LOW-LEVEL DESIGN DOCUMENT

**Designing a Streaming Data Pipeline with Dataflow and visualization dashboard using Looker**

Created Date: 10/04/2023

TABLE OF CONTENTS

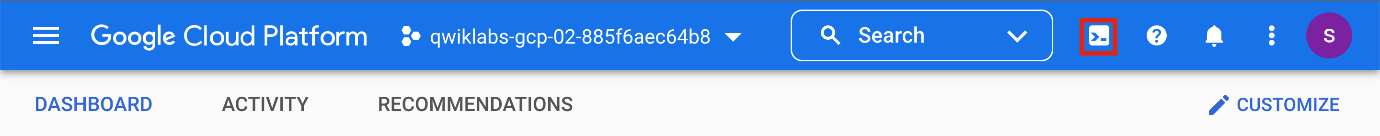
1. Setting up of Prerequisites
2. Create Pub/Sub topic
3. create a BigQuery dataset and Table
4. Create a Cloud Storage bucket
5. Python Code to Publish messages to Pub/Sub topic
6. Data in Pub/Sub
7. Set up a Dataflow Pipeline
8. Analyze the data using BigQuery
9. Create a real-time dashboard using Looker

**Task 1.** **Setting up of Prerequisites**

1. **Activate Cloud Shell**

Google Cloud Shell is a virtual machine that is loaded with development tools. Google Cloud Shell provides command-line access to your Google Cloud resources.

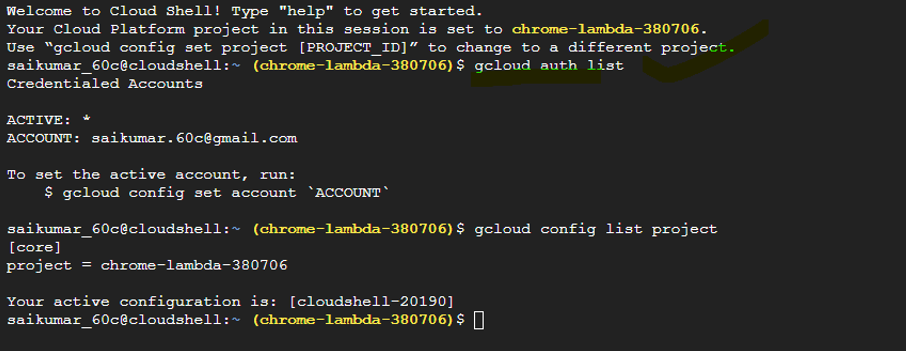
* In Cloud console, on the top right toolbar, click the Open Cloud Shell button.



You can list the active account name with this command:

gcloud auth list

gcloud config list project



## **Confirm that needed APIs are Enabled**

In the GCP Console, in the Navigation menu, click **APIs & Services** and ensure that the following APIs are enabled:

* Cloud Pub/Sub API

Graphical user interface, text, application, email

Description automatically generated

* Dataflow API

Graphical user interface, text, application, email

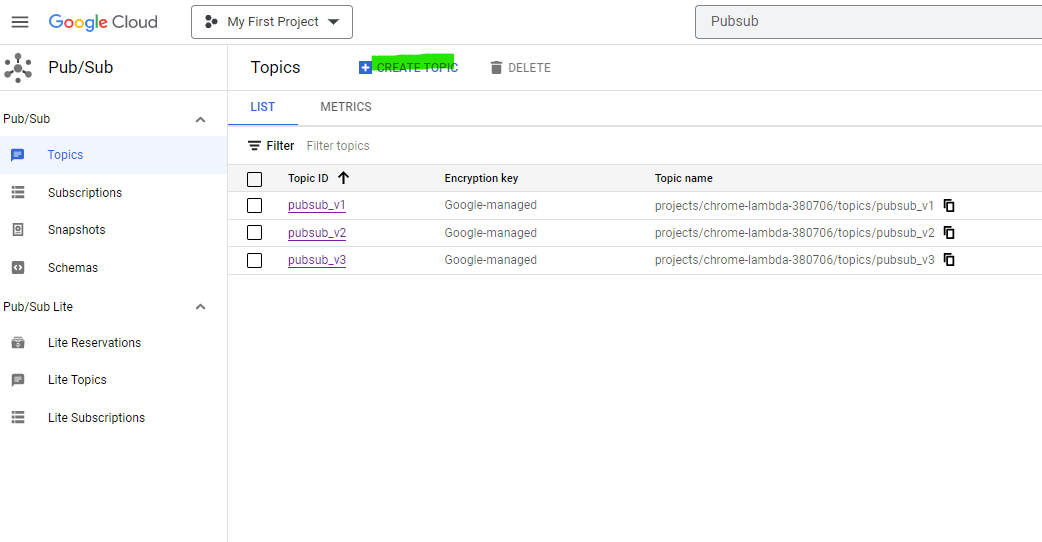
Description automatically generated

**Task 2.** **Create Pub/Sub topic**

**#Approach 1**

**Creating Topic in PubSub**

* In the Cloud console, in the Navigation menu , click on PubSub > .



