

Lab 10.1g

- How long did the job take to execute?

About 2.5 minutes

```
Gemini CLI is available in Cloud Shell terminal! Type gemini to try it. Learn more Don't show again Dismiss

- deleted:serviceAccount:gcs-lab@cloud-arthur-emmanart.iam.gserviceaccount.com?uid=117969495140847958021
  role: roles/storage.objectViewer
  etag: BwZEveNB6g=
  version: 1
emmanart@cloudshell:~ (cloud-arthur-emmanart)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME} ; date
Listed 0 items.
Sat Nov 29 04:03:47 PM UTC 2025
emmanart@cloudshell:~ (cloud-arthur-emmanart)$ date

Sat Nov 29 04:05:03 PM UTC 2025
gcloud dataproc jobs submit spark --cluster ${CLUSTERNAME} \
  --class org.apache.spark.examples.SparkPi \
  --jars file:///usr/lib/spark/examples/jars/spark-examples.jar -- 1000 \
  >& output.txt &
[1] 2268
emmanart@cloudshell:~ (cloud-arthur-emmanart)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME} ; date
JOB ID: bdd94c19a7a44ecca15bfe1000bbe8e6
TYPE: spark
STATUS: RUNNING
Sat Nov 29 04:05:17 PM UTC 2025
emmanart@cloudshell:~ (cloud-arthur-emmanart)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME} ; date
JOB ID: bdd94c19a7a44ecca15bfe1000bbe8e6
TYPE: spark
STATUS: RUNNING
Sat Nov 29 04:05:31 PM UTC 2025
emmanart@cloudshell:~ (cloud-arthur-emmanart)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME} ; date
JOB ID: bdd94c19a7a44ecca15bfe1000bbe8e6
TYPE: spark
STATUS: RUNNING
Sat Nov 29 04:05:56 PM UTC 2025
emmanart@cloudshell:~ (cloud-arthur-emmanart)$ gcloud dataproc jobs list --cluster ${CLUSTERNAME} ; date
JOB ID: bdd94c19a7a44ecca15bfe1000bbe8e6
TYPE: spark
STATUS: DONE
[1]+  Done                  gcloud dataproc jobs submit spark --cluster ${CLUSTERNAME} --class org.apache.spark.examples.SparkPi --jars file:///u
sr/lib/spark/examples/jars/spark-examples.jar -- 1000 >&output.txt
Sat Nov 29 04:07:43 PM UTC 2025
emmanart@cloudshell:~ (cloud-arthur-emmanart)$
```

Odin ID: emmanart

- Examine `output.txt` and show the estimate of π calculated.

Directly from output.txt, there is a line that says "Pi is roughly 3.1417748314177483"

- How long did the job take to execute? How much faster did it take?

A little over a limit from examining output.txt

- Examine `output2.txt` and show the estimate of π calculated.

Directly from output.txt, there is a line that says "Pi is roughly 3.1416026714160266"

Lab 10.2g

- Where is the input taken from by default?

'../javahelp/src/main/java/com/google/cloud/training/dataanalyst/javahelp/'

- Where does the output go by default? /tmp/output
- Examine both the `getPackages()` function and the `splitPackageName()` function. What operation does the `'PackageUse()'` transform implement?

It takes each Java import statement and breaks it down into all of its parent packages.

- Look up Beam's `CombinePerKey`. What operation does the `TotalUse` operation implement?

It counts the number of times each package is being used by counting the frequency of the occurrence of each package in the dataset

- Which operations correspond to a "Map"?

```
'GetImports' >> beam.FlatMap(lambda line: startsWith(line, keyword))
```

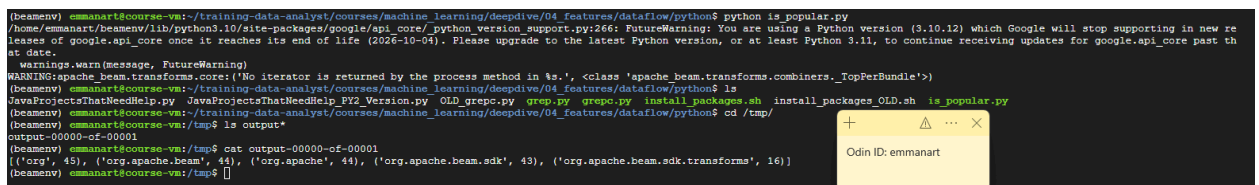
- Which operation corresponds to a "Shuffle-Reduce"?

```
'TotalUse' >> beam.CombinePerKey(sum)
```

