Streaming data from Pubsub to bigquery

Prerequisites: GCP account

Flow: 1 Create a cloud storage bucket

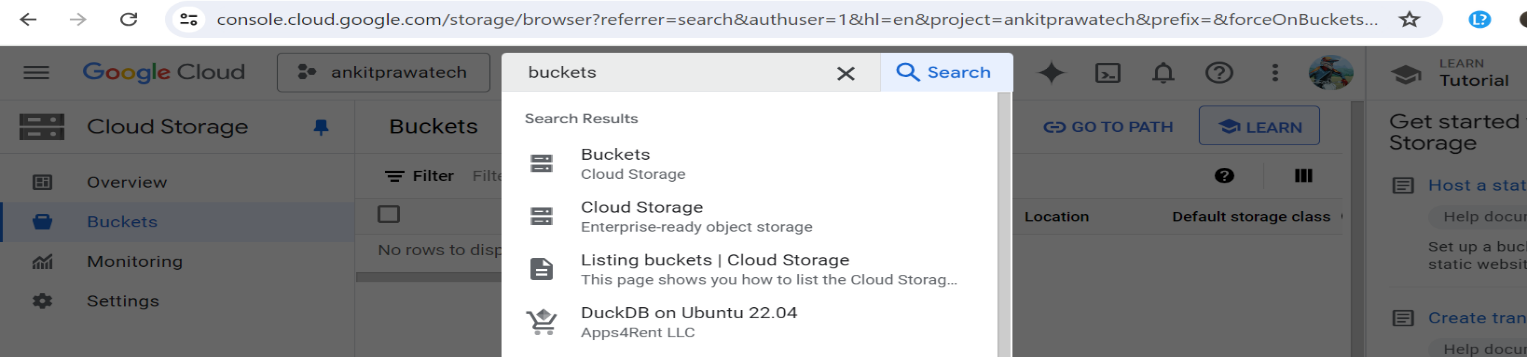
2 Create BigQuery dataset and Table

3 Create Pub/Sub Topic or subscription

4 Run Dataflow

**Step 1 Create cloud Storage bucket**

On the home