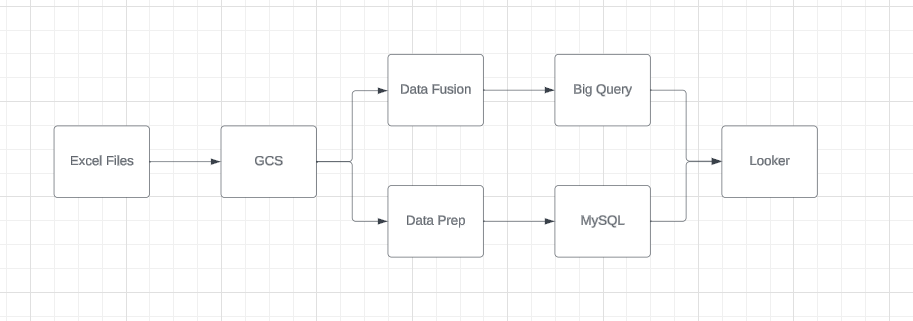
**Data Analytics Pipeline 2**

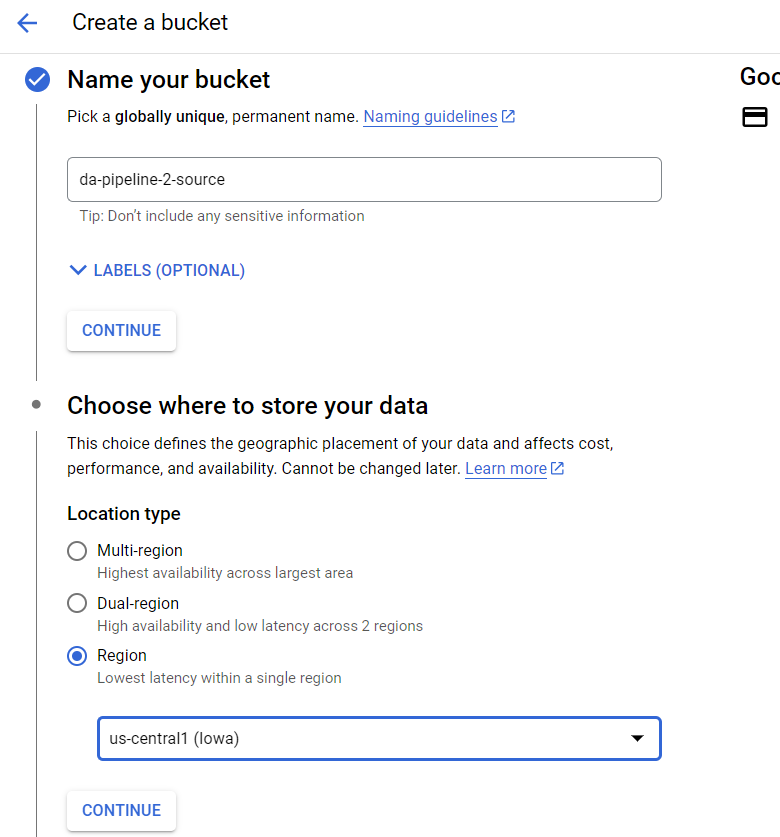
Flow of Pipeline:

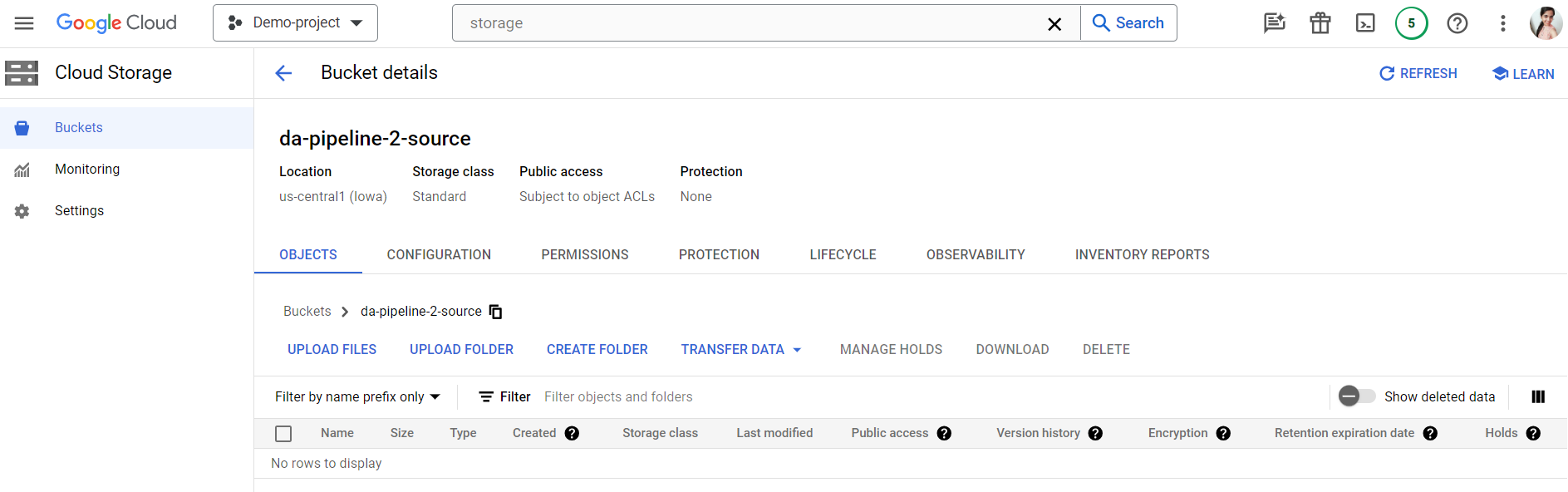


Steps:

