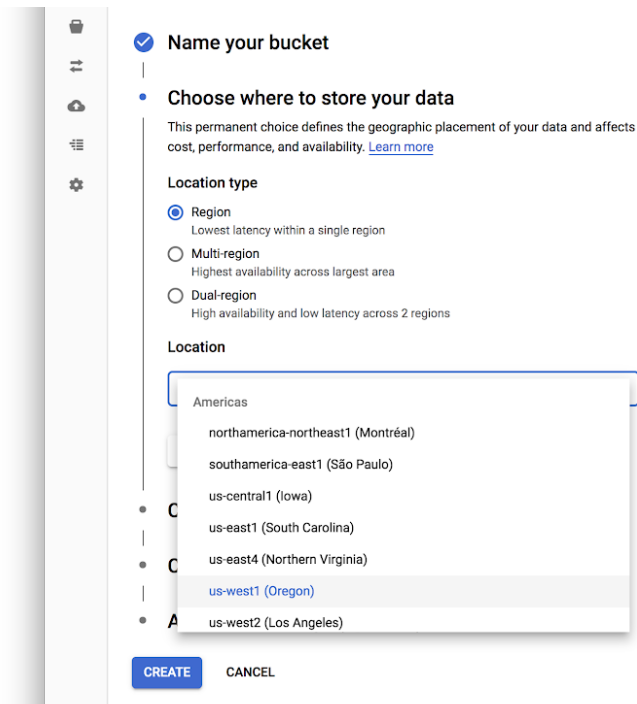
This will be used for ALL of your future labs & projects. You can create *many* buckets, so do not worry about bucket names, etc.

1. Visit <https://console.cloud.google.com/storage/>
2. Click on “Create a Bucket” button, and **CAREFULLY** name your bucket (I’m kidding; it doesn’t matter).



The screenshot shows the Google Cloud Platform console with the 'Create a bucket' wizard. The top navigation bar includes the Google Cloud Platform logo, the user 'beer-spear', and a search icon. The left sidebar contains navigation icons. The main content area is titled 'Create a bucket' and shows the first step: 'Name your bucket'. It prompts the user to 'Pick a globally unique, permanent name' and provides a text input field containing 'name-your-buckets-carefully'. A tip below the field says 'Tip: Don't include any sensitive information'. A 'CONTINUE' button is visible. Below the input field, a list of steps is shown: 'Name your bucket' (current), 'Choose where to store your data', 'Choose a default storage class for your data', 'Choose how to control access to objects', and 'Advanced settings (optional)'. At the bottom, there are 'CREATE' and 'CANCEL' buttons.

3. Choose the “**Region**” radio button, and then choose the “**us-west1**” from Dropdown menu.
(Why? Region is cheapest, and “us-west1” is what we’ll be using to reduce latency between your machine & GCP data center.)



The screenshot shows the second step of the 'Create a bucket' wizard: 'Choose where to store your data'. It explains that this choice defines the geographic placement of data and affects cost, performance, and availability. Under 'Location type', three options are listed: 'Region' (selected with a radio button), 'Multi-region', and 'Dual-region'. Below this, the 'Location' dropdown menu is open, showing a list of regions. The 'us-west1 (Oregon)' region is highlighted. At the bottom, there are 'CREATE' and 'CANCEL' buttons.

4. Choose “Standard” for the storage option (feel free to check out other options, BUT you should use “Standard” because it’s most reasonable for our purposes).

- **Choose a default storage class for your data**

A storage class sets costs for storage, retrieval, and operations. Pick a default storage class based on how long you plan to store your data and how often it will be accessed. [Learn more](#)

☒ **Standard** 

Best for short-term storage and frequently accessed data

☐ **Nearline**

Best for backups and data accessed less than once a month

☐ **Coldline**

Best for disaster recovery and data accessed less than once a quarter

☐ **Archive**

Best for long-term digital preservation of data accessed less than once a year

[CONTINUE](#)

5. Choose “Fine-grained” and “Google-managed key” next. This won’t matter as much.

- ✓ **Choose a default storage class for your data**

- **Choose how to control access to objects**

Access control

☒ **Fine-grained**

Specify access to individual objects by using object-level permissions (ACLs) in addition to your bucket-level permissions (IAM). [Learn more](#)

☐ **Uniform**

Ensure uniform access to all objects in the bucket by using only bucket-level permissions (IAM). This option becomes permanent after 90 days. [Learn more](#)

