Documentation

Arooba jamil khokhar

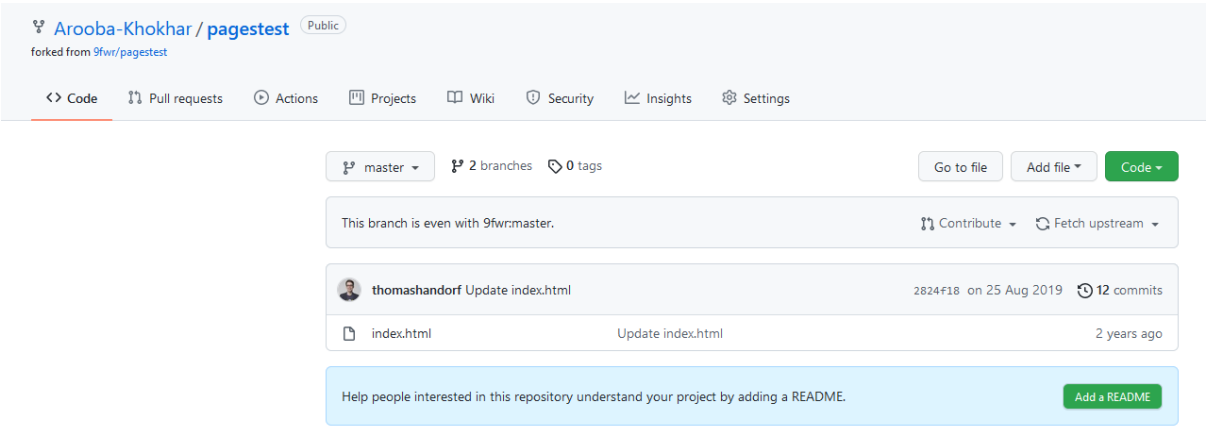
Prerequisite:

Create accounts on following sites

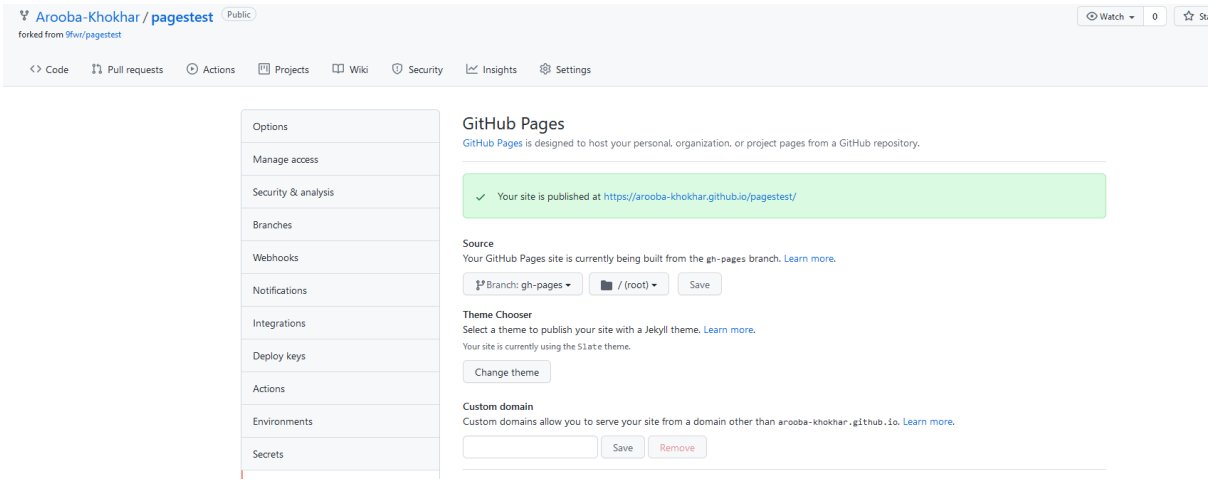
- Create account on Segment
- Create Google Analytics account
- Create Service account
- Create Bigquery account
- Configure Cloud Platform
- google analytics
- Installing Cloud SDK
- Google Data Studio

1. Data Acquisition

- Fork segment project



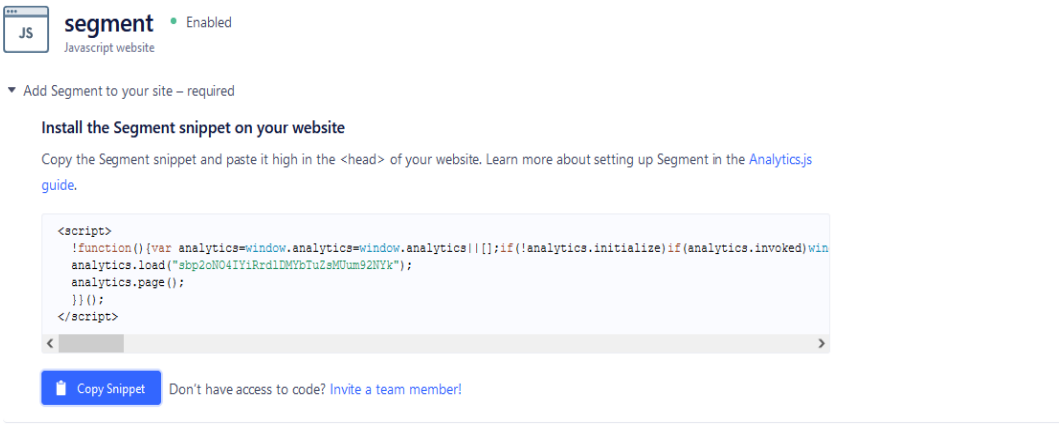
- created github page (https://arooba-khokhar.github.io/pagestest/)



