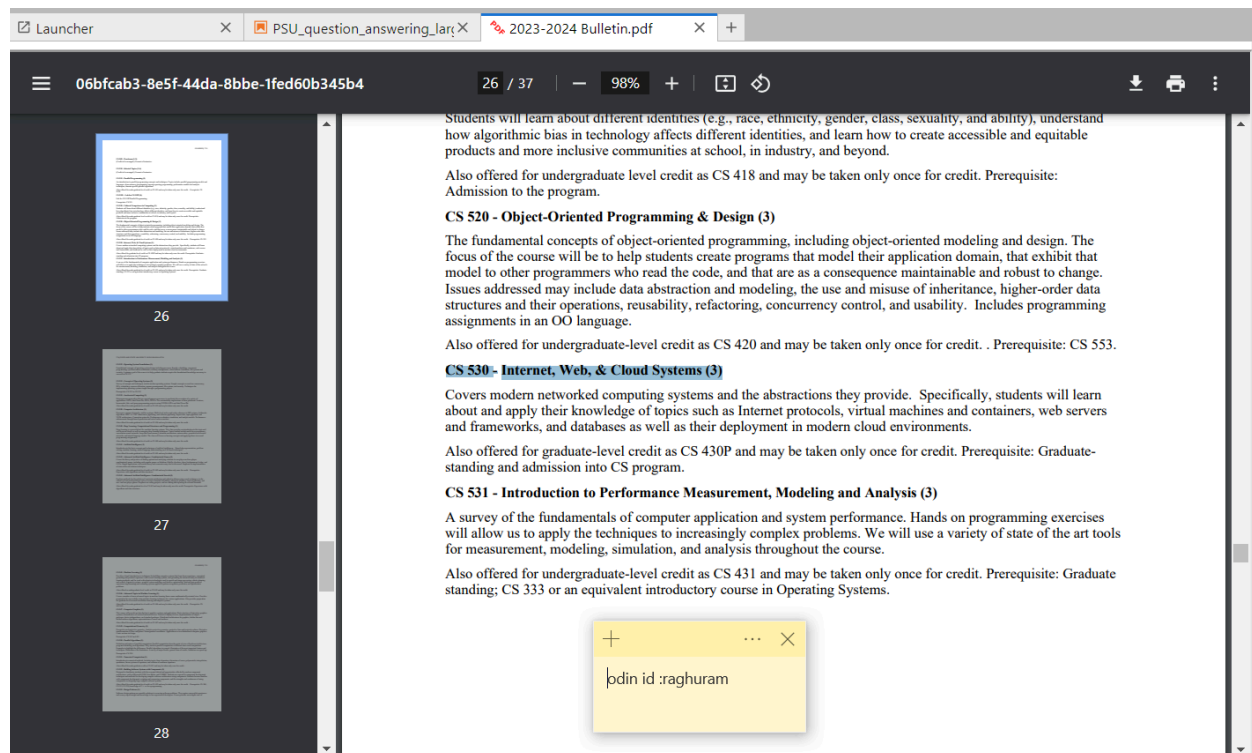


10.1g: LLMs.....	2
4. Walk through notebook.....	3
5. Final questions and clean-up.....	6
10.2g: CDN.....	6
6. Deployment.....	6
8. Update deployment.....	9
16. Test groups.....	11
19. Test load balancer.....	12
20. Siege! (Part 1).....	13
21. Siege! (Part 2).....	14

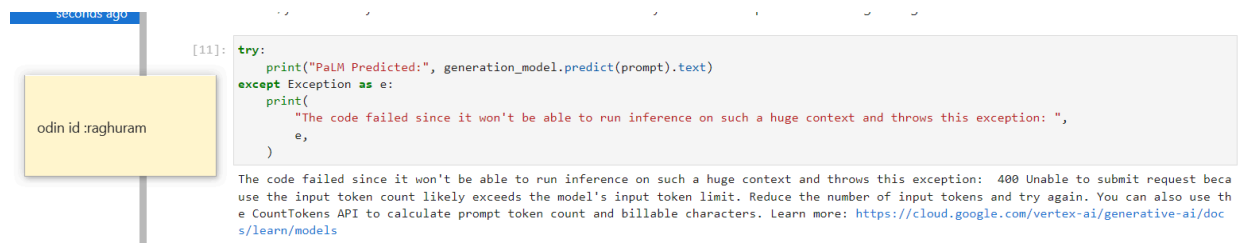
10.1g: LLMs

4. Walk through notebook

- Take a screenshot that includes your OdinID showing the page number and the description of the class for your lab notebook