screen/dashboard search bucket – select the first option

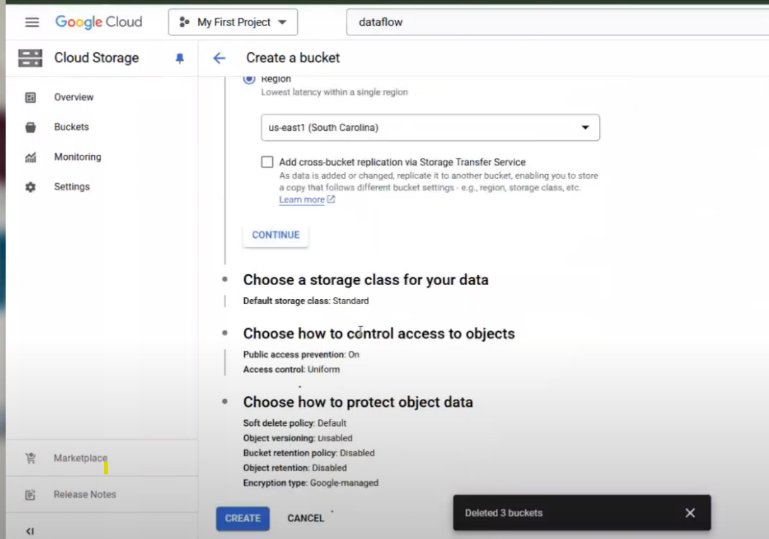


-Provide the name of the bucket which is globally unique

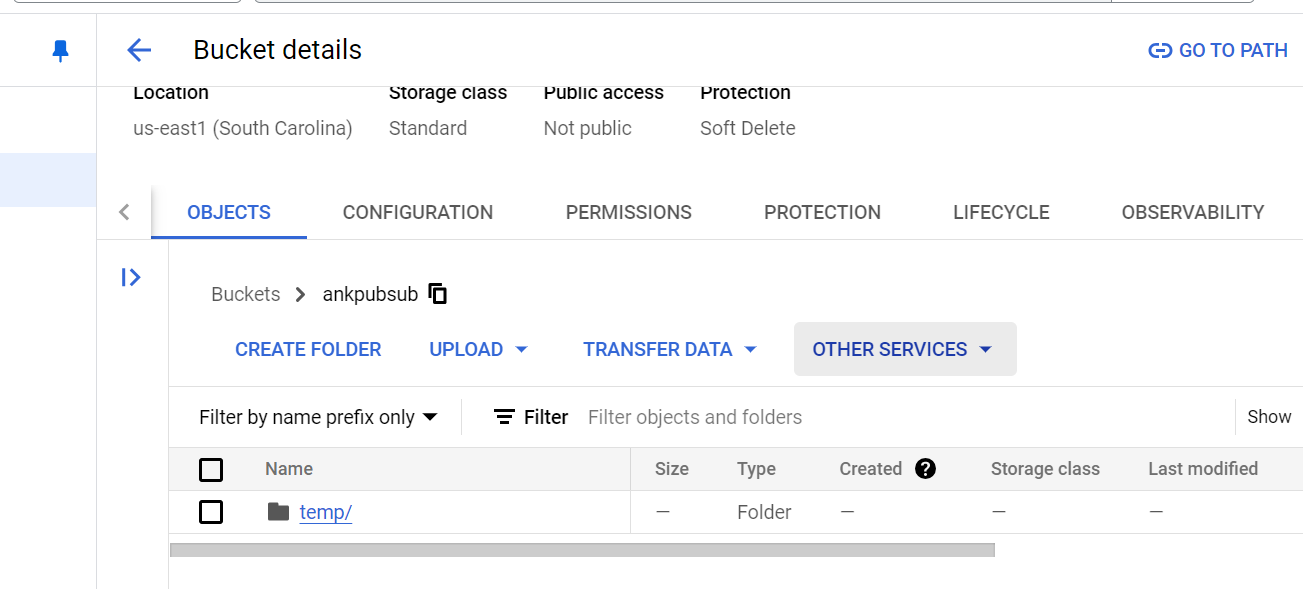
-Then chose where to store your data

-Keep the other settings to default and proceed to the next step.

