CS 5408 – Data Management, Warehousing and Analytics

Assignment – 2

Problem 2

Task 1

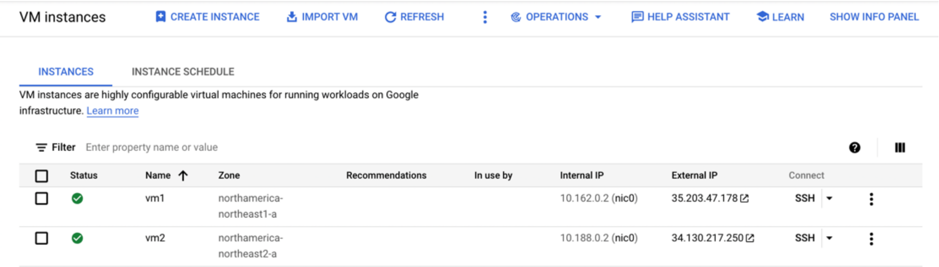
Prashit Patel - B00896717

- **Data Cleaning and Decomposition**

1. **olist_order_review_dataset:**
   - For title attribute - blanks, numbers and special characters are replaced with NA as it would represent that data is not available for this field.
   - For comments attribute – blanks are replaced with NA as it would represent that data is not available for this field.
2. **olist_orders_dataset:**
   - For order_approved_at, order_delivered_carrier_date, order_delivered_customer_date – blanks are replaced with 0000-00-00 00:00 as it would represent that no date is available for this field.
3. **olist_products_dataset:**
   - For product_category_name – blanks are replaced with NA as it would represent that no data available for this field.
   - For product_name_length, production_description_length, product_photos_qty, product_weight_g, product_length_cm, and product_width_cm attributes – blanks are replaced with 0 as it would represent no data values are available for this fields.
4. **olist_sellers_dataset:**
   - seller_city_name and seller_state attributes are deleted as they are redundant, and the same information can be obtained from olist_geolocation_dataset using seller_zip_code_prefix attribute.
5. **olist_customer_dataset:**
   - customer_city_name and customer_state attributes can be deleted as they are redundant, and the same information can be obtained from olist_geolocation_dataset using customer_zip_code_prefix attribute but are kept in same table as it would be required for task2.

- **Data Fragmentation and Transparency**
  - olist_customer, olist_geolocation and olist_seller datasets will be converted to tables and stored on VM1 site as those are independent tables other than orders and products.
  - olist_orders, olist_order_items, olist_order_reviews, olist_order_payments, olist_products and product_category_name_translation datasets will be converted to tables and stored on VM2 site as those tables will contain all orders and products related information and will be stored together for better performance.
  - Transaction transparency will be considered to maintain the consistency and integrity of the database as order, items and payment information is critical for all user orders.
  - Distribution transparency will be considered to hide the details of the distributed database and users will think that they are working with a single database system.

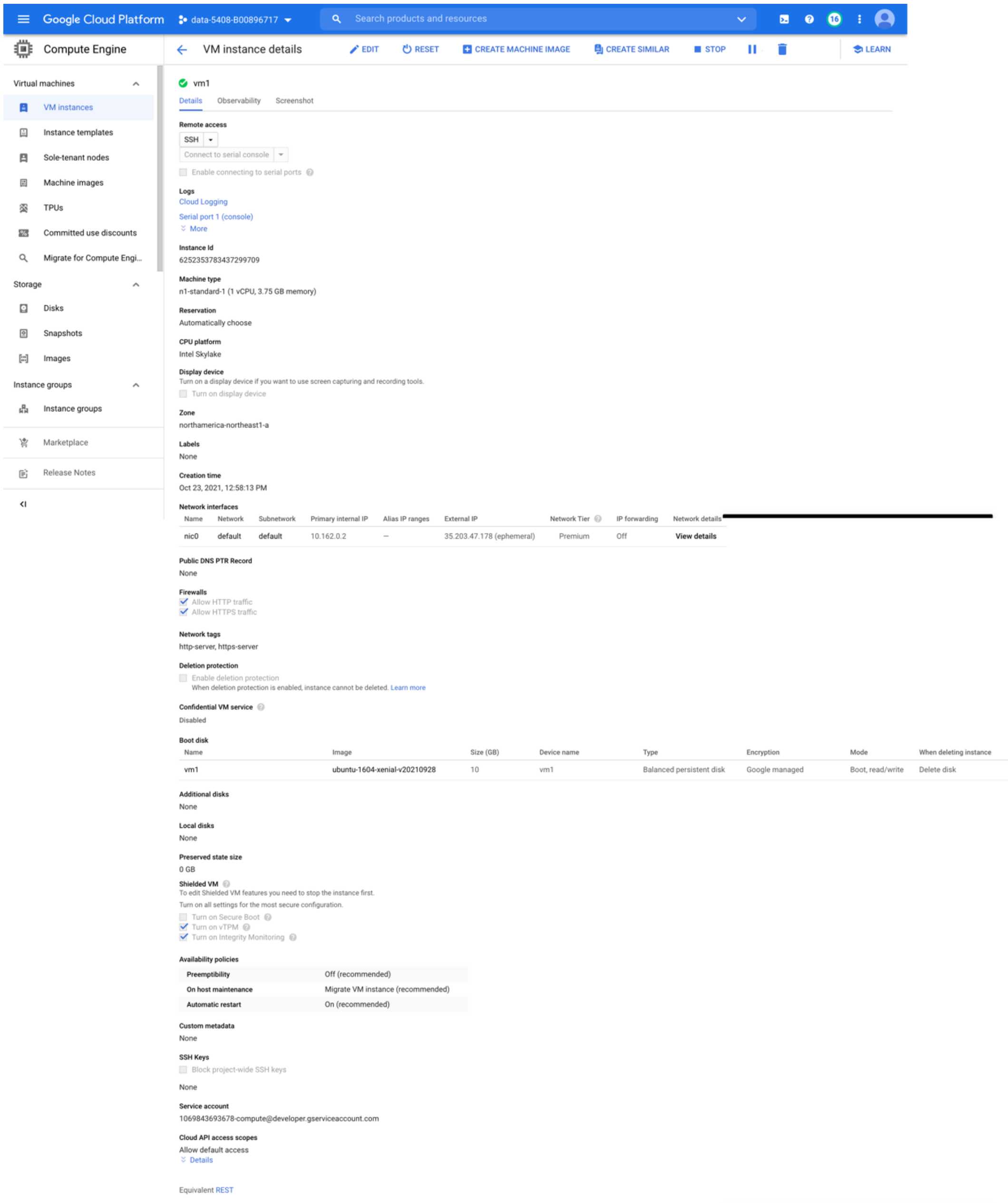
- ## Distributed database setup
    - ### VM Screenshots