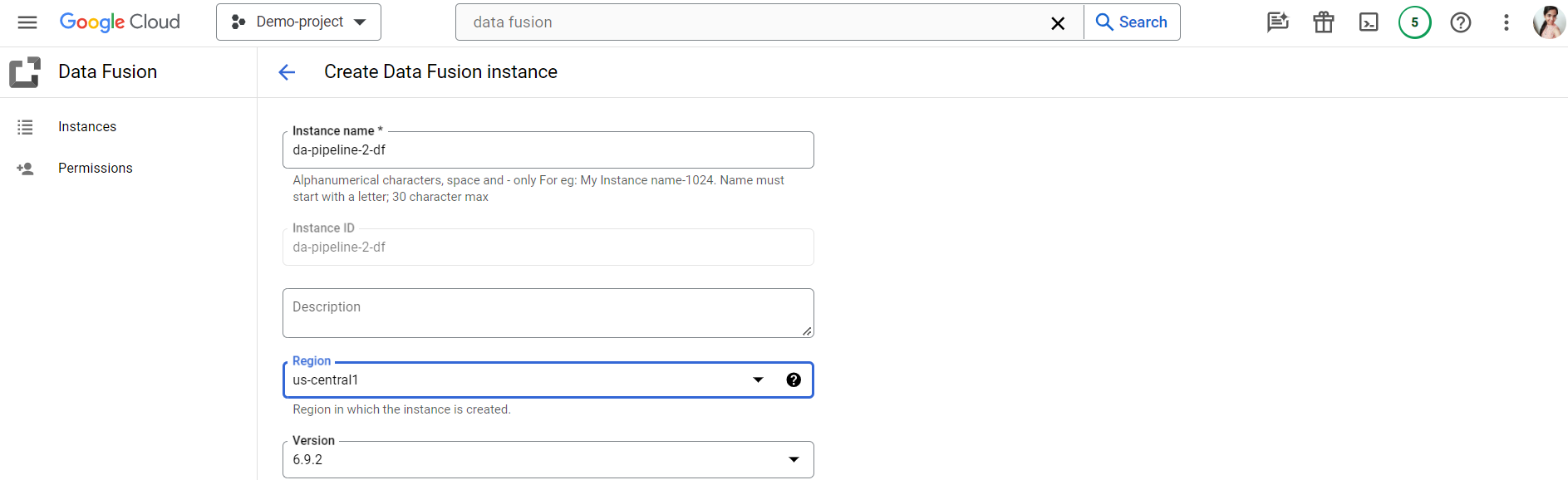
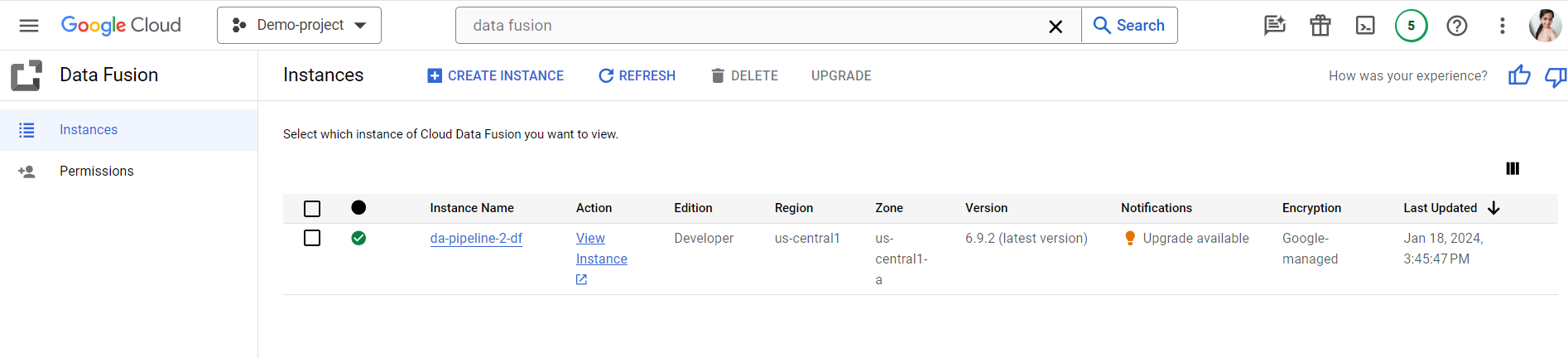
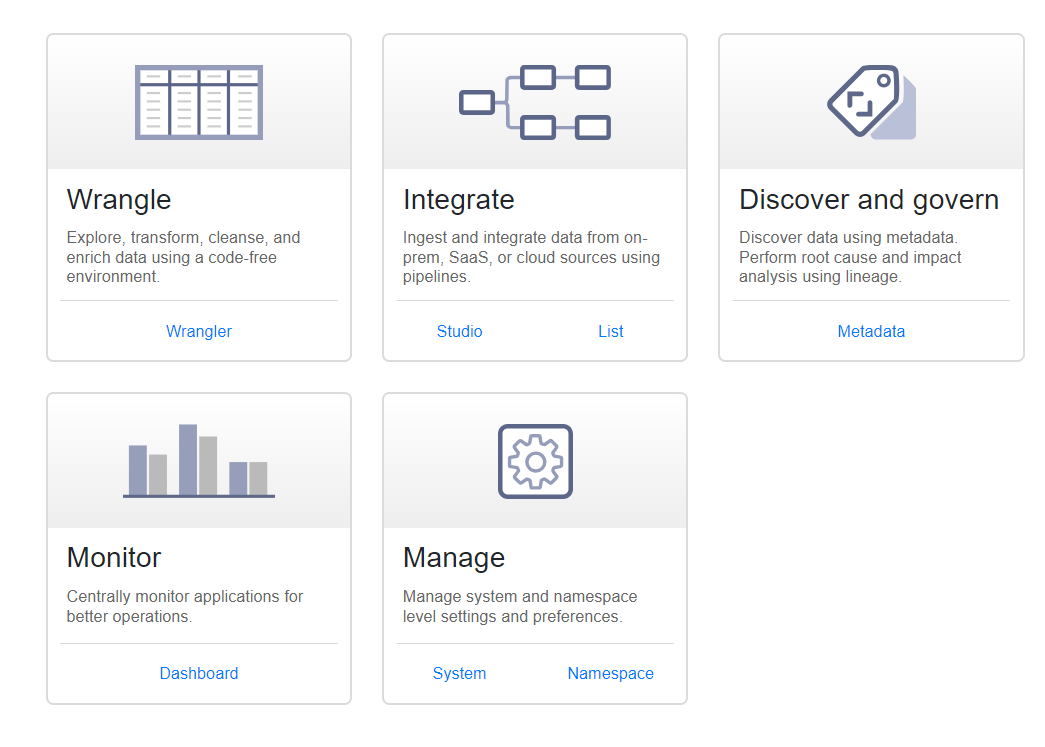
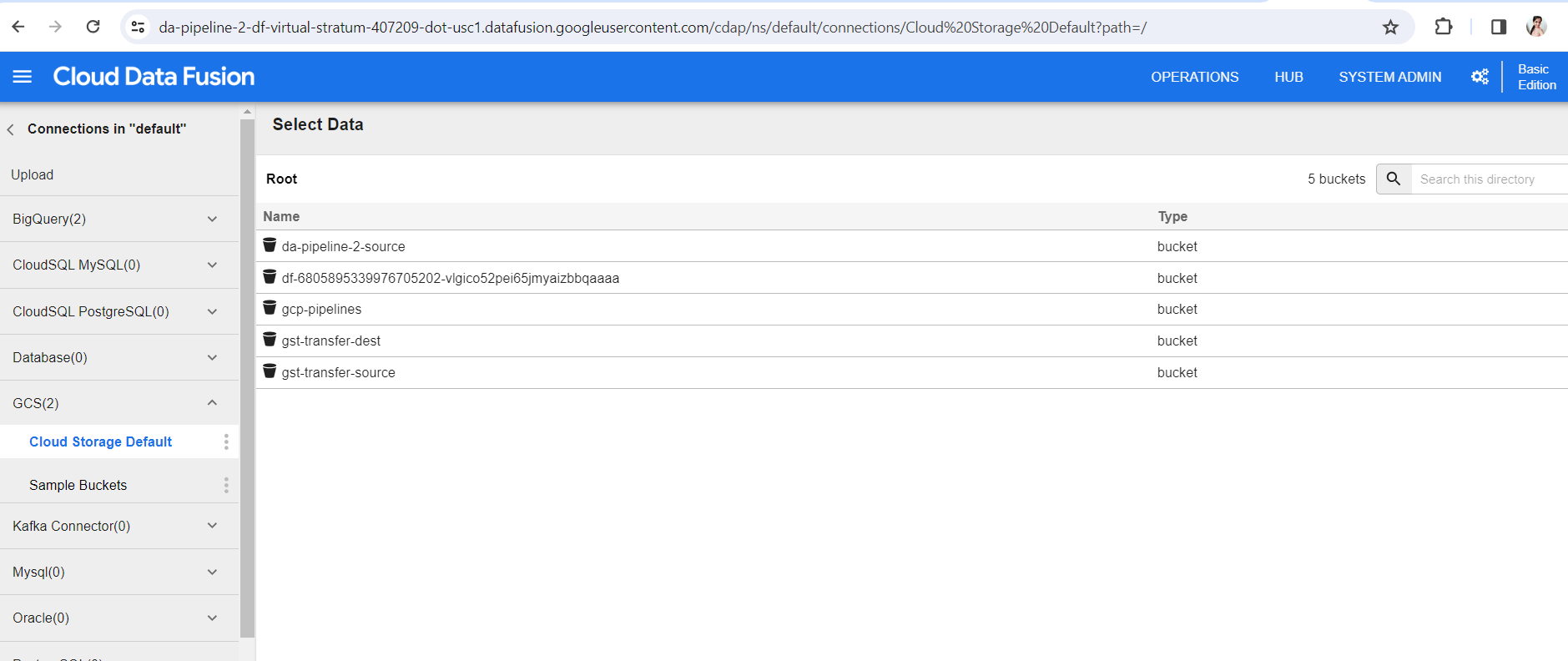
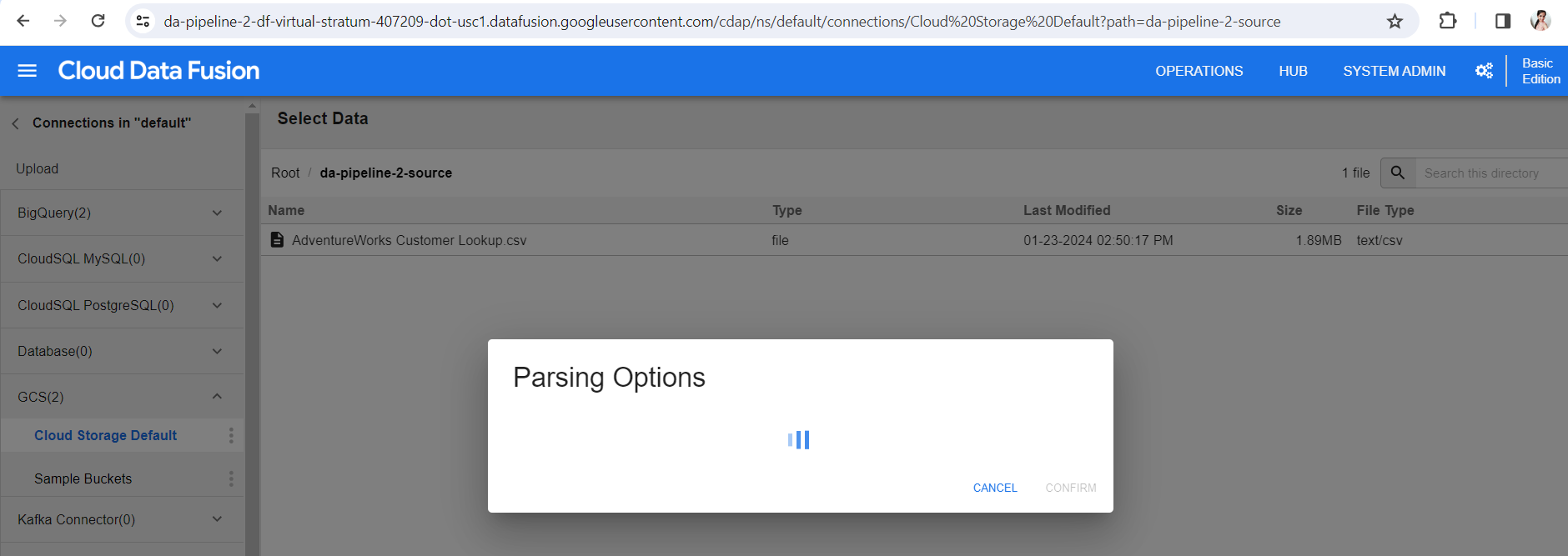
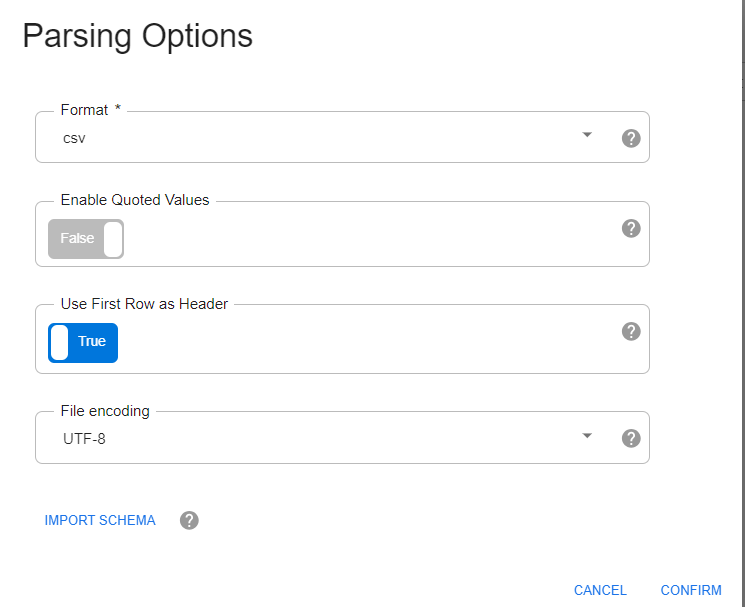
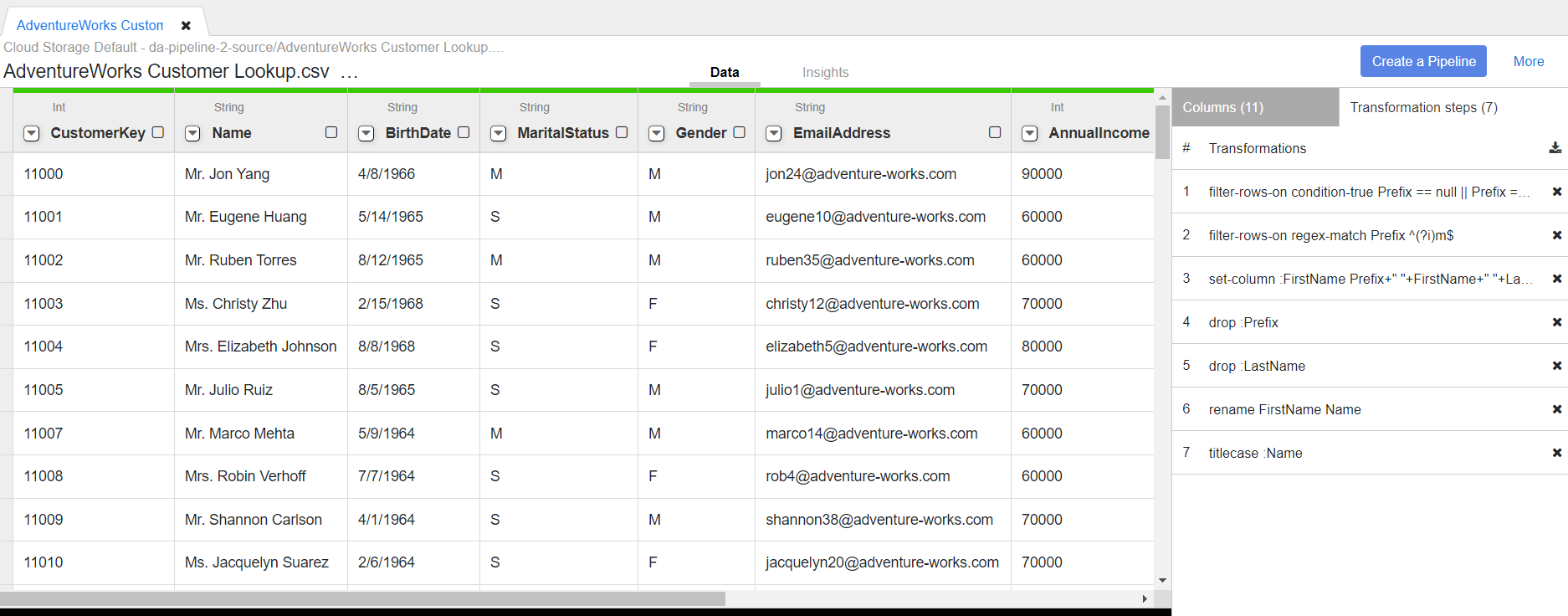