-Finally lick on the create bucket option



- create a folder inside the bucket named “ temp” which should look like this

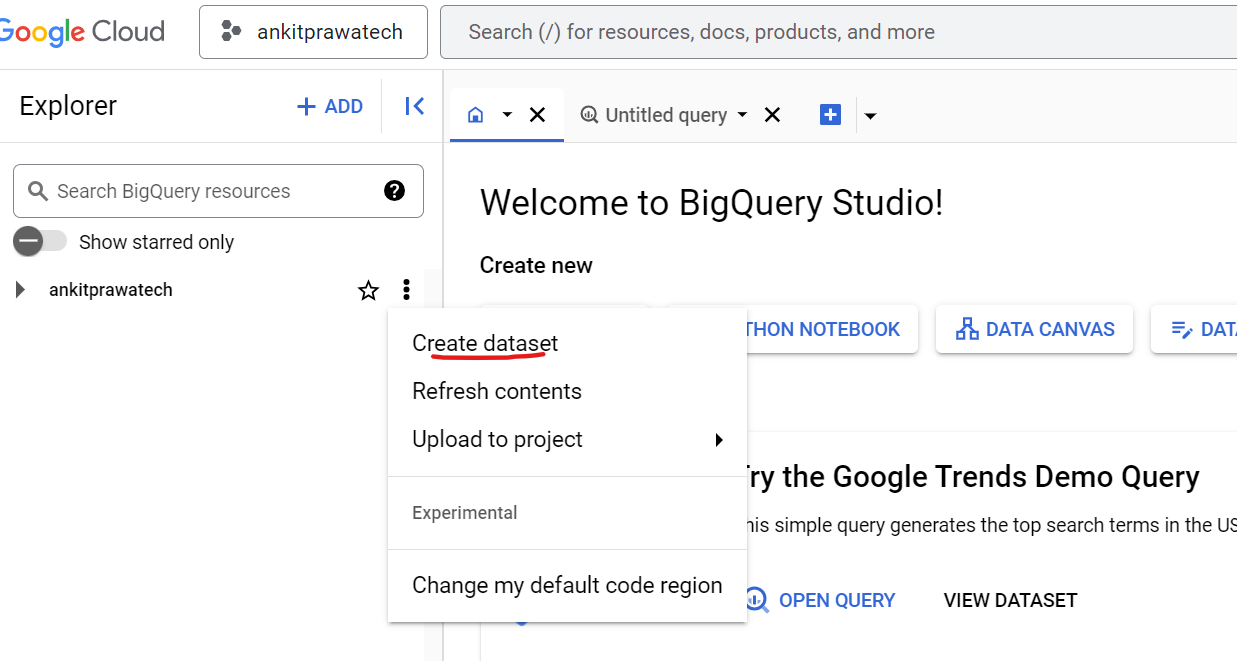


**Step 2: Create BigQuery Dataset**

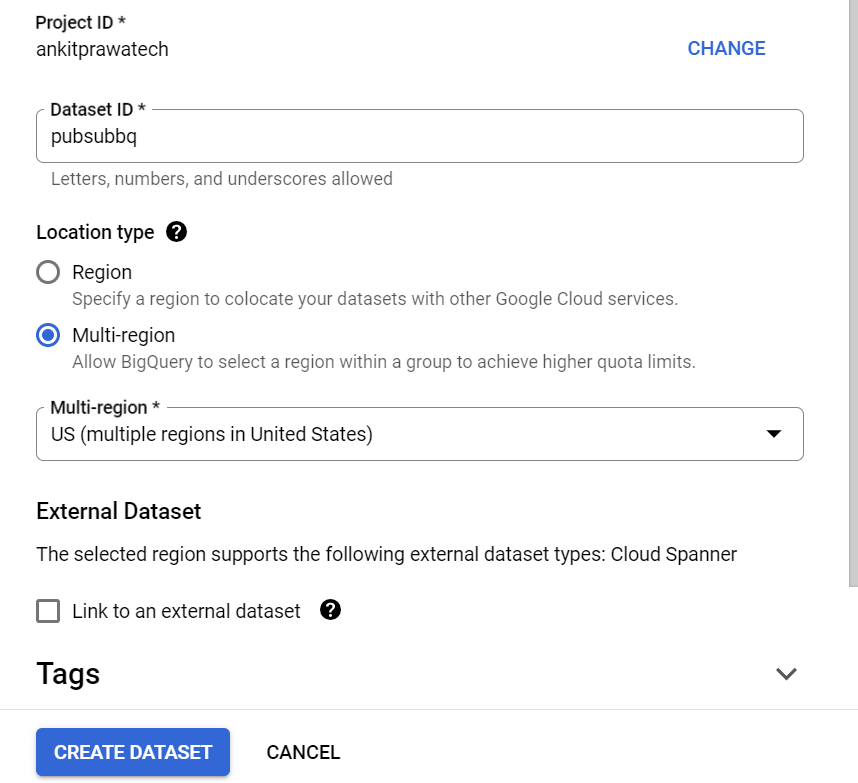
-search bigquery on the search bar and select the first option

-You can open it in new tab for easy access

- click on the 3 dots option beside the project name as displayed on the left hand side of the screenshot and select the “create dataset “ option



- Name your dataset keeping the remaining settings as it is



-Now we need to create a table inside this dataset

- On the left hand side of the screen we an see the project name , click on the 3 dots icon and select “create Table” option