VM instances are highly configurable virtual machines for running workloads on Google infrastructure. Learn more

☐	Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect
☐	✓	vm1	northamerica-northeast1-a			10.162.0.2 (nic0)	35.203.47.178	SSH
☐	✓	vm2	northamerica-northeast2-a			10.188.0.2 (nic0)	34.130.217.250	SSH

VM1 Configuration



✓ vm1

Remote access

SSH

Connect to serial console

☐ Enable connecting to serial ports

Logs

Cloud Logging

Serial port 1 (console)

More

Instance Id

6252353783437299709

Machine type

n1-standard-1 (1 vCPU, 3.75 GB memory)

Reservation

Automatically choose

CPU platform

Intel Skylake

Display device

Turn on a display device if you want to use screen capturing and recording tools.

☐ Turn on display device

Zone

northamerica-northeast1-a

Labels

None

Creation time

Oct 23, 2021, 12:58:13 PM

Network interfaces

Name	Network	Subnetwork	Primary internal IP	Alias IP ranges	External IP	Network Tier	IP forwarding	Network details
nic0	default	default	10.162.0.2	–	35.203.47.178 (ephemeral)	Premium	Off	View details

Public DNS PTR Record

None

Firewalls

☑ Allow HTTP traffic

☑ Allow HTTPS traffic

Network tags

http-server, https-server

Deletion protection

☐ Enable deletion protection

When deletion protection is enabled, instance cannot be deleted. Learn more

Confidential VM service

Disabled

Boot disk

Name	Image	Size (GB)	Device name	Type	Encryption	Mode	When deleting instance
vm1	ubuntu-1604-xenial-v20210928	10	vm1	Balanced persistent disk	Google managed	Boot, read/write	Delete disk

Additional disks

None

Local disks

None

Preserved state size

0 GB

Shielded VM

To edit Shielded VM features you need to stop the instance first.

Turn on all settings for the most secure configuration.

☐ Turn on Secure Boot

☑ Turn on vTPM

☑ Turn on Integrity Monitoring

Availability policies

Preemptibility	Off (recommended)
On host maintenance	Migrate VM instance (recommended)
Automatic restart	On (recommended)

Custom metadata

None

SSH Keys

☐ Block project-wide SSH keys

None

Service account

1069843693678-compute@developer.gserviceaccount.com

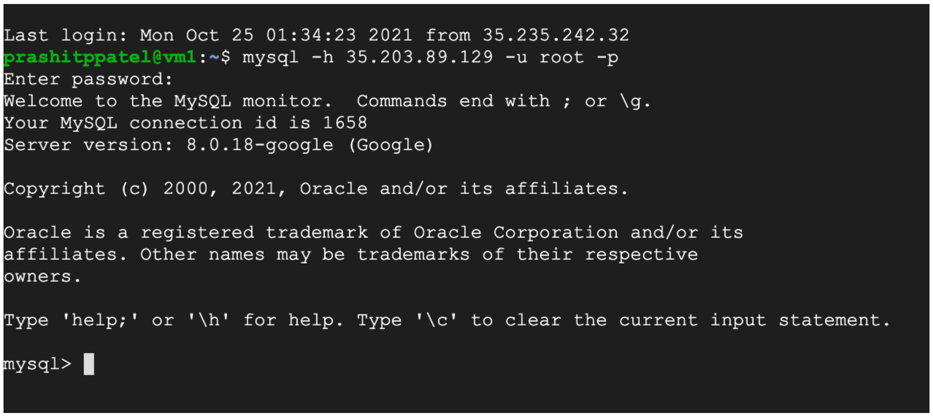
Cloud API access scopes

Allow default access

Details

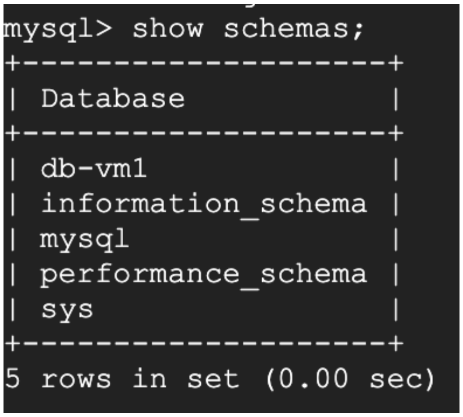
Equivalent REST

VM1 connection with SQL Instance 1



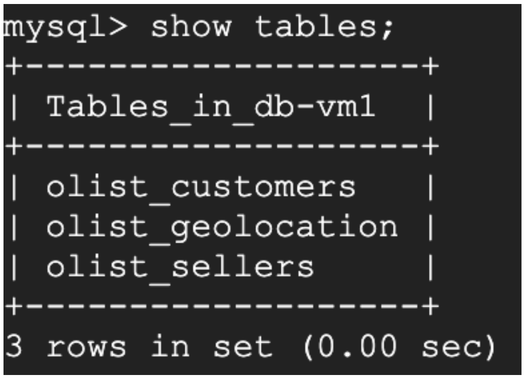
```
Last login: Mon Oct 25 01:34:23 2021 from 35.235.242.32
prashitppatel@vm1:~$ mysql -h 35.203.89.129 -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 1658
Server version: 8.0.18-google (Google)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> 
```

