Pelago - Data Engineer Assignment

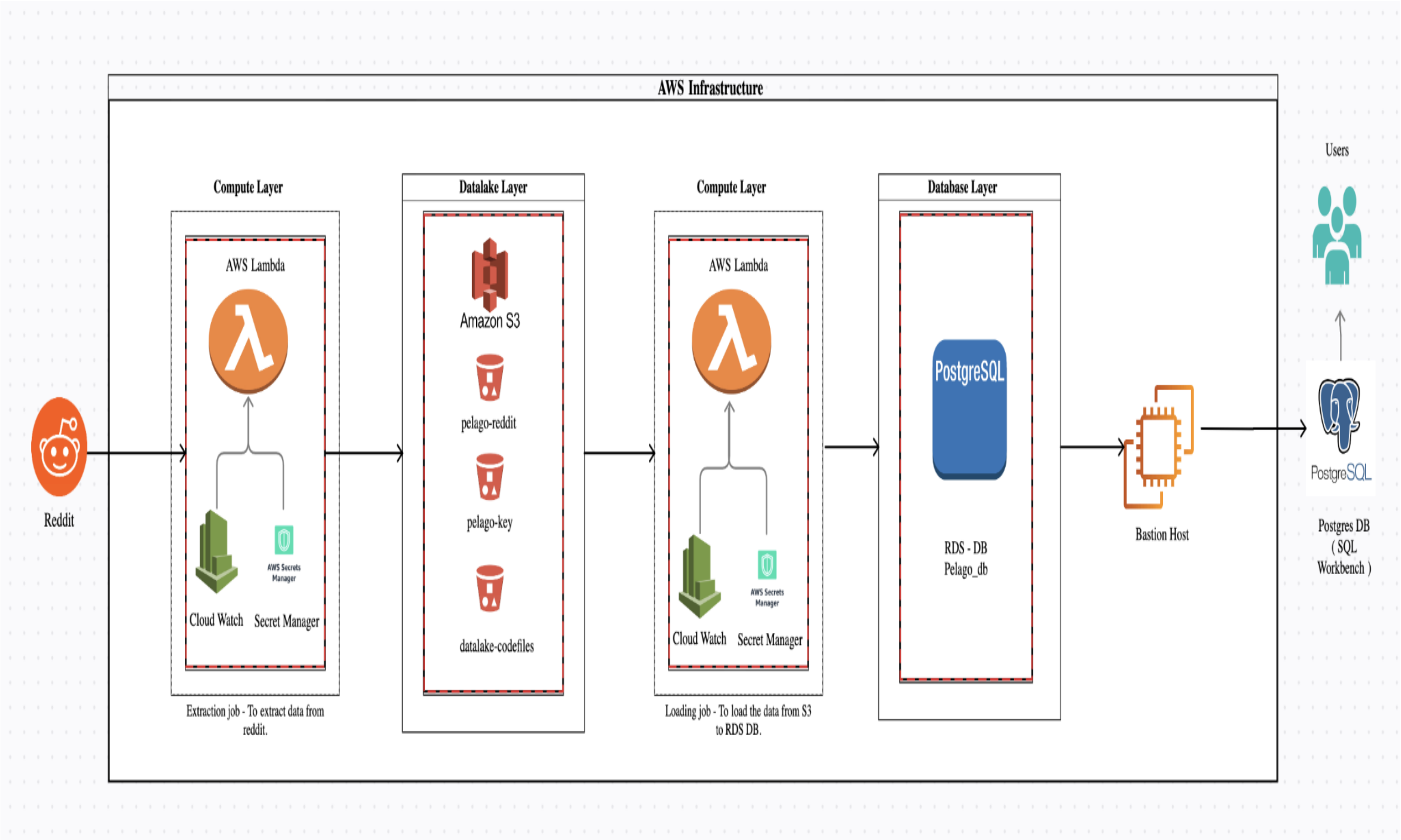
1. Assignment Description and Tasks:

Use the Reddit API to read posts from a **subreddit** and store them in a database on an hourly schedule

Task

1. Create a data schema in any database of your choice
2. Create any AWS service of your choice to read data from the API
3. Process and clean the data as required
4. Insert top 100 HOT posts into the database table(s) on an hourly schedule

2. High level architecture of the design:



2.1 Explanation on choosing the AWS resources -

a. **Lambda :**

   Two Lambda jobs (python script) are used in the project.
   Extraction Lambda job: To extract the top 100 hot posts to S3 Bucket (pelago-reddit) from reddit thru Praw API .
   Loading