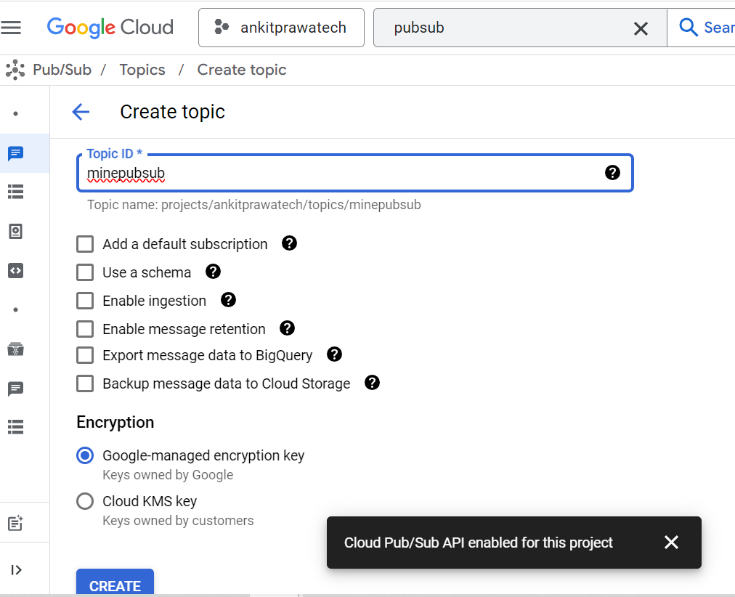
-Provide the table name and schema details accordingly

- Finally lick on the create Table option and now we have our Table ready

**Step 3: Create Pub/Sub Topic**

- search pubsub in the search bar and select the first option.

- click the “Create Topic” option , input the name and select the create button.



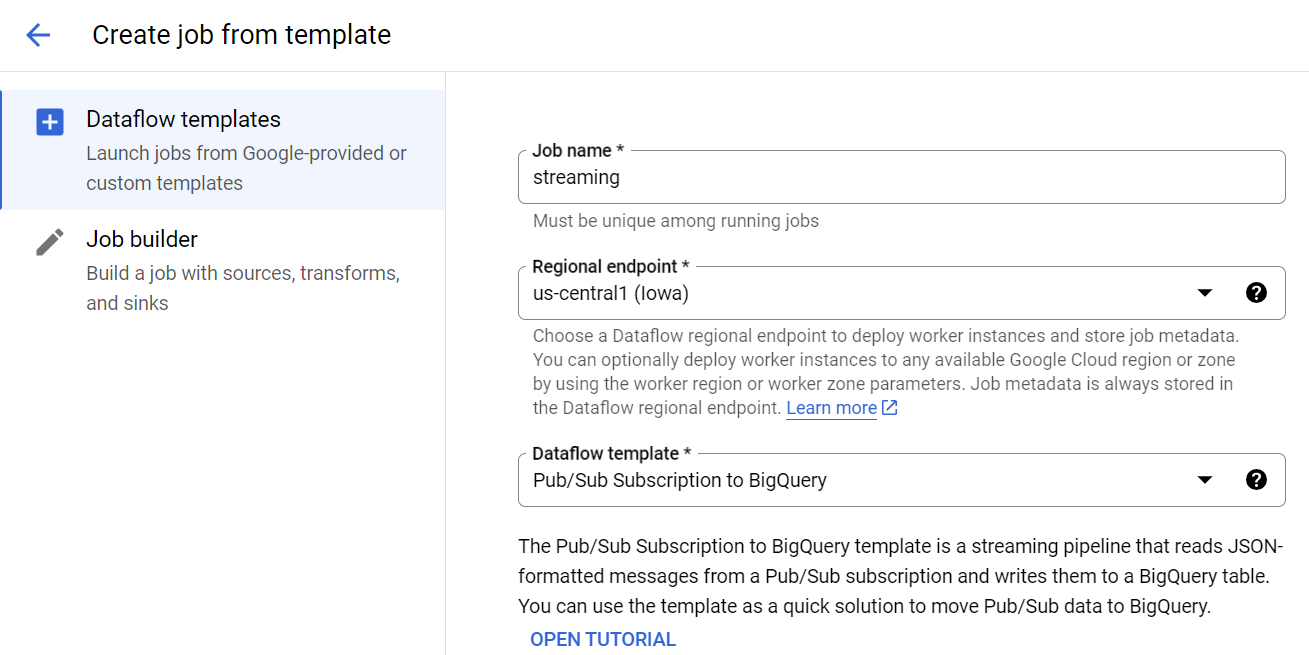
- then select the option “create subscription” from the left menu and provide the name under “Subscription ID”

