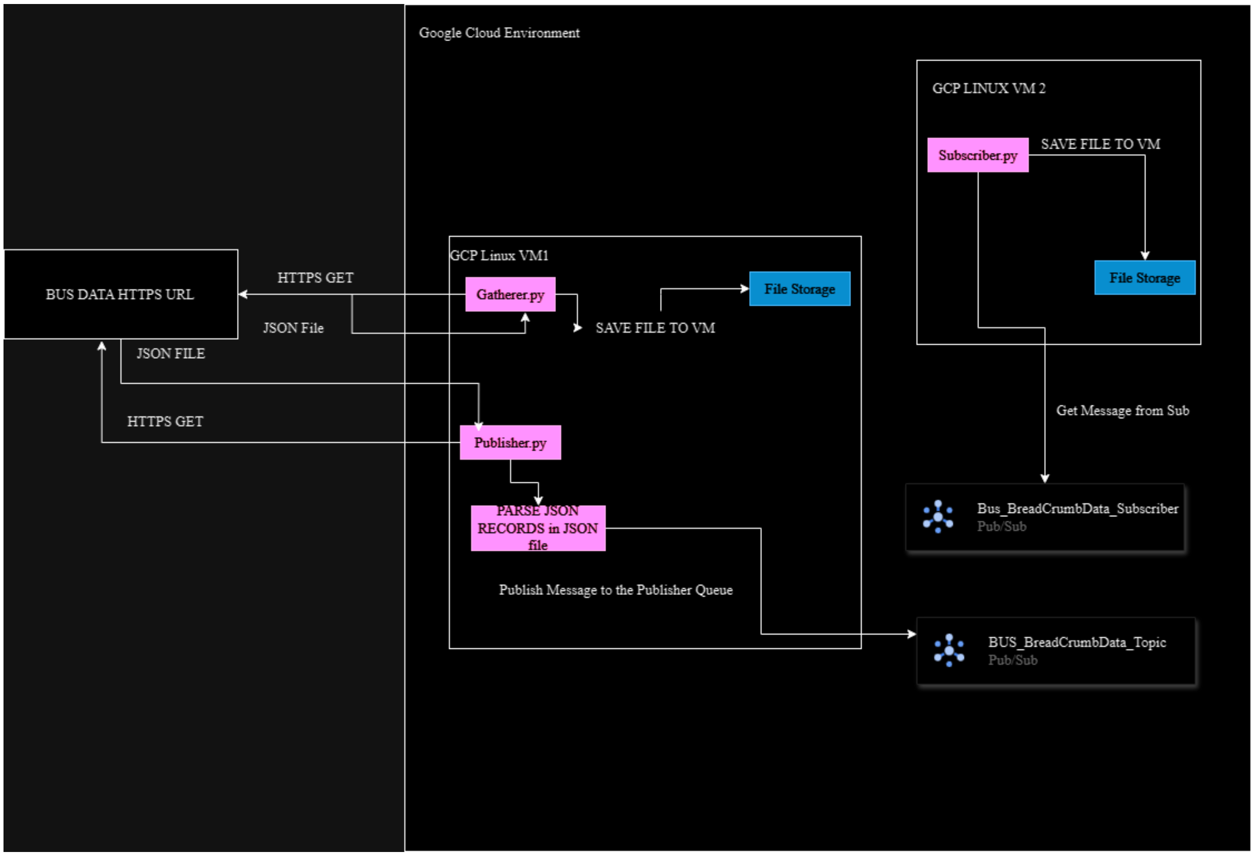
DataEng S24: Project Assignment 1

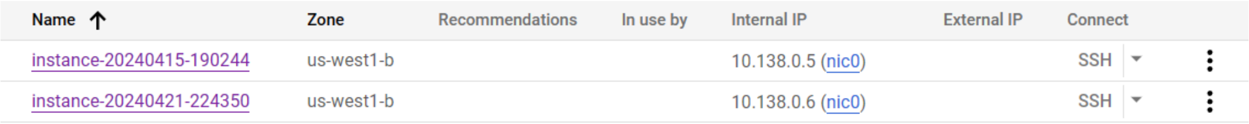
Gather and Transport

Current Planned Architecture Diagram:



A. Create, Configure and Run Your Virtual Machine

VMs created for our project in GCP is as follows :