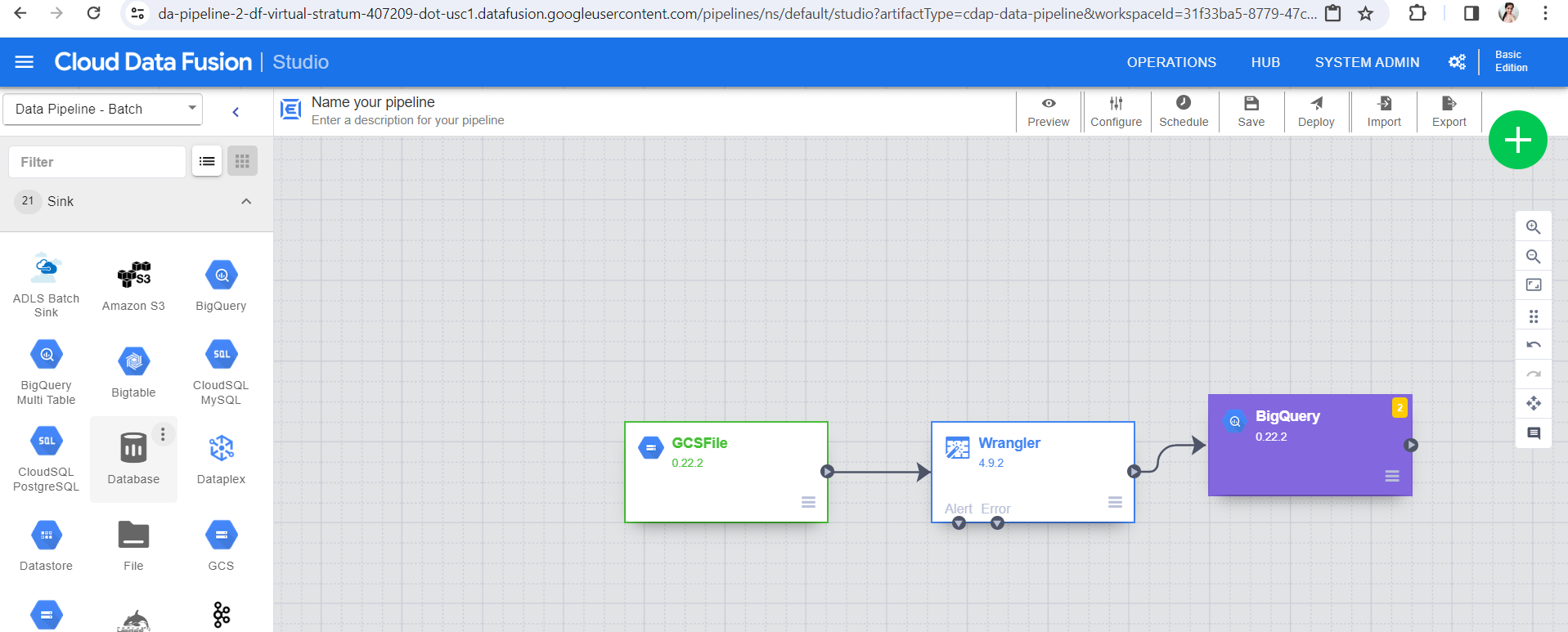
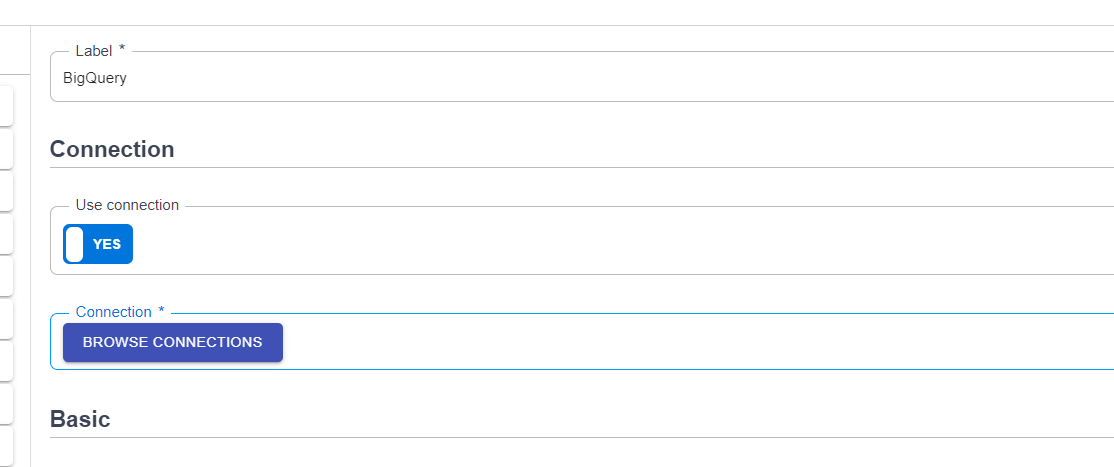
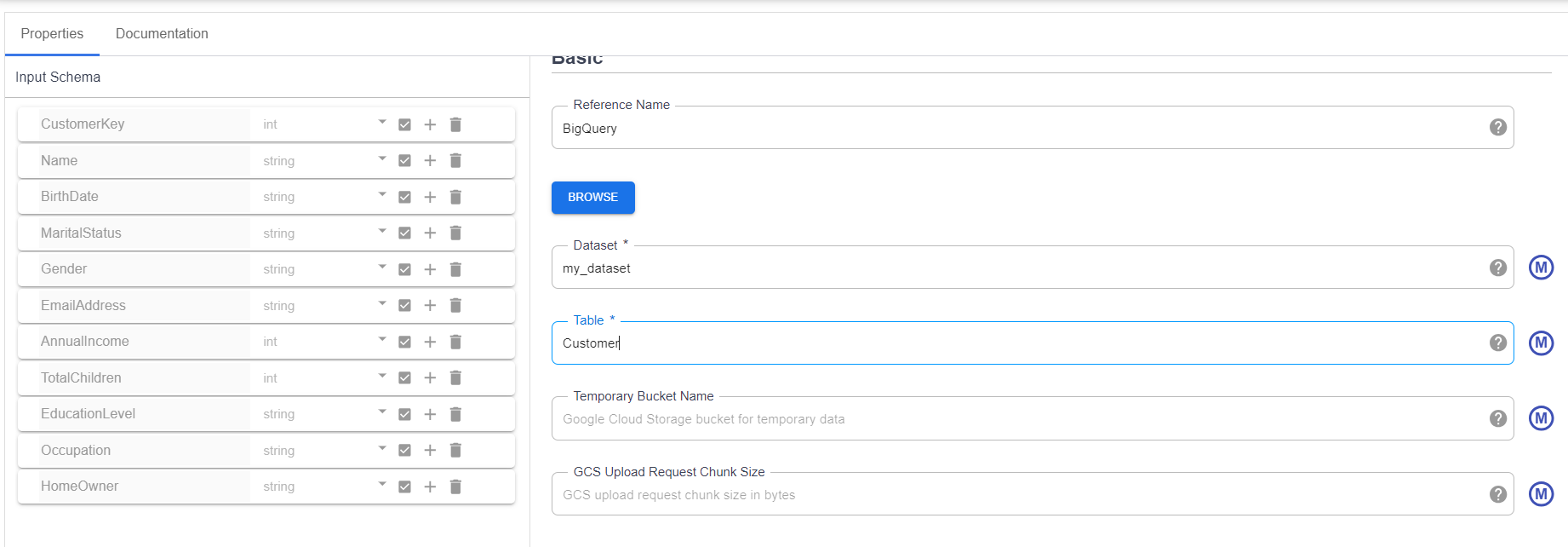
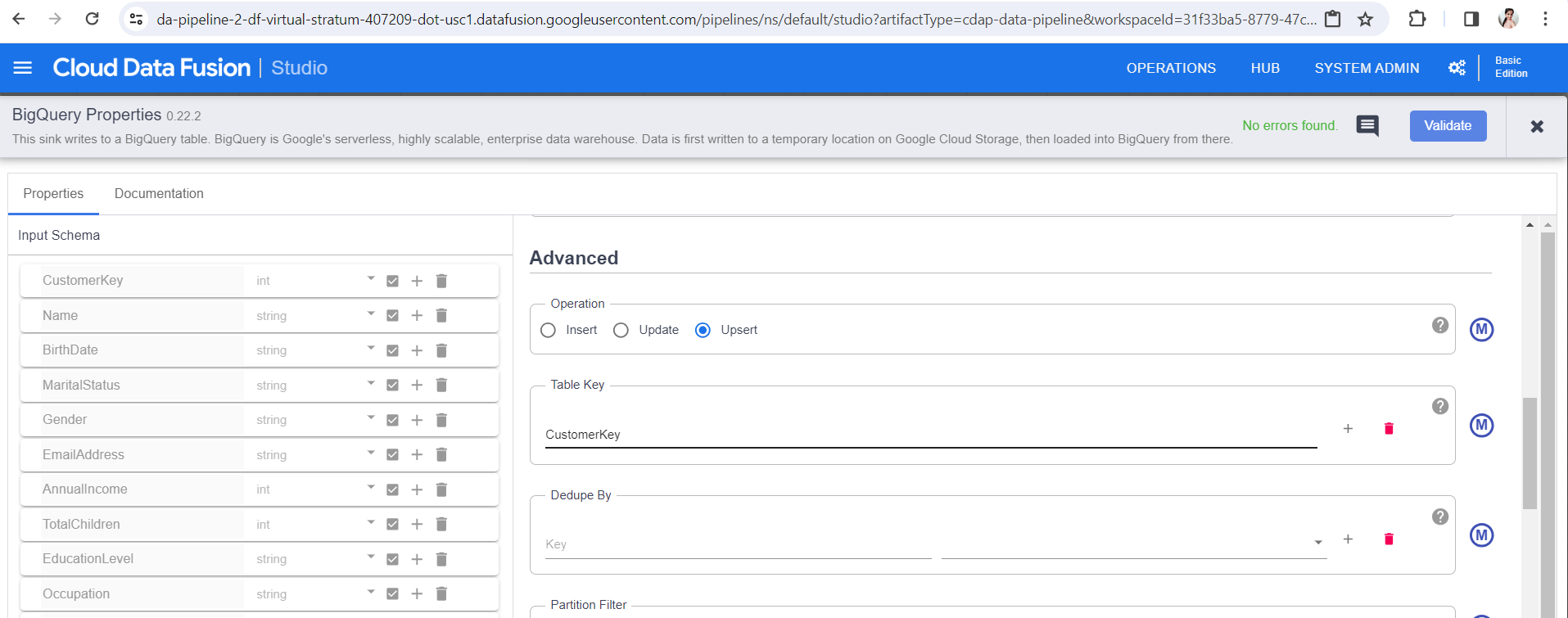
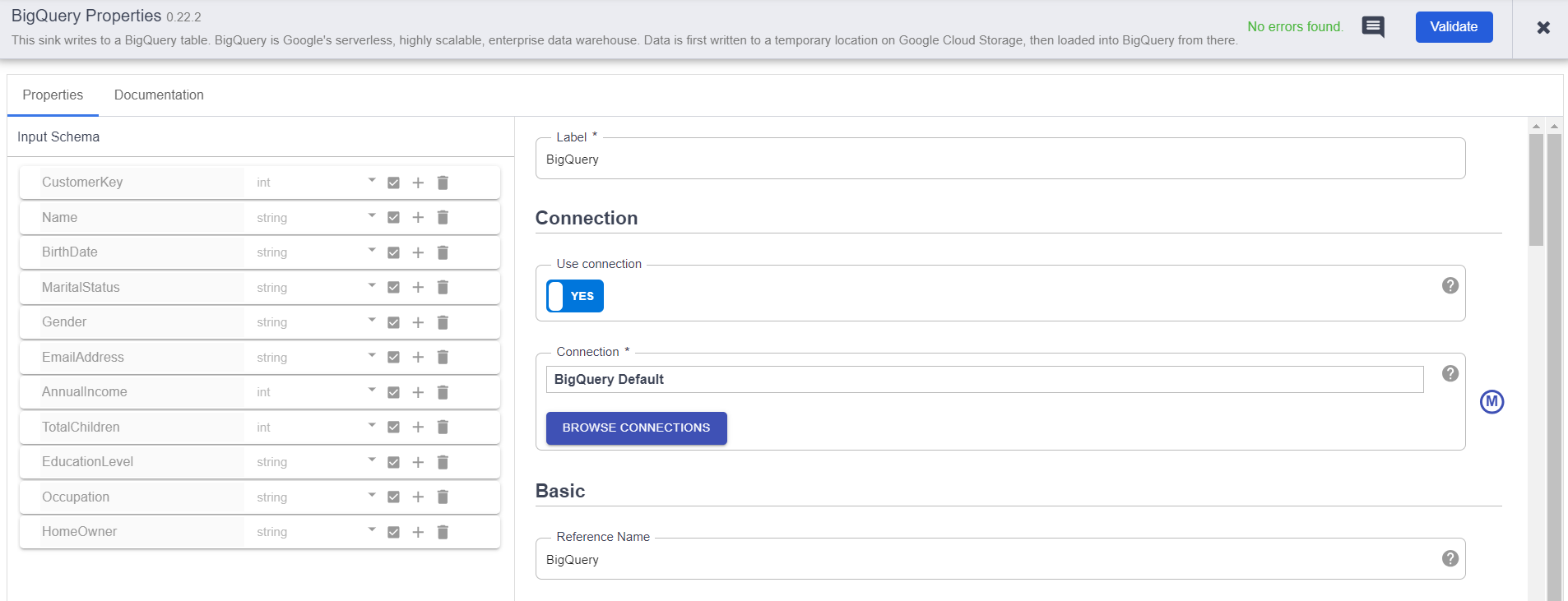
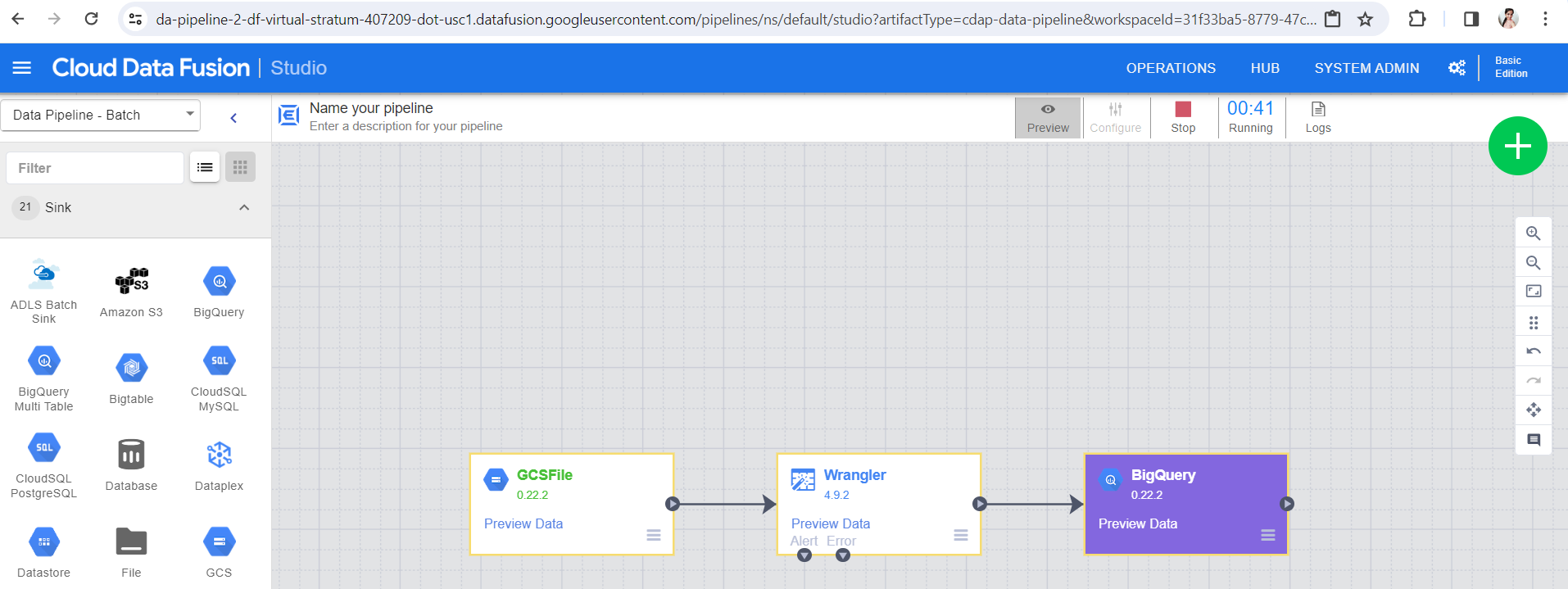
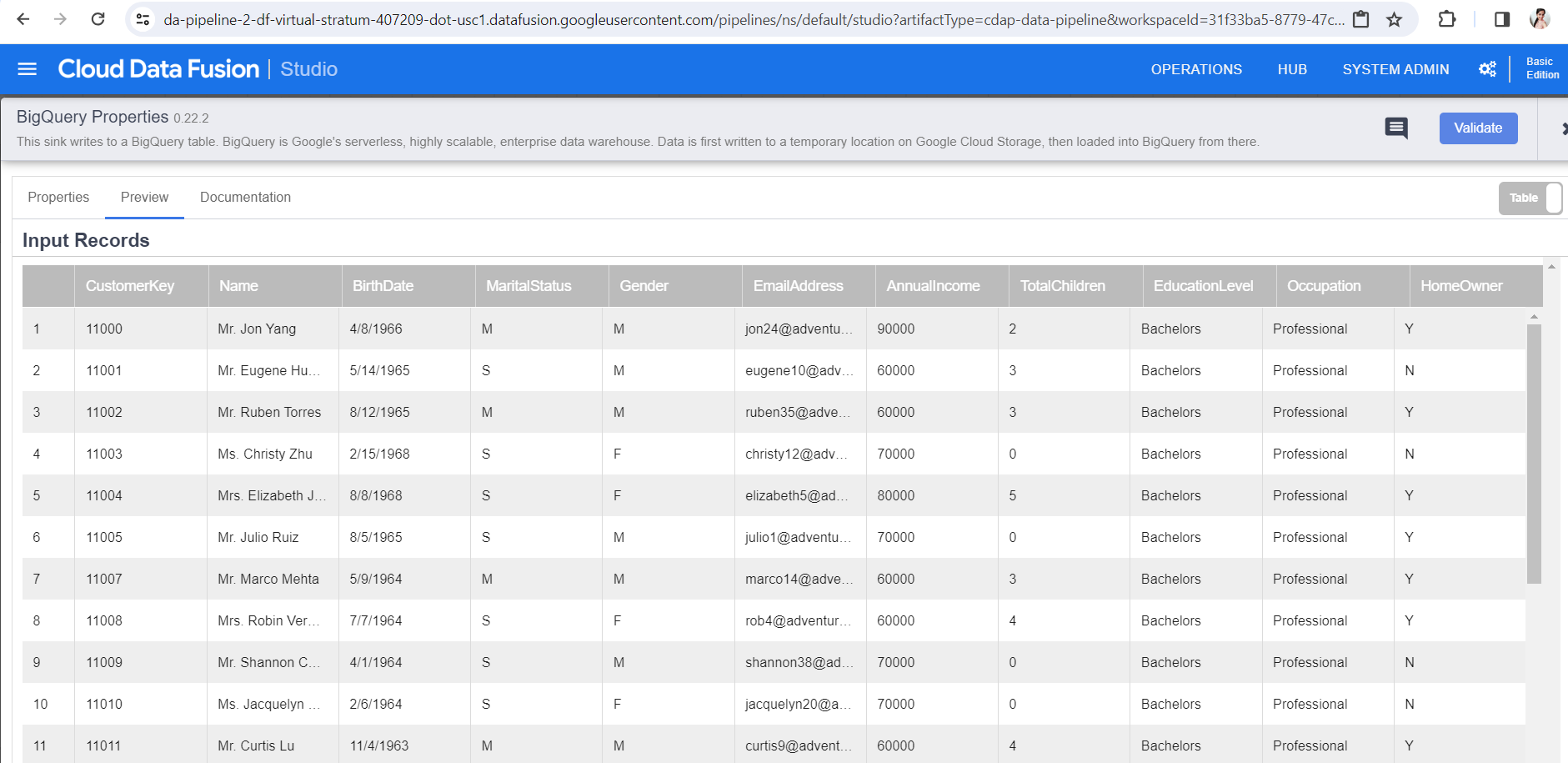