- select the “cloudPub/Sub topic” which will appear by default

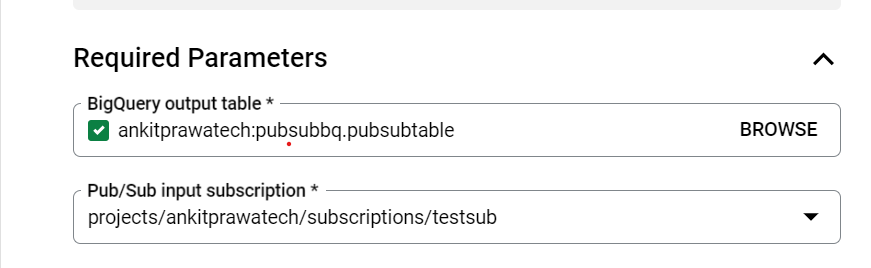
- Finally select the create subscription button

**Step 4: Run Dataflow**

-search for Dataflow to open it

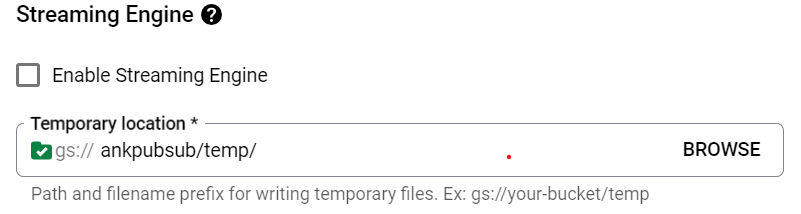
-under the jobs select “Create jobs from Template” 

-select the Data flow template as “Pub/sub Subsription to Bigquery

-under the required Parameters browse the table that you created

- select the streaming node as “exactly one”

-under the pun/sub input subscription select the one you want

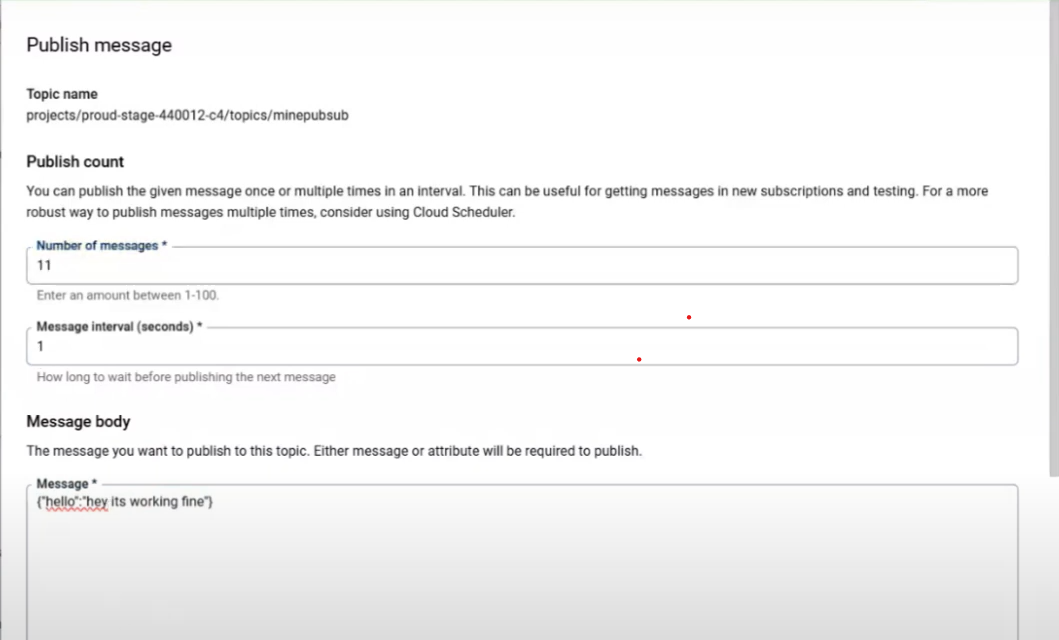
-Below the streaming engine option browse the bucket location which we created in the Step 1 