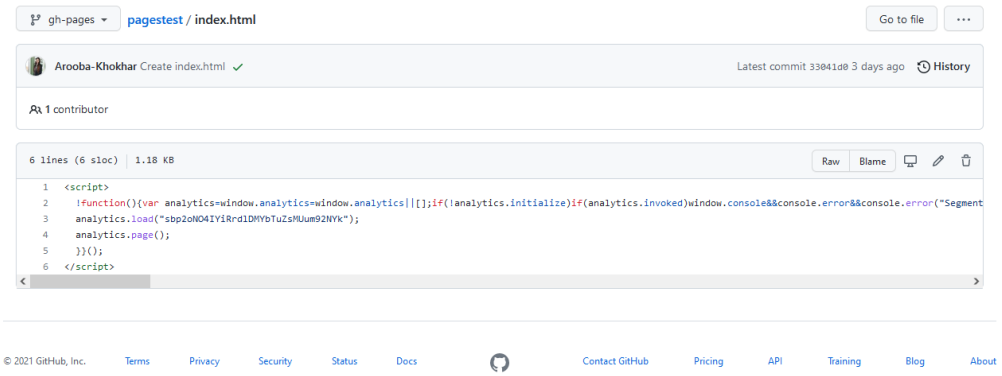
- Github page

2. Data Warehouse

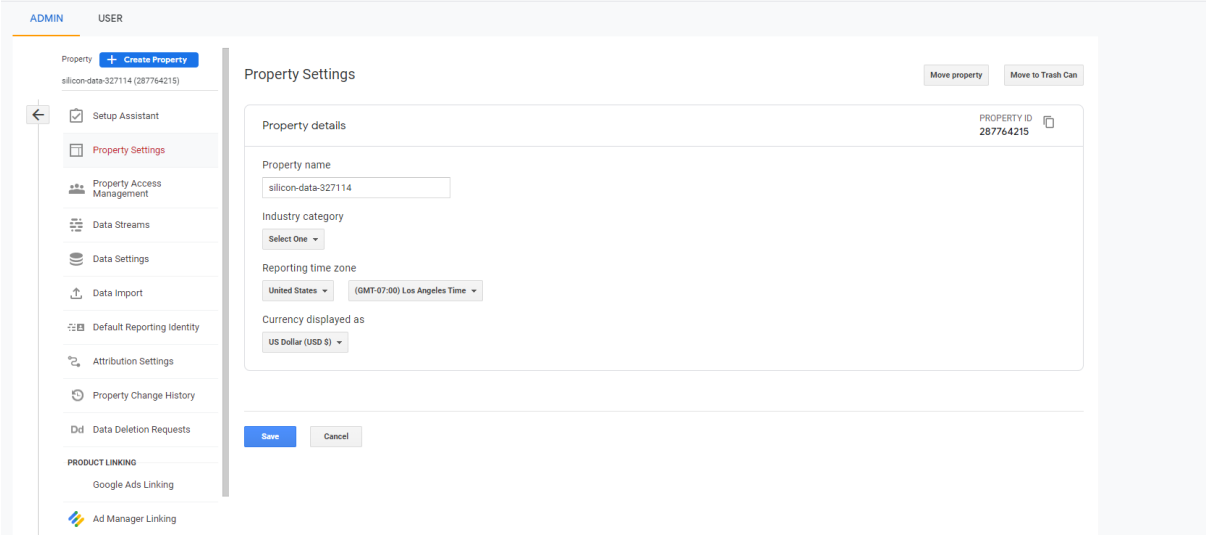
- Create account on Segment
- Add source
- Copy the Segment snippet



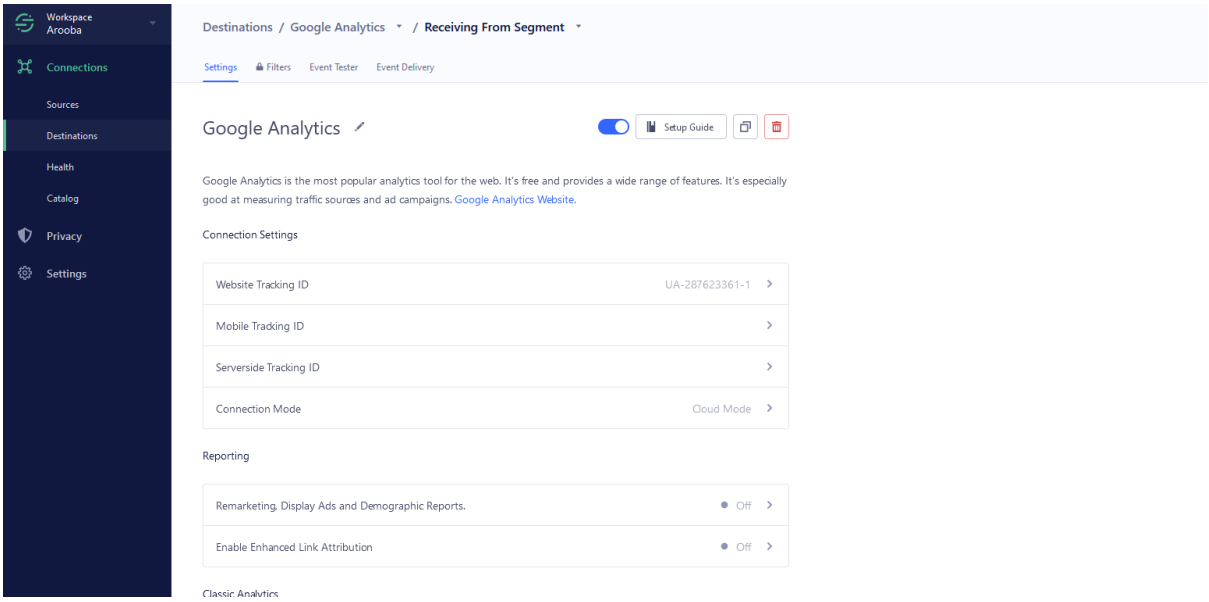
- I pasted it high in the <head> of my website.



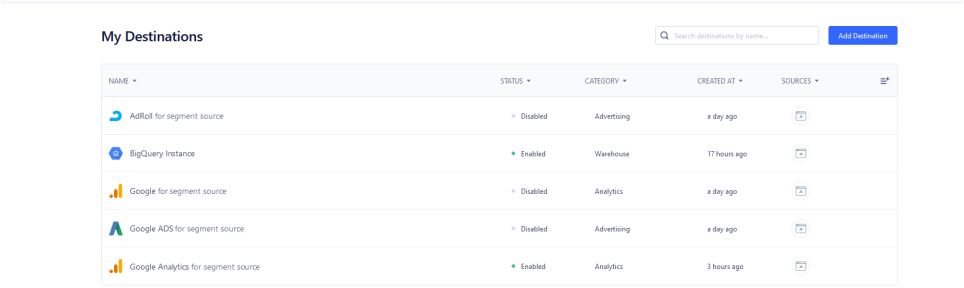
- I created account on google analytics



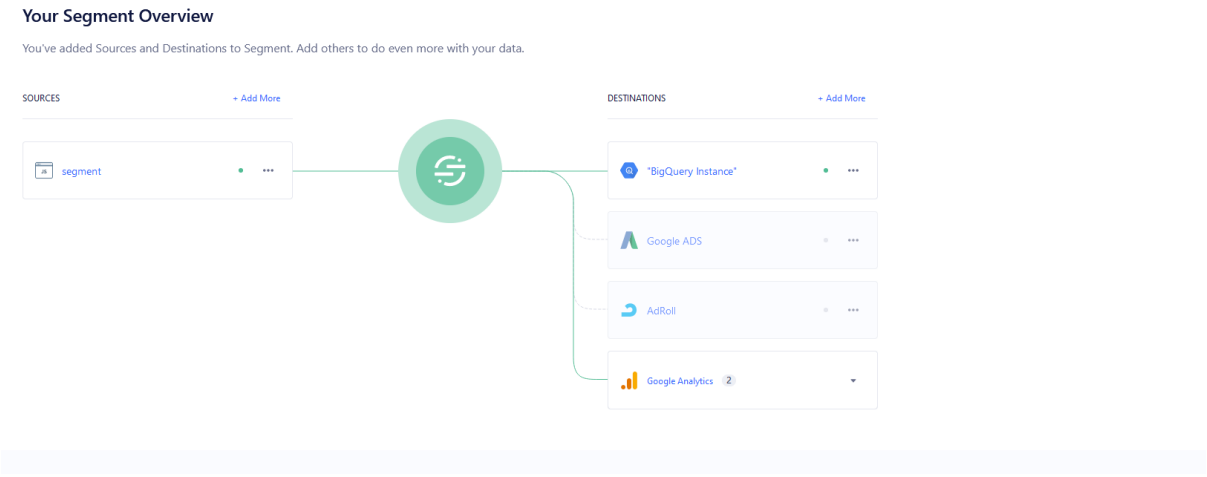
- Connected destination to google Analytics