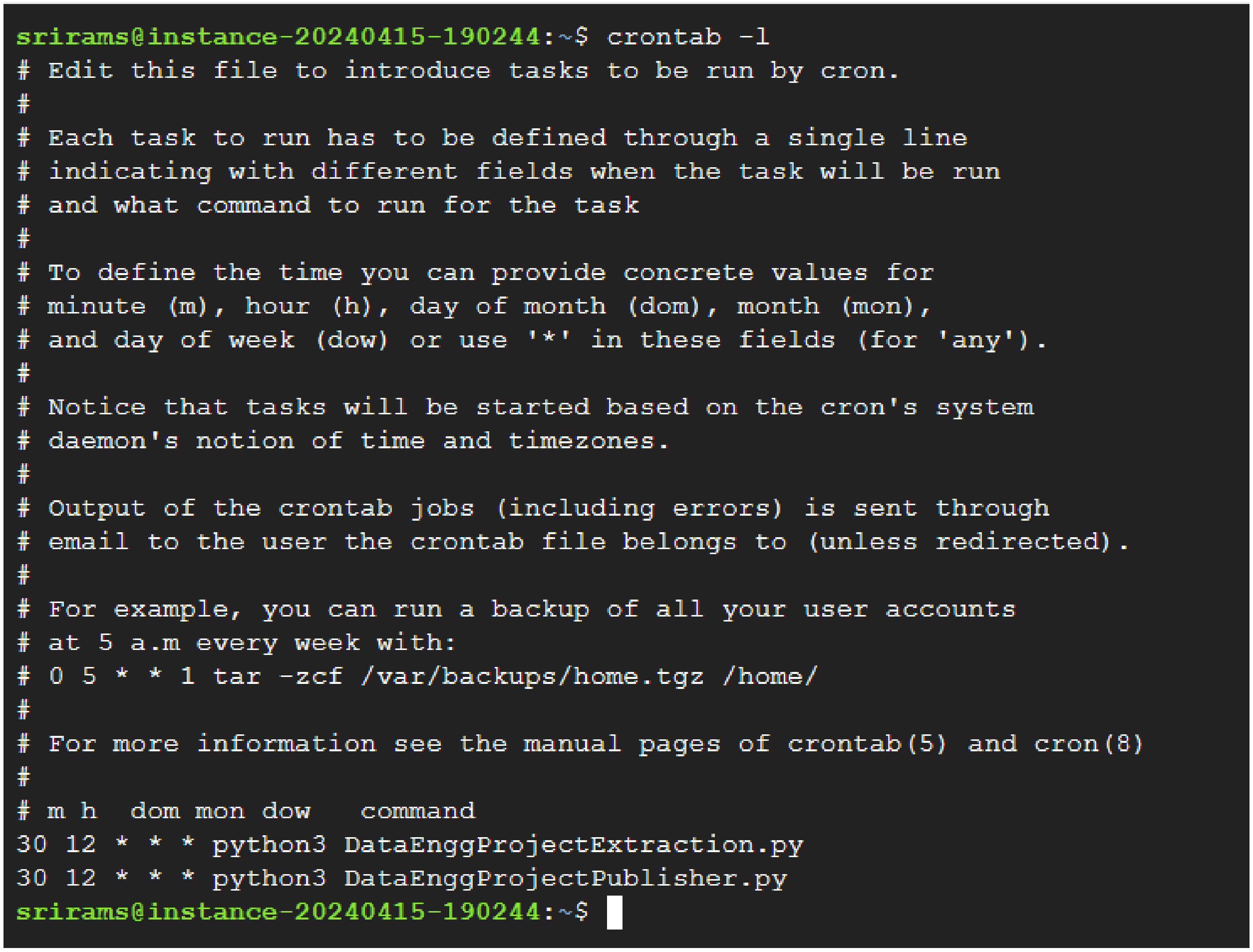


Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect	
instance-20240415-190244	us-west1-b			10.138.0.5 (nic0)		SSH ▾	⋮
instance-20240421-224350	us-west1-b			10.138.0.6 (nic0)		SSH ▾	⋮

