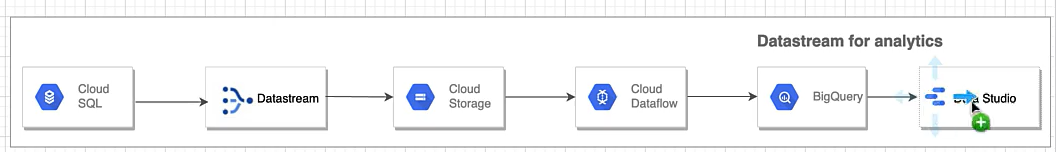
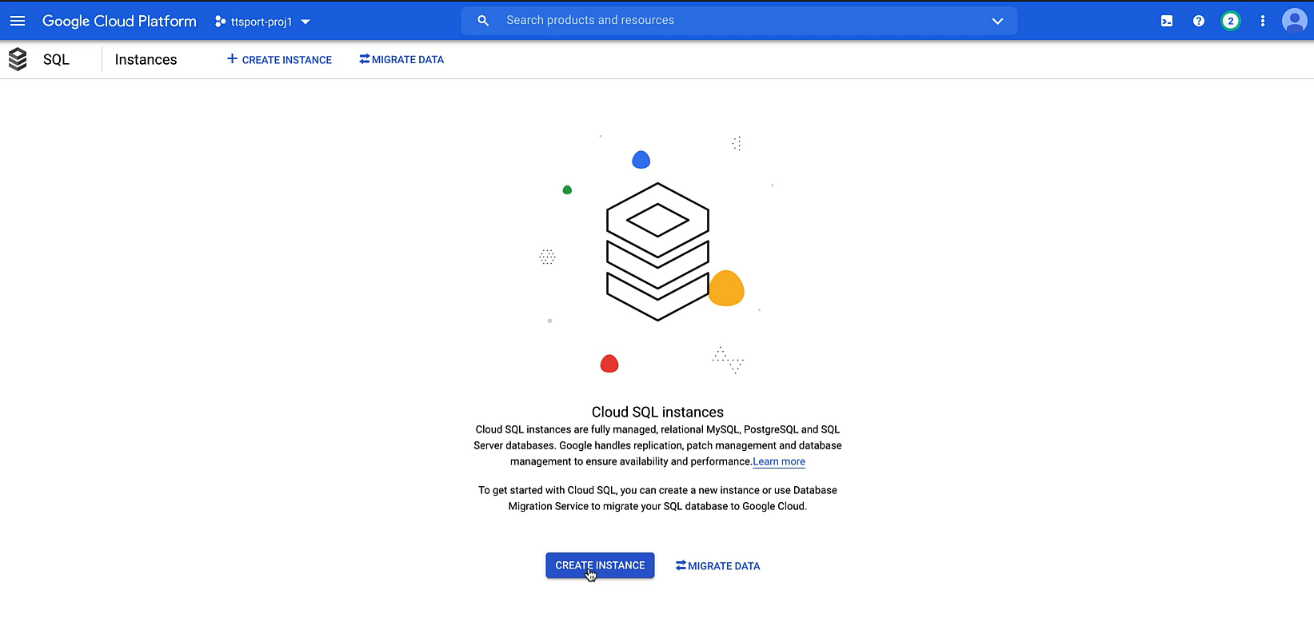
**Data Analytics Pipeline 1**

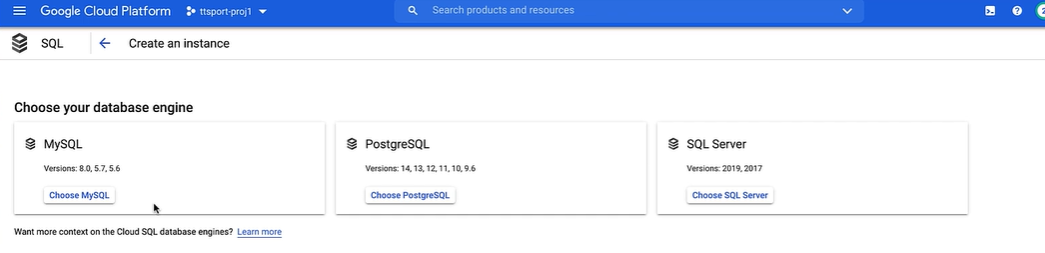
Flow of Pipeline:



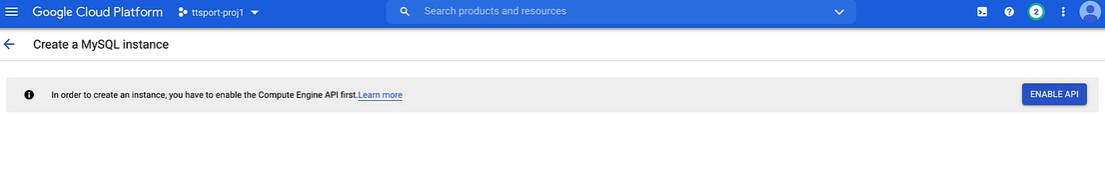
Creating My SQL database in Cloud SQL



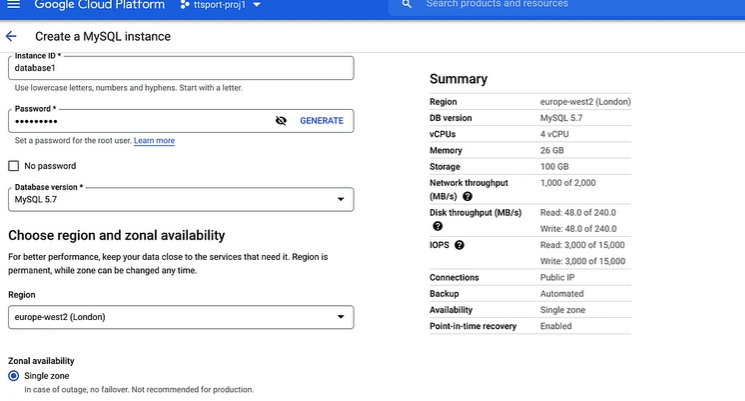
There will be three types of databases available as below:

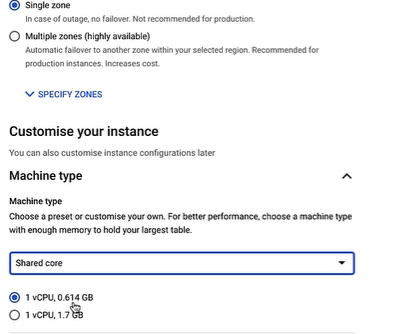


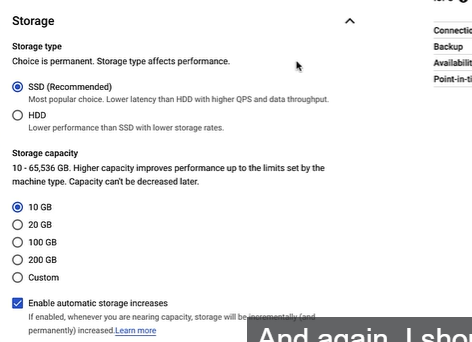
Click on My SQL and enable My SQL API:

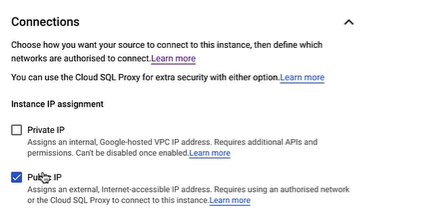


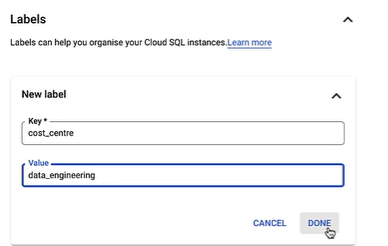
Fill up the details for creating a MySQL instance:





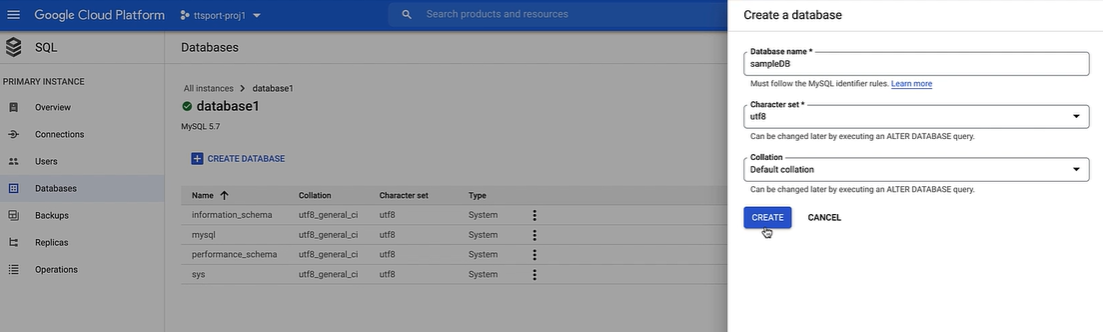




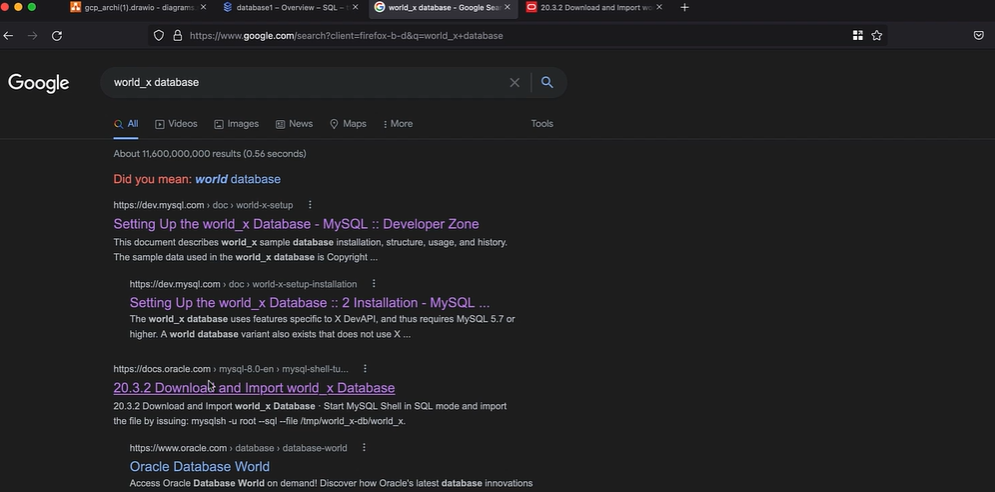


Rest will remain as default and then click on create instance.

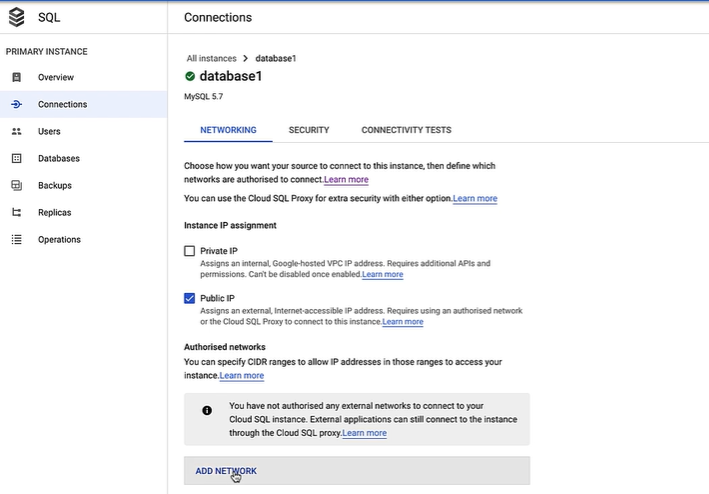
In Database section, Create a SampleDB database:



Now we will import data on our database and for that we need to first download the data from internet.



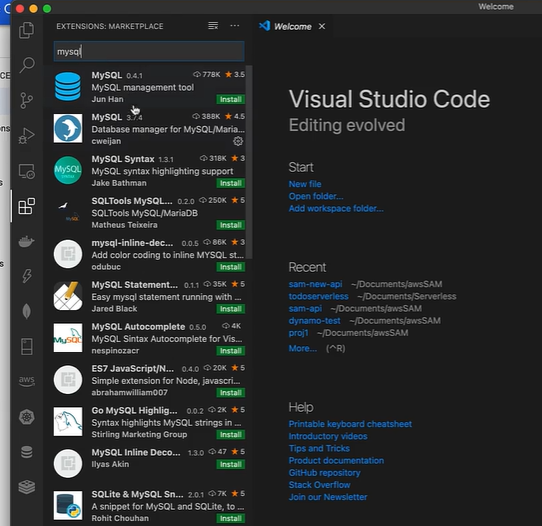
Now before we connect to the database, first whitelist the IP addresses. This means we can connect to the database only with the specified IP addresses. Follow below steps:



Give Name as Home let say and provide your IP address.

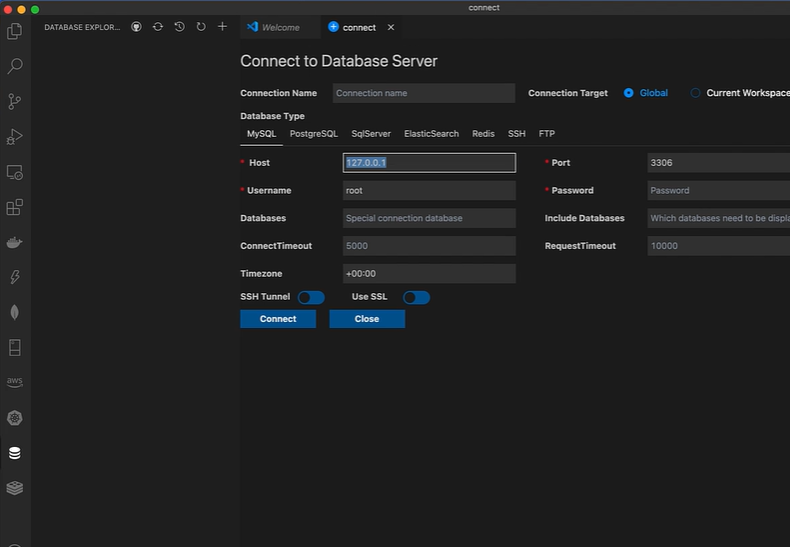
To connect we will use visual studio code