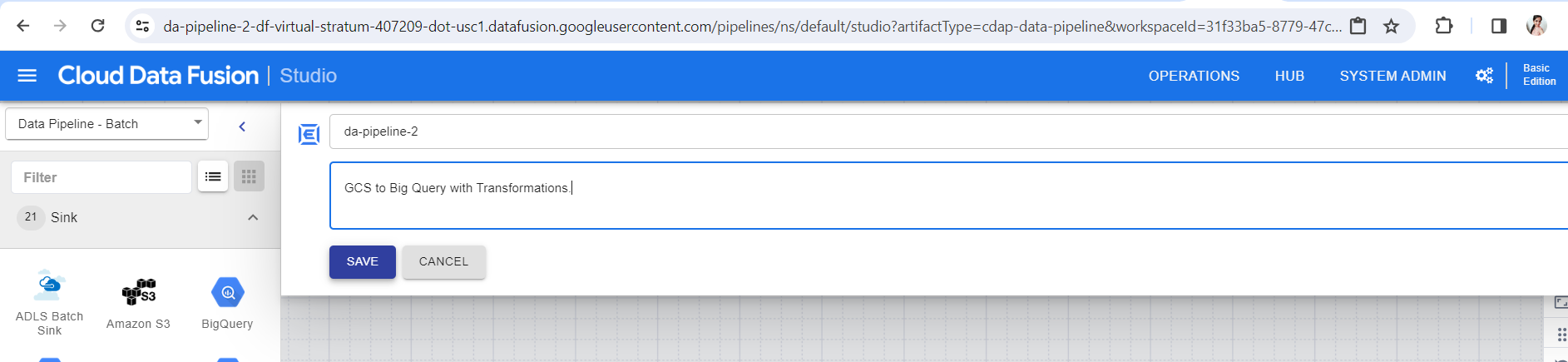
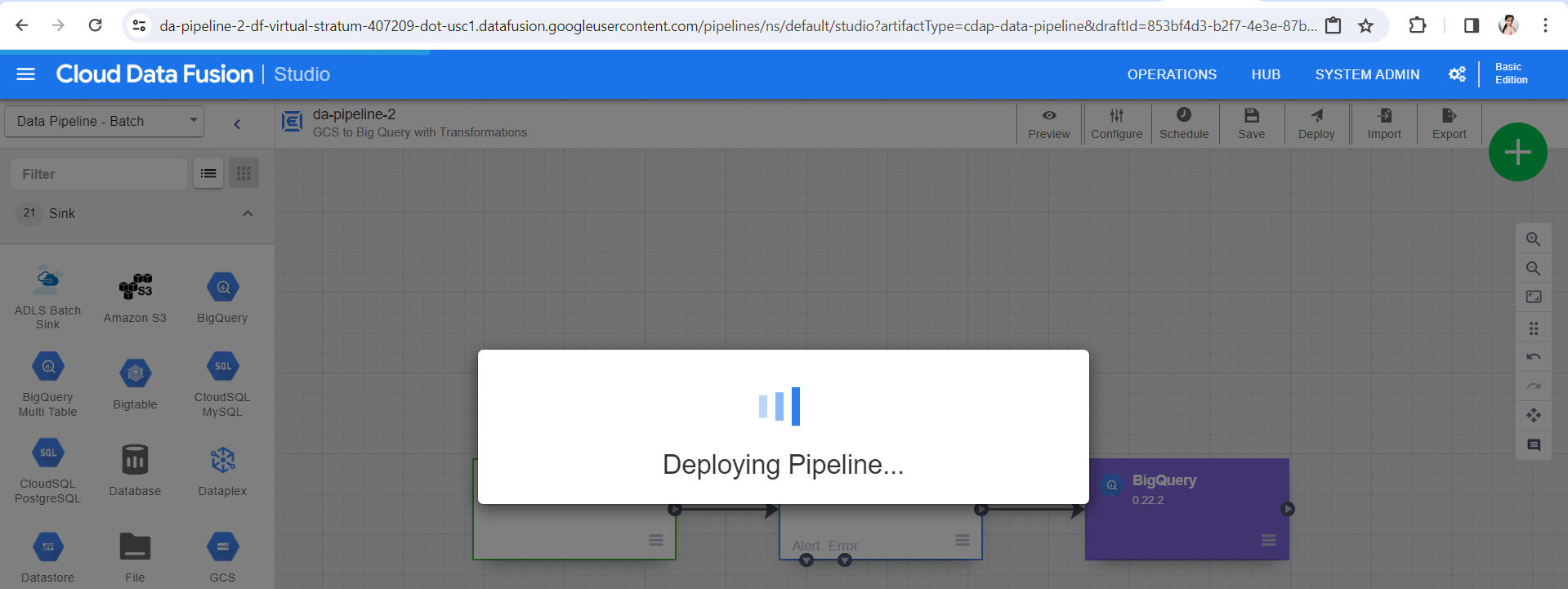
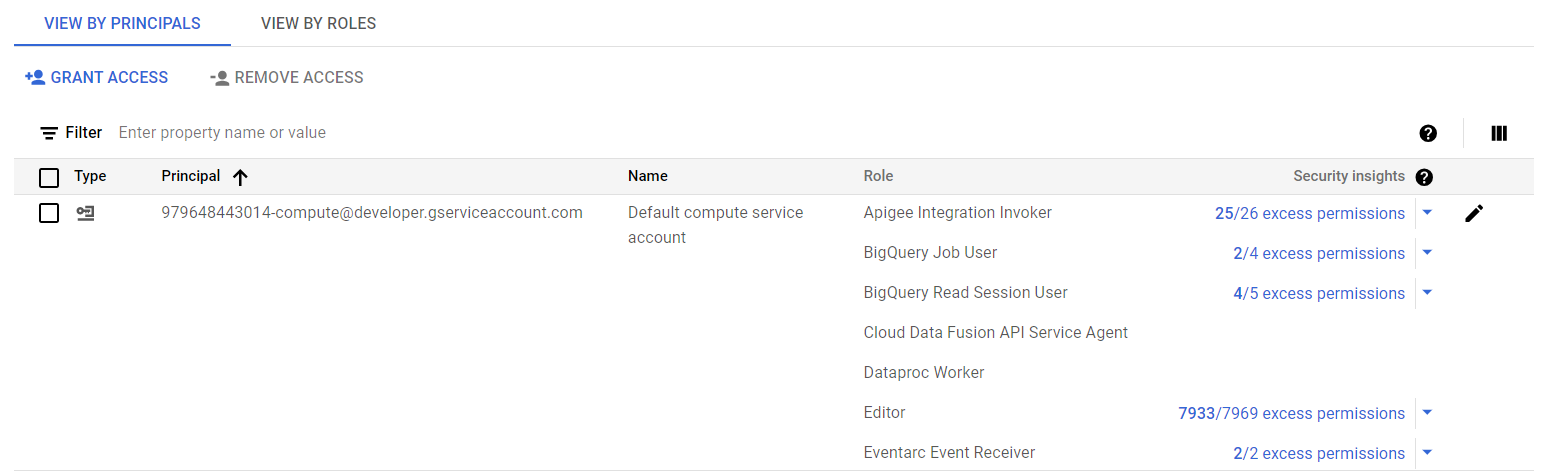
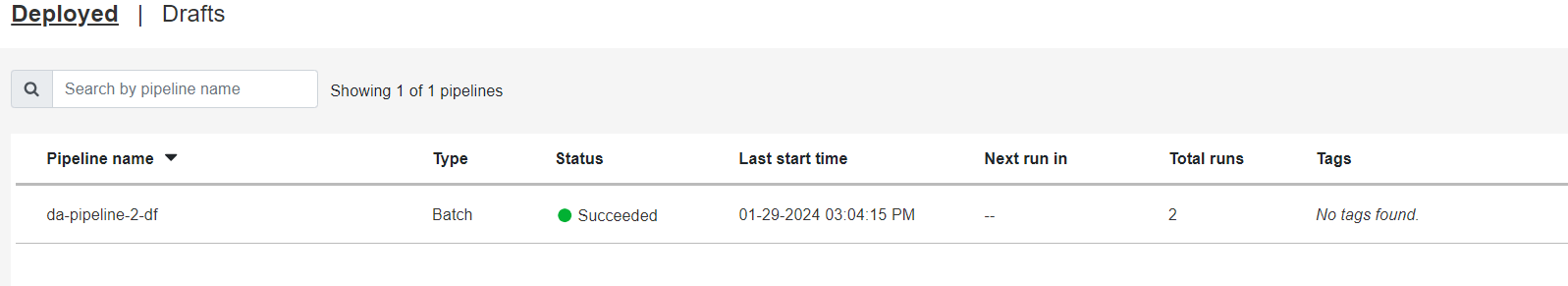
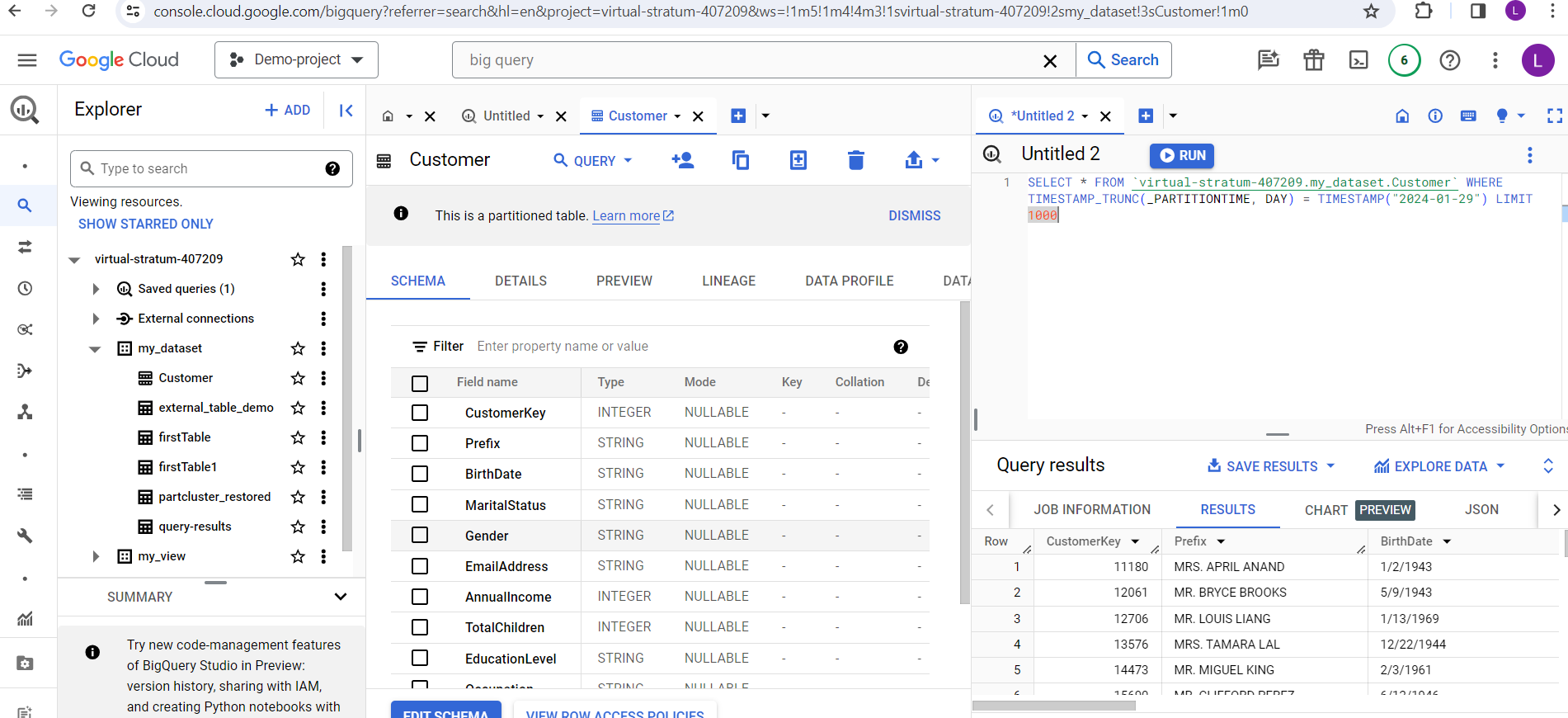
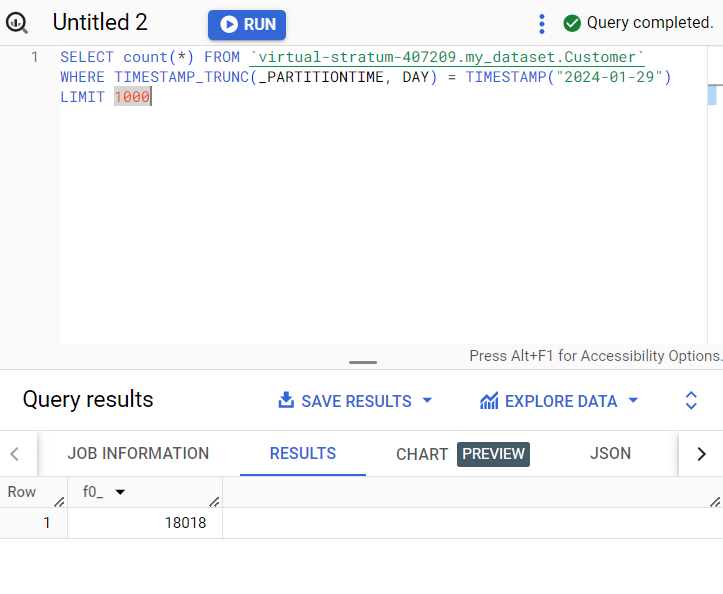
Creating GCS Bucket:

* Create a GCS bucket names as da-pipeline-2-source and select the required region.

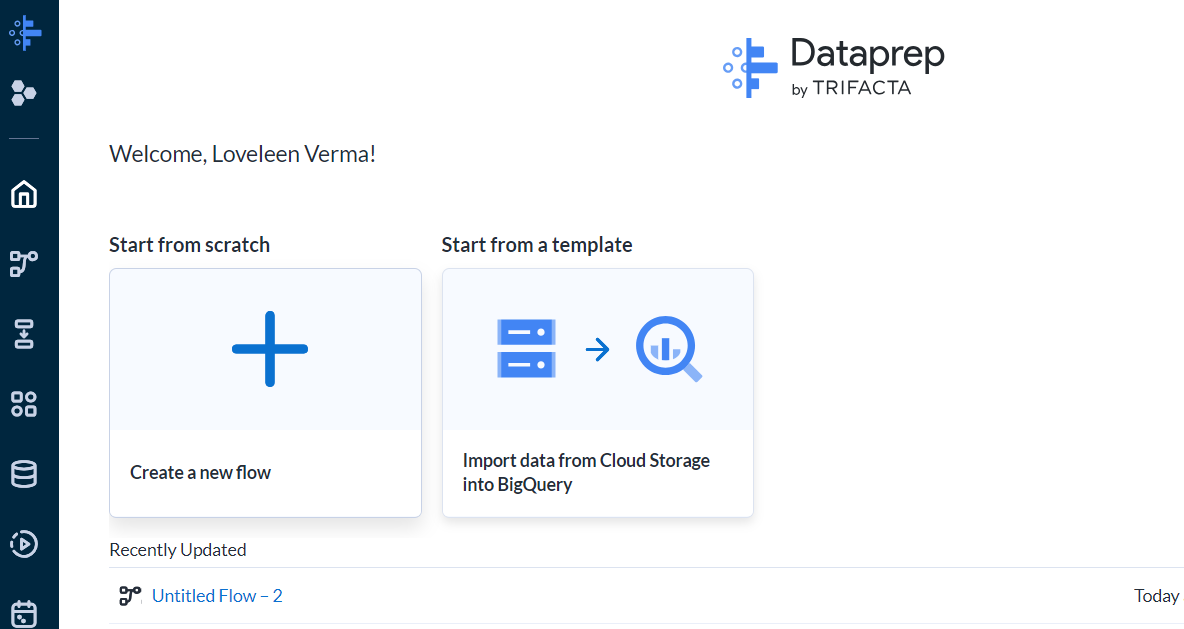
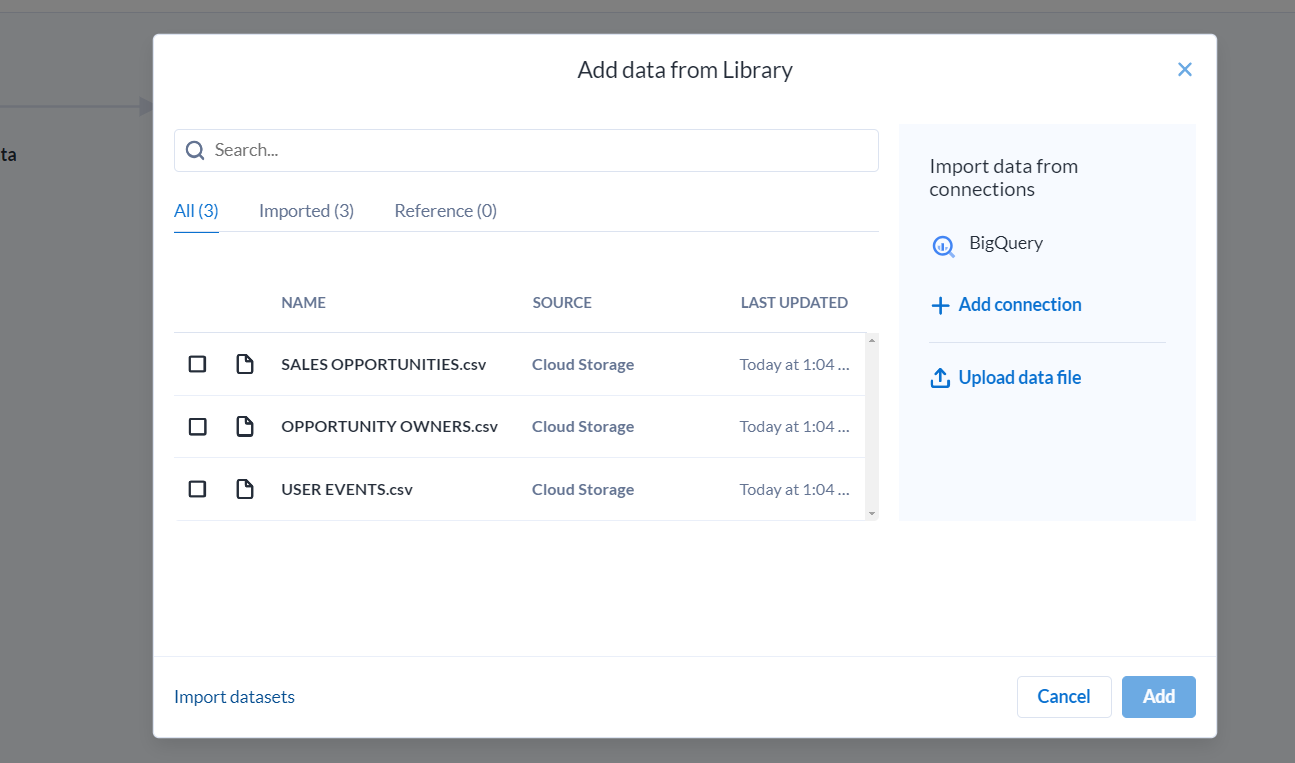
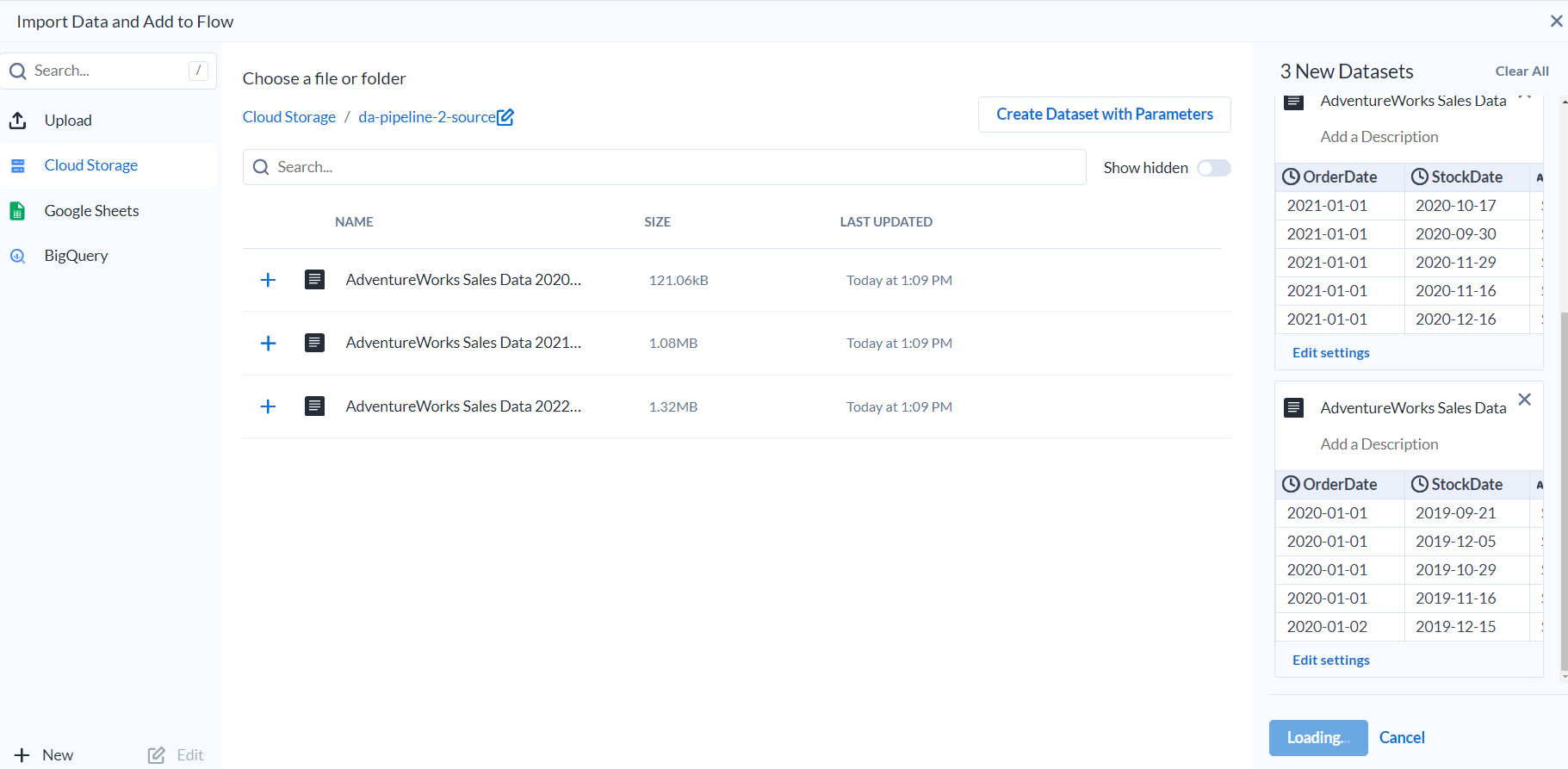
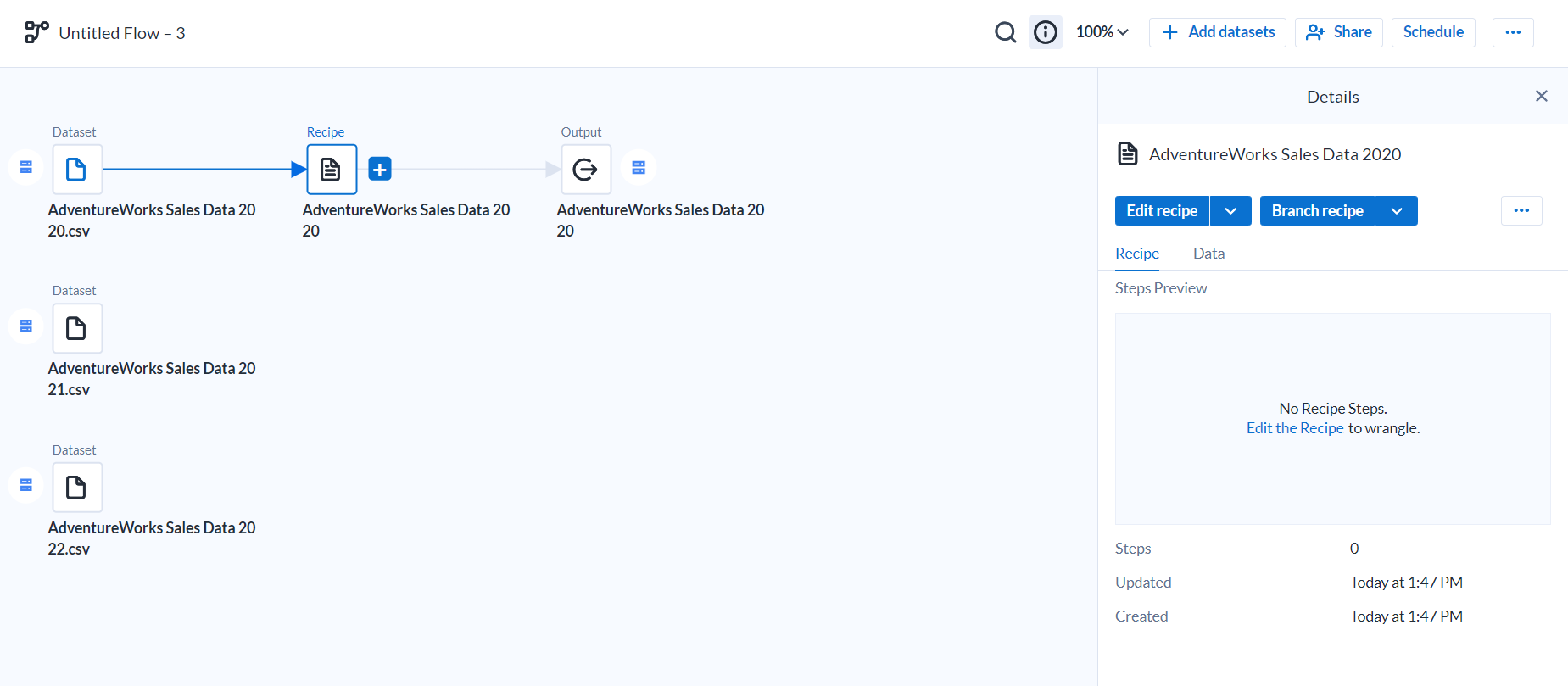