- Which operation corresponds to a "Reduce"?

```
'Top_5' >> beam.transforms.combiners.Top.Of(5, key=lambda kv: kv[1])
```

- Take a screenshot of its contents



```
(beamenv) emmanart@course-vsi:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python$ python is_popular.py
/home/emmanart/beamenv/lib/python3.10/site-packages/google/api_core/python_version_support.py:266: FutureWarning: You are using a Python version (3.10.12) which Google will stop supporting in new releases of google.api_core once it reaches its end of life (2026-10-04). Please upgrade to the latest Python version, or at least Python 3.11, to continue receiving updates for google.api_core past that date.
  warnings.warn(message, FutureWarning)
WARNING:apache_beam.transforms.core:('No iterator is returned by the process method in %s.', <class 'apache_beam.transforms.combiners.TopPerBundle'>)
(beamenv) emmanart@course-vsi:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python$ ls
JavaProjectThatNeedHelp.py  JavaProjectThatNeedHelp.PY2_Version.py  OLD_grep.py  grep.py  grep.py  install_packages.sh  install_packages_OLD.sh  is_popular.py
(beamenv) emmanart@course-vsi:~/training-data-analyst/courses/machine_learning/deepdive/04_features/dataflow/python$ cd /tmp/
(beamenv) emmanart@course-vsi:~/tmp$ ls output*
output-00000-of-00001
(beamenv) emmanart@course-vsi:~/tmp$ cat output-00000-of-00001
[('org', 45), ('org.apache.beam', 44), ('org.apache.beam.sdk', 43), ('org.apache.beam.sdk.transforms', 16)]
(beamenv) emmanart@course-vsi:~/tmp$
```

- Explain what the data in this output file corresponds to based on your understanding of the program.
This output shows the top 5 most frequently imported Java packages (including parent packages) from the java files in that were given as input
- What are the names of the stages in the pipeline?

The stages are : Split, PairWithOne, GroupAndSum

- Describe what each stage does.

Split breaks input text into individual words; PairWithOne attaches the number 1 to each word.; GroupAndSum counts how many times each word appears.

- Use `wc` with an appropriate flag to determine the number of different words in King Lear.
- Use sort with appropriate flags to perform a *numeric* sort on the *key field* containing the count for each word in *descending* order. Pipe the output into `head` to show the top 3 words in King Lear and the number of times they appear

This screenshot is for both requests above

```
(beamenv) emmanart@course-vm:~$ ls outputs*
outputs-00000-of-00001
(beamenv) emmanart@course-vm:~$ word: count
Command 'word:' not found, did you mean:
  command 'wordo' from snap wordo (1.0)
See 'snap info <snapname>' for additional versions.
(beamenv) emmanart@course-vm:~$ wc -l outputs*
4784 outputs-00000-of-00001
(beamenv) emmanart@course-vm:~$ sort -t: -k2 -nr outputs* | head -3
the: 786
I: 622
and: 594
(beamenv) emmanart@course-vm:~$
```

