DataEng Project - Assignment 1 Submission Document

Chethana Muppalam
Srivashinie Dhamodharasamy
Vaishnavi Srinath

Github Link

https://github.com/chethana613/TriMet.git

Screenshots

Crontab:
The data gatherer that we have used is scheduled to run at 12:00 pm everyday.

```
sridhamo@ragnarok-vm:~$ crontab -l
00 12 * * *  /usr/bin/python3 /home/sridhamo/bus.py
#
# 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
```

systemctl/systemd:

We have used two separate VM for the publisher and the subscriber. Therefore we have set up two systemd services,one for the publisher and one for the subscriber.

Publisher:

```
  GNU nano 7.2                      /etc/systemd/system/publisher.service
[Unit]
Description=Publisher program for bus data
After=multi-user.target

[Service]
Type=simple
Restart=no
ExecStart=/home/sridhamo/env/bin/python /home/sridhamo/publisher.py
WorkingDirectory=/home/sridhamo
User=sridhamo
Group=sridhamo

[Install]
WantedBy=multi-user.target
```

```
sridhamo@ragnarok-vm:~$ sudo systemctl status publisher.service
o publisher.service - Publisher program for bus data
     Loaded: loaded (/etc/systemd/system/publisher.service; enabled; preset: enabled)
     Active: inactive (dead) since Sun 2024-04-21 13:43:49 PDT; 4h 52min ago
   Duration: 40.173s
    Process: 56616 ExecStart=/home/sridhamo/env/bin/python /home/sridhamo/publisher.py (code=exited, status=0/SUCCESS)
   Main PID: 56616 (code=exited, status=0/SUCCESS)
        CPU: 34.502s

Apr 21 13:43:09 ragnarok-vm.c.data-engineering-420102.internal systemd[1]: Started publisher.service - Publisher program for bus data.
```

Subscriber:

```
  GNU nano 7.2                      /etc/systemd/system/subscriber.service
[Unit]
Description=Receiver program for bus data
After=multi-user.target

[Service]
Type=simple
Restart=no
ExecStart=/home/sridhamo/env/bin/python /home/sridhamo/reciever.py
WorkingDirectory=/home/sridhamo
User=sridhamo
Group=sridhamo

[Install]
WantedBy=multi-user.target
```
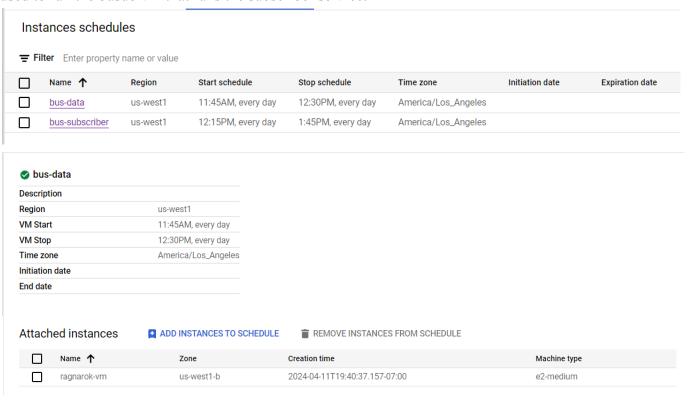
```
sridhamo@busub:~$ sudo systemctl status subscriber.service
● subscriber.service - Receiver program for bus data
     Loaded: loaded (/etc/systemd/system/subscriber.service; enabled; preset: enabled)
     Active: active (running) since Sun 2024-04-21 18:59:03 PDT; 2min 1s ago
   Main PID: 135678 (python)
      Tasks: 26 (limit: 4686)
     Memory: 103.7M
        CPU: 34.887s
     CGroup: /system.slice/subscriber.service
             └─135678 /home/sridhamo/env/bin/python /home/sridhamo/reciever.py

Apr 21 18:59:03 busub.c.data-engineering-420102.internal systemd[1]: Started subscriber.service - Receiver program for bus data.
```

GCP Instance Schedule:

We have two instance schedules. The bus-data schedule is used to schedule the ragnarok-vm which is used to run the gatherer file and the publisher service. The bus-subscriber schedule is used to run the busub vm that runs the subscriber service.

Instances schedules

≡ Filter   Enter property name or value

	Name ↑	Region	Start schedule	Stop schedule	Time zone	Initiation date	Expiration date
☐	bus-data	us-west1	11:45AM, every day	12:30PM, every day	America/Los_Angeles		
☐	bus-subscriber	us-west1	12:15PM, every day	1:45PM, every day	America/Los_Angeles		

✅ bus-data

Description	
Region	us-west1
VM Start	11:45AM, every day
VM Stop	12:30PM, every day
Time zone	America/Los_Angeles
Initiation date	
End date	

Attached instances   ➕ ADD INSTANCES TO SCHEDULE   🗑 REMOVE INSTANCES FROM SCHEDULE

	Name ↑	Zone	Creation time	Machine type
☐	ragnarok-vm	us-west1-b	2024-04-11T19:40:37.157-07:00	e2-medium

bus-subscriber

Description	
Region	us-west1
VM Start	12:15PM, every day
VM Stop	1:45PM, every day
Time zone	America/Los_Angeles
Initiation date	
End date	

Attached instances   ⊕ ADD INSTANCES TO SCHEDULE   🗑 REMOVE INSTANCES FROM SCHEDULE

	Name ↑	Zone	Creation time	Machine type
☐	busub	us-west1-b	2024-04-17T21:20:03.611-07:00	e2-medium

Tracking Table:

Date	Day of Week	Approximate Time of day for your data access	# Sensor Readings	Total Data Saved (KBs)	# Pub/Sub messages published and received
04/16/2024	Tuesday	12 pm	227963	88154	227963
04/17/2024	Wednesday	12 pm	237596	91871	237596
04/18/2024	Thursday	12 pm	365451	141353	365451
04/19/2024	Friday	12 pm	339754	133123	339754
04/20/2024	Saturday	12 pm	369321	138377	369321
04/21/2024	Sunday	12 pm	346518	134037	346518