Download the extension



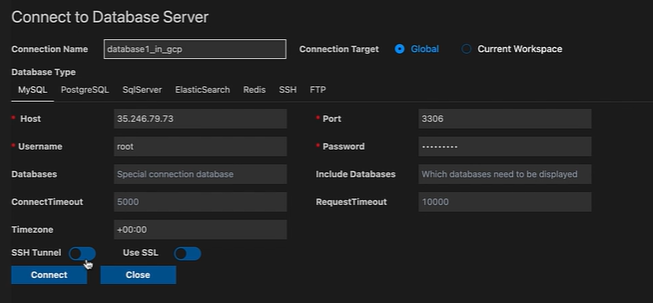


After that a database icon will come

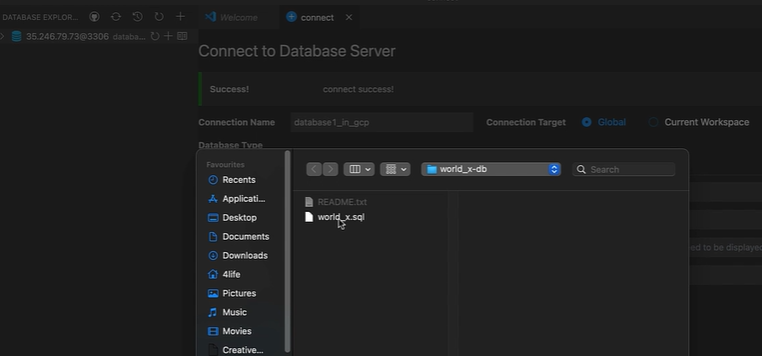


Click on + to make a new connection

Provide hostname which is the public IP address where the database is hosted



Then right click on connection name and click import SQL and select the world\_x database.

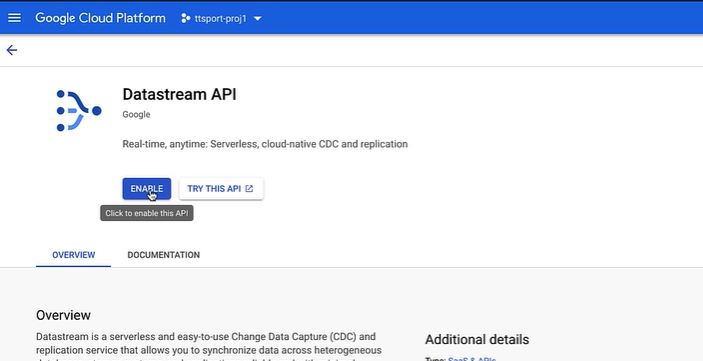


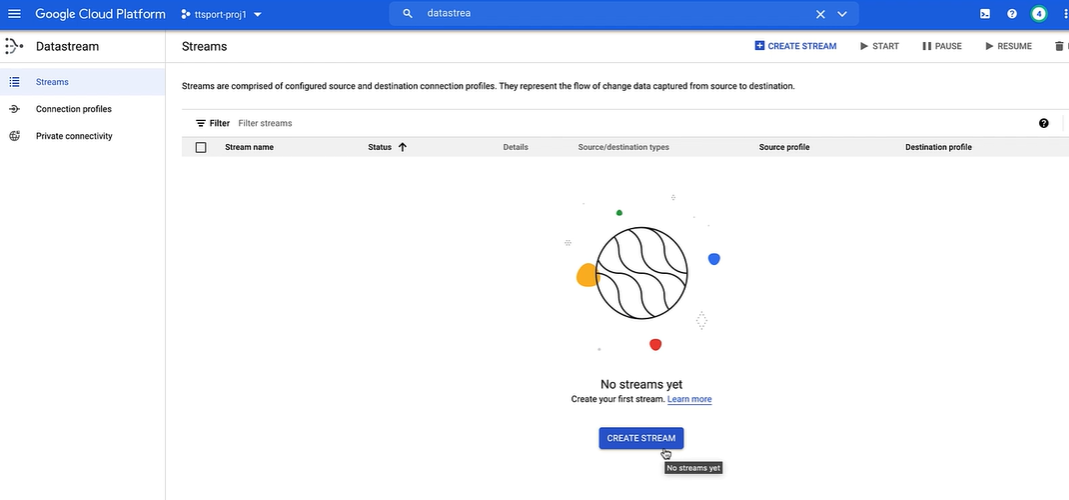
After this the database will get reflected in Cloud SQL.

Creating DataStream Instance

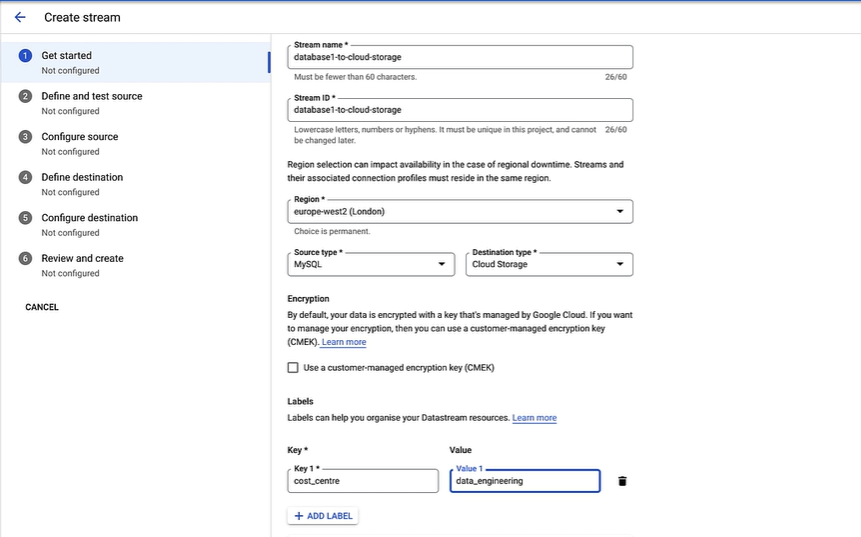
Data Stream provides step by step guide on how to create a stream.