VM1 Databases



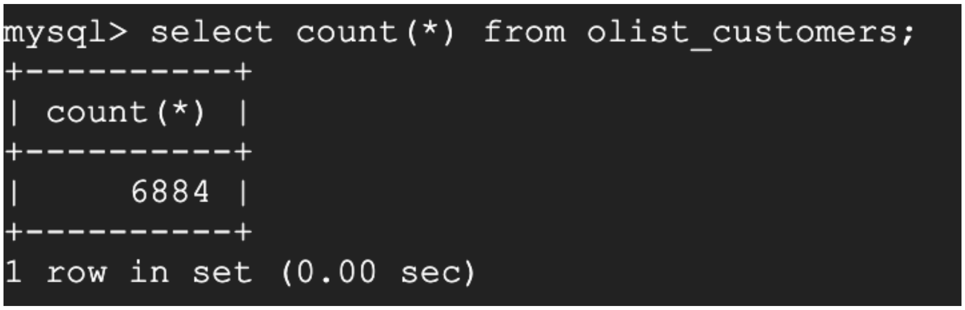
```
mysql> show schemas;
+--------------------+
| Database           |
+--------------------+
| db-vm1             |
| information_schema |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
5 rows in set (0.00 sec)
```

Tables in db-vm1

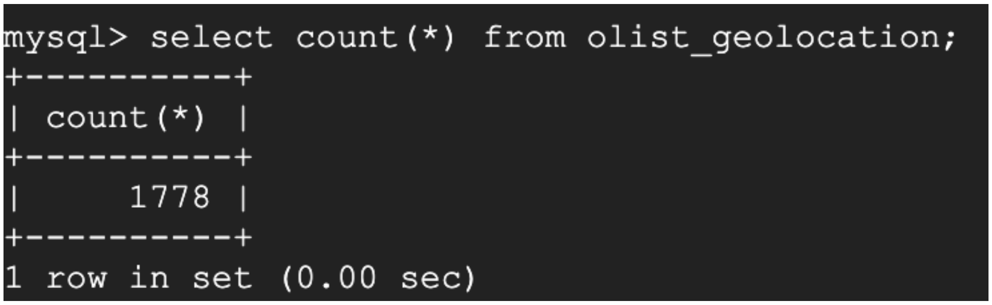


```
mysql> show tables;
+-------------------+
| Tables_in_db-vm1  |
+-------------------+
| olist_customers   |
| olist_geolocation |
| olist_sellers     |
+-------------------+
3 rows in set (0.00 sec)
```

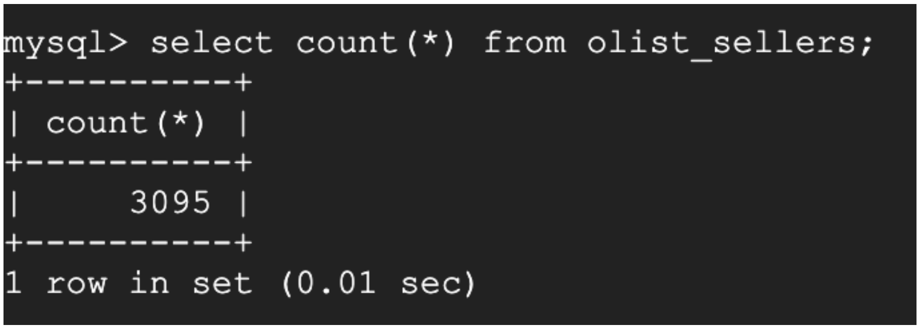
Data count for olist_customers, olist_geolocation and olist_sellers tables



```
mysql> select count(*) from olist_customers;
+----------+
| count(*) |
+----------+
|     6884 |
+----------+
1 row in set (0.00 sec)
```