- Enable the Dataflow prime

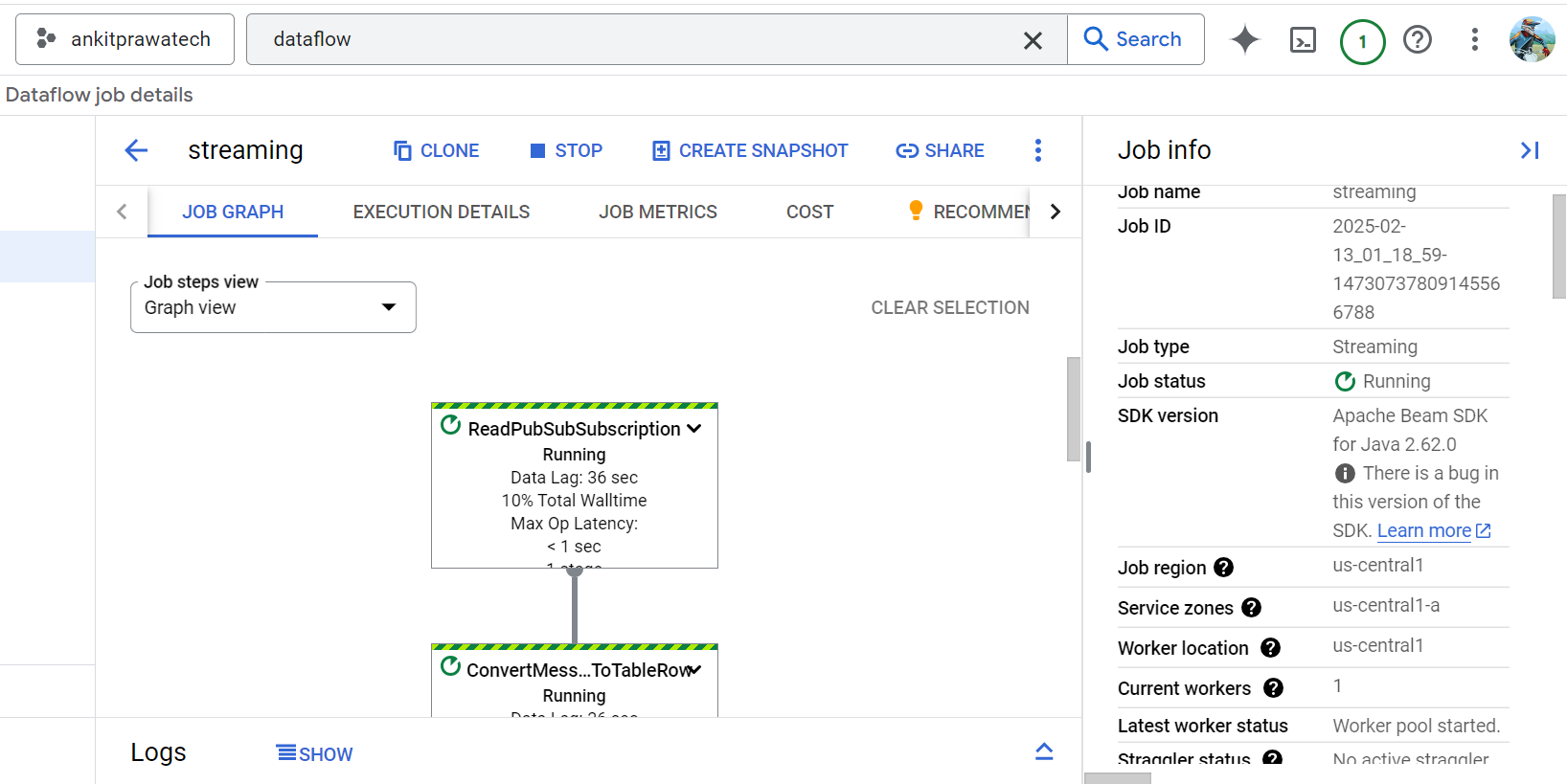
-Run the job

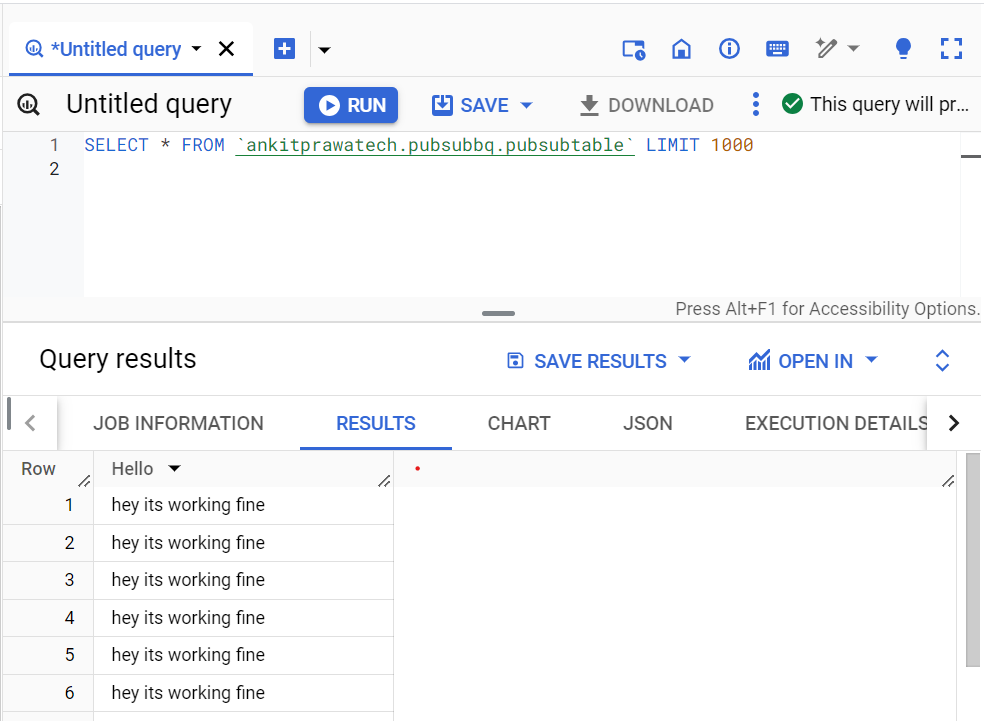
-the job will take some time to execute depending upon the resource allocation

- Now to pass the input go to pubsub->topic->message then to publish message



-then under the pubsub->topic->message for the step 2 select the pubsub subscription to pull the messages

- to validate go to dataflow to check if its streaming and cross check through bigquery 

- check under the bigquery if the data has been inserted into the dataset’s table , for it use the query SELECT \* FROM `ankitprawatech.pubsubbq.pubsubtable` LIMIT 100 to view the result/ table details 

-The job will keep on running until its stopped .