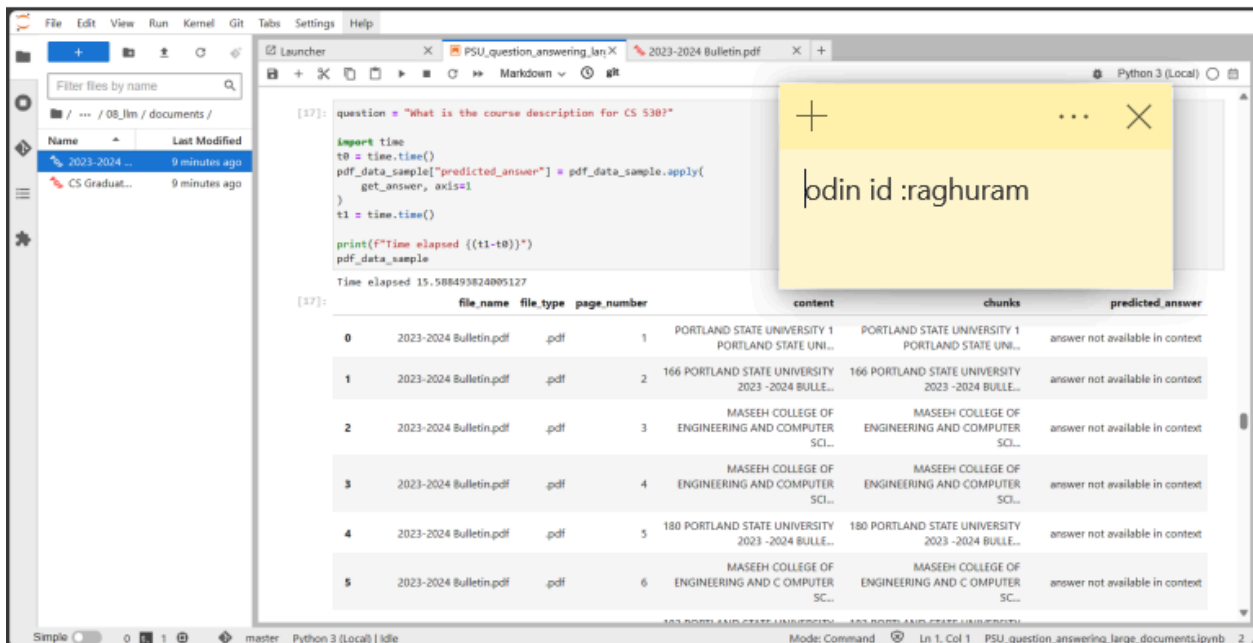
- Take a screenshot that includes your OdinID showing the error that is returned for your lab notebook



- Provide an explanation as to why the description is not returned for your lab notebook

The lab notebook description cannot be provided because the necessary context, referred to as `context[:5000]`, is unavailable in the provided code. Without this information, it's not possible to generate a meaningful response

- Take a screenshot including your OdinID that shows how long it took to perform the prediction across every chunk



The screenshot shows a Jupyter Notebook interface. The code in the cell [17]: is as follows:

```
[17]: question = "What is the course description for CS 530?"

import time
t0 = time.time()
pdf_data_sample["predicted_answer"] = pdf_data_sample.apply(
    get_answer, axis=1
)
t1 = time.time()

print(f"Time elapsed {(t1-t0)}")
pdf_data_sample
```

Below the code, the output shows the time elapsed: 15.588493824005127. Then, a table displays the results for 6 chunks of the PDF file '2023-2024 Bulletin.pdf'.

	file_name	file_type	page_number	content	chunks	predicted_answer
0	2023-2024 Bulletin.pdf	pdf	1	PORTLAND STATE UNIVERSITY 1 PORTLAND STATE UNI...	PORTLAND STATE UNIVERSITY 1 PORTLAND STATE UNI...	answer not available in context
1	2023-2024 Bulletin.pdf	pdf	2	166 PORTLAND STATE UNIVERSITY 2023 -2024 BULLE...	166 PORTLAND STATE UNIVERSITY 2023 -2024 BULLE...	answer not available in context
2	2023-2024 Bulletin.pdf	pdf	3	MASEEH COLLEGE OF ENGINEERING AND COMPUTER SCI...	MASEEH COLLEGE OF ENGINEERING AND COMPUTER SCI...	answer not available in context
3	2023-2024 Bulletin.pdf	pdf	4	MASEEH COLLEGE OF ENGINEERING AND COMPUTER SCI...	MASEEH COLLEGE OF ENGINEERING AND COMPUTER SCI...	answer not available in context
4	2023-2024 Bulletin.pdf	pdf	5	180 PORTLAND STATE UNIVERSITY 2023 -2024 BULLE...	180 PORTLAND STATE UNIVERSITY 2023 -2024 BULLE...	answer not available in context
5	2023-2024 Bulletin.pdf	pdf	6	MASEEH COLLEGE OF ENGINEERING AND C OMPUTER SC...	MASEEH COLLEGE OF ENGINEERING AND C OMPUTER SC...	answer not available in context