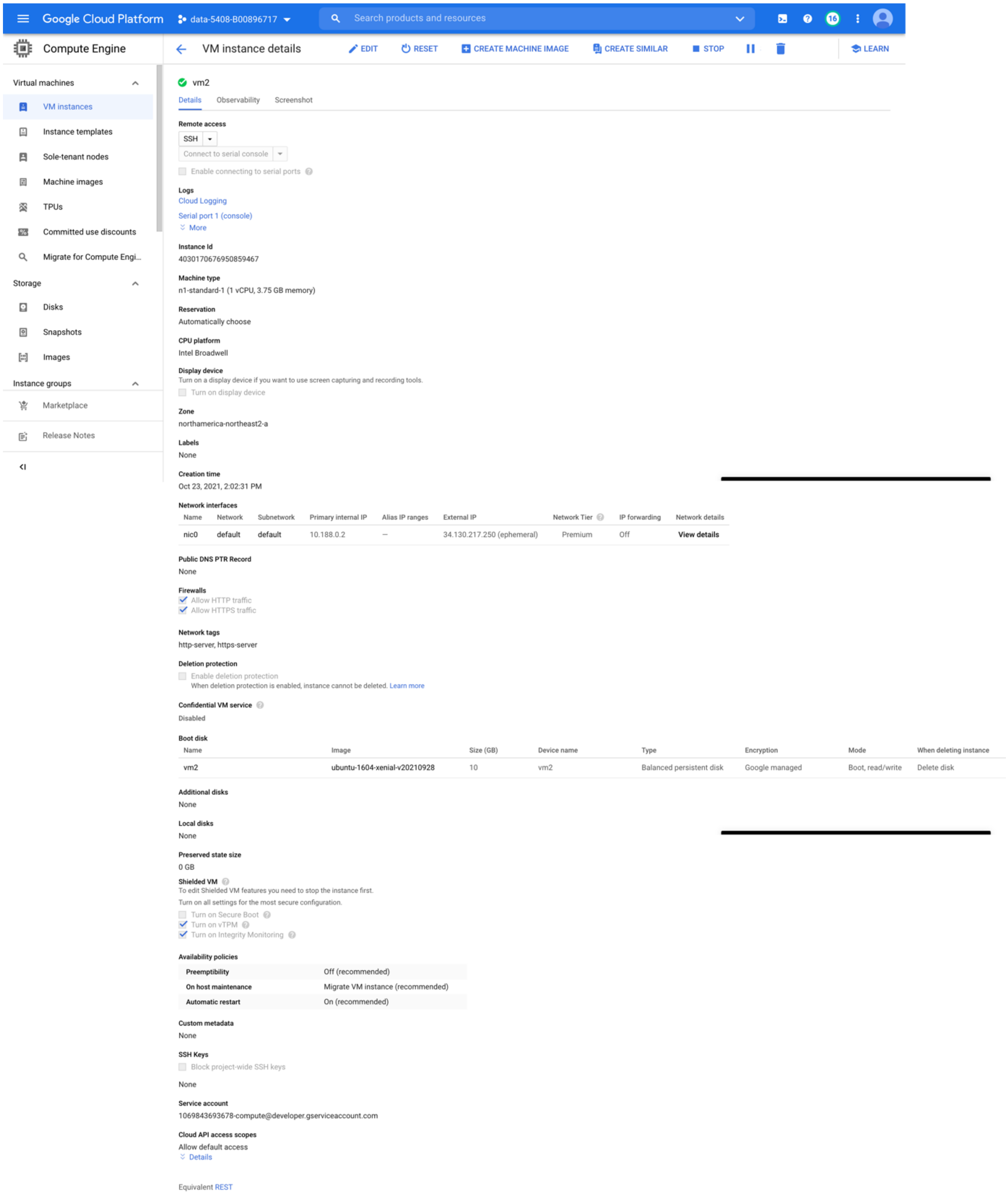


```
mysql> select count(*) from olist_geolocation;
+----------+
| count(*) |
+----------+
|     1778 |
+----------+
1 row in set (0.00 sec)
```



```
mysql> select count(*) from olist_sellers;
+----------+
| count(*) |
+----------+
|     3095 |
+----------+
1 row in set (0.01 sec)
```

VM2 Configuation



vm2

Remote access

SSH

Connect to serial console

□ Enable connecting to serial ports

Logs

Cloud Logging

Serial port 1 (console)

More

Instance Id

4030170676950859467

Machine type

n1-standard-1 (1 vCPU, 3.75 GB memory)

Reservation

Automatically choose

CPU platform

Intel Broadwell

Display device

Turn on a display device if you want to use screen capturing and recording tools.

□ Turn on display device

Zone

northamerica-northeast2-a

Labels

None

Creation time

Oct 23, 2021, 2:02:31 PM

Network interfaces

Name	Network	Subnetwork	Primary internal IP	Alias IP ranges	External IP	Network Tier	IP forwarding	Network details
nic0	default	default	10.188.0.2	—	34.130.217.250 (ephemeral)	Premium	Off	View details

Public DNS PTR Record

None

Firewalls

☑ Allow HTTP traffic
☑ Allow HTTPS traffic

Network tags

http-server, https-server

Deletion protection

□ Enable deletion protection
When deletion protection is enabled, instance cannot be deleted. Learn more

Confidential VM service

Disabled

Boot disk

Name	Image	Size (GB)	Device name	Type	Encryption	Mode	When deleting instance
vm2	ubuntu-1604-xenial-v20210928	10	vm2	Balanced persistent disk	Google managed	Boot, read/write	Delete disk

Additional disks

None

Local disks

None

Preserved state size

0 GB

Shielded VM

To edit Shielded VM features you need to stop the instance first.

Turn on all settings for the most secure configuration.

□ Turn on Secure Boot
☑ Turn on vTPM
☑ Turn on Integrity Monitoring

Availability policies

Preemptibility	Off (recommended)
On host maintenance	Migrate VM instance (recommended)
Automatic restart	On (recommended)