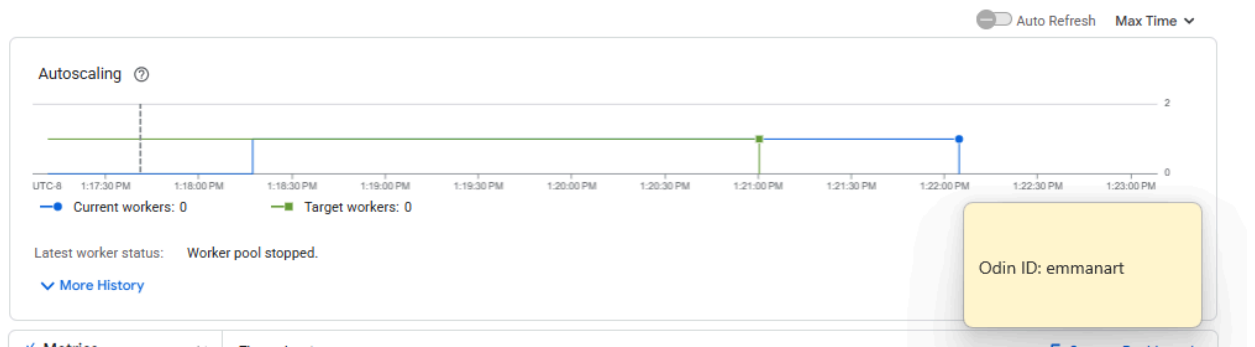
- Use the previous method to show the top 3 words in King Lear, case-insensitive, and the number of times they appear.

```
INFO:apache_beam.runners.worker.sdk_worker:Closing all cached gRPC state handlers.
INFO:apache_beam.runners.worker.sdk_worker:Done consuming work.
INFO:apache_beam.runners.portability.portable_runner:Job state changed to DONE
(beamenv) emmanart@course-vm:~$ sort -t: -k2 -nr outputs_lower* | head -3
the: 908
and: 738
i: 622
(beamenv) emmanart@course-vm:~$
```

- The part of the job graph that has taken the longest time to complete.

Read part

- The autoscaling graph showing when the worker was created and stopped.



- Examine the output directory in Cloud Storage. How many files has the final write stage in the pipeline created?

It created 3 files

Buckets > cloud-arthur-emmanart > tmp > beamapp-emmanart-1129211704-929992-1001f6j.1764451024.930183

Create folder Upload Transfer data Other services

Filter by name prefix only Filter Filter objects and folders Show Live objects only

<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption
<input type="checkbox"/>	pipeline.pb	48.8 KB	application/octet-stream	Nov 29, 2025, 1:17:05 PM	Standard	Nov 29, 2025, 1:17:05 PM	Not public	—	Google-ma
<input type="checkbox"/>	submission_environment_depend_	2.8 KB	application/octet-stream	Nov 29, 2025, 1:17:05 PM	Standard	Nov 29, 2025, 1:17:05 PM	Not public	—	Google-ma
<input type="checkbox"/>	tmp-11f7e76fd53bf997-00000-of-...	1.2 KB		Nov 29, 2025, 1:20:48 PM	Standard	Nov 29, 2025, 1:20:48 PM	Not public	—	Google-ma

Odin ID: emmanart

- Take a screenshot listing the different fields of this object.

```
emmanart@cloudshell:~ (cloud-arthur-emmanart) $ gcloud pubsub subscriptions pull taxiub --auto-ack
Data: [{"ride_id": "7026d995-e678-40f1-9c81-9212edd4995b", "point_id": 50, "latitude": 40.7652800000000004, "longitude": -73.98765, "timestamp": "2025-11-29T18:00:40.14645-05:00", "meter_reading": 2.4426080, "meter_increment": 0.04608696, "ride_status": "enroute", "p
MESSAGE ID: 6710952292060870
ORDERING KEY:
ATTRIBUTES: ee2025-11-29T18:00:40.14645-05:00
DELIVERY ATTEMPT:
ACK STATUS: SUCCESS
emmanart@cloudshell:~ (cloud-arthur-emmanart) $
```

Odin ID: emmanart

- Take a screenshot of the pipeline that includes its stages and the number of elements per second being handled by individual stages.



- Take a screenshot showing the number of passengers and the amount paid for the first ride

The screenshot shows a Google Cloud console interface with a table of taxi ride data. The table has columns: Row, ride_id, point_idx, latitude, longitude, timestamp, meter_reading, meter_increment, ride_status, and passenger_count. The first row shows a ride with 1 passenger and a meter reading of 4.4809666.