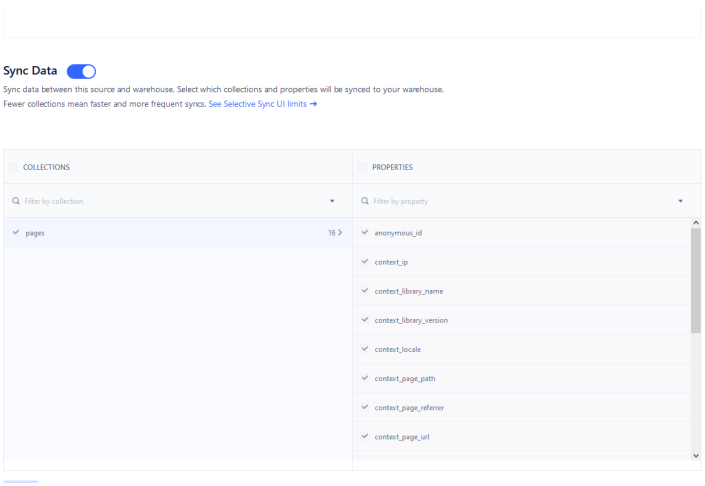


- My Destination

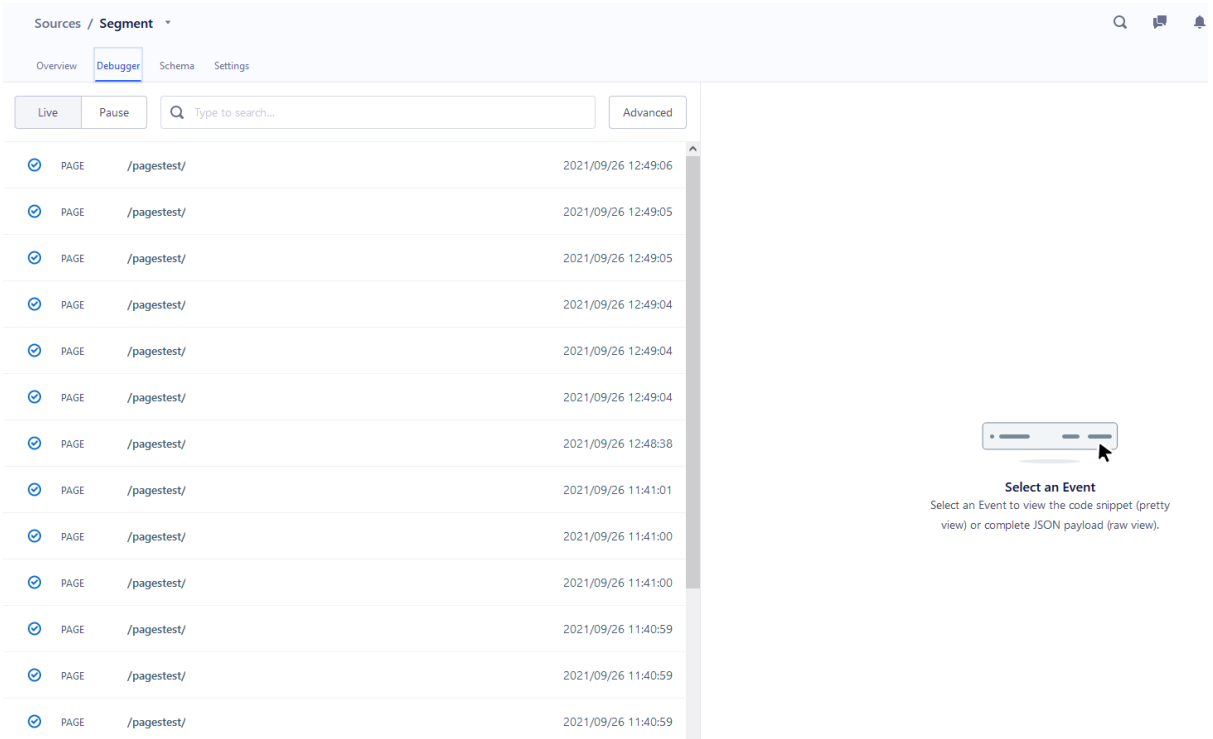


- connected source to destination

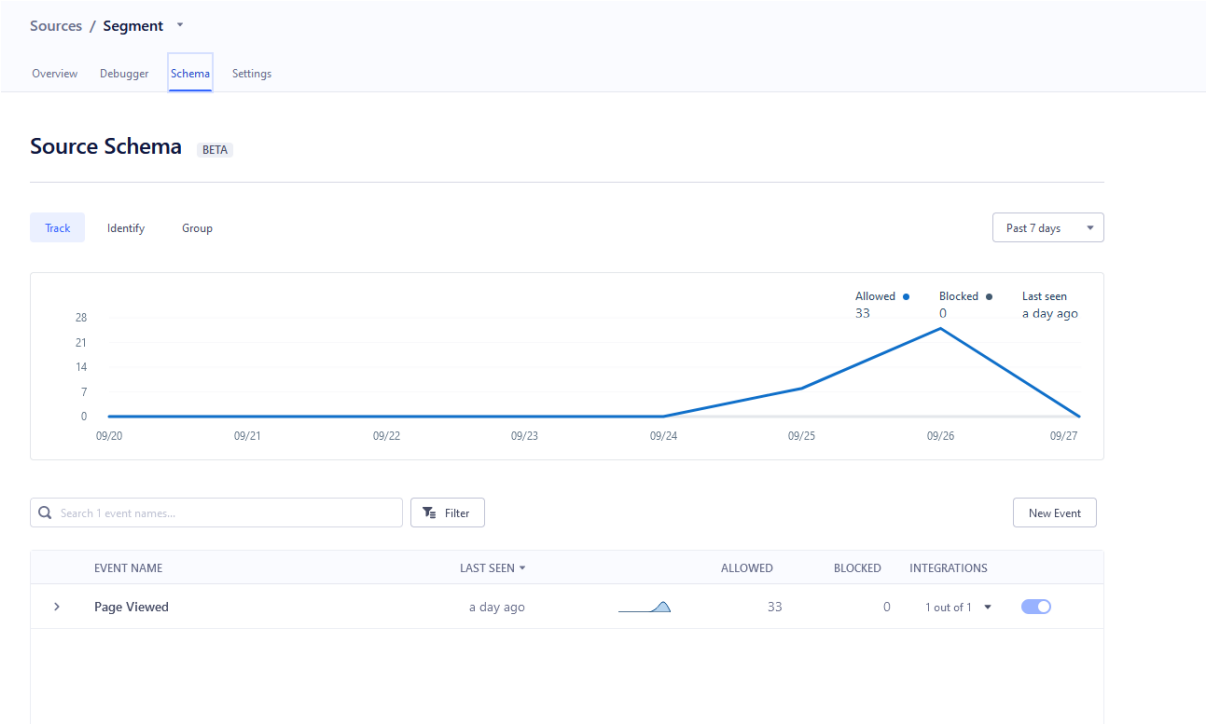




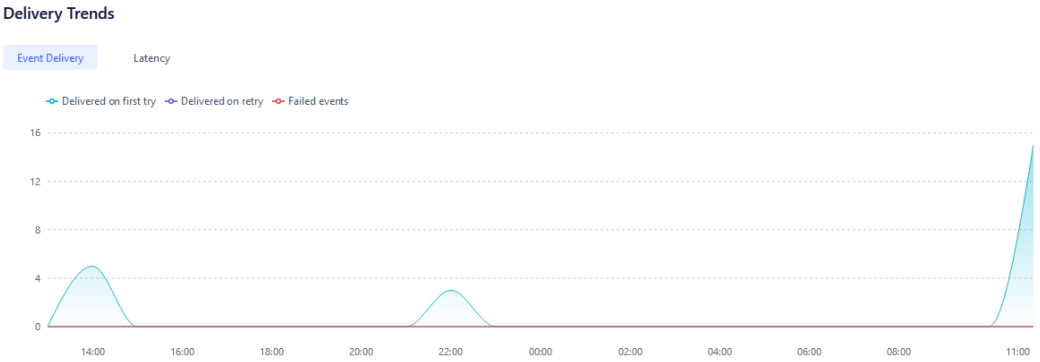
- We need to track the pageview itself (every time the website loads) as well as a conversion event.