[CONTINUE](#)

- **Advanced settings (optional)**

Encryption

☒ **Google-managed key**

No configuration required

☐ **Customer-managed key**

Manage via Google Cloud Key Management Service

Retention policy

Set a retention policy to specify the minimum duration that this bucket's objects must be protected from deletion or modification after they're uploaded. You might set a policy to address industry-specific retention challenges. [Learn more](#)

☐ Set a retention policy

Labels

Labels are key:value pairs that allow you to group related buckets together or with other Cloud Platform resources. [Learn more](#)

[+ ADD LABEL](#)

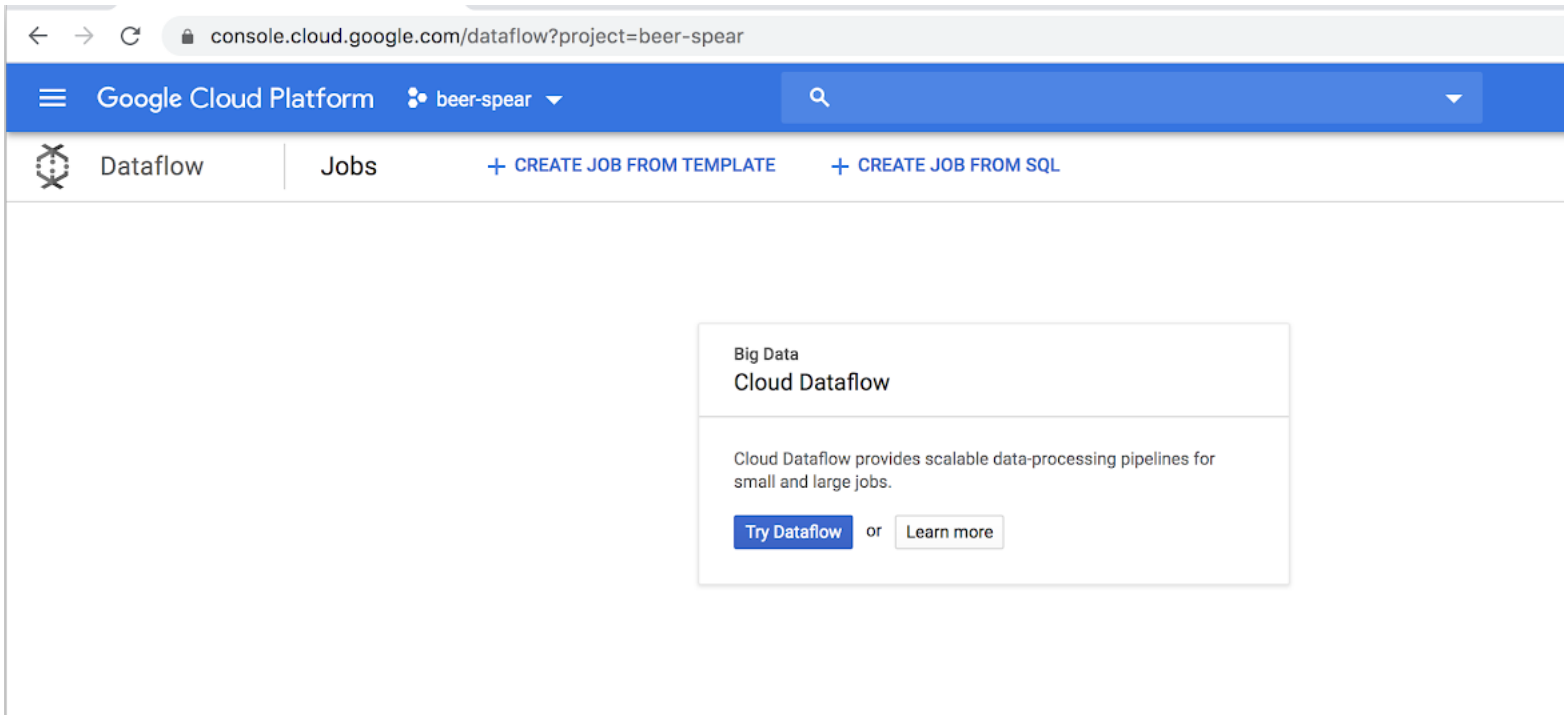
[CREATE](#)

[CANCEL](#)