Row	ride_id	point_idx	latitude	longitude	timestamp	meter_reading	meter_increment	ride_status	passenger_count
1	b1b2f41-09fc-42be-a7c7-af2c...	187	40.6447200...	-73.795340...	2025-11-29 23:13:10.274580 U...	4.4809666	0.023962388	enroute	1
2	6799748-be94-487b-a007-acf...	234	40.75441	-73.98443	2025-11-29 23:13:31.598070 U...	9.042586	0.03864353	enroute	1
3	ea305a8-657d-44ca-b20f-2ec...	560	40.73834	-73.84985	2025-11-29 23:13:31.809760 U...	12.812339	0.022879176	enroute	5
4	e01cac80-d5d4-4bf1-b2d7-9a1...	2105	40.76941	-73.91182	2025-11-29 23:13:31.795810 U...	42.11595	0.020007579	enroute	1
5	9e68a81-5339-44c6-9af7-f80...	155	40.7718000...	-73.95618	2025-11-29 23:13:29.667770 U...	6.35022	0.040969163	enroute	1
6	67e7b4be-4a85-4097-a079-25...	42	40.7414600...	-73.97829	2025-11-29 23:13:29.711520 U...	1.586262	0.03776824	enroute	2
7	6fcc92bf-7408-4159-b782-f7db...	439	40.78987	-73.982120...	2025-11-29 23:13:31.671940 U...	11.223803	0.02556675	enroute	2
8	24975703-041f-4b3b-8b3d-8f3...	25	40.7483600...	-73.982520...	2025-11-29 23:13:31.822070 U...	1.0309278	0.041237112	enroute	1
9	2cd6582e-bdb4-4313-b34d-7da...	31	40.6598300...	-73.79331	2025-11-29 23:13:29.883410 U...	0.21896203	0.007063291	enroute	1
10	9da965b3-7e44-4161-bbda-f02...	334	40.75798	-74.004370...	2025-11-29 23:13:31.629690 U...	9.979189	0.029877812	enroute	1
11	bcc0ccbb-ad9c-4797-697e-351...	605	40.73814	-74.01008	2025-11-29 23:13:31.741540 U...	14.3185215	0.023666978	enroute	5
12	f6958c49-3848-4d17-972a-5cc...	184	40.7042800...	-73.994740...	2025-11-29 23:13:31.603030 U...	4.386498	0.023839664	enroute	1
13	74f77e86-71f4-481d-bdfe-0d70...	1160	40.7716900...	-73.871140...	2025-11-29 23:13:31.846240 U...	32.951897	0.028406806	enroute	1
14	ae39f667-c19b-4963-8011-4ed...	226	40.76339	-73.96526	2025-11-29 23:13:32.033410 U...	7.105303	0.031439394	enroute	5
15	837004be-61dd-4072-8b48-a5...	430	40.7619000...	-73.99682	2025-11-29 23:13:31.924970 U...	10.583477	0.024612738	enroute	1
16	02d0033-dbef-4f6c-9198-6188...	452	40.76044	-73.97148	2025-11-29 23:13:29.853080 U...	14.217623	0.031454917	enroute	1
17	65c490fe-3623-4d22-b6e7-00b...	673	40.7484900...	-73.96895	2025-11-29 23:13:30.033410 U...	17.175203	0.025520362	enroute	1
18	b0c86316-908b-4524-92f5-00b...	175	40.73004	-74.010470...	2025-11-29 23:13:31.581700 U...	4.7585435	0.027191678	enroute	1
19	556af9b2-7bba-4c9e-85d1-73a...	205	40.78195	-73.97548	2025-11-29 23:13:31.683890 U...	6.76699	0.033009708	enroute	1
20	e2b0c4ac-26a2-4520-b69b-99e...	170	40.76189	-73.99382	2025-11-29 23:13:31.684360 U...	5.0824738	0.029896906	enroute	1
21	ceb6703c-10ae-4f2d-b51ecc8...	336	40.7648	-73.98429	2025-11-29 23:13:31.892910 U...	11.385124	0.033884298	enroute	1
22	b4c45705-0bb1-47e7-8b8f-a49...	914	40.7116700...	-73.822250...	2025-11-29 23:13:31.815900 U...	17.172121	0.018787878	enroute	1

- Take a screenshot showing the estimated number of rows in the table.

cloud-arthur-emmanart / Datasets / taxirides / Tables / realtime

realtime [Query](#) [Open in](#) [Share](#) [Copy](#) [Snapshot](#) [Delete](#) [Export](#) [Refresh](#)