- Source schema

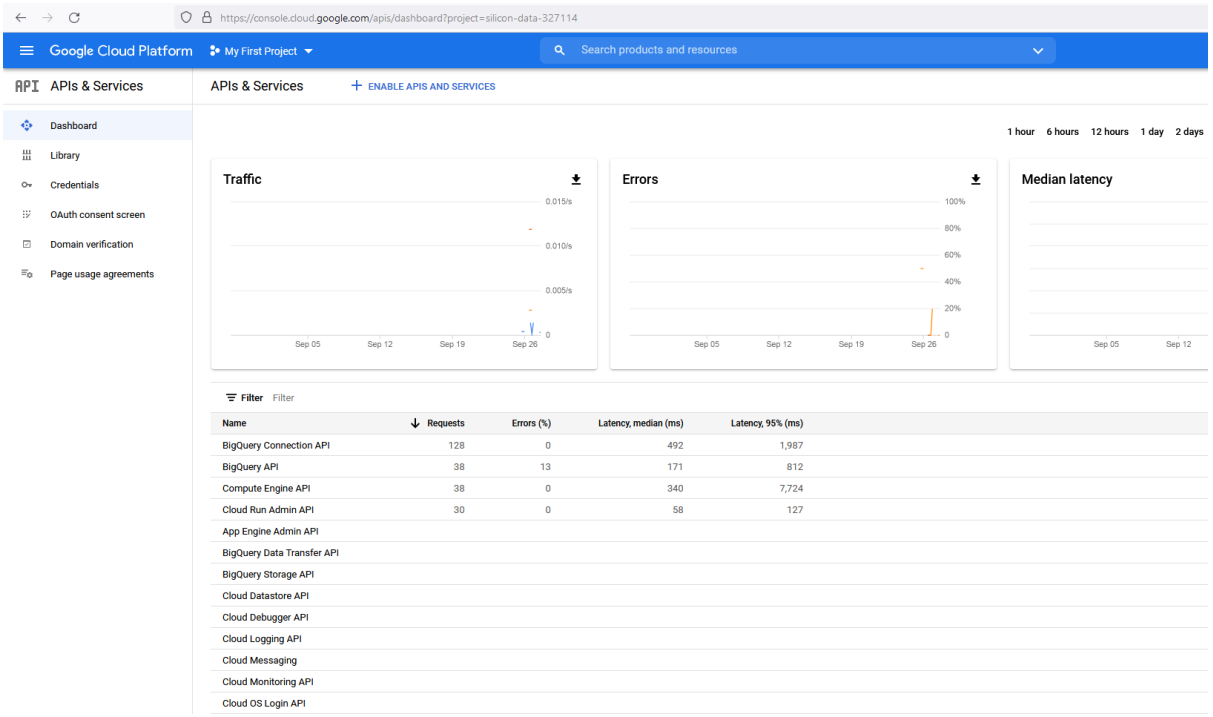


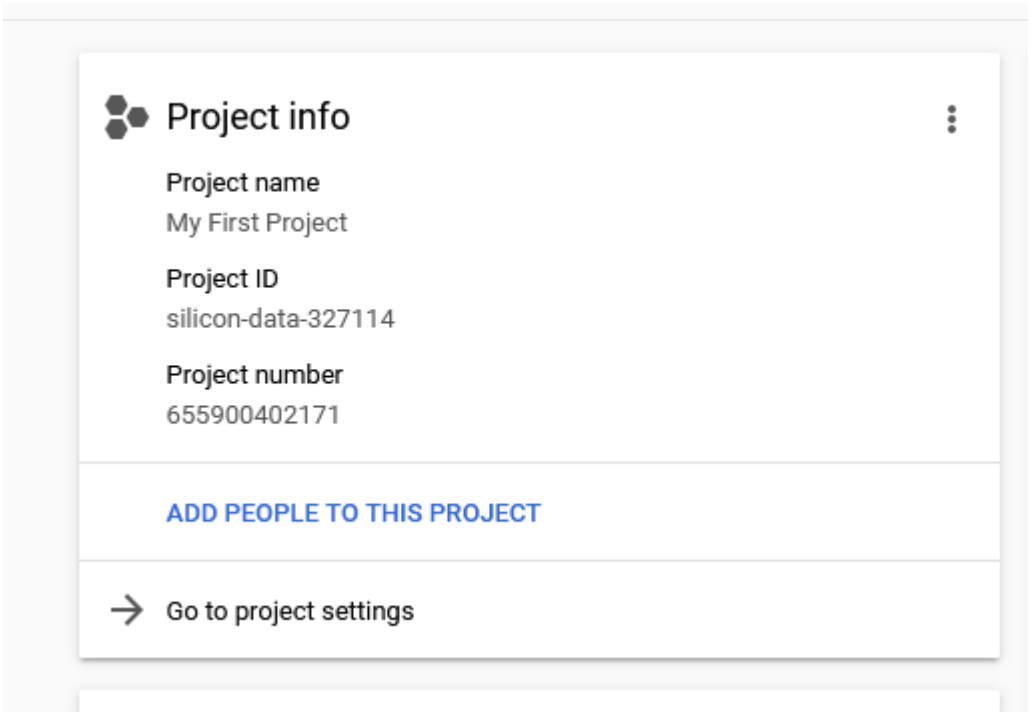
- Event Delivery



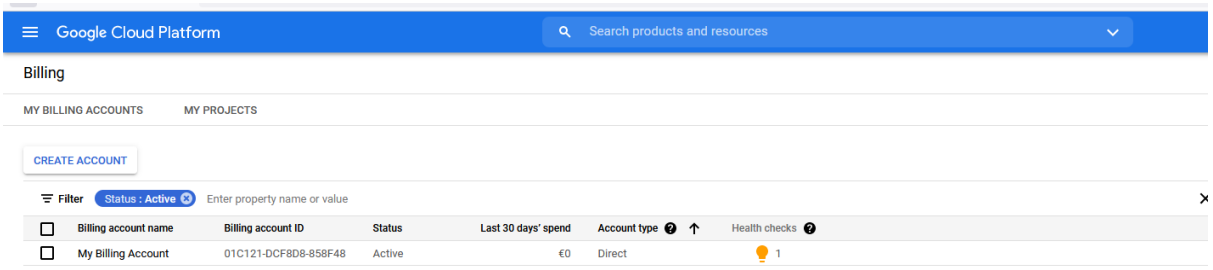
BigQuery Destination

- Create account on BigQuery
- Create a Project and Enable BigQuery

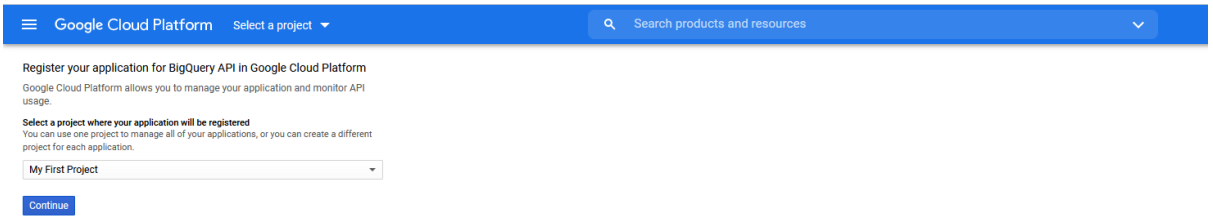