- How many chunks returned predictions?

3 chunks returned

- Take a screenshot that includes your OdinID showing the result that is returned for your lab notebook

```
[19]: prompt = f"""Answer the question as precise as possible using the provided context. If the answer is
      not contained in the context, say "answer not available in context" \n\n
      Context: \n {context_map_reduce}\n
      Question: \n {question} \n
      Answer:
      """

      print("the prompt: ", prompt)
      print("the number of words in the prompt: ", len(prompt))

      print("PaLM Predicted:", generation_model.predict(prompt).text)

      the prompt: Answer the question as precise as possible using the provided context. If the answer is
      not contained in the context, say "answer not available in context"

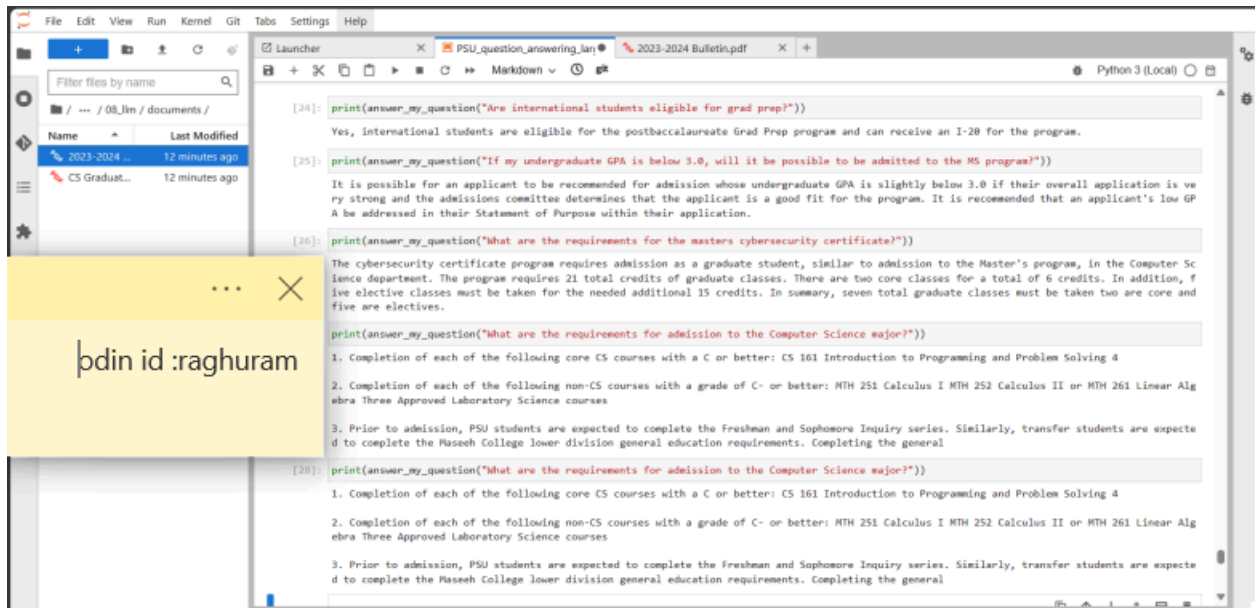
      Context:
      ['Internet, Web, Cloud Systems', 'Internet, Web, Cloud Systems', 'Covers modern networked computing systems and the abstractions they provide S
      pecifically, students will learn about and apply their knowledge of topics such as Internet protocols, virtual machines and containers, web serv
      ers and frameworks, and databases as well as their deployment in modern cloud environments', 'Covers modern networked computing systems and the
      abstractions they provide Specifically, students will learn about and apply their knowledge of topics such as Internet protocols, virtual machin
      es and containers, web servers and frameworks, and databases as well as their deployment in modern cloud environments Also offered for graduate
      -level credit as CS 430P and may be taken only once for credit Prerequisite: Graduate - standing and admission into CS program', 'Advanced softw
      are design patterns using Java as the presentation language Course is suitable to software architects and developers who are already well -verse
      d in this language In addition, it offers continuous opportunities for learning the most advanced featur es of the Java language and understandi
      ng some principles behind the design of its fundamental libraries Also offered as CS 653 and may be taken only once for credit Prerequisite: pro
      gramming in Java and CS 520']?

      Question:
      What is the course description for CS 530?

      Answer:

      the number of words in the prompt: 1623
      PaLM Predicted: Covers modern networked computing systems and the abstractions they provide Specifically, students will learn about and apply th
      eir knowledge of topics such as Internet protocols, virtual machines and containers, web servers and frameworks, and databases as well as their
```