This is a partitioned table. [Learn more](#) [Dismiss](#)

Schema **Details** Preview Table Explorer **Preview** Insights Lineage Data Profile Data Quality

Long term logical bytes	0 B
Current physical bytes	0 B
Total physical bytes	0 B
Active physical bytes	0 B
Long term physical bytes	0 B
Time travel physical bytes	0 B

Streaming buffer statistics

Estimated size	211.24 MB
Estimated rows	1,272,319
Earliest entry time	Nov 29, 2025, 3:15:24 PM UTC-8

Job history [Show](#)

- Take a screenshot showing the per-minute number of rides, passengers, and revenue for the data collected

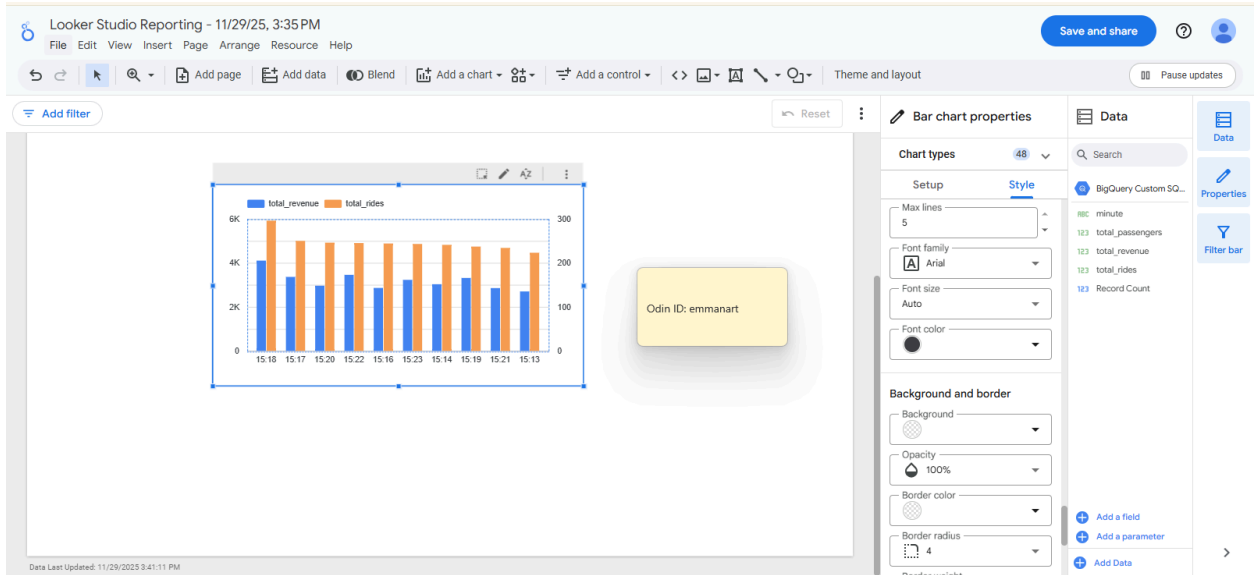
✓ Query completed

Query results [Save results](#) [Open in](#) [↕](#)

Job information **Results** Visualization JSON Execution details Execution graph

Row	minute	total_rides	total_passengers	total_revenue
1	15:13	223	360	2701.640009600...
2	15:14	241	364	3028.9199975
3	15:15	222	365	2792.340001300...
4	15:16	244	360	2861.379982600...
5	15:17	250	373	3362.459995599...
6	15:18	296	468	4103.050007100...
7	15:19	237	381	3314.509993899...
8	15:20	246	372	2962.240006099...
9	15:21	234	377	2853.929991899...
10	15:22	245	408	3456.599993199...
11	15:23	243	392	3227.9599978
12	15:24	147	261	1760.769996099...