Enable Data stream API

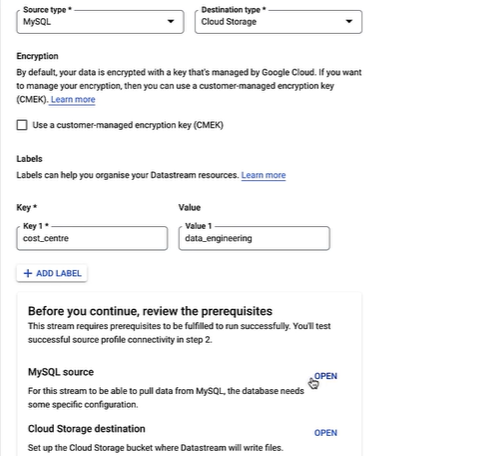




Fill Up the details:

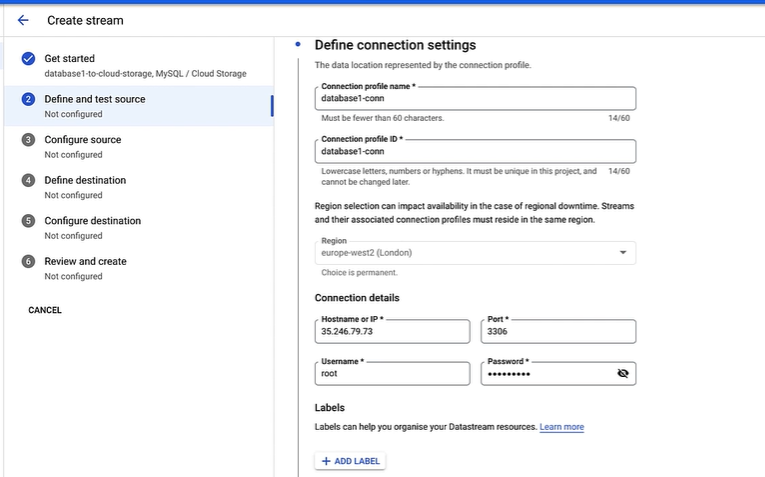


After this we need to see and review some pre-requisite.

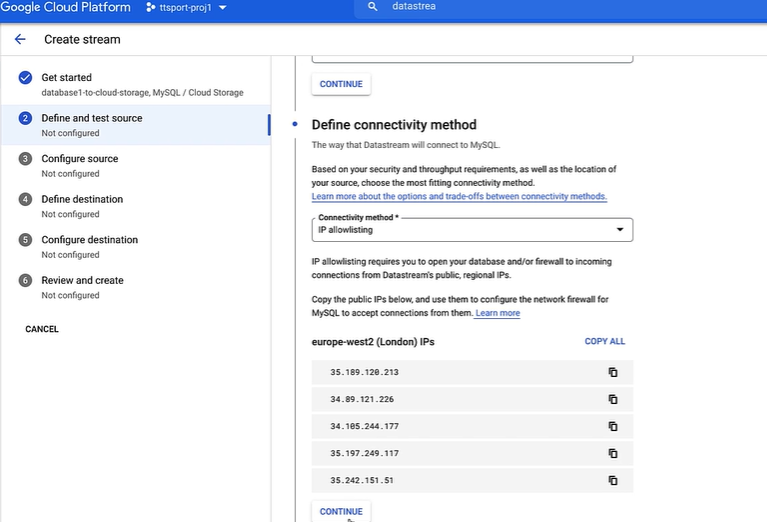


Leave them as it is if not required. As of now we will not require to do that.

Then we need to define the source connectivity details.



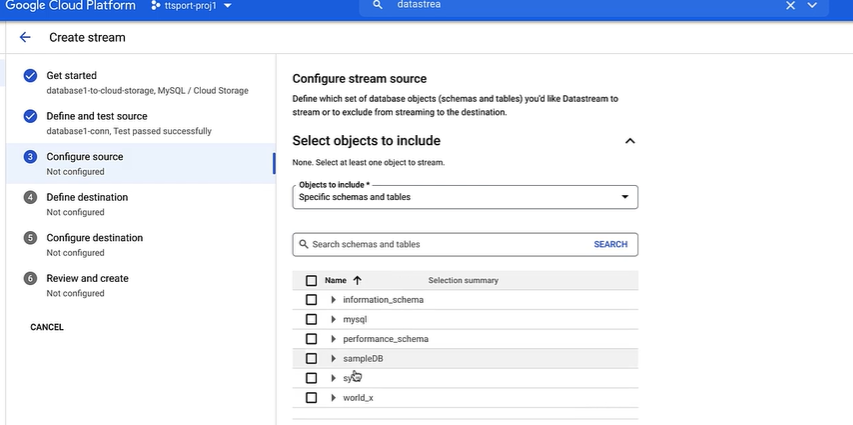
Then you need to select connectivity method like how this data stream will connect to the My SQL instance. For simplification we can click on IP Address Allowance.