- Take a screenshot including your OdinID that shows the results of the queries



5. Final questions and clean-up

- Which of the approaches described would have issues with token limits on LLMs?

Token limits on LLMs can cause issues with the "Stuffing" technique, which involves consolidating all data into one prompt. This might exceed the token limit, especially with substantial data.

- Which of the approaches would result in the most queries for the LLM to handle? How many LLM requests are performed from a single user query in this approach?

Multiple requests are made

- Which of the approaches requires one to search a vector database for an appropriate context that is then sent to the LLM?

The Map Reduce with embeddings

10.2g: CDN

6. Deployment

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

```
raghuram@cloudshell:~/networking101 (cl.CloseTab #jsh:raghuram)$ gcloud deployment-manager deployments create networking101 --config networking-lab.yaml
The fingerprint of the deployment is b'771gfjJSBMZLDQzNE9MR-g=='
Waiting for create [operation-1718066612628-61a928e256b5a-0dfebe44-bd62ee95]...done.
Create operation operation-1718066612628-61a928e256b5a-0dfebe44-bd62ee95 completed successfully.
NAME: asia-east1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: asial-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: e1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: eu1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: europe-west1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: networking101
TYPE: compute.v1.network
STATE: COMPLETED
ERRORS: []
INTENT:
```

```

NAME: networking101
TYPE: compute.v1.network
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: us-east5
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: us-west-s1
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: us-west-s2
TYPE: compute.v1.subnetwork
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: w1-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:

NAME: w2-vm
TYPE: compute.v1.instance
STATE: COMPLETED
ERRORS: []
INTENT:
raghuram@cloudshell:~/networking101 (cloud-nataraja-raghuram) $

```

5-subnetworks

1-network

5-instances

- 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

network

Google Cloud

cloud-nataraja-raghuram

vpc networks

Search

4

VPC network

VPC networks

IP addresses

Internal ranges

Bring your own IP

Firewall

Routes

VPC network peering

Shared VPC

Serverless VPC access

CREATE VPC NETWORK

REFRESH

NETWORKS IN CURRENT PROJECT

SUBNETS IN CURRENT PROJECT

SMTP port 25 disallowed in this project. [Learn more](#)

VPC networks

Filter Enter property name or value

Name	Subnets	MTU	Mode	IPv6 ULA range	Gateways	Firewall rules	Global dynamic routing
custom-network1	2	1460	Custom			0	Off
default	42	1460	Auto			10	Off
networking101	5	1460	Custom			0	Off

subnetwork

Google Cloud

cloud-nataraja-raghuram

vpc networks

Search

4

VPC network

VPC network details

DELETE VPC NETWORK

SHOW INFO P

VPC networks

IP addresses

Internal ranges

Bring your own IP

Firewall

Routes

VPC network peering

Shared VPC

ADD SUBNET

FLOW LOGS

Filter Enter property name or value

Name	Region	Stack Type	Primary IPv4 range	Secondary IPv4 ranges	IPv6 ranges	Reserved internal ranges	Gateway	Private Go
asia-east1	asia-east1	IPv4	10.40.0.0/16			None	10.40.0.1	Off
europa-west1	europa-west1	IPv4	10.30.0.0/16			None	10.30.0.1	Off
us-east5	us-east5	IPv4	10.20.0.0/16			None	10.20.0.1	Off
us-west-s1	us-west1	IPv4	10.10.0.0/16			None	10.10.0.1	Off
us-west-s2	us-west1	IPv4	10.11.0.0/16			None	10.11.0.1	Off

- 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.

cloud-nataraja-raghuram comput Search 4 ? ⋮

VM instances [CREATE INSTANCE](#) [IMPORT VM](#) [REFRESH](#) [LEA](#)