* Give topic ID pubsub\_v2, then click Create Topic.

Graphical user interface, text, application, email

Description automatically generated

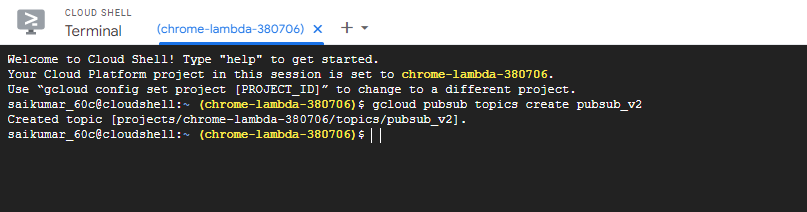
Creating topic can also be achieved using cloud shell

**#Approach 2**

Go to the Cloud Shell

Use below gcloud command create topic from cloud shell

* gcloud pubsub topics create pubsub\_v2



**Task 3) Creating BQ Dataset & Table**

* In the Google Cloud console, Navigateto BigQuery.
* In the explorer, next to your Project ID, Click on View actions and then click Create dataset.

Graphical user interface, text, application

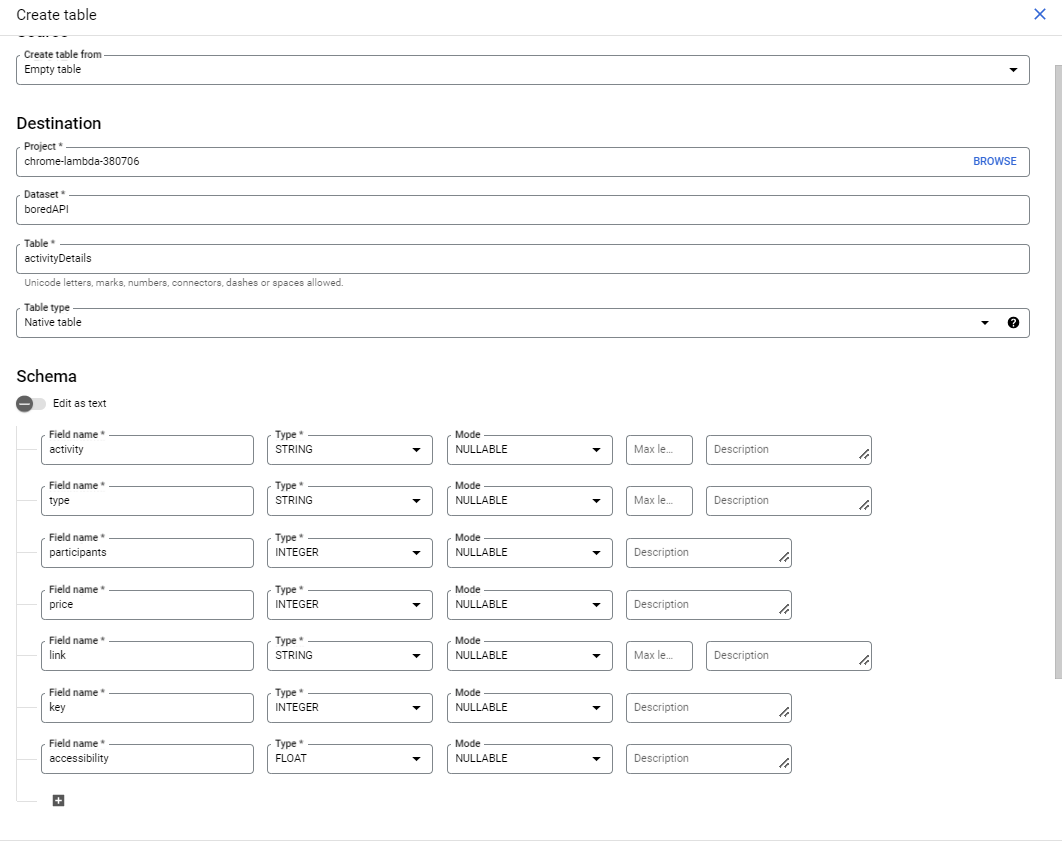
Description automatically generated

* In Dataset ID, type boredAPI and select Location Type as shown below.