Custom metadata

None

SSH Keys

□ Block project-wide SSH keys

None

Service account

1069843693678-compute@developer.gserviceaccount.com

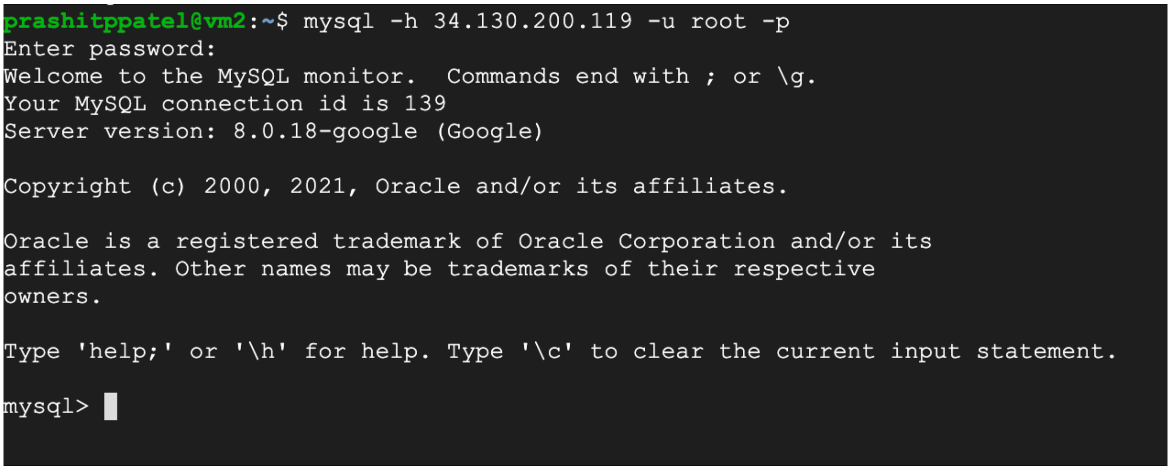
Cloud API access scopes

Allow default access

Details

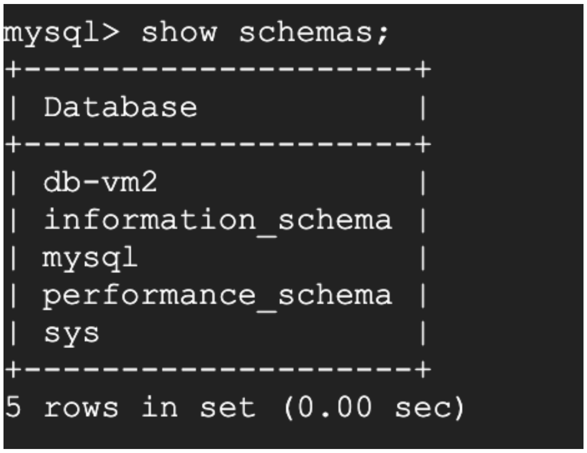
Equivalent REST

VM2 connection with SQL Instance 2



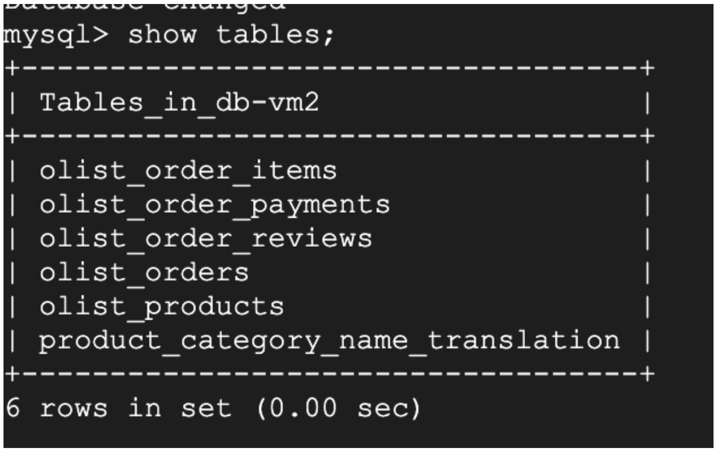
```
prashitppatel@vm2:~$ mysql -h 34.130.200.119 -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 139
Server version: 8.0.18-google (Google)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> 
```

VM2 databases



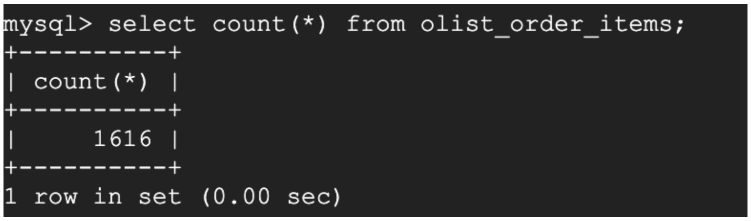
```
mysql> show schemas;
+--------------------+
| Database           |
+--------------------+
| db-vm2             |
| information_schema |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
5 rows in set (0.00 sec)
```

Tables in db-vm2

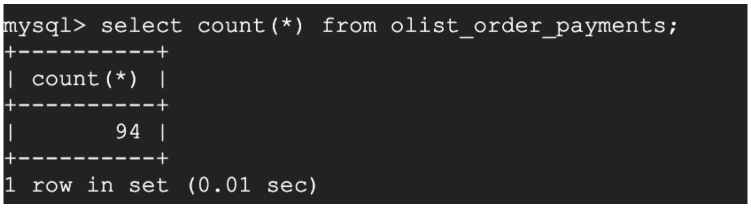


```
Database changed
mysql> show tables;
+-----------------------------------+
| Tables_in_db-vm2                  |
+-----------------------------------+
| olist_order_items                 |
| olist_order_payments              |
| olist_order_reviews               |
| olist_orders                      |
| olist_products                    |
| product_category_name_translation |
+-----------------------------------+
6 rows in set (0.00 sec)
```

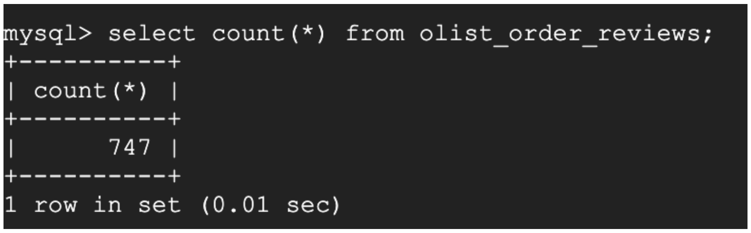
Data count for olist_order_items, olist_ order_payments, olist_order_reviews, olist_orders, olist_products and product_category_name_translation tables