- Take a screenshot showing the plot for your data for your lab notebook



Lab 10.3g

- Take a screenshot of the output to include in your lab notebook. How many networks, subnetworks, and VM instances have been created?

```

Google Cloud cloud-arthur-emmanart Cloud Dataflow
CLOUD SHELL Terminal (cloud-arthur-emmanart)
google_compute_subnetwork.us-west-s1: Creation complete after 21s [id-projects/cloud-arthur-emmanart/regions/us-west1/subnetworks/us-west-s1]
google_compute_instance.v1.vm: Creating...
google_compute_subnetwork.europe-west1: Creation complete after 23s [id-projects/cloud-arthur-emmanart/regions/europe-west1/subnetworks/europe-west1]
google_compute_instance.eu1.vm: Creating...
google_compute_instance.us-east5: Creation complete after 24s [id-projects/cloud-arthur-emmanart/regions/us-east5/subnetworks/us-east5]
google_compute_instance.eu1.vm: Creating...
google_compute_subnetwork.asia-east1: Still creating... [30s elapsed]
google_compute_instance.w2.vm: Still creating... [10s elapsed]
google_compute_instance.w1.vm: Still creating... [10s elapsed]
google_compute_subnetwork.asia-east1: Creation complete after 32s [id-projects/cloud-arthur-emmanart/regions/asia-east1/subnetworks/asia-east1]
google_compute_instance.asia1.vm: Creating...
google_compute_instance.eu1.vm: Still creating... [10s elapsed]
google_compute_instance.eu1.vm: Still creating... [10s elapsed]
google_compute_instance.w2.vm: Still creating... [20s elapsed]
google_compute_instance.w1.vm: Still creating... [20s elapsed]
google_compute_instance.asia1.vm: Still creating... [10s elapsed]
google_compute_instance.eu1.vm: Still creating... [20s elapsed]
google_compute_instance.eu1.vm: Still creating... [20s elapsed]
google_compute_instance.w1.vm: Creation complete after 26s [id-projects/cloud-arthur-emmanart/zones/us-west1-b/instances/w1-vm]
google_compute_instance.w2.vm: Creation complete after 27s [id-projects/cloud-arthur-emmanart/zones/europe-west1-d/instances/eu1-vm]
google_compute_instance.eu1.vm: Still creating... [20s elapsed]
google_compute_instance.asia1.vm: Creation complete after 29s [id-projects/cloud-arthur-emmanart/zones/us-east5-a/instances/e1-vm]
google_compute_instance.asia1.vm: Creation complete after 30s [id-projects/cloud-arthur-emmanart/zones/asia-east1-b/instances/asial-vm]

Apply complete! Resources: 11 added, 0 changed, 0 destroyed.

Outputs:
instance_internal_ips = {
  "asia1" = "10.40.0.2"
  "e1" = "10.20.0.2"
  "eu1" = "10.30.0.2"
  "w1" = "10.10.0.2"
  "w2" = "10.11.0.100"
}
instance_nat_ips = {
  "asia1" = "104.199.130.203"
  "e1" = "34.162.221.221"
  "eu1" = "35.205.244.124"
  "w1" = "34.187.193.61"
  "w2" = "34.19.35.39"
}
emmanart@cloudshell:~/cs430-src/10_CDN (cloud-arthur-emmanart)

```

1 network 5 subnets and 5 vms were created

- Visit the web console for VPC network and show the network and the subnetworks that have been created. Validate that it has created the infrastructure in the initial figure. Note the lack of firewall rules that have been created.

Google Cloud console showing VPC network details for 'networking101'. The page displays a list of subnets with the following data:

Name	Region	Stack Type	Primary IPv4 range	Secondary IPv4 ranges	IPv6 ranges	Reserved internal ranges	PDP	Gateway	Private Google Access
asia-east1	asia-east1	IPv4 (single-stack)	10.40.0.0/16			None	None	10.40.0.1	On
eu-west1	eu-west1	IPv4 (single-stack)	10.30.0.0/16			None	None	10.30.0.1	On
us-east5	us-east5	IPv4 (single-stack)	10.20.0.0/16			None	None	10.20.0.1	On
us-west-1	us-west1	IPv4 (single-stack)	10.10.0.0/16			None	None	10.10.0.1	On
us-west-2	us-west1	IPv4 (single-stack)	10.11.0.0/16			None	None	10.11.0.1	On

Below the subnets table, there is a section for 'Reserved proxy-only subnets for load balancing' which currently shows 'No rows to display'.