Graphical user interface, text, application, email

Description automatically generated

* In the Explorer pane, click expand node to reveal the new boredAPI dataset.
* Click on View actions next to the boredAPI dataset
* Click Create Table.
* In Table, type table name as activityDetails
* For adding the table schema, click on add fields
* Click Create Table



Note: Make sure add schema to the table as per website data

Sample data in the website - <https://www.boredapi.com/api/activity/>

{"**activity**":"Memorize a favorite quote or poem","**type**":"education","**participants**":1,"**price**":0,"**link**":"","**key**":"9008639","**accessibility**":0.8}

**Task 4. Create a Cloud Storage bucket**

In this task, you create a Cloud Storage bucket to provide working space for your Dataflow pipeline.

* In the Cloud console, Navigate to Cloud Storage > Buckets.
* Click Create Bucket.
* Give the bucket name same as Project ID, and then click Continue.
* Click Create.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Task 5. Python Code to Publish messages to Pub/Sub topic**

**a)** Use Visual Studio code in your system to run the below pip commands in the terminal to Install the required libraries

* pip install google.cloud
* pip install google-cloud-pubsub
* pip install -U requests

Text

Description automatically generated

Text

Description automatically generated

**b) create a service account to access the Google cloud from Visual Studio code**

* In the Cloud console, navigate to IAM & Admin > Service accounts.
* In the top menu bar, click Create Service Account
* In Service Account name, type Streaming Data
* Click on Create and Continue
* In select role Search for Pub/Sub Publisher and click on continue and Done Graphical user interface, text, application, email

  Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

* Click on that service account you have created
* Click on **Keys** and **ADD KEY** and **Create New Key**
* Choose Key Type as Json and this will download the .json file to your Local system

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

**c) Python code to extract the website data to Pubsub topic**

Run the below code in Visual Studio Code and this code will trigger/refresh website -https://www.boredapi.com/api/activity/ and store the data into Pub/Sub Topic(pubsub\_v2)

In the credentials path give the path of .json key that you have downloaded in the previous step.

from google.cloud import pubsub\_v1

from pip.\_vendor import requests

#import pip.\_vendor.requests

import time

import os,ssl

credentials\_path=r'C:\Users\Saikumar Cheppa\Downloads\chrome-lambda-380706-76cc3c25caf2.json'

os.environ['GOOGLE\_APPLICATION\_CREDENTIAL'] =credentials\_path

api="https://www.boredapi.com/api/activity/"

publisher=pubsub\_v1.PublisherClient()

topic\_path='projects/chrome-lambda-380706/topics/pubsub\_v2'

while True:

    response=requests.request("GET", api, verify=False).text

    publisher.publish(topic\_path, data=response.encode('utf-8'))

    time.sleep(15)