B. Initial Python Data Gatherer

Refer to the file **DataEnggProjectExtractionInitial.py** in our Project Repo

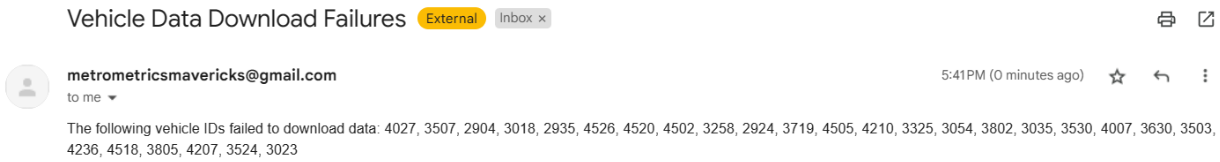
C. Run the Data Gatherer Daily



```
srirams@instance-20240415-190244:~$ crontab -l
# Edit this file to introduce tasks to be run by cron.
#
# Each task to run has to be defined through a single line
# indicating with different fields when the task will be run
# and what command to run for the task
#
# To define the time you can provide concrete values for
# minute (m), hour (h), day of month (dom), month (mon),
# and day of week (dow) or use '*' in these fields (for 'any').
#
# Notice that tasks will be started based on the cron's system
# daemon's notion of time and timezones.
#
# Output of the crontab jobs (including errors) is sent through
# email to the user the crontab file belongs to (unless redirected).
#
# For example, you can run a backup of all your user accounts
# at 5 a.m every week with:
# 0 5 * * 1 tar -zcf /var/backups/home.tgz /home/
#
# For more information see the manual pages of crontab(5) and cron(8)
#
# m h  dom mon dow   command
30 12 * * * python3 DataEnggProjectExtraction.py
30 12 * * * python3 DataEnggProjectPublisher.py
srirams@instance-20240415-190244:~$ 
```

Extra Credit:

Updated my Code to make it check for Storage space and for IDs where download fails due to any issue, one Notification email will be sent with the list of vehicle IDs.



Vehicle Data Download Failures

metrometricsmavericks@gmail.com

to me

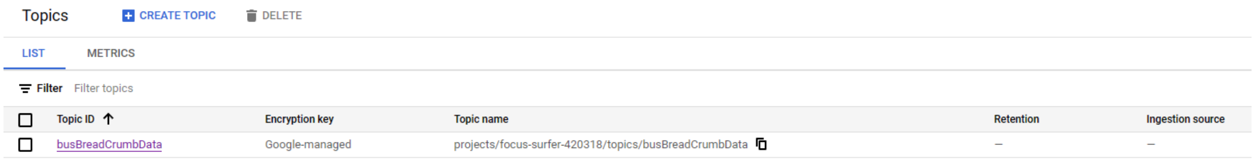
5:41 PM (0 minutes ago)

The following vehicle IDs failed to download data: 4027, 3507, 2904, 3018, 2935, 4526, 4520, 4502, 3258, 2924, 3719, 4505, 4210, 3325, 3054, 3802, 3035, 3530, 4007, 3630, 3503, 4236, 4518, 3805, 4207, 3524, 3023

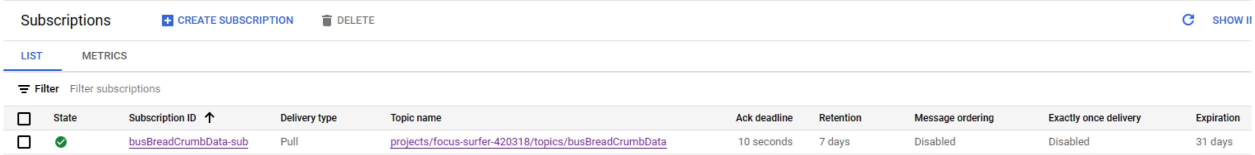
Refer to file **DataEnggProjectExtraction.py** in our project Repo.