Filter Enter property name or value

<input type="checkbox"/>	Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Network	Connect	
<input type="checkbox"/>	✓	asia1-vm	asia-east1-b			10.40.0.2 (nic0)	104.155.230.56 (nic0)	networking101	SSH	⋮
<input type="checkbox"/>	⊛	course-vm	us-west1-b			10.138.0.10 (nic0)		default	SSH	⋮
<input type="checkbox"/>	✓	e1-vm	us-east5-a			10.20.0.2 (nic0)	34.162.44.196 (nic0)	networking101	SSH	⋮
<input type="checkbox"/>	✓	eu1-vm	europa-west1-d			10.30.0.2 (nic0)	35.187.6.97 (nic0)	networking101	SSH	⋮
<input type="checkbox"/>	✓	w1-vm	us-west1-b			10.10.0.2 (nic0)	34.19.18.19 (nic0)	networking101	SSH	⋮
<input type="checkbox"/>	✓	w2-vm	us-west1-b			10.11.0.100 (nic0)	34.19.100.145 (nic0)	networking101	SSH	⋮

Related actions [HIDE](#)

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

No it did not connect.

ERROR: (gcloud.compute.ssh) [/usr/bin/ssh] exited with return code [255].

8. Update deployment

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

[←](#) VPC network details [DELETE VPC NETWORK](#)

networking101

[OVERVIEW](#) [SUBNETS](#) [STATIC INTERNAL IP ADDRESSES](#) [FIREWALLS](#) [FIREWALL ENDPOINTS](#) [ROUTES](#) [VPC NETWORK PEERING](#) [PRIVATE SE](#)

[ADD FIREWALL RULE](#) [DELETE](#)

Filter Enter property name or value

<input type="checkbox"/>	Name	Enforcement order ↑	Type	Deployment scope	Rule priority	Targets	Source	Destination	Pre
<input type="checkbox"/>	vpc-firewall-rules	1	VPC firewall rules	Global					
<input type="checkbox"/>	networking-firewall-allow-icmp		Ingress firewall rule	Global	1000	Appl...	IPv4 range	—	icr
<input type="checkbox"/>	networking-firewall-allow-internal		Ingress firewall rule	Global	1000	Appl...	IPv4 range	—	tcp, ud, icr
<input type="checkbox"/>	networking-firewall-allow-ssh		Ingress firewall rule	Global	1000	Appl...	IPv4 range	—	tcp

- 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	49.7 ms
us-west1 europe-west1	~93 ms	134ms
us-west1 asia-east1	~114 ms	118.6 ms
us-east5 europe-west1	~76 ms	87.4ms
us-east5 asia-east1	~141 ms	166.9 ms
europe-west1 asia-east1	~110 ms	250.5 ms

16. Test groups

- Are the instances in the same availability zone or in different ones?

They are all in different zones

- List all availability zones that your servers show up in for your lab notebook.

<input type="checkbox"/>	<input checked="" type="checkbox"/>	e1-vm	us-east5-a	(nic0)	10.20.0.2	34.162.44.196	networking101	SS
<input type="checkbox"/>	<input checked="" type="checkbox"/>	eu1-vm	europa-west1-d	(nic0)	10.30.0.2	35.187.6.97	networking101	SS
<input type="checkbox"/>	<input checked="" type="checkbox"/>	europa-west1-mig-2fgw	europa-west1-c	europa-west1-mig	10.30.0.3	34.38.145.242	networking101	SS
<input type="checkbox"/>	<input checked="" type="checkbox"/>	europa-west1-mig-cltw	europa-west1-b	europa-west1-mig	10.30.0.5	35.195.51.136	networking101	SS
<input type="checkbox"/>	<input checked="" type="checkbox"/>	europa-west1-mig-w8b1	europa-west1-d	europa-west1-mig	10.30.0.4	34.38.137.163	networking101	SS