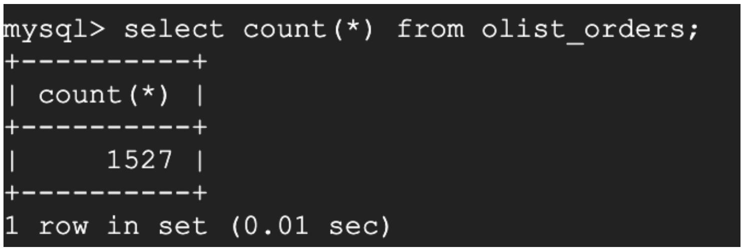
```
mysql> select count(*) from olist_order_items;
+----------+
| count(*) |
+----------+
|     1616 |
+----------+
1 row in set (0.00 sec)
```



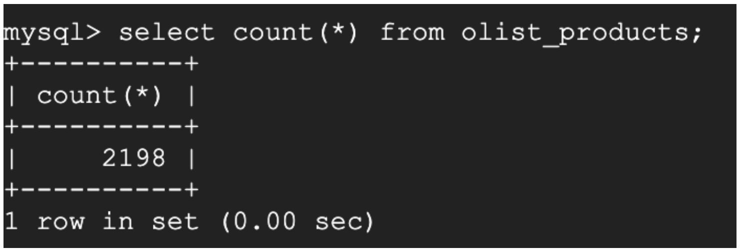
```
mysql> select count(*) from olist_order_payments;
+----------+
| count(*) |
+----------+
|       94 |
+----------+
1 row in set (0.01 sec)
```



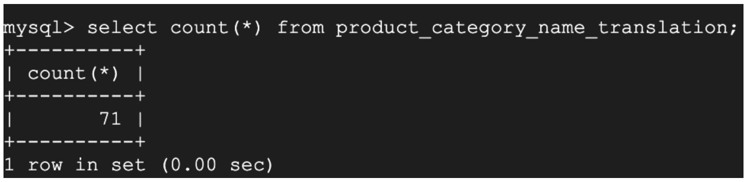
```
mysql> select count(*) from olist_order_reviews;
+----------+
| count(*) |
+----------+
|      747 |
+----------+
1 row in set (0.01 sec)
```



```
mysql> select count(*) from olist_orders;
+----------+
| count(*) |
+----------+
|     1527 |
+----------+
1 row in set (0.01 sec)
```

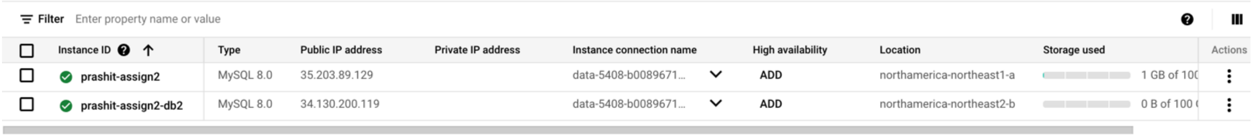


```
mysql> select count(*) from olist_products;
+----------+
| count(*) |
+----------+
|     2198 |
+----------+
1 row in set (0.00 sec)
```



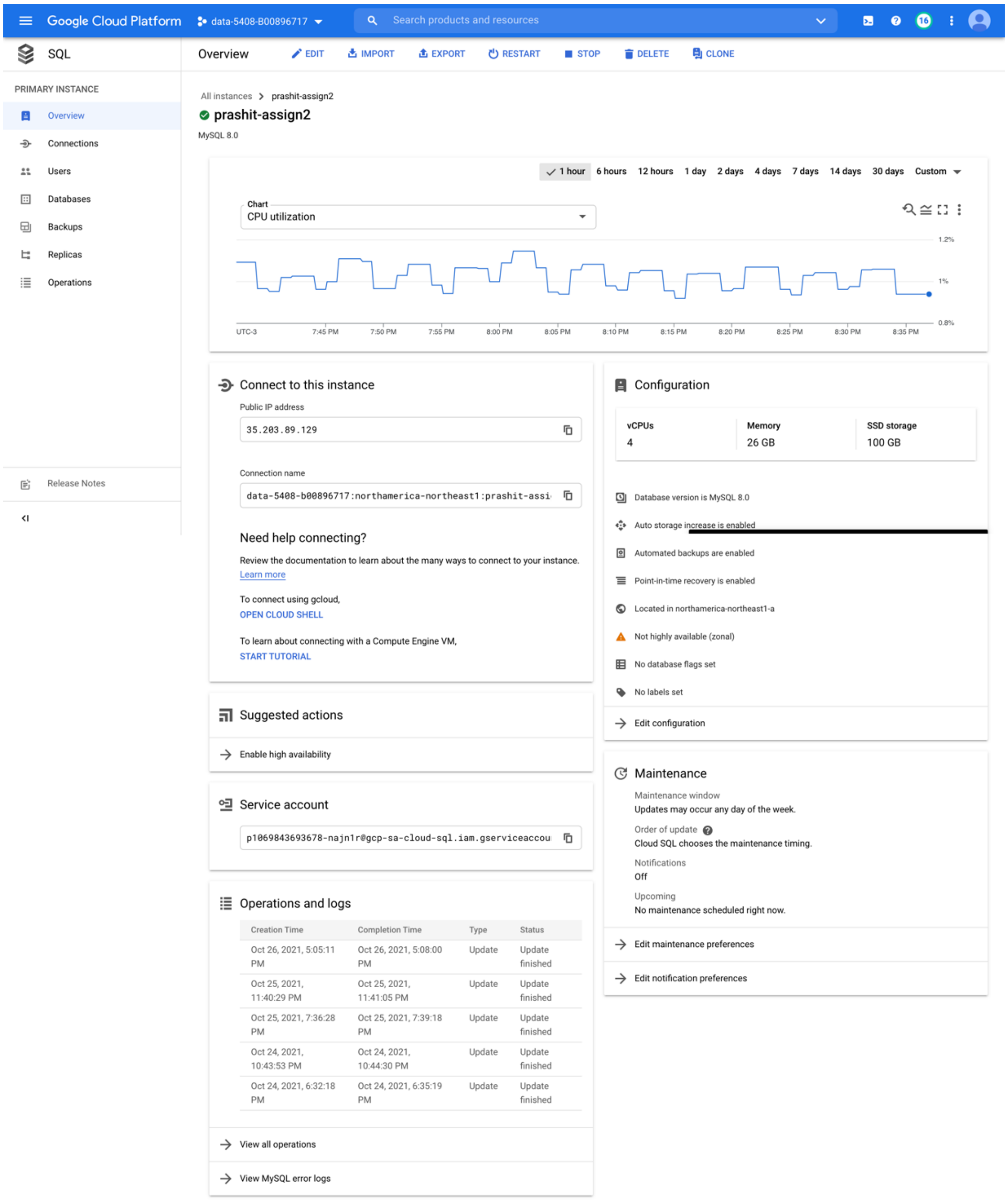
```
mysql> select count(*) from product_category_name_translation;
+----------+
| count(*) |
+----------+
|       71 |
+----------+
1 row in set (0.00 sec)
```