>To tun the code right click on the code and select ‘Run Python File in Terminal

And check the logs in the terminal Text

Description automatically generated

**Task 6) Verify the data in Pub/Sub**

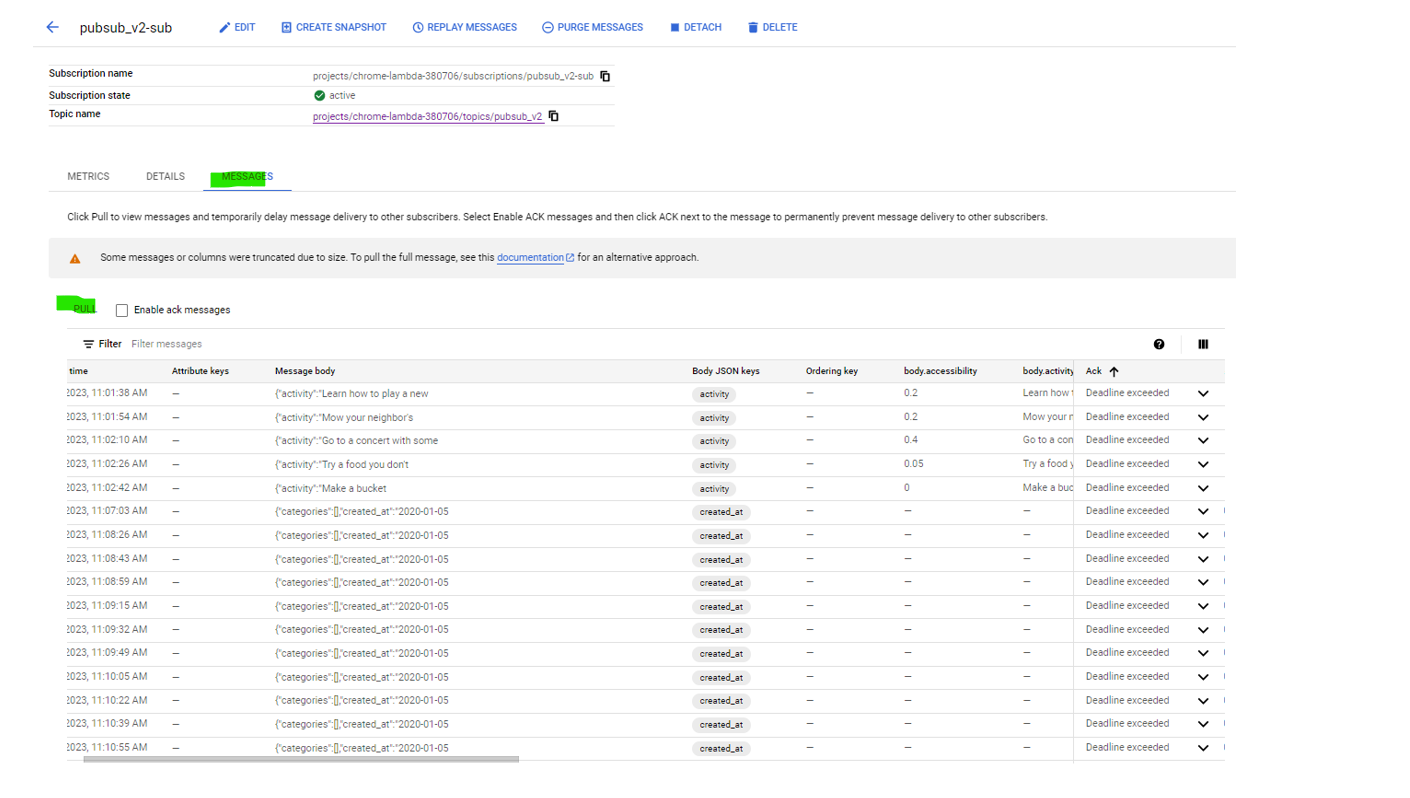
To check Streaming data in Pub/Sub

* In the Cloud console, navigate PubSub >
* In Topics click on **pubsub\_v2**(topic created in task 1)and thenclick on **pubsub\_v2-sub**

Graphical user interface, text, application, email

Description automatically generated

* Click on Messages and then click on PULL to check If data is streaming with latest timestamp or not

x

**Task 7. Set up a Dataflow Pipeline**

Once the data available in Pub/Sub and now create Dataflow Template

**Create a new streaming pipeline:**

* In the Cloud console, Navigateto Dataflow.
* In the top menu bar, click Create Job from Template.
* Type streaming-activityDetails-pipeline-Bored API as the Job name for your Dataflow job.
* In Regional endpoint, select us-central1 (Iowa).
* In Dataflow template, select the Pub/Sub Topic to BigQuery template.
* In Input Pub/Sub topic, select the topic that already exists in your project from the dropdown list . It will appear like the following:

**projects/<myprojectid>/topics/pubsub\_v2**

* In BigQuery output table, Click on browse and select table as activityDetails
* In Temporary location, click Browse.
* Click view child resources(view child resources).
* Click Create new folderand then type the name tmp.
* Click Create, and then click Select tmp folder.
* Click Show Optional Parameters.
* In Max workers, type 2
* In Number of workers, type 1
* Click Run Job.

­ Graphical user interface

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

A new streaming job has started! You can now see a visual representation of the data pipeline. It will take 3 to 5 minutes for data to begin moving into BigQuery.

**Task 8. Analyze the data using BigQuery**

In this task, you analyze the data as it is streaming.

* In the Cloud console, navigate to BigQuery.
* In the Query Editor, type the following, and then click Run:

select \* from chrome-lambda-380706.boredAPI.activityDetails

Graphical user interface, table

Description automatically generated

**Note:** If no records are returned, wait another minute and re-run the above query (Dataflow takes 3-5 minutes to setup the stream).

**Task 9. Create a real-time dashboard using Looker**

* In the Query Editor, clear the current query.
* Copy and paste the following query, and then click Run

select

activity, type,participants, link, accessibility

 from chrome-lambda-380706.boredAPI.activityDetails

* Click Save > Save query.
* In the Save query dialog, in the Name field, type **My Saved Query**.
* Click Save.

Table

Description automatically generated

* In BigQuery, click Explore Data > Explore with Looker Studio.

Graphical user interface, application

Description automatically generated

* Looker Studio Opens as below

Graphical user interface

Description automatically generated

* In the Looker Studio window, customize your reporting dashboard as per requirement

Chart, bar chart

Description automatically generated