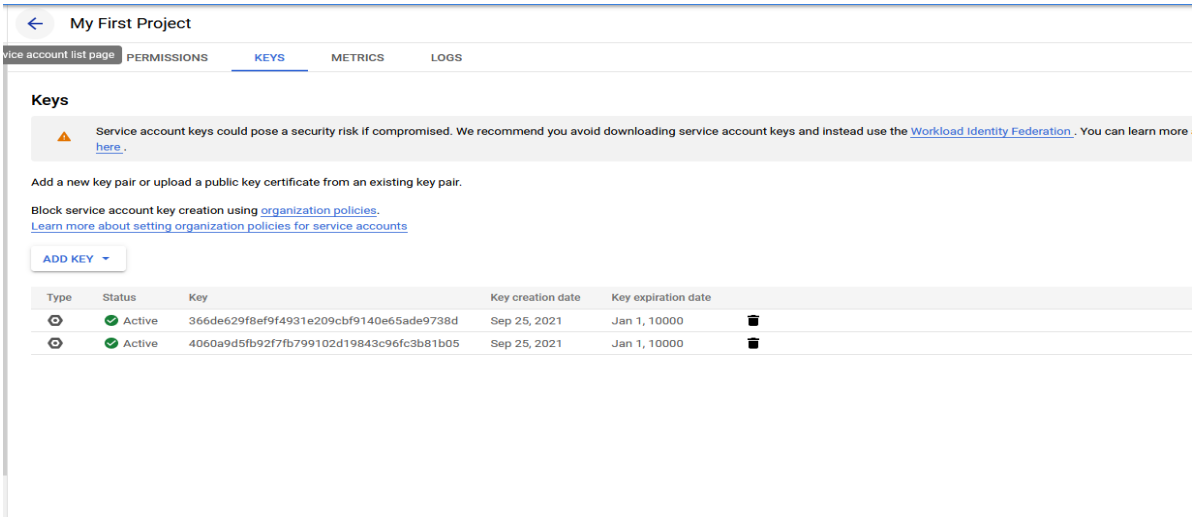
- To re-enable billing on a project, do the following.

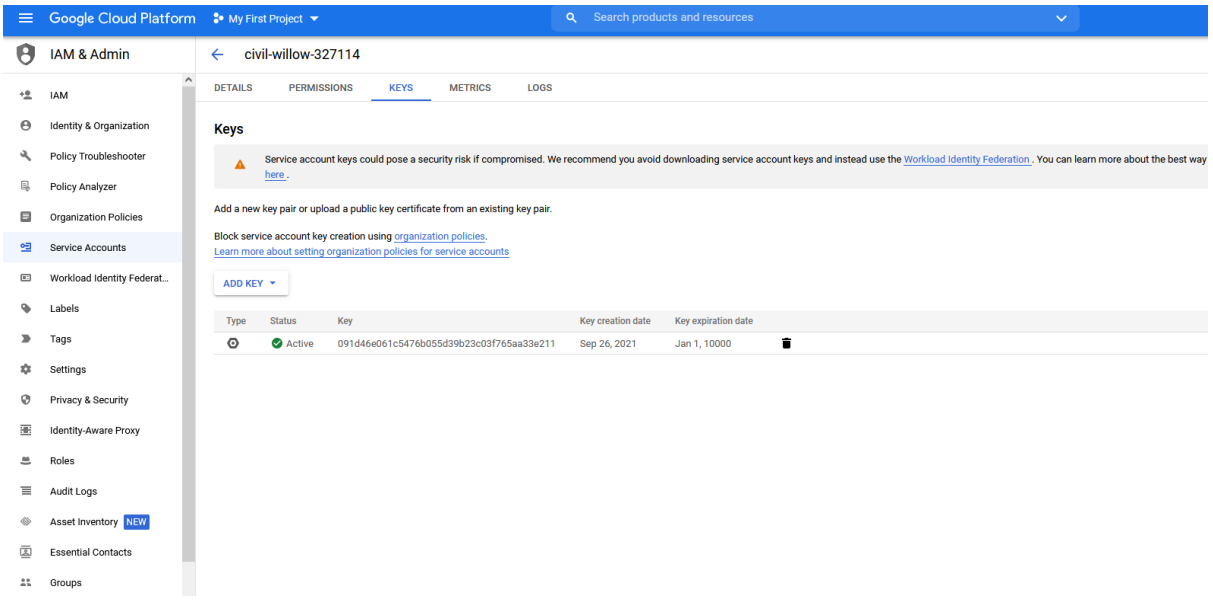


- Enable the BigQuery API

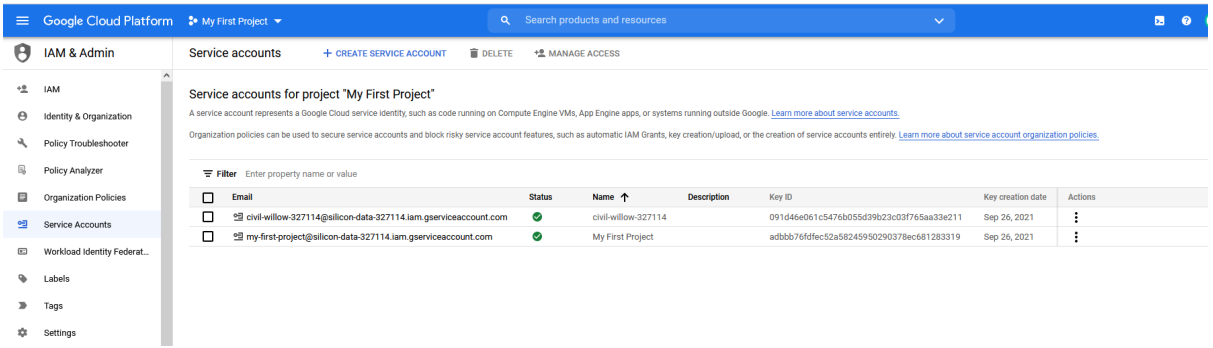


- Create a Service Account for Segment
- From the Navigation panel on the left, go to IAM & admin > Service accounts