Networking 101 Lab

Client IP

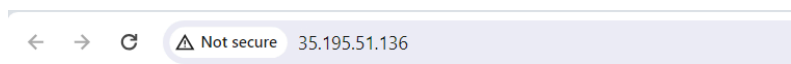
Your IP address : 71.59.145.43

Hostname

Server Hostname: europa-west1-mig-w8b1

Server Location

Region and Zone: europa-west1-d



Networking 101 Lab

Client IP

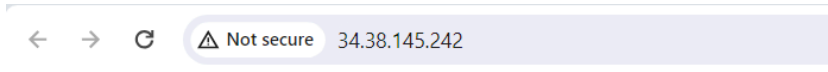
Your IP address : 71.59.145.43

Hostname

Server Hostname: europa-west1-mig-cltw

Server Location

Region and Zone: europa-west1-b



Networking 101 Lab

Client IP

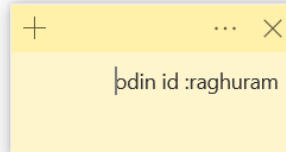
Your IP address : 71.59.145.43

Hostname

Server Hostname: europe-west1-mig-2fgw

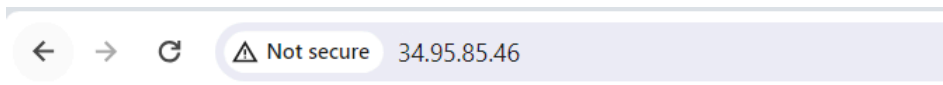
Server Location

Region and Zone: europe-west1-c



19. Test load balancer

- Show a screenshot of the page that is returned.



Networking 101 Lab

Client IP

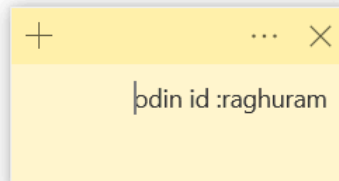
Your IP address : 35.191.42.176

Hostname

Server Hostname: us-east5-mig-83vm

Server Location

Region and Zone: us-east5-c

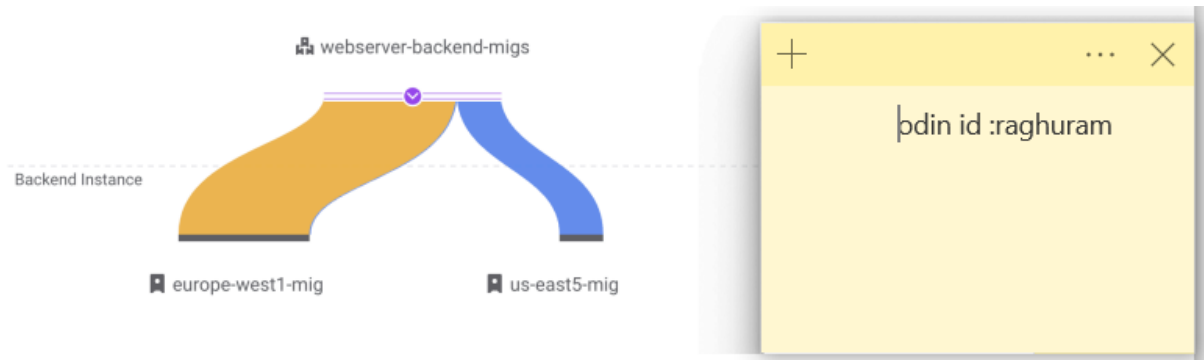


- Which availability zone does the server handling your request reside in?

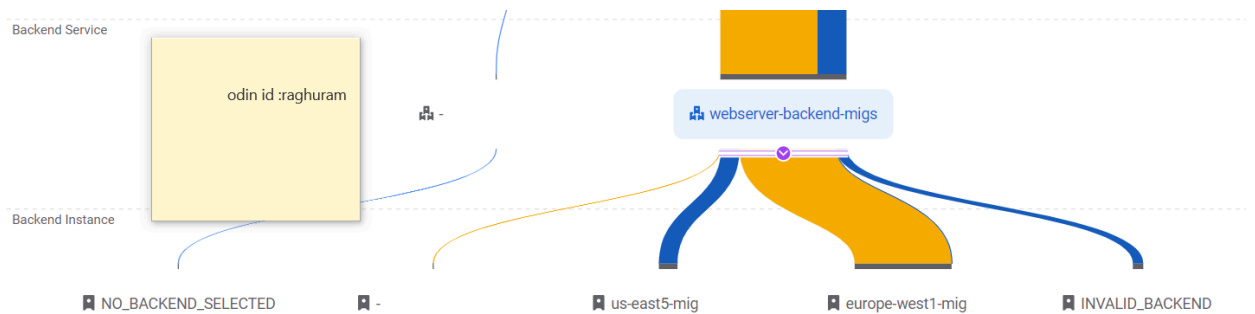
Region and Zone: us-east5-c

20. Siege! (Part 1)

- Take a screenshot of the initial traffic distribution



- Take a screenshot of the UI as additional instances are brought up and show that the traffic distribution shifts



21. Siege! (Part 2)

- Show a screenshot of the final traffic distribution.