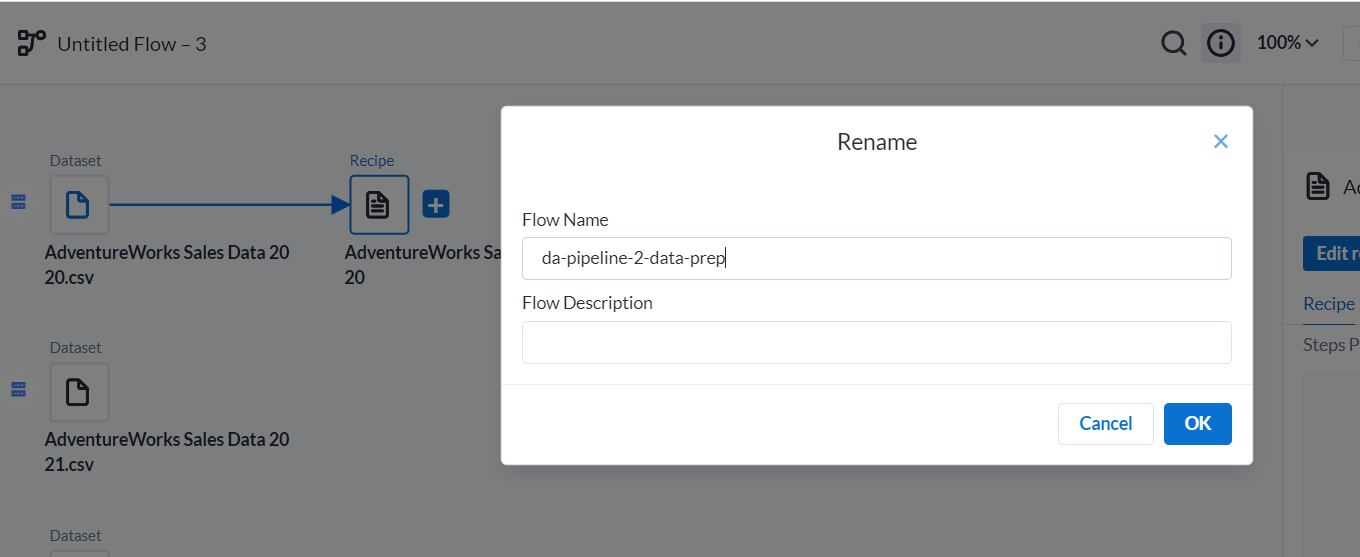
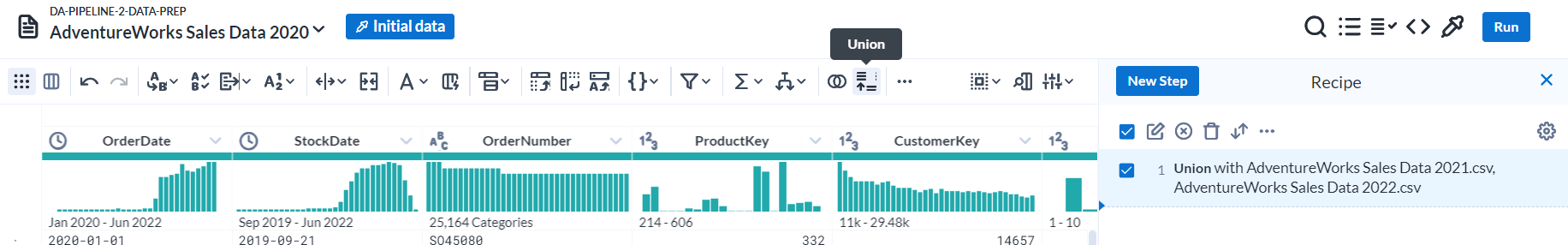
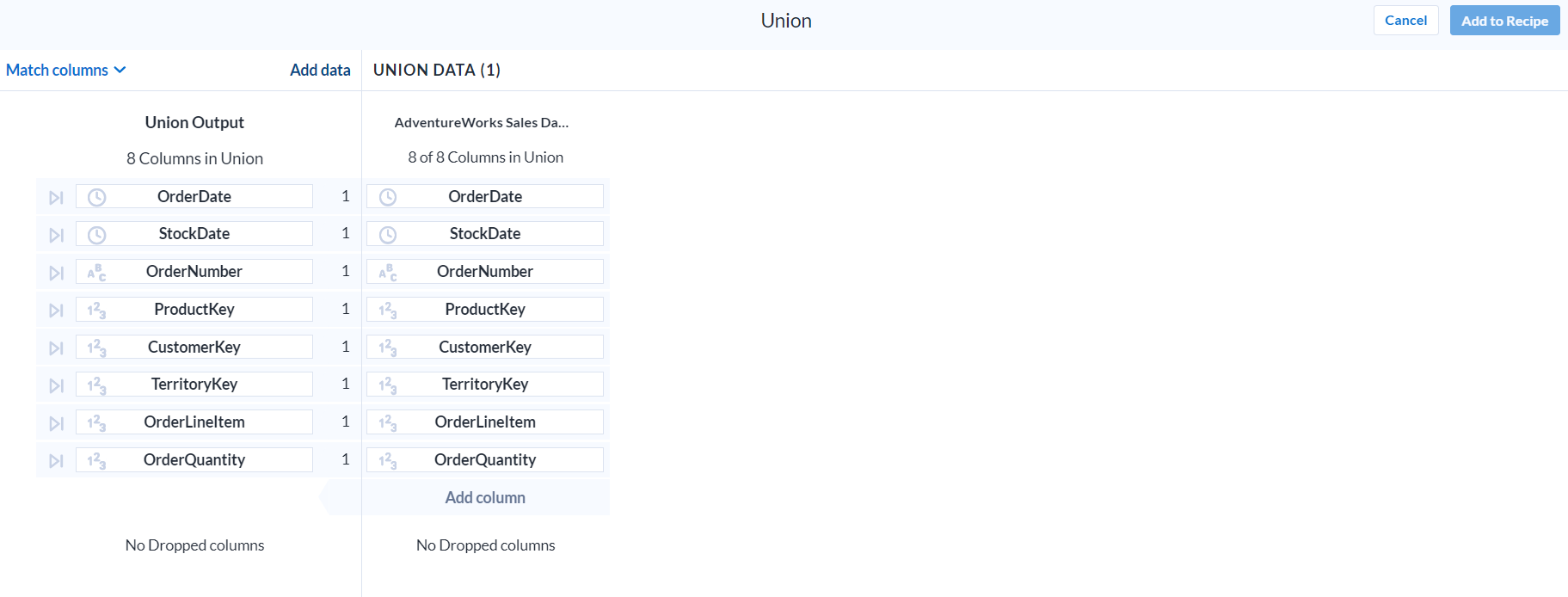
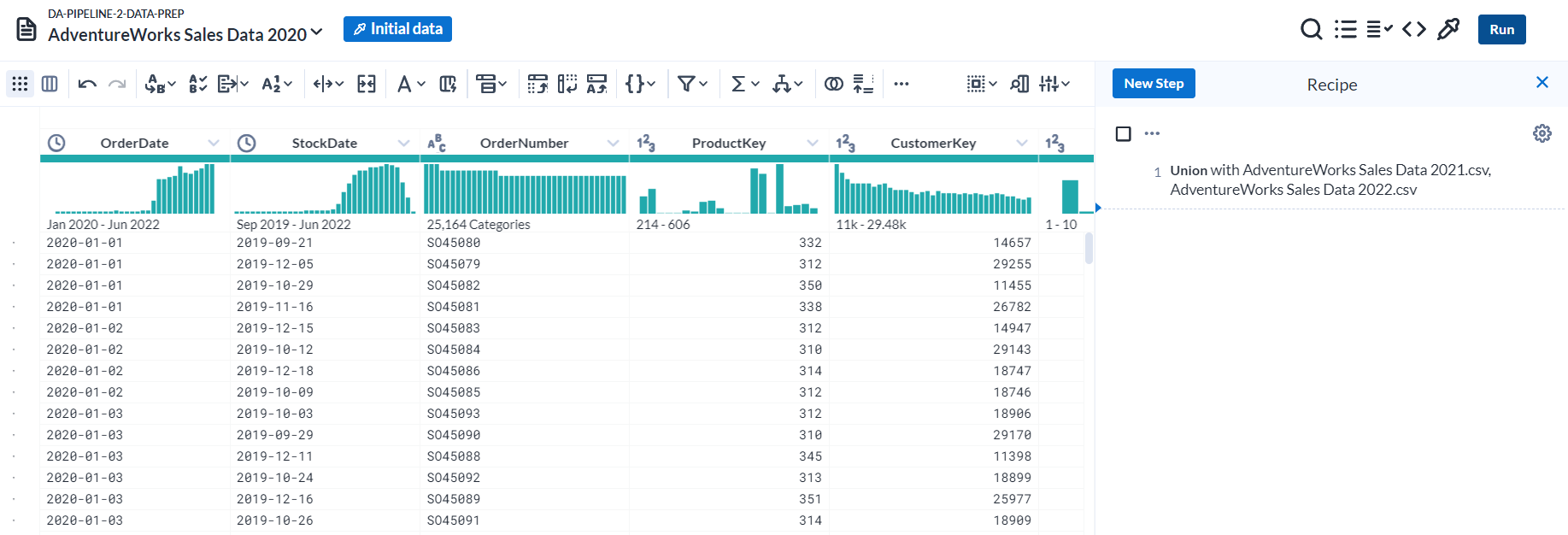
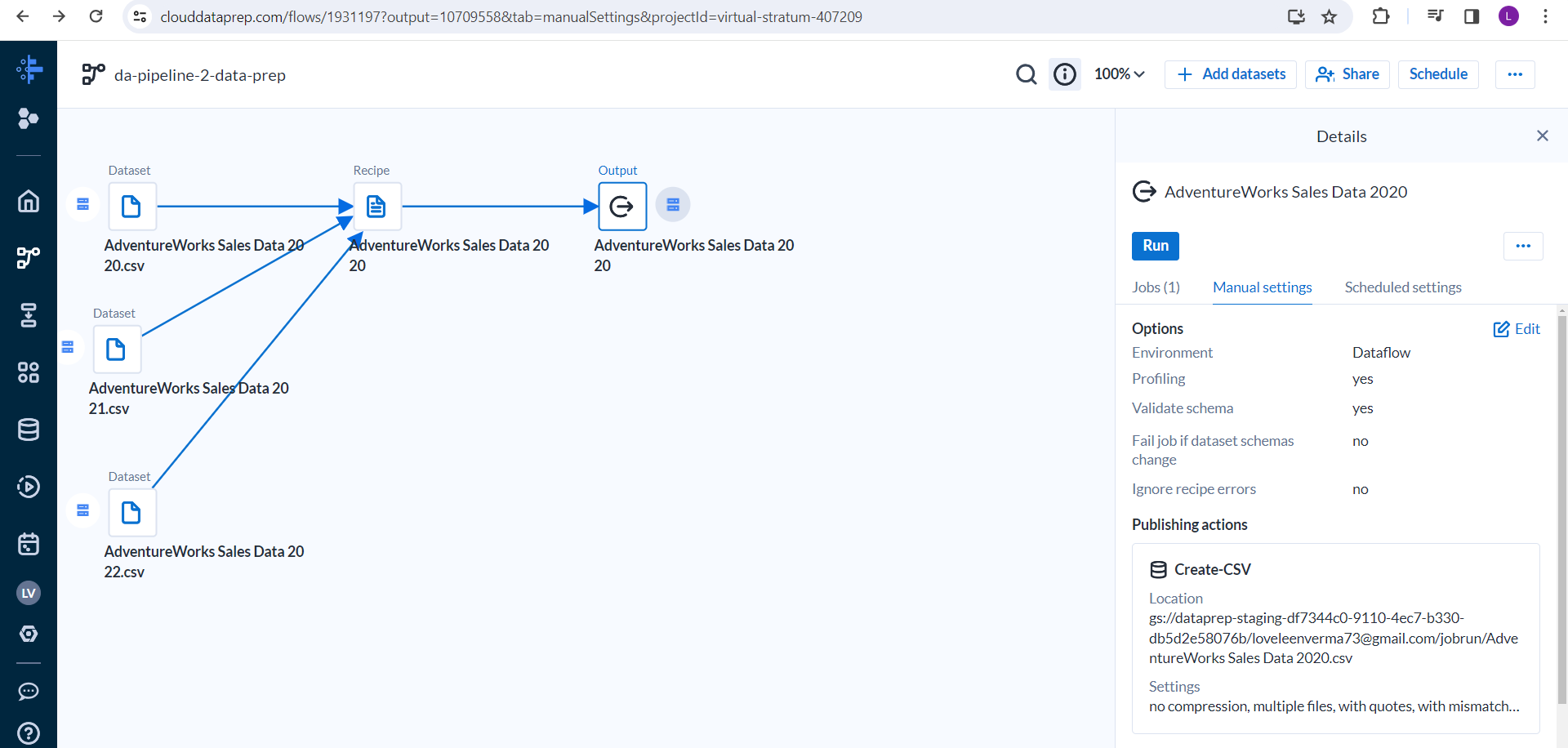
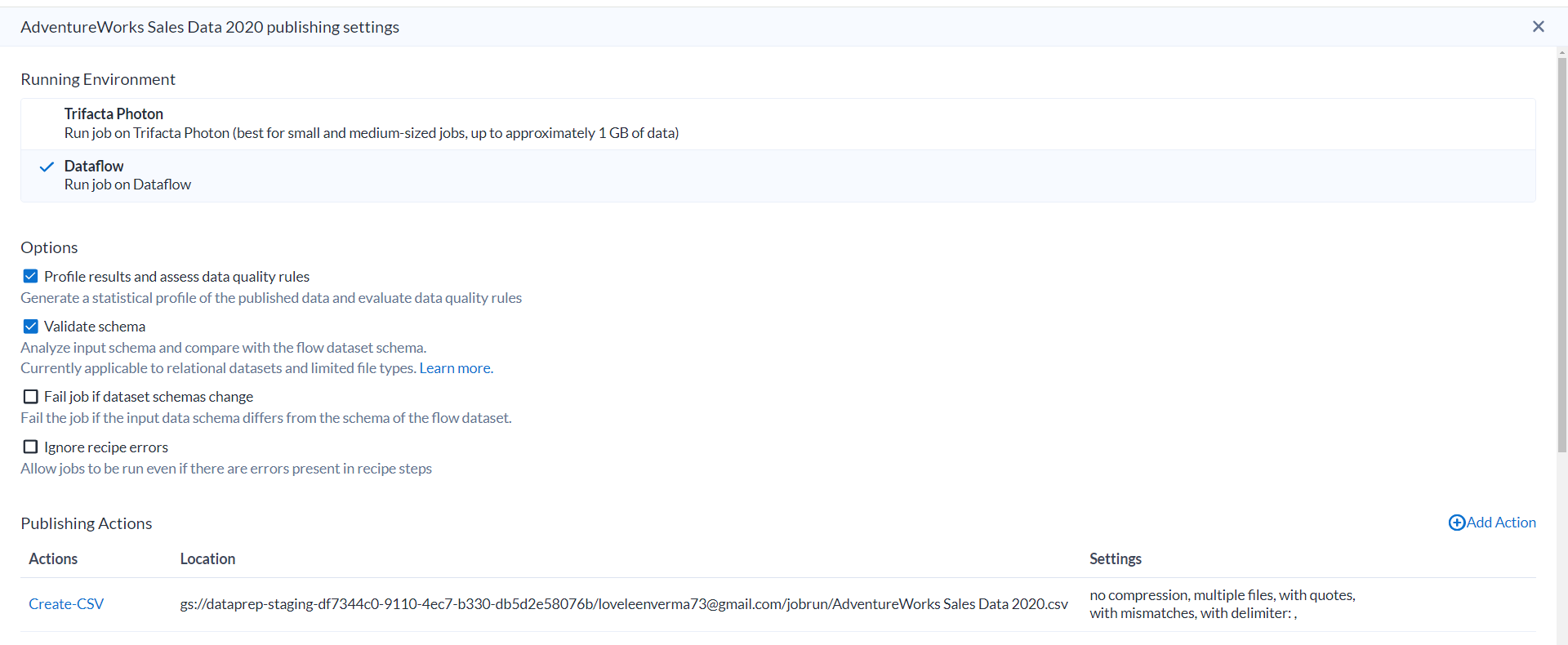
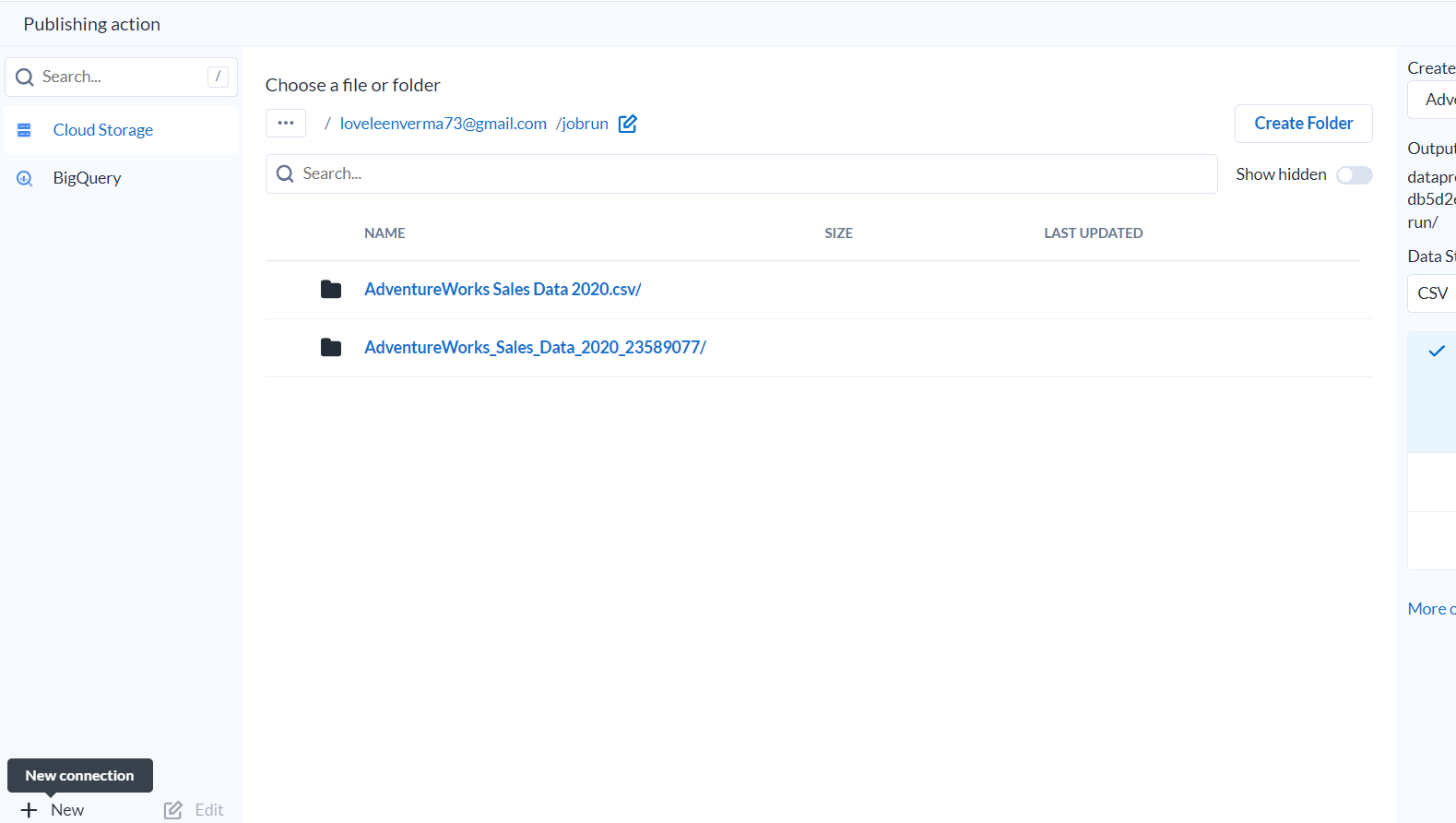