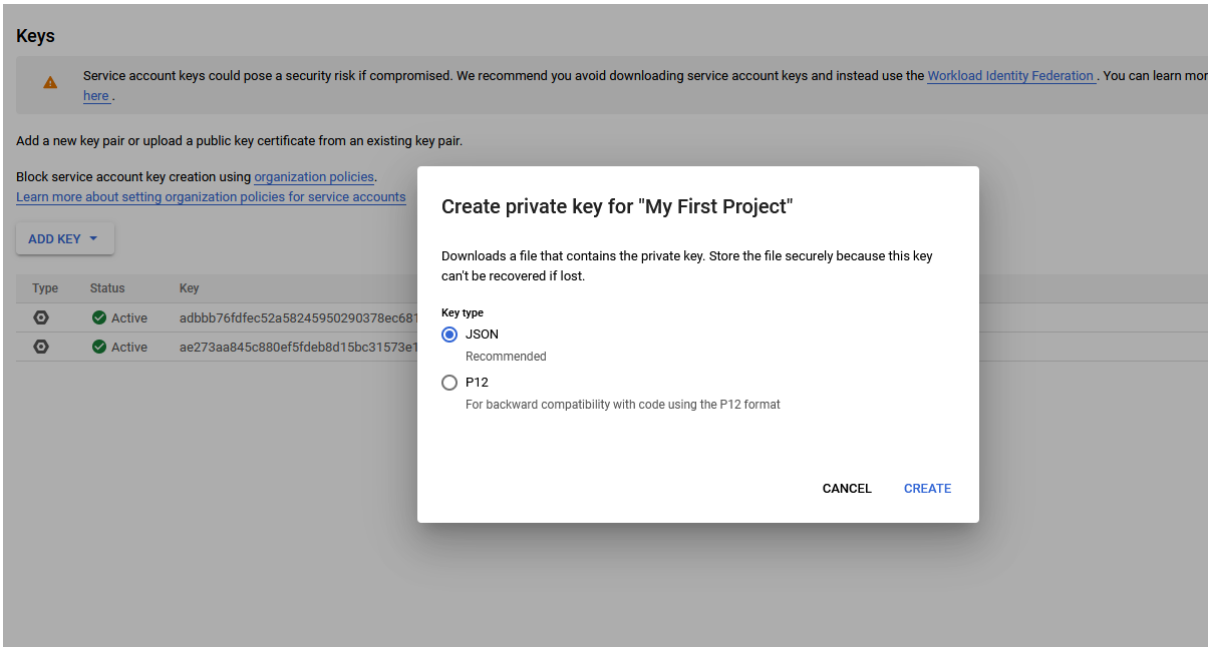




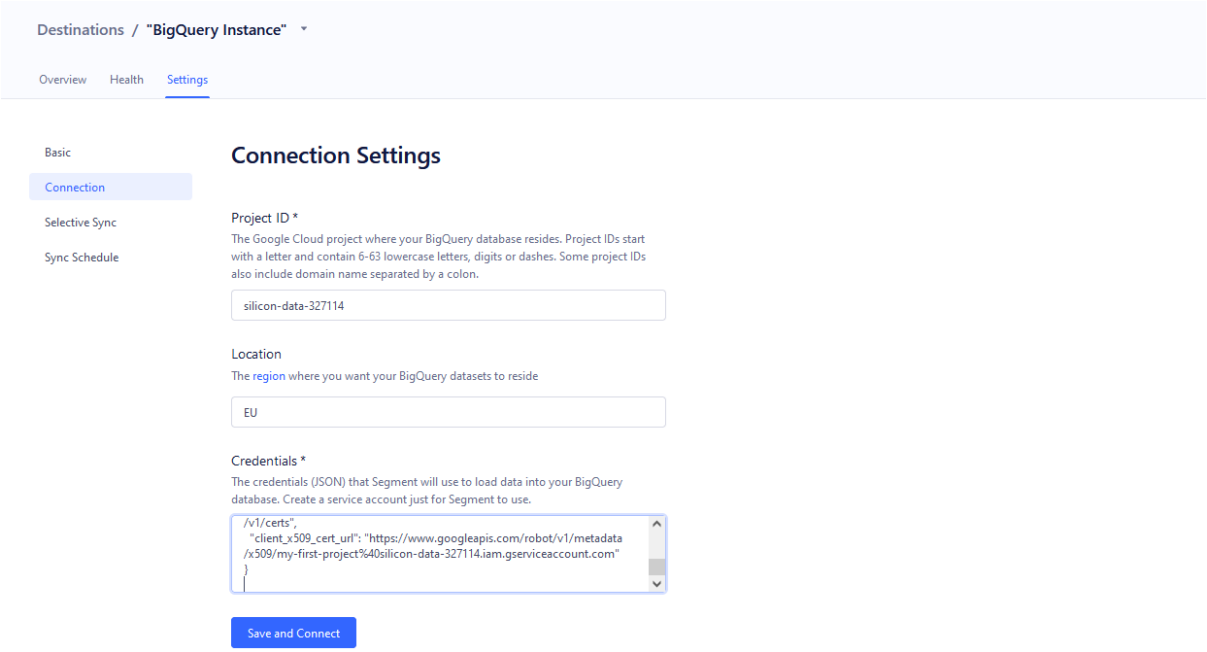
- Create Service Account

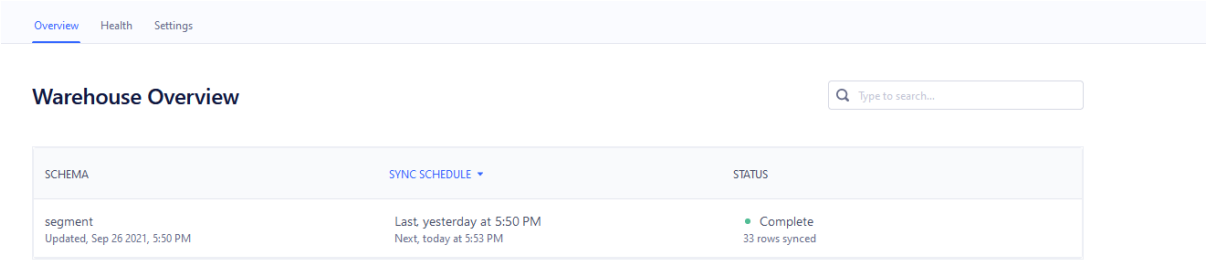


- Create a JSON key



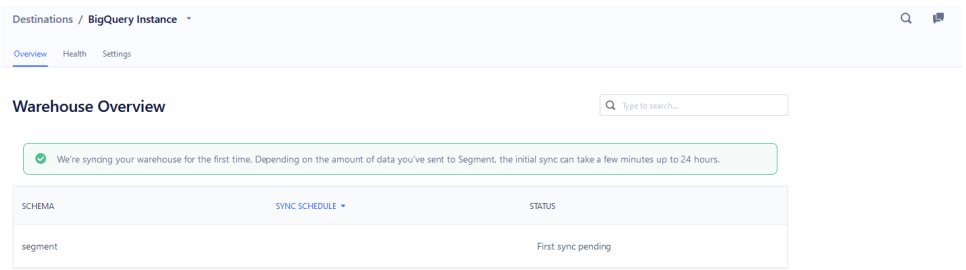
- Create the Warehouse in Segment

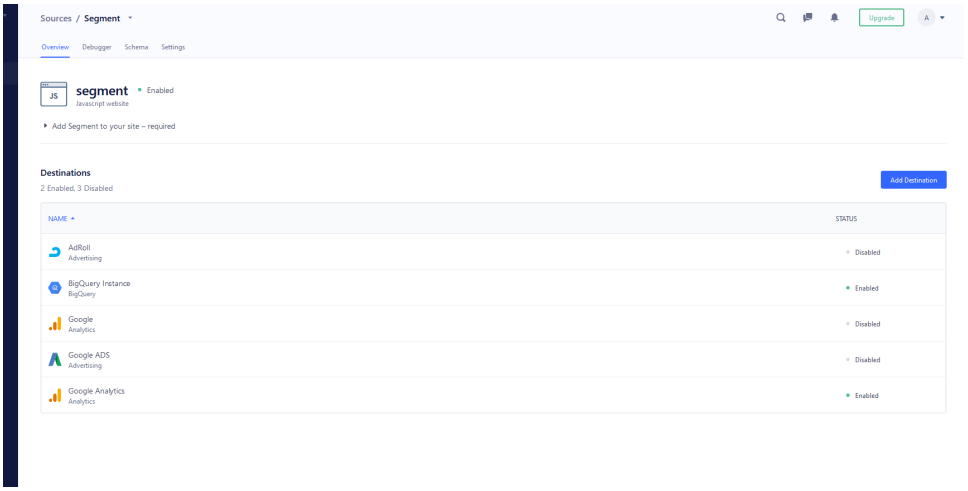


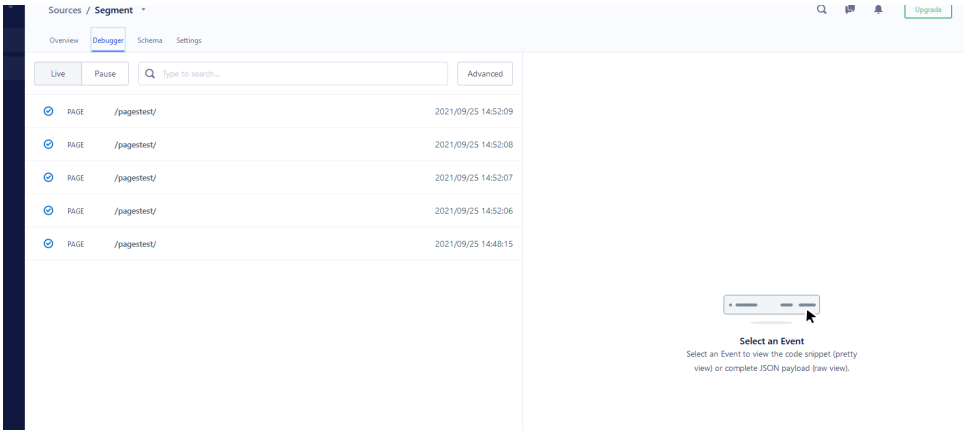


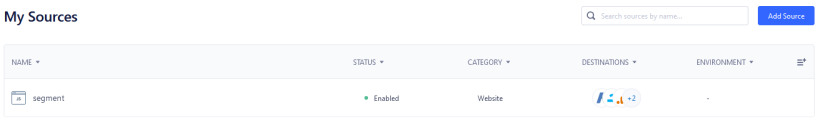
- Warehouse Overview

We’re syncing your warehouse for the first time. Depending on the amount of data you’ve sent to Segment, the initial sync can take a few minutes up to 24 hours.


Destinations / BigQuery Instance
Overview
Health
Settings
Warehouse Overview
Type to search...
We're syncing your warehouse for the first time. Depending on the amount of data you've sent to Segment, the initial sync can take a few minutes up to 24 hours.
SCHEMA
SYNC SCHEDULE
STATUS
segment
First sync pending

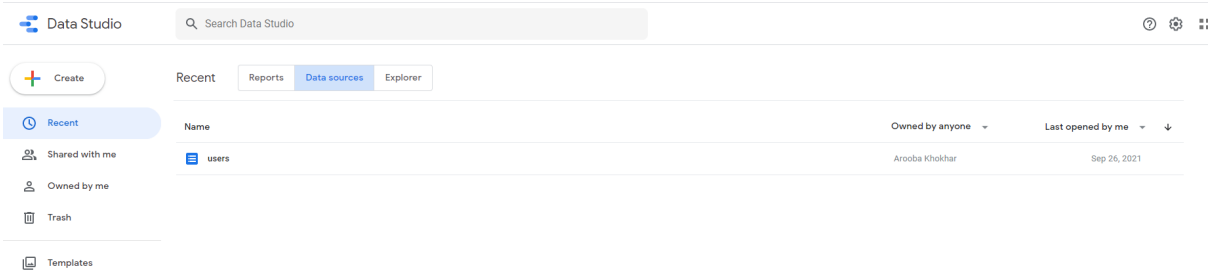

Sources / Segment
Upgrade
Overview
Debugger