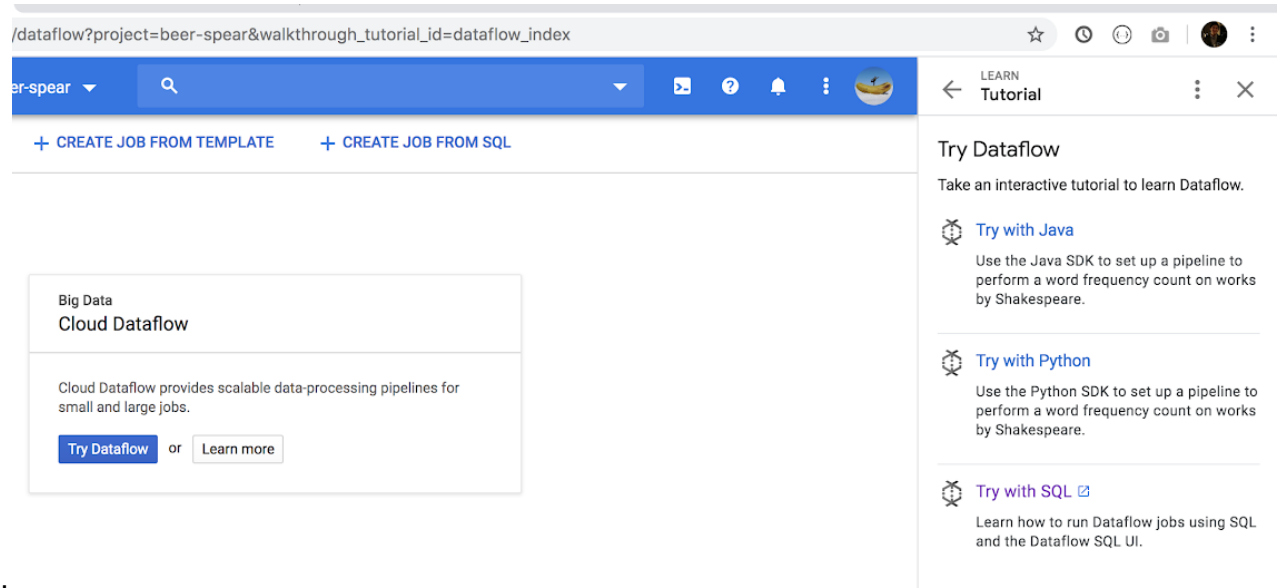
6. Create it! Now, upload the “[sample.file](#)” that was provided along with this guide. That way, I can also verify (on my end) that your setup is correct. (Only after you grant me the Owner access, though).

TASK C: Enable Dataflow API on GCP.

1. Visit <https://console.cloud.google.com/dataflow?project=beer-spear> (replace with your project)



2. Click on “Try Dataflow” and click on “Try with Java”.



3. It will ask you to “Enable APIs” which you must do.

The screenshot shows the Google Cloud Dataflow console interface. The main content area on the left features a 'Big Data Cloud Dataflow' card with the text 'Cloud Dataflow provides scalable data-processing pipelines for small and large jobs.' and a 'Try Dataflow' button. The right sidebar contains a 'LEARN Tutorial' section with the following steps:

- Set up Cloud Dataflow**
To use Dataflow, turn on the Cloud Dataflow APIs and open the Cloud Shell.
- Turn on Google Cloud APIs**
Dataflow processes data in many GCP data stores and messaging services, including BigQuery, Google Cloud Storage, and Cloud Pub/Sub. Enable the APIs for these services to take advantage of Dataflow's data processing capabilities.
[Enable APIs](#)
- Open the Cloud Shell**
Cloud Shell is a built-in command line tool for the console. You're going to use Cloud Shell to deploy your app.
Open Cloud Shell by clicking the [Activate Cloud Shell](#) button in the navigation bar in the upper-right corner of the console.

4. Once you click on “Enable APIs”, it will take a few seconds to enable all of those APIs. Once the check marks are all green, proceed to the next steps to complete the tutorial (you don't have to follow all of their instructions; just click on “next” several times).
5. Once this step is completed, check out the code from github, and execute the main method.
6. If everything works fine, then you should be able to see your job up and running on this website (Dataflow Web Console). Otherwise, you'll probably see warnings/errors on your console (where you executed your Main).

Feel free to share on Piazza screenshots & error messages you get along the way; also, make your posts public so that the teaching staff do not have to answer the same question multiple times.

Once you are done with all Tasks A-C, then at some point it'll be verified automatically by the grading system (it'll notify you, so don't worry).