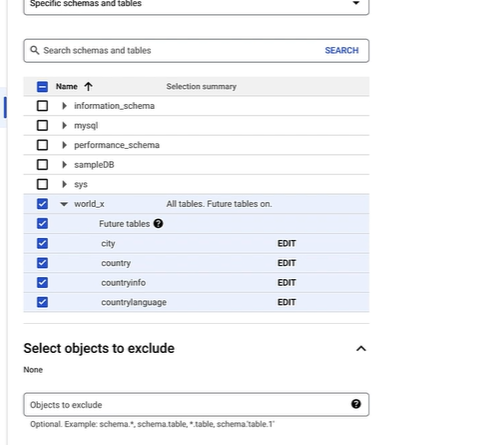
Then it will provide a list of IP Addresses, we need to copy that and then add them in network section under connections in My SQL instance one by one like we have done for our own IP address previously.

After that click continue.

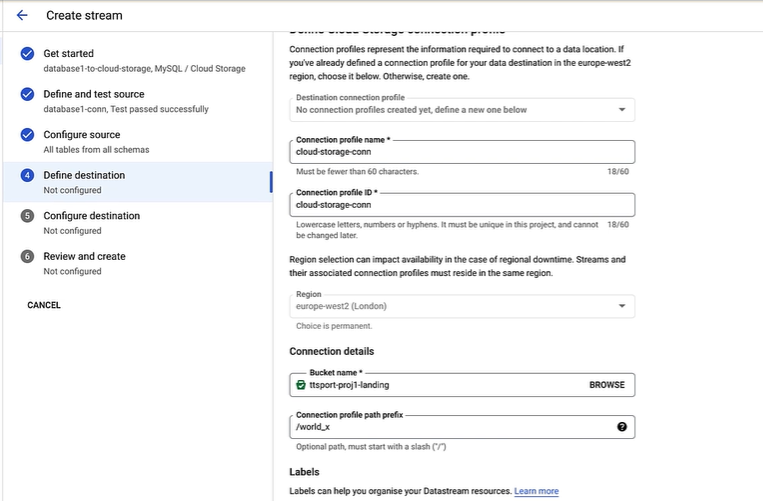
Then we need to select the data we need to replicate. Select specific schema or tables.



Select the tables



Now we need to specify target connectivity details.

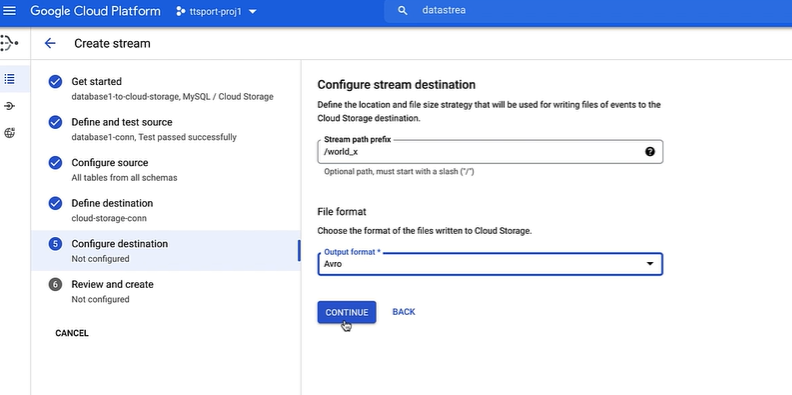


Specify any connection profile name select the bucket.

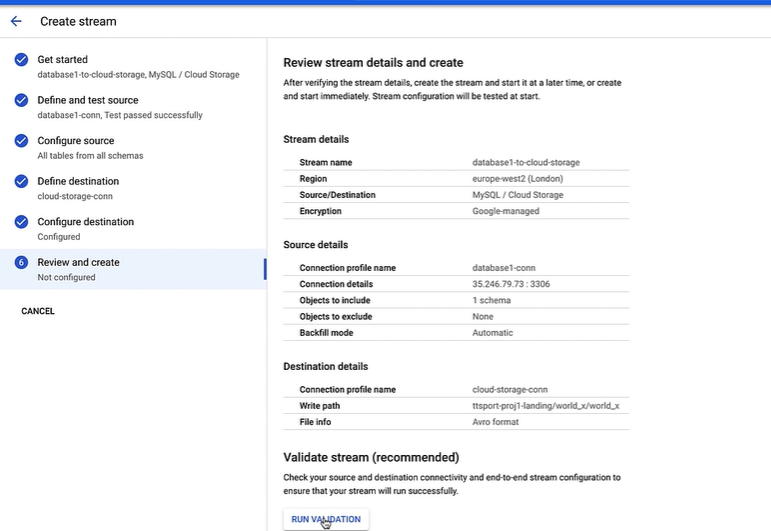
In connection profile path prefix, specify the folder name you want to create.