* Keep rest as default and click create and now the bucket is created.  
  
* Upload the file Adventure Works Customer Lookup.csv in this bucket.

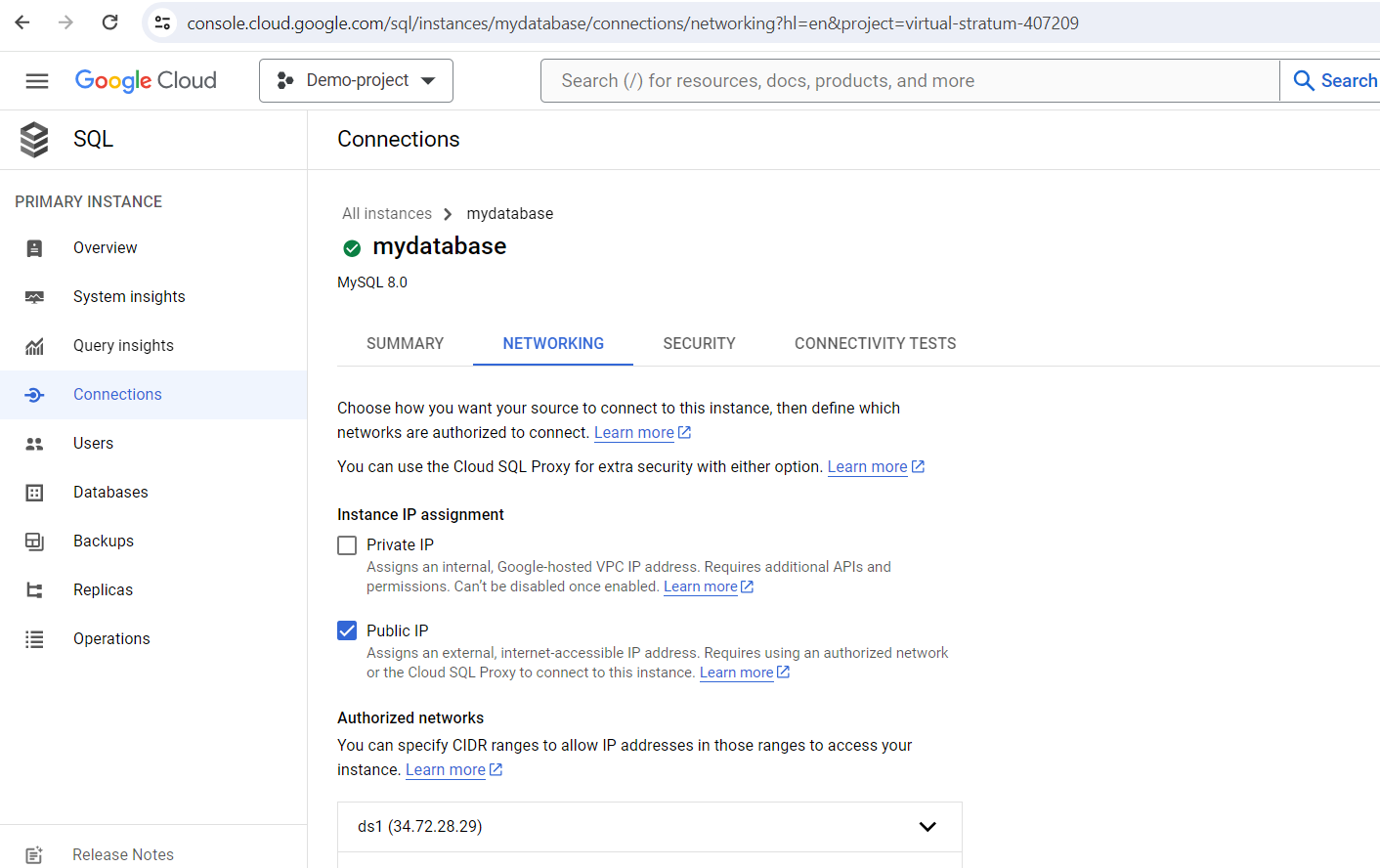
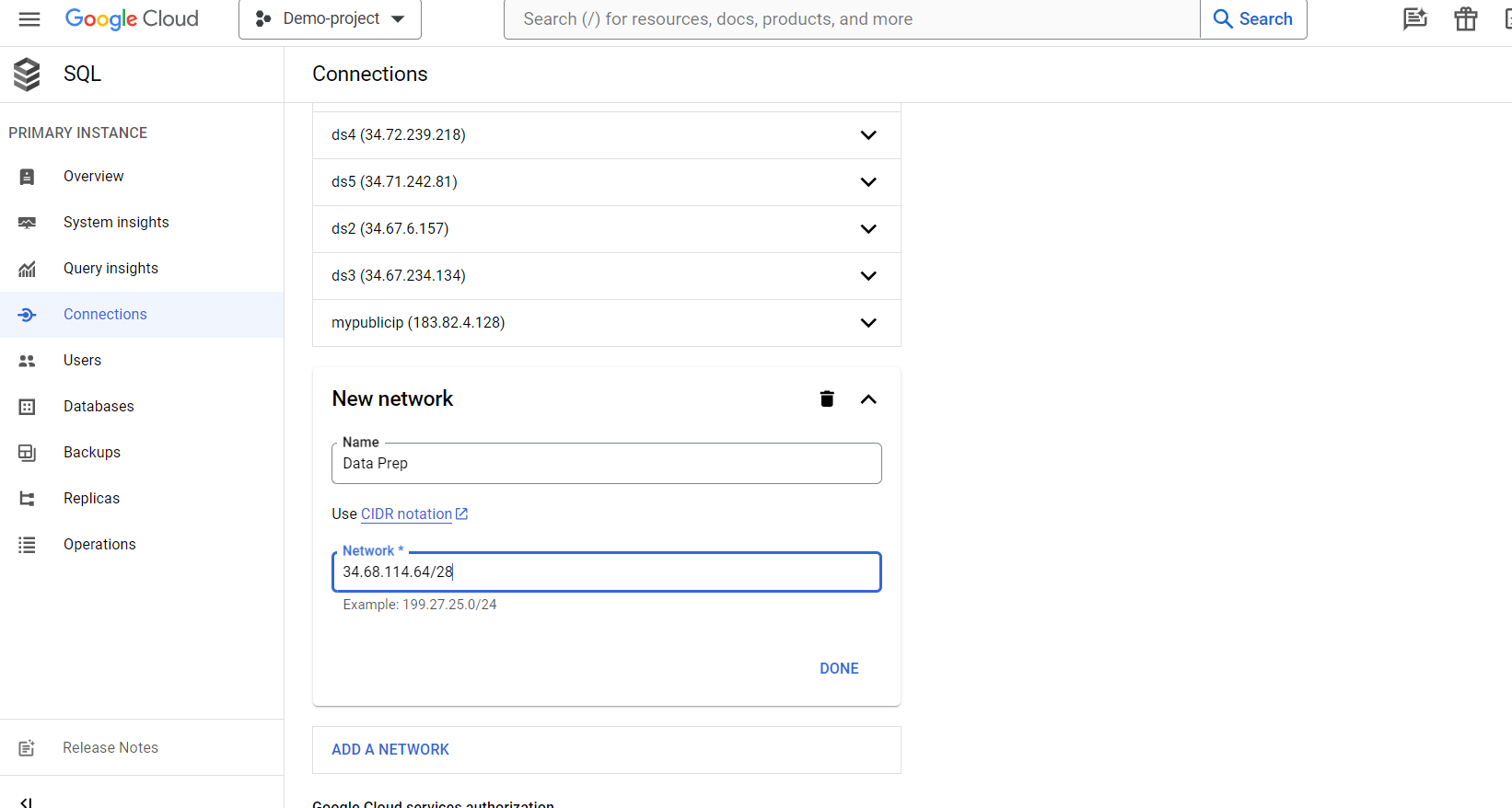
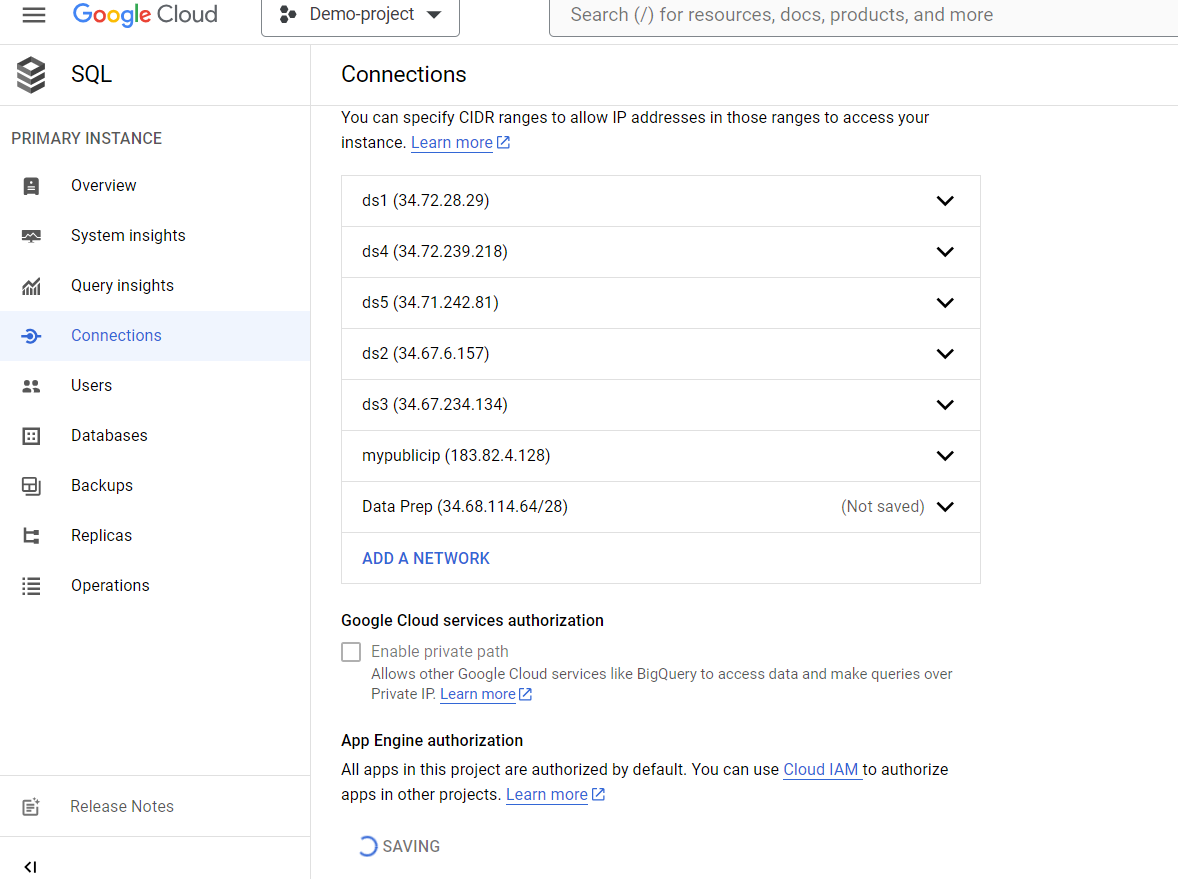
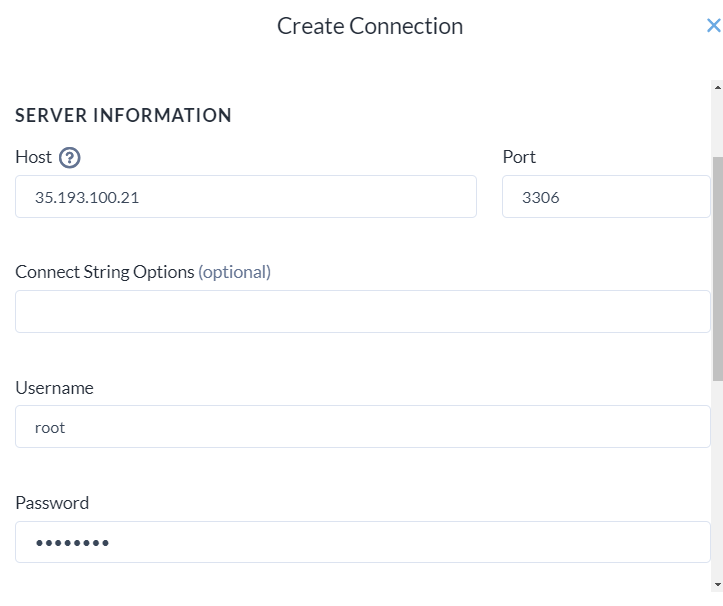
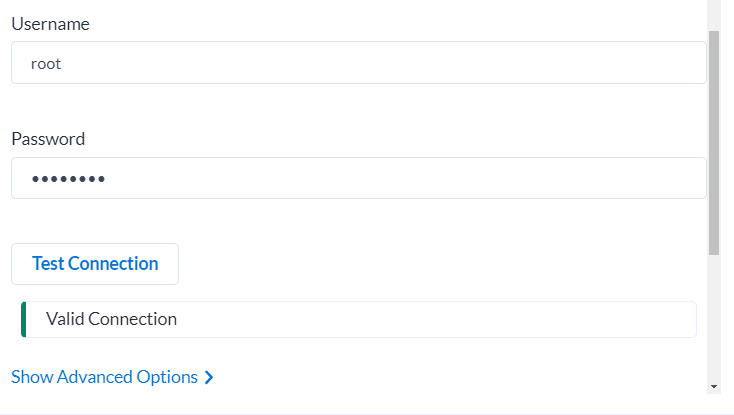
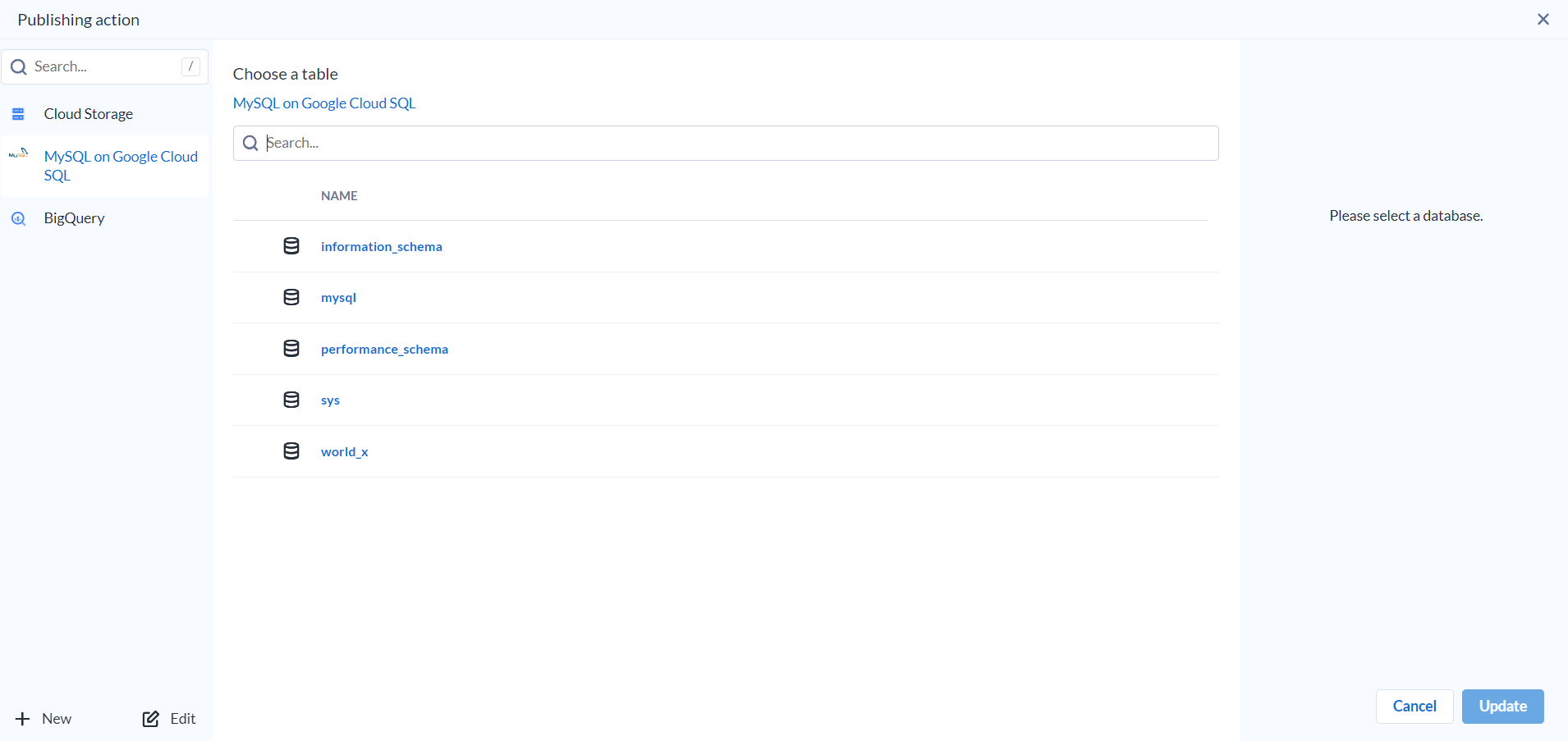
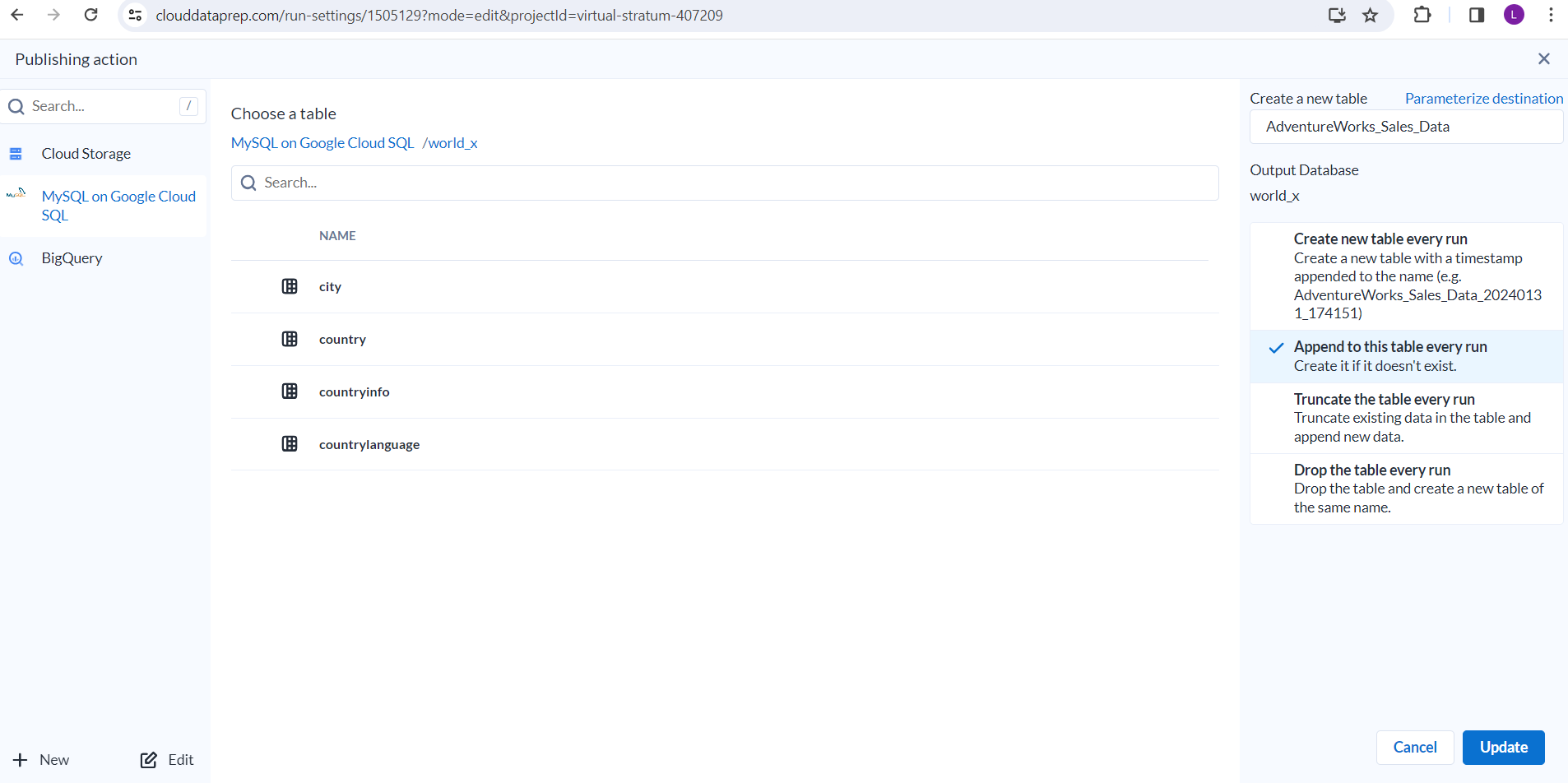
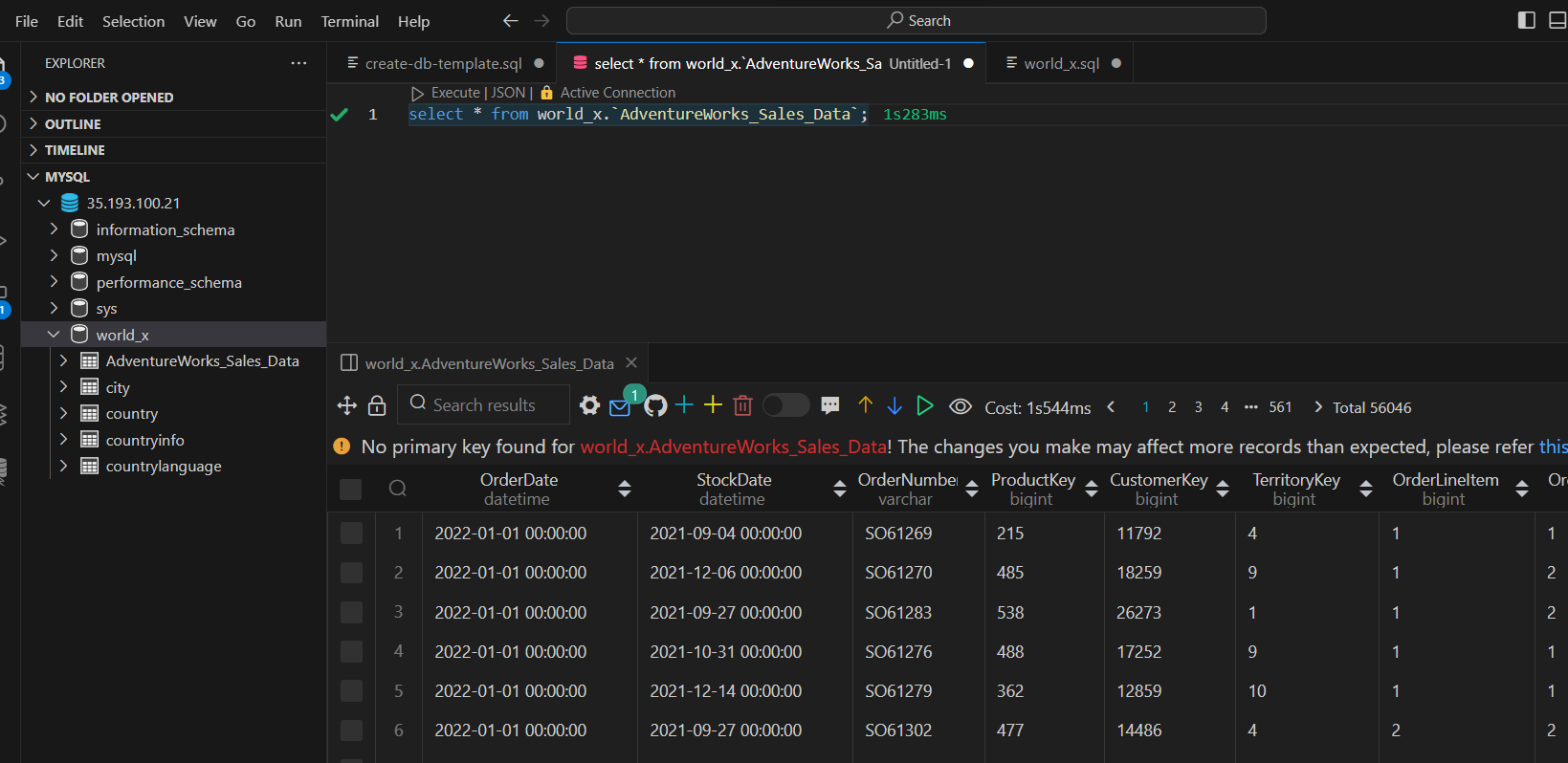
Data Fusion Instance Creation:

* Enable Data Fusion API.
* Click on create Data Fusion Instance.
* Provide instance name, region and select the latest instance version.  
  
* Click create and the instance will get created.  
  
* Open Data Fusion Studio -> Select Wrangler.  
  
* Click on Data Source from where you want to take up the data. As we have our file in GCS, we will click GCS.  
  
* Select the file.  
  
* As we need our first row as Header. So, click on that.  
  
* Click Confirm.
* Perform the Transformation and click on Create Pipeline.  
  
* Go to sink, Click on Big Query from sink plugins.  
  
* Click on Properties in Big Query plugin and browse for your connection.  
  
* Provide the Data Set and Table Name.  
  
* In Advanced options click on Upsert and provide a key-> CustomerKey.  
  
* Click on Validate.  
  
* Click on Preview and then Run to check the sample data at every step.  
  
* Once its done, click on preview on Big query and we can see the data after transformation there.  
  
* Provide name of the pipeline and save.  
  
* Click on Deploy.  
  
* Click run. 
* Before run make sure you have given Cloud Data Fusion Service Agent role to Data Proc service account.  
  
* After run you can see the status of pipeline as succeeded.  
  
* Now we will check in Big Query whether the data has come by querying.  
  
* For verification, we can check the row count.  
  
* Now we will move to Data Prep.

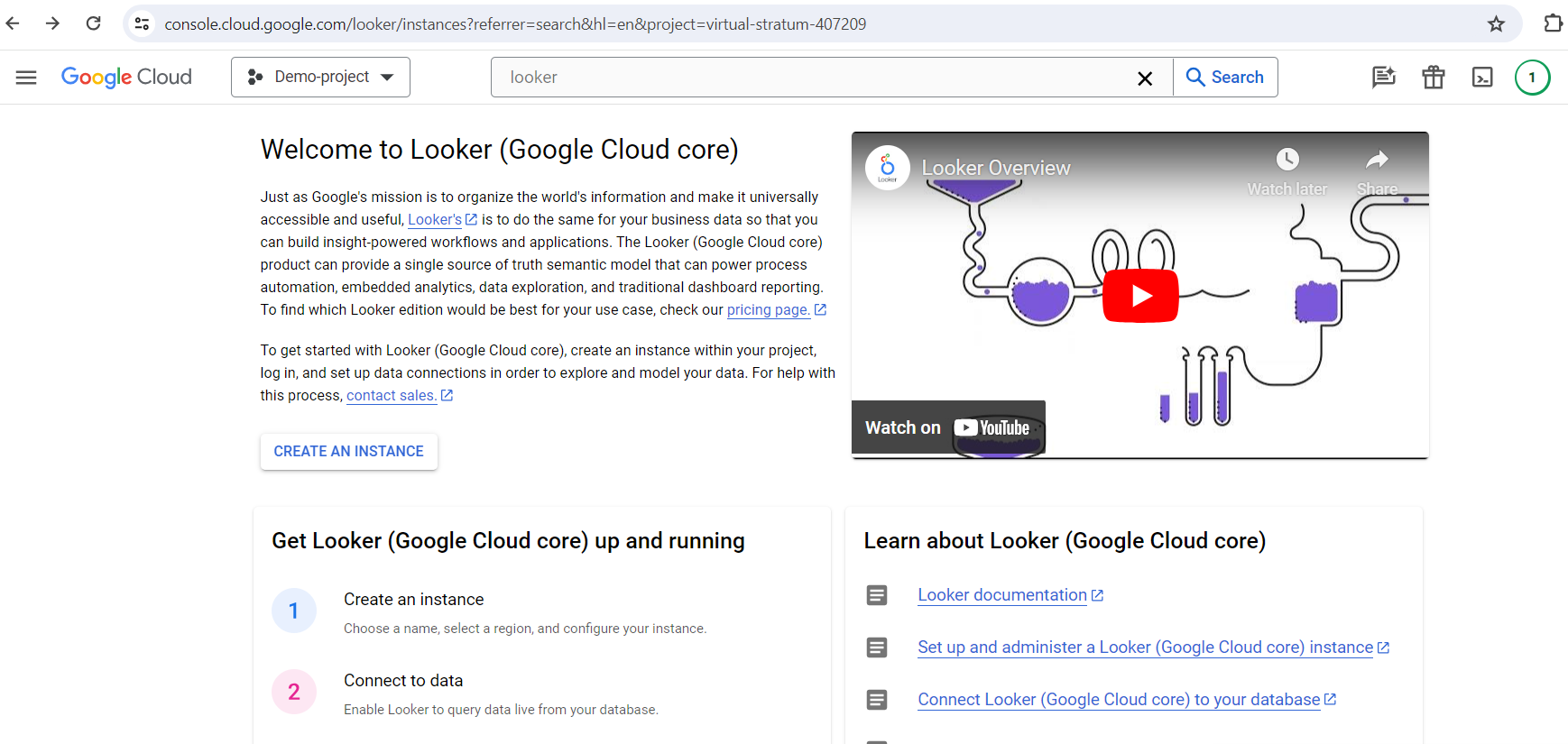
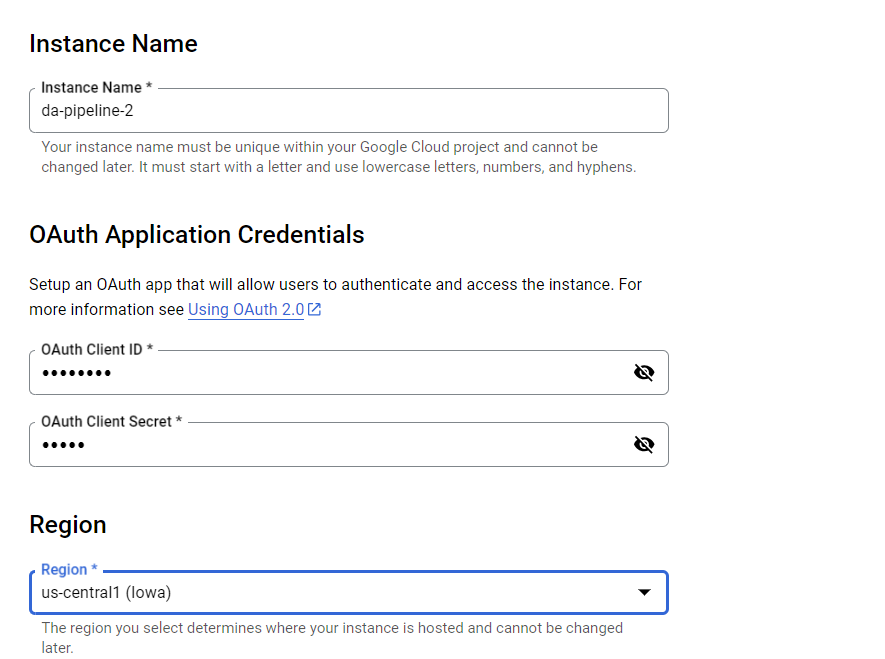
Data Prep Transformation

* To create a data prep instance, follow the steps:
* Open Data Prep Instance.
* Click on create a new Flow:  
  
* Now, we need to import the datasets that we want to transform.  
  
* Click on Cloud Storage files you want to add. And then click on Import and Add to Flow.  
  
* The following window will appear.  
  
* Rename the pipeline.  
  
* Click on recipe.
* Now we will union these three files so as to accommodate the data completely in one file.  
  
* A new window will appear. Click on Add Data.  
  
* Select the files you want to union and then click Apply.
* Now you can see the data from 2020 to 2022.  
  
* Now as I want to dump this data to MySQL. So, need to configure that in Output.
* Click on Output. Then Click on Edit.  
  
* Click on Actions under Publishing Actions.  
  
* Click on New Connection:  
  
* Select MySQL on Google Cloud SQL there, but before starting with this you need to first whitelist the Platform Service in your database.

For Data Prep it is: 34.68.114.64/28

* For this go to Cloud MySQL instance.
* Go to Connections and then networking.  
  
* Click on Add Network. Give any name and the Platform Service IP.  
  
* Click on Save.  
    
  
* Then on Data Prep, provide the server information as follows:  
    
  
* Test the connection.  
  
* Click Create.
* Now your connection will appear in the list. Select your database.  
  
* Select the table if you already had or click on create new table.  
  
* Click Update.
* Then Click Save Settings.
* Click on Run.
* Once done you can see the status in the job overview.  
  
* Now you can query the database to see whether data is there or not.  
  
* Its done.

Visualize data in Looker

* At First, we need to create one Looker instance.
* Go to looker. Click on Create instance.  
  
* Give Name, Provide Outh Client ID and Secret and select the Region.  
  
* Click Create.