Schema
Settings
segment
Enabled
Javascript website
Add Segment to your site – required
Destinations
2 Enabled, 3 Disabled
Add Destination
NAME
STATUS
AdRoll
Advertising
Disabled
BigQuery Instance
BigQuery
Enabled
Google
Analytics
Disabled
Google ADS
Advertising
Disabled
Google Analytics
Analytics
Enabled


Sources / Segment
Upgrade
Overview
Debugger
Schema
Settings
Live
Pause
Type to search...
Advanced
PAGE /pagestest/ 2021/09/25 14:52:09
PAGE /pagestest/ 2021/09/25 14:52:08
PAGE /pagestest/ 2021/09/25 14:52:07
PAGE /pagestest/ 2021/09/25 14:52:06
PAGE /pagestest/ 2021/09/25 14:48:15
Select an Event
Select an Event to view the code snippet (pretty view) or complete JSON payload (raw view).

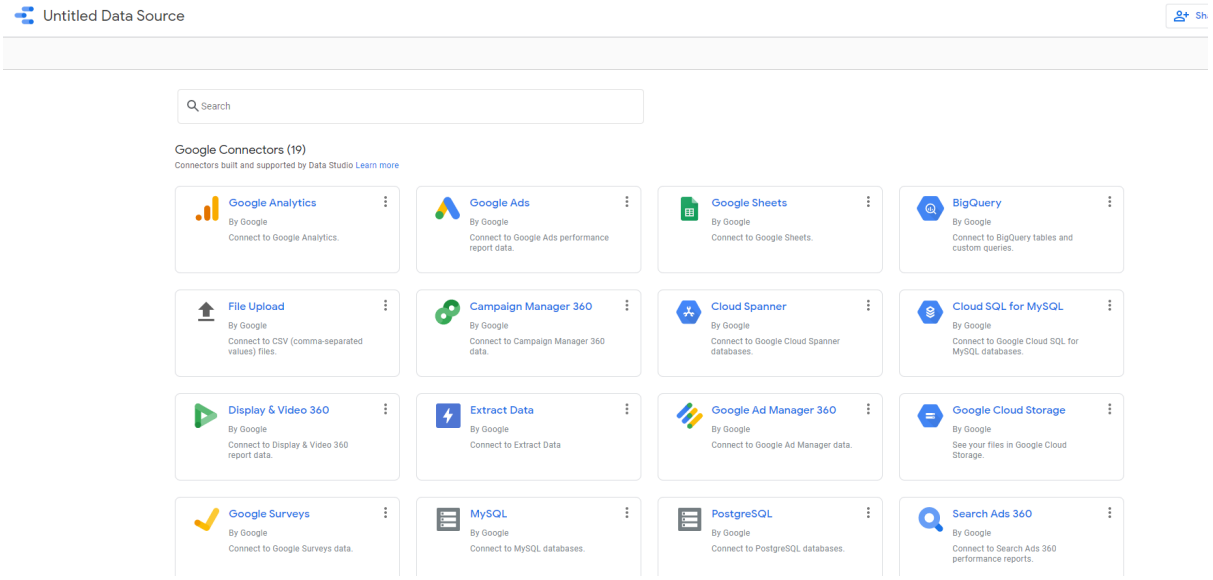

My Sources
Search sources by name...
Add Source
NAME
STATUS
CATEGORY
DESTINATIONS
ENVIRONMENT
segment
Enabled
Website
+2

3. Visualize

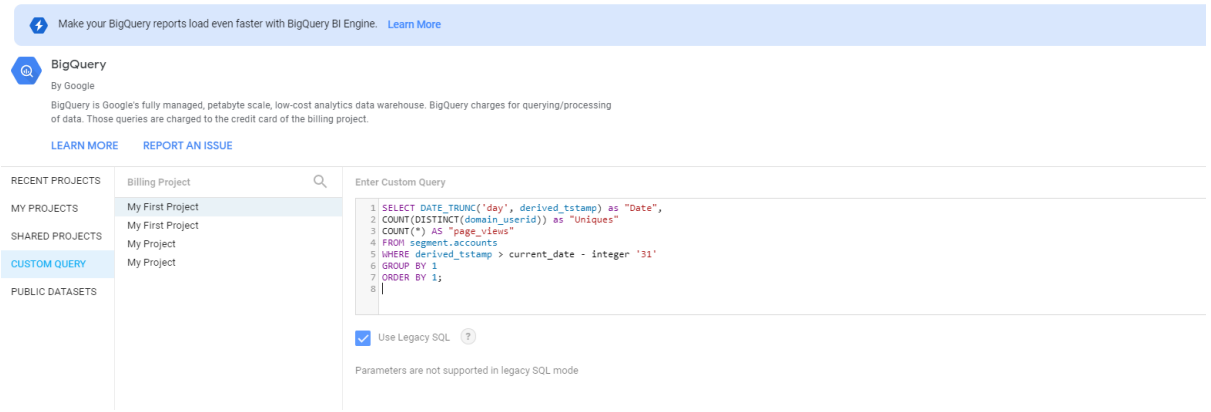
- Create an account with Google Data Studio



- Select “Bigquery”



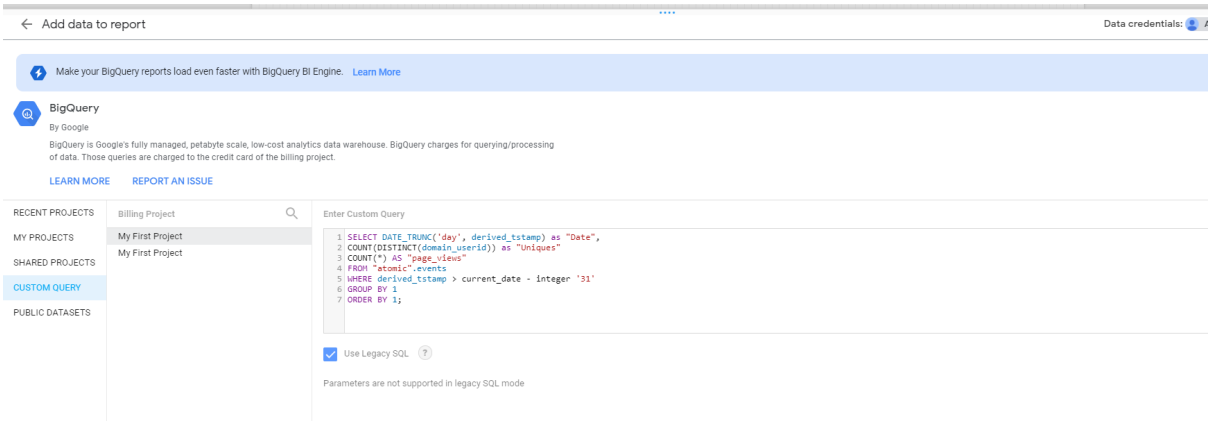
- the custom query, I chose my GCP project and then type my query



Create a dashboard with

a) a graph showing the number of unique users per day and pageviews per day

b) show the user based daily conversion rate over time. i.e. the percentage of users that clicked to button for which you created the event on each day



Part B

SQL :

- Created Customer table

```
CREATE DATABASE Customers;
CREATE TABLE Customers
( Emp_ID INT,
 Name   VARCHAR(20),
 City   VARCHAR(20) ,
 Age      INT,
 Profession VARCHAR(20),
);
```

- Add fields (NAME,CITY,PROFESSION,AGE)

```
INSERT INTO Customers
(Name, City, Age,Profession )
VALUES
```

('Russell', 'berlin', '12', 'doctor'),

('Richard', 'munich', '13', 'lawyer'),

('Rusty', 'hamburg', '15', 'actor'),

('Lexi', 'halle', '19', 'director');

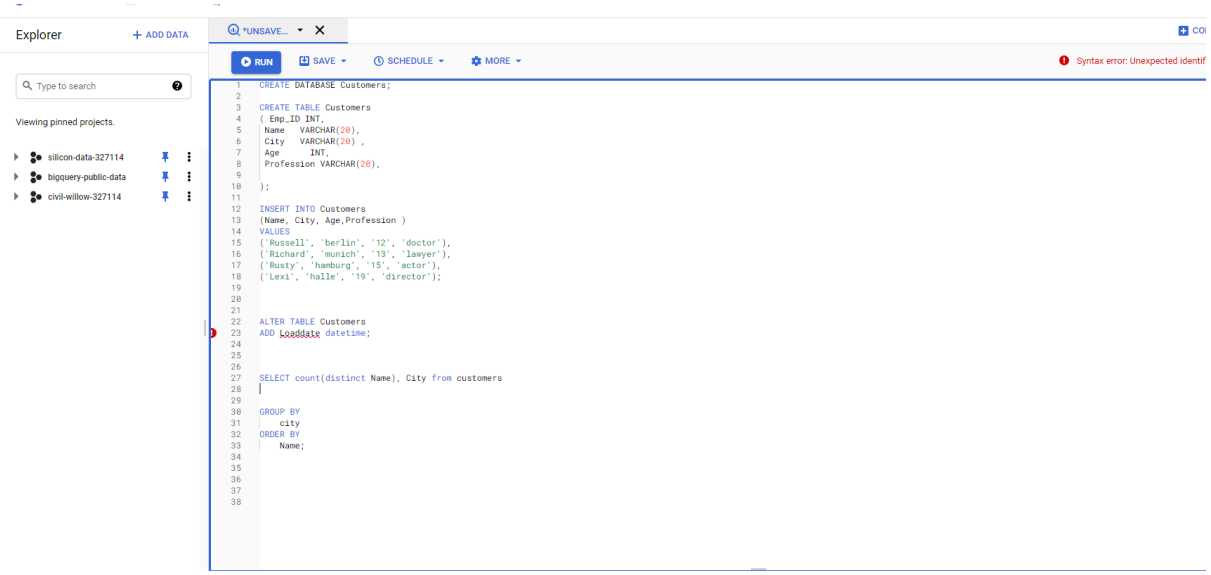
- Add new column “LOAD DATE” with the current timestamp

```
ALTER TABLE Customers
ADD LoadDate datetime;
```

- Return list of cities and the number of customers

```
SELECT count(distinct Name), City from customers

GROUP BY
   city
ORDER BY
   Name;
```



ETL/python

I unpivoted form where rows are facts with the dimensions, country and date.

```
pd.drop(['Province/State', 'Lat', 'Long'],axis=1,inplace=True)
pd=pd.rename(columns={"Country/Region": "Country"})
# do the unpivoting here ...


pd = df.melt(id_vars=['Country'], var_name='Date')
pd


# ignore BigQuery Loading for this test
# client = bigquery.Client()
# table_id='f-26914.test_covid.confirmed'
# job_config = bigquery.LoadJobConfig(
# write_disposition="WRITE_TRUNCATE"
# )
# job = client.load_table_from_dataframe(
# pd,
# table_id,
# job_config=job_config,
# location="europe-west3",
# )
# job.result()
main()
```

Console 1/A

```
In [19]: pd = df.melt(id_vars=['Country'], var_name='Date')
Traceback (most recent call last):

  File "<ipython-input-19-553b53d75c0a>", line 1, in <module>
    pd = df.melt(id_vars=['Country'], var_name='Date')

NameError: name 'df' is not defined


In [20]: pd = pd.melt(id_vars=['Country'], var_name='Date')

In [21]: pd
Out[21]:
                  Country     Date   value
0             Afghanistan  1/22/20       0
1                 Albania  1/22/20       0
2                 Algeria  1/22/20       0
3                 Andorra  1/22/20       0
4                  Angola  1/22/20       0
...                   ...      ...     ...
171022            Vietnam  9/25/21  746678
171023  West Bank and Gaza  9/25/21  395677
171024               Yemen  9/25/21    8891
171025              Zambia  9/25/21  208829
171026            Zimbabwe  9/25/21  129505

[171027 rows x 3 columns]

In [22]:
```