○ SQL Instance Screenshots



□	Instance ID	Type	Public IP address	Private IP address	Instance connection name	High availability	Location	Storage used
□	✓ prashit-assign2	MySQL 8.0	35.203.89.129		data-5408-b0089671...	ADD	northamerica-northeast1-a	1 GB of 100
□	✓ prashit-assign2-db2	MySQL 8.0	34.130.200.119		data-5408-b0089671...	ADD	northamerica-northeast2-b	0 B of 100

SQL Instance 1 Configuration



All instances > prashit-assign2

✓ prashit-assign2

MySQL 8.0

Chart: CPU utilization

Connect to this instance

Public IP address: 35.203.89.129

Connection name: data-5408-b00896717:northamerica-northeast1:prashit-assi

Need help connecting?

Review the documentation to learn about the many ways to connect to your instance. Learn more

To connect using gcloud,
OPEN CLOUD SHELL

To learn about connecting with a Compute Engine VM,
START TUTORIAL

Suggested actions

→ Enable high availability

Service account

p1069843693678-najn1r@gcp-sa-cloud-sql.iam.gserviceaccou

Operations and logs

Creation Time	Completion Time	Type	Status
Oct 26, 2021, 5:05:11 PM	Oct 26, 2021, 5:08:00 PM	Update	Update finished
Oct 25, 2021, 11:40:29 PM	Oct 25, 2021, 11:41:05 PM	Update	Update finished
Oct 25, 2021, 7:36:28 PM	Oct 25, 2021, 7:39:18 PM	Update	Update finished
Oct 24, 2021, 10:43:53 PM	Oct 24, 2021, 10:44:30 PM	Update	Update finished
Oct 24, 2021, 6:32:18 PM	Oct 24, 2021, 6:35:19 PM	Update	Update finished

→ View all operations

→ View MySQL error logs

Configuration

vCPUs	Memory	SSD storage
4	26 GB	100 GB

Database version is MySQL 8.0

Auto storage increase is enabled

Automated backups are enabled

Point-in-time recovery is enabled

Located in northamerica-northeast1-a

Not highly available (zonal)

No database flags set

No labels set

→ Edit configuration

Maintenance

Maintenance window
Updates may occur any day of the week.

Order of update
Cloud SQL chooses the maintenance timing.