- Visit the web console for Compute Engine and show all VMs that have been created, their internal IP addresses and the subnetworks they have been instantiated on. Validate that it has created the infrastructure shown in the initial figure.

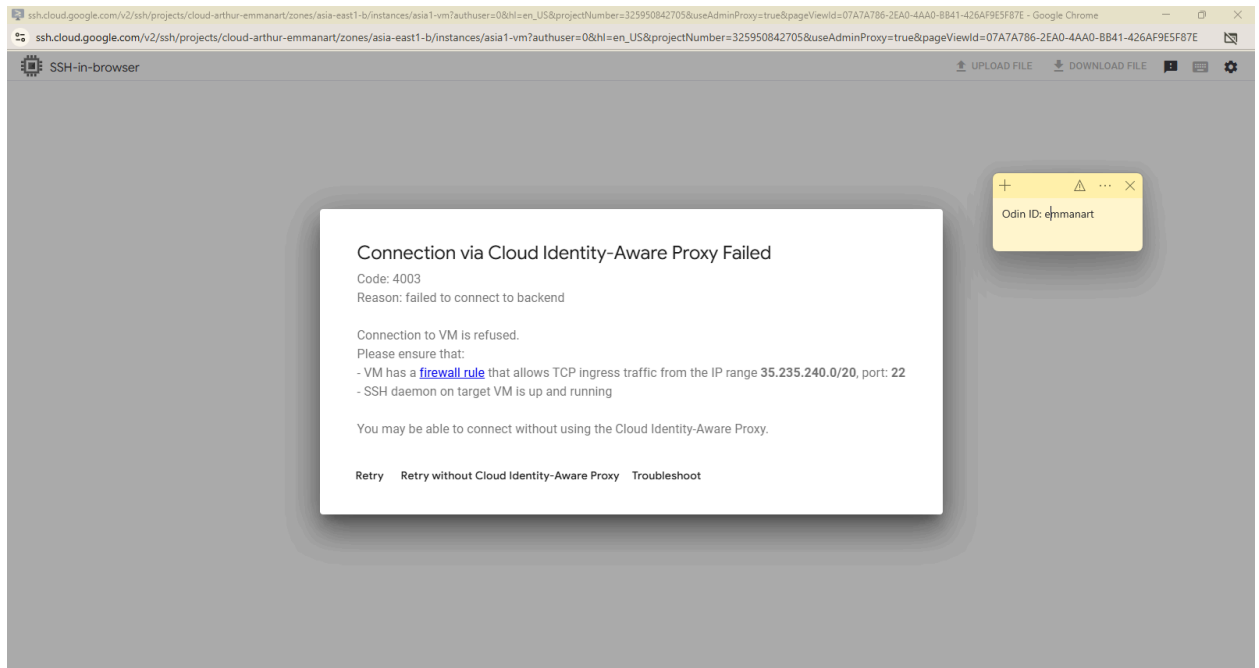
Google Cloud console showing Compute Engine VM instances. The page displays a list of VM instances with the following data:

Status	Name	Zone	Creation time	Recommendations	In use by	Internal IP	External IP	Network	Connect
✓	asia1-vm	asia-east1-b	Nov 29, 2025, 6:17:44 PM			10.40.0.2 (nic0)	104.199.130.203 (nic0)	networking101	SSH
✓	e1-vm	us-east5-a	Nov 29, 2025, 6:17:35 PM			10.20.0.2 (nic0)	34.162.221.221 (nic0)	networking101	SSH
✓	eu1-vm	eu-west1-d	Nov 29, 2025, 6:17:32 PM			10.30.0.2 (nic0)	35.205.244.124 (nic0)	networking101	SSH
✓	w1-vm	us-west1-b	Nov 29, 2025, 6:17:30 PM			10.10.0.2 (nic0)	34.187.193.61 (nic0)	networking101	SSH
✓	w2-vm	us-west1-b	Nov 29, 2025, 6:17:30 PM			10.11.0.100 (nic0)	34.19.35.39 (nic0)	networking101	SSH