D. Configure Google Cloud Pub/Sub

We have Configured Pub/Sub for our project as follows:



□	Topic ID ↑	Encryption key	Topic name	Retention	Ingestion source
□	busBreadCrumbData	Google-managed	projects/focus-surfer-420318/topics/busBreadCrumbData	–	–



□	State	Subscription ID ↑	Delivery type	Topic name	Ack deadline	Retention	Message ordering	Exactly once delivery	Expiration
□	✓	busBreadCrumbData-sub	Pull	projects/focus-surfer-420318/topics/busBreadCrumbData	10 seconds	7 days	Disabled	Disabled	31 days

E. Parse JSON into Individual Breadcrumb Records

F. Send Breadcrumb Records to Pub/Sub Topic

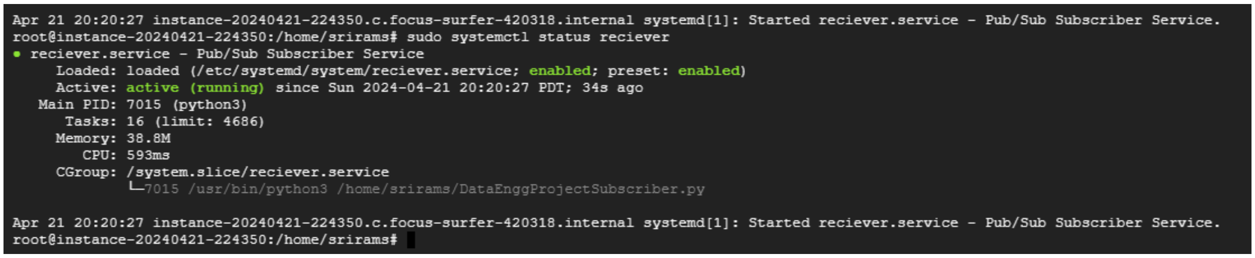
Refer to the file **DataEnggProjectPublisher.py** in our Project Repo

G. Pub/Sub Receiver

Refer to the file **DataEnggProjectSubscriber.py** in our Project Repo

H. Configure Linux to Run Pub/Sub Receiver Continually

Service is enabled for Receiver and we have configured the service to restart on failures and this is the status screenshot.

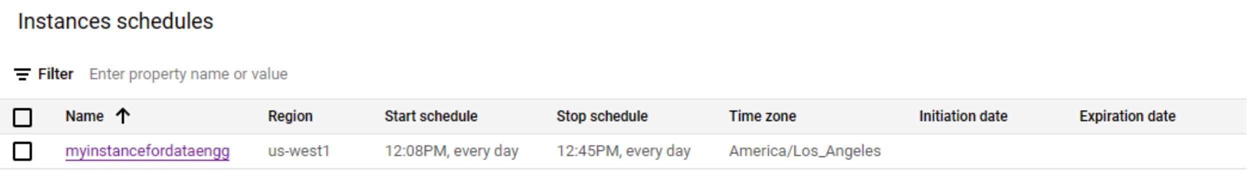


```
Apr 21 20:20:27 instance-20240421-224350.c.focus-surfer-420318.internal systemd[1]: Started reciever.service - Pub/Sub Subscriber Service.
root@instance-20240421-224350:/home/srirams# sudo systemctl status reciever
● reciever.service - Pub/Sub Subscriber Service
     Loaded: loaded (/etc/systemd/system/reciever.service; enabled; preset: enabled)
     Active: active (running) since Sun 2024-04-21 20:20:27 PDT; 34s ago
   Main PID: 7015 (python3)
      Tasks: 16 (limit: 4686)
     Memory: 38.8M
        CPU: 593ms
     CGroup: /system.slice/reciever.service
             └─7015 /usr/bin/python3 /home/srirams/DataEnggProjectSubscriber.py

Apr 21 20:20:27 instance-20240421-224350.c.focus-surfer-420318.internal systemd[1]: Started reciever.service - Pub/Sub Subscriber Service.
root@instance-20240421-224350:/home/srirams# 
```

I. Schedule your VM to start and stop automatically

We have setup the Instance for our project as follows



Instances schedules

Filter Enter property name or value

□	Name ↑	Region	Start schedule	Stop schedule	Time zone	Initiation date	Expiration date
□	myinstancefordataengg	us-west1	12:08PM, every day	12:45PM, every day	America/Los_Angeles		

DataEng Project Assignment 1 Submission Document

Construct a table showing each day for which your pipeline successfully, automatically processed one complete day's worth of sensor readings.