Click continue.

Then we can configure the stream path prefix we have provided.



At last run a validation.

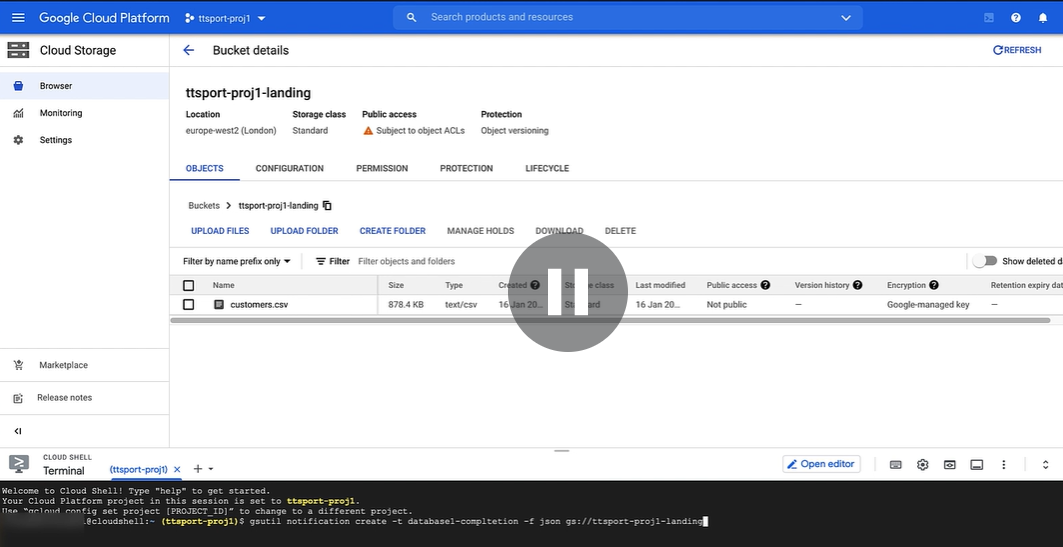


Then we can either Create or Create and start. For now, click on Create only.

Configure Pub Sub Notifications:

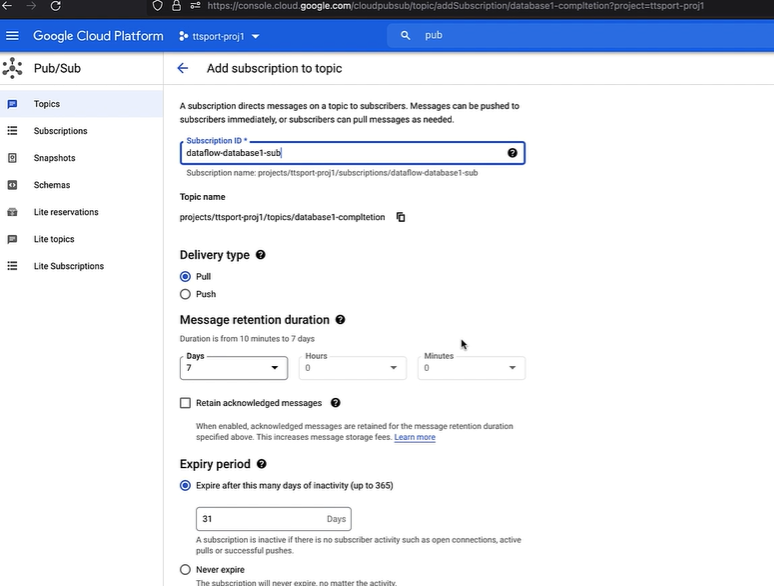
Now we will create a Pub Sub Notification that will notify the Data Flow that we will create to copy data from GCS to Big Query after the Data stream has copied the database in GCS.

Go to your GCS bucket, click on cloud shell and create notification.



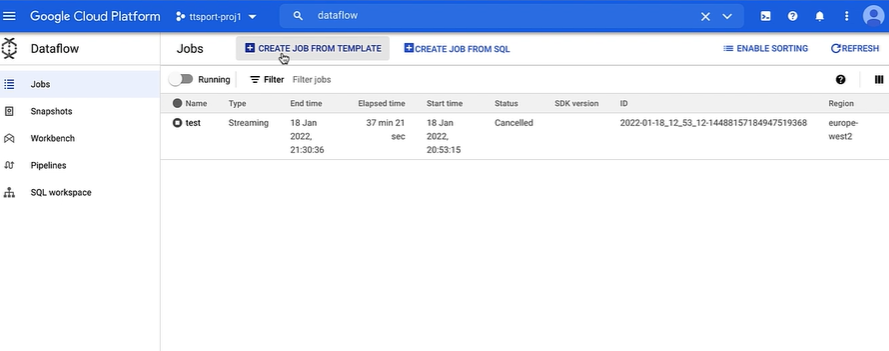
After this one pub sub topic will get created automatically that receive notification from GCS.