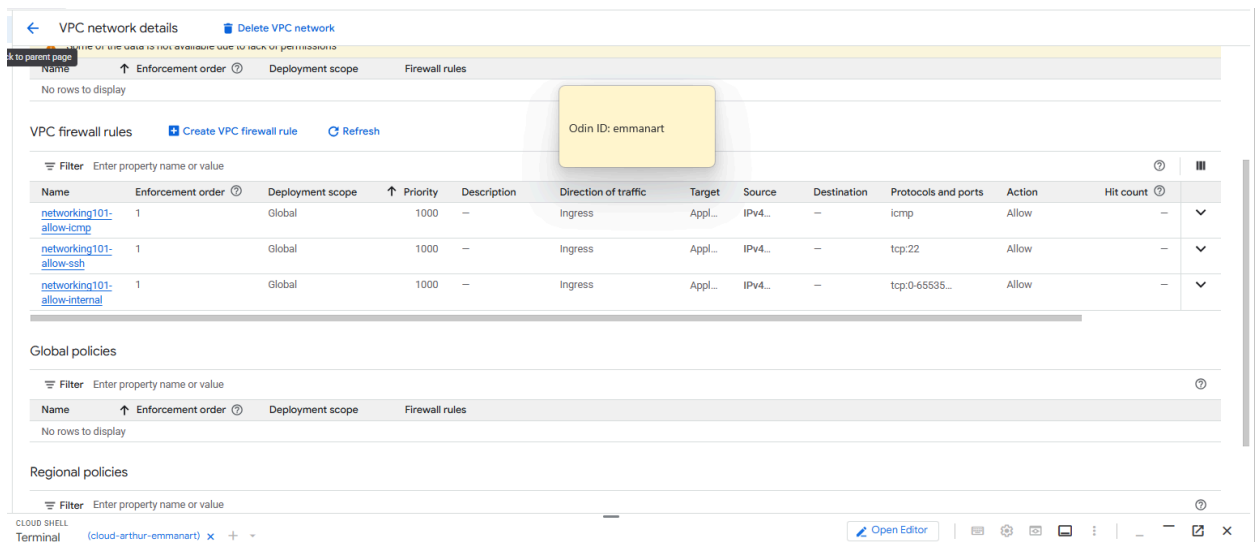
Below the VM instances table, there is a 'Related actions' section with several cards for managing VMs, including 'Fix data protection gaps', 'View billing report', 'Monitor VMs', 'Explore VM logs', 'Set up firewall rules', 'Patch management', and 'Load balance between VMs'.

- Click on the **ssh** button for one of the VMs and attempt to connect. Did it succeed?

No



- Take a screenshot that indicates the new rules have been deployed



- Given this, fill in the table with the measured latencies between the 6 pairs and include it in your lab notebook. Use the shortest latency measured for each pair.

Location pair	Ideal latency	Measured latency
us-west1 us-east5	~45 ms	56.6
us-west1 europe-west1	~93 ms	133.1
us-west1 asia-east1	~114 ms	116.2
us-east5 europe-west1	~76 ms	98.5
us-east5 asia-east1	~141 ms	163.6
europe-west1 asia-east1	~110 ms	248.8

- Are the instances in the same availability zone or in different ones?
- List all availability zones that your servers show up in for your lab notebook.
- Show a screenshot of the page that is returned.

console.cloud.google.com/net-services/loadbalancing/details/http/webserver-front-end-lb?project=cloud-arthur-emmanart&cloudshell=true

Google Cloud cloud-arthur-emmanart Search (/) for resources, docs, products, and more

Load balancer details

Edit Delete View in Network Topology

URL Rule

Backend Service

webserver-backend-migs

Odin ID: emmanart

CLOUD SHELL

Terminal (cloud-arthur-emmanart) x (cloud-arthur-emmanart) x (cloud-arthur-emmanart) x +

Open Editor

Gemini CLI is available in Cloud Shell terminal! Type gemini to try it. [Learn more](#) Don't show again Dismiss

```
{
  "response_time": 14.25,
  "transaction_rate": 13.53,
  "throughput": 0.01,
  "concurrency": 192.74,
  "successful_transactions": 577,
  "failed_transactions": 1020,
  "longest_transaction": 19.28,
  "shortest_transaction": 0.00
}
```

emmanart@eul-vm1-4 |