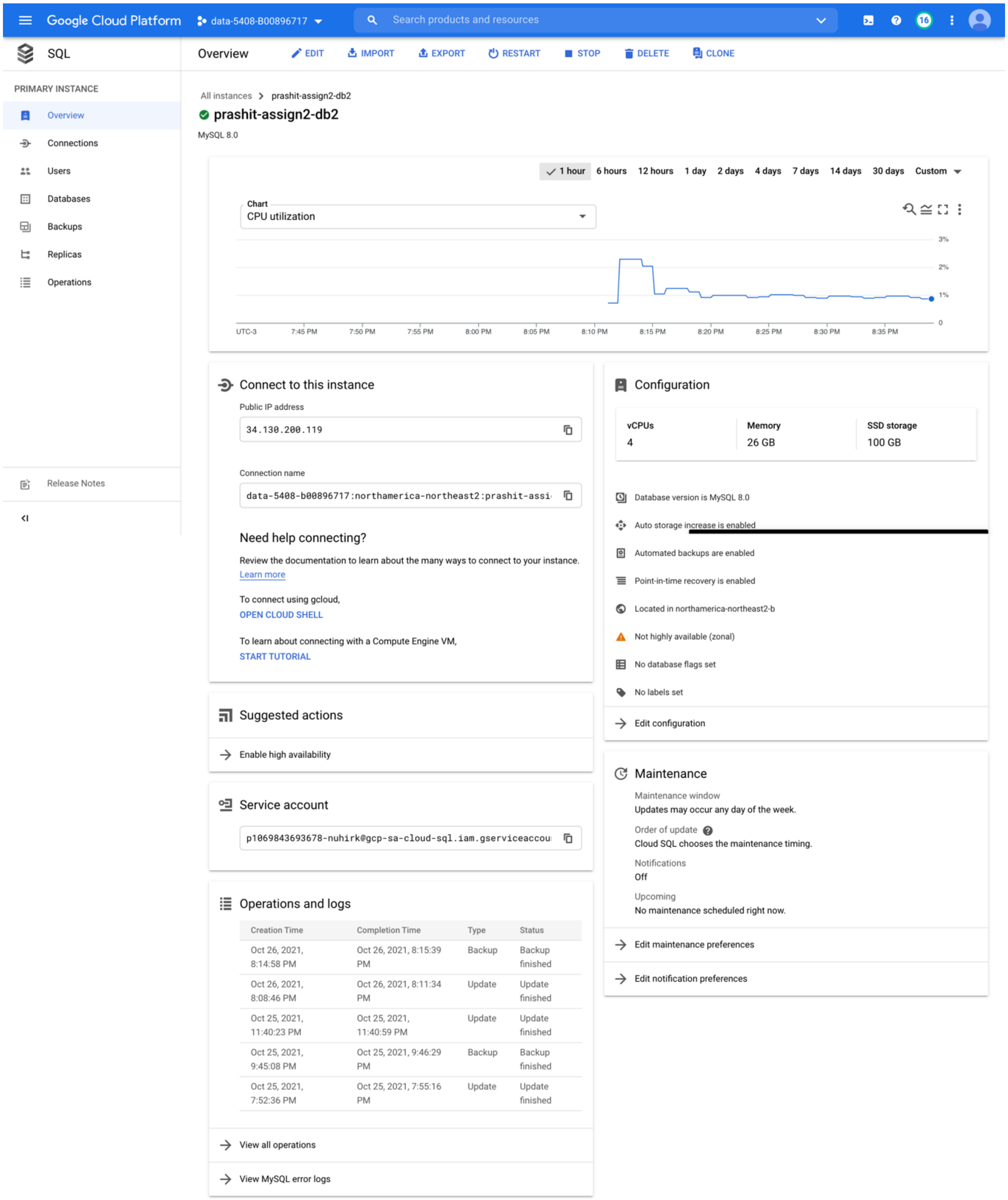
Notifications
Off

Upcoming
No maintenance scheduled right now.

→ Edit maintenance preferences

→ Edit notification preferences

SQL Instance 2 Configuration



All instances > prashit-assign2-db2

prashit-assign2-db2

MySQL 8.0

Chart: CPU utilization

Connect to this instance

Public IP address

```
34.130.200.119
```

Connection name

```
data-5408-b00896717:northamerica-northeast2:prashit-assi
```

Need help connecting?

Review the documentation to learn about the many ways to connect to your instance. Learn more

To connect using gcloud,
OPEN CLOUD SHELL

To learn about connecting with a Compute Engine VM,
START TUTORIAL

Suggested actions

→ Enable high availability

Service account

```
p1069843693678-nuhirk@gcp-sa-cloud-sql.iam.gserviceaccou
```

Operations and logs

Creation Time	Completion Time	Type	Status
Oct 26, 2021, 8:14:58 PM	Oct 26, 2021, 8:15:39 PM	Backup	Backup finished
Oct 26, 2021, 8:08:46 PM	Oct 26, 2021, 8:11:34 PM	Update	Update finished
Oct 25, 2021, 11:40:23 PM	Oct 25, 2021, 11:40:59 PM	Update	Update finished
Oct 25, 2021, 9:45:08 PM	Oct 25, 2021, 9:46:29 PM	Backup	Backup finished
Oct 25, 2021, 7:52:36 PM	Oct 25, 2021, 7:55:16 PM	Update	Update finished

→ View all operations

→ View MySQL error logs

Configuration

vCPUs	Memory	SSD storage
4	26 GB	100 GB

- Database version is MySQL 8.0
- Auto storage increase is enabled
- Automated backups are enabled
- Point-in-time recovery is enabled
- Located in northamerica-northeast2-b
- Not highly available (zonal)
- No database flags set
- No labels set

→ Edit configuration

Maintenance

Maintenance window
Updates may occur any day of the week.

Order of update
Cloud SQL chooses the maintenance timing.

Notifications
Off

Upcoming
No maintenance scheduled right now.

→ Edit maintenance preferences

→ Edit notification preferences

Global Data Dictionary

- I would put global data dictionary (GDD) file in both instances, VM1 as well as VM2. The reason for considering the same is we only have 2 servers for distributed database and so we can easily manage both the data dictionaries. Also, we can keep GDD of VM1 for primary use and use GDD in VM2 as a backup for the one in VM1. We can update the changes in GDD of VM1 when any updates are needed for the tables and later, we can update GDD of VM2 as it serves as a backup only. Due to this there will be less overhead of updating both the GDDs at the same time.

- I would create GDD using Excel and keep it as a csv file in VM instances. All the operations will be first passed to VM1 and will be further redirected based on GDD.

- GDD would include all the table information such as table names, attribute names, types, and constraints such as primary keys and foreign keys.

- Also, along with the GDD, local data dictionary will also be maintained in both VM instances which will contain information about the tables stored in respective instances.

- Please find the GDD file along with the current file in the zip attached.

References:

[1] Digital Ocean [Online].
Available: https://www.digitalocean.com/community/tutorials/how-to-install-mysql-on-ubuntu-20-04

[2] Google Cloud Docs [Online]
Available: https://cloud.google.com/sql/docs/mysql/connect-compute-engine