Date	Day of Week	Approximate Time of day for your data access	# Sensor Readings	Total Data Saved (KBs)	# Pub/Sub messages published and received

04/14/2024	Sunday	12:30 PM PDT	{
			"EVENT_NO_TRIP":220385292,
			"EVENT_NO_STOP":220385302,
			"OPD_DATE":"15DEC2022:00:00:00",
			"VEHICLE_ID":2902,
			"METERS":19021,
			"ACT_TIME":23580,
			"GPS_LONGITUDE":-122.69081,
			"GPS_LATITUDE":45.535613,
			"GPS_SATELLITES":8.0,
			"GPS_HDOP":1.1
			}
				78985328	311034

04/15/2024	Monday	12:30 PM PDT	{	78472540	325960

			"EVENT_NO_TRIP":221431889, "EVENT_NO_STOP":221431906, "OPD_DATE":"16DEC2022:00:00:00", "VEHICLE_ID":2902, "METERS":8487, "ACT_TIME":47927, "GPS_LONGITUDE":-122.802103, "GPS_LATITUDE":45.463523, "GPS_SATELLITES":12.0, "GPS_HDOP":0.8 }		

04/16/2024	Tuesday	12:30 PM PDT	{
			"EVENT_NO_TRIP":221510274,
			"EVENT_NO_STOP":221510352,
			"OPD_DATE":"17DEC2022:00:00:00",
			"VEHICLE_ID":3023,
			"METERS":27683,
			"ACT_TIME":66544,
			"GPS_LONGITUDE":-122.58922,
			"GPS_LATITUDE":45.523058,
			"GPS_SATELLITES":12.0,
			"GPS_HDOP":0.7
			}
				76143201	290505

04/17/2024	Wednesday	12:30 PM PDT	{ "EVENT_NO_TRIP":222064319, "EVENT_NO_STOP":222064321, "OPD_DATE":"18DEC2022:00:00:00", "VEHICLE_ID":3059, "METERS":104866, "ACT_TIME":47402, "GPS_LONGITUDE":-122.572777, "GPS_LATITUDE":45.43827, "GPS_SATELLITES":12.0, "GPS_HDOP":0.7	78085328	310035

			}		

04/18/2024	Thursday	12:30 PM PDT	{
			"EVENT_NO_TRIP":222781137,
			"EVENT_NO_STOP":222781139,
			"OPD_DATE":"19DEC2022:00:00:00",
			"VEHICLE_ID":2902,
			"METERS":887,
			"ACT_TIME":21816,
			"GPS_LONGITUDE":-122.84017,
			"GPS_LATITUDE":45.509218,
			"GPS_SATELLITES":12.0,
			"GPS_HDOP":0.7
			}
				81485328	310250

04/19/2024	Friday	12:30 PM PDT	{
			"EVENT_NO_TRIP":223461587,
			"EVENT_NO_STOP":223461601,
			"OPD_DATE":"20DEC2022:00:00:00",
			"VEHICLE_ID":2935,
			"METERS":12758,
			"ACT_TIME":21665,
			"GPS_LONGITUDE":-122.988017,
			"GPS_LATITUDE":45.521155,
			"GPS_SATELLITES":12.0,
			"GPS_HDOP":0.8
			}
				82688306	330365

04/20/2024	Saturday	12:30 PM PDT	{ "EVENT_NO_TRIP":224100636, "EVENT_NO_STOP":224100637, "OPD_DATE":"21DEC2022:00:00:00", "VEHICLE_ID":2902, "METERS":73, "ACT_TIME":25211, "GPS_LONGITUDE":-	82372950	329324

			122.844165, "GPS_LATITUDE":45.503545, "GPS_SATELLITES":12.0, "GPS_HDOP":0.9 }		

04/21/2024	Sunday	12:30 PM PDT	{
			"EVENT_NO_TRIP":224411073,
			"EVENT_NO_STOP":224411153,
			"OPD_DATE":"22DEC2022:00:00:00",
			"VEHICLE_ID":2902,
			"METERS":56812,
			"ACT_TIME":30645,
			"GPS_LONGITUDE":-122.78018,
			"GPS_LATITUDE":45.427715,
			"GPS_SATELLITES":10.0,
			"GPS_HDOP":1.0
			}
				83373964	339476

Additionally, include screenshots for the parts C, H and I

1. Output of crontab -l: Your scheduled cron jobs.
2. systemctl status: This will show the status of your receiver program.
3. VM instance schedule: This will display the schedule settings for your GCP VM instance.