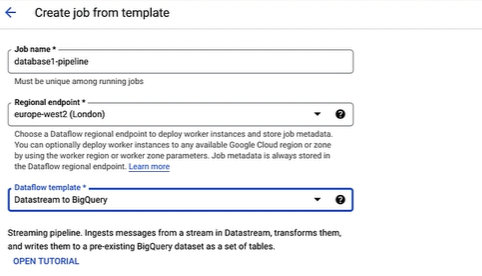
Then create a pub sub pull subscription that will be subscribed by dataflow that we will create further.



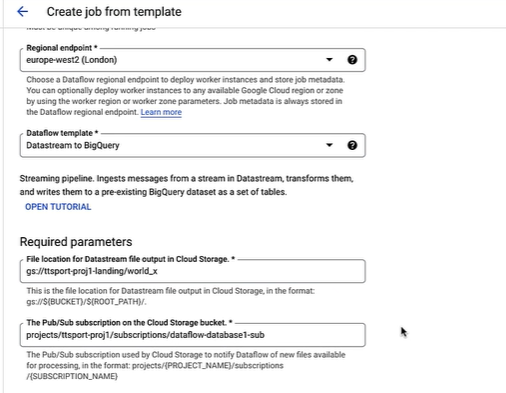
Set up Cloud Dataflow.

We will create the dataflow with the predefined templates.

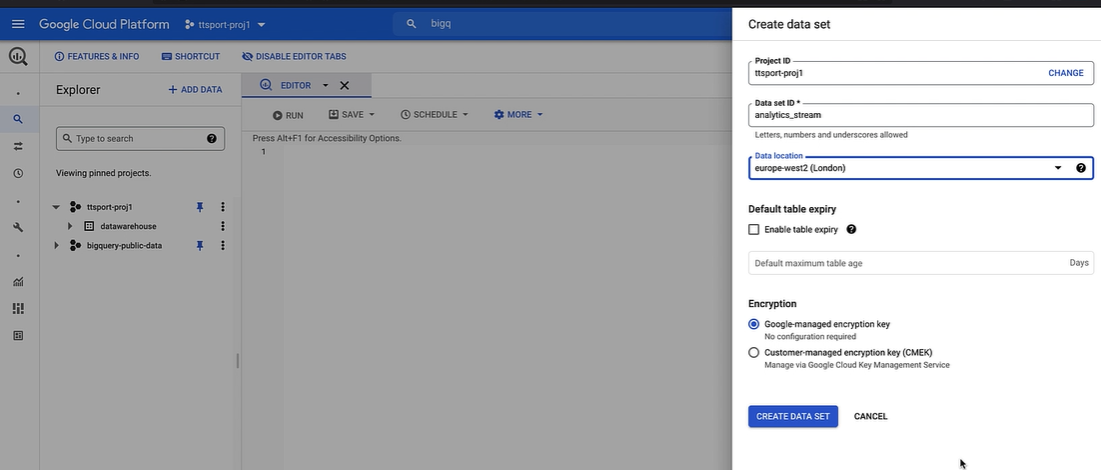




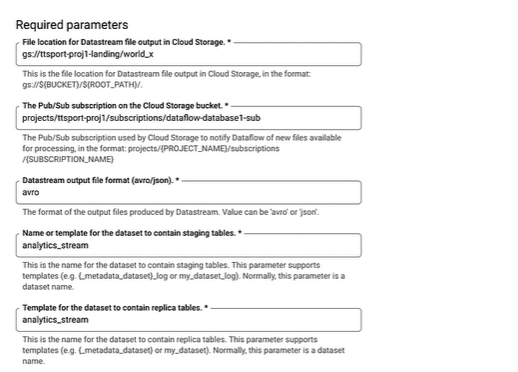
Then we need to provide the location of data Stream files in GCS and the subscription to which this dataflow job will subscribe so that it will run whenever that data Stream file got created.



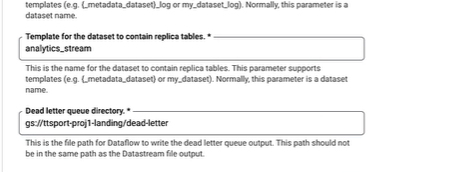
Now first create a dataset in Big Query



Provide that dataset name in the required details in dataflow creation.



Give location of one GCS folder etc. to contain the output whenever this dataflow fails.



This should not be same as data Stream file.

Then click run job.

This will not start as it will wait for pub sub notification.

So go to data Stream and run that. Then the files get copied to GCS and once that is done the notification will publish msg to pub sub topic and then as dataflow has subscribed to that topic it will run and copied the AVRO files to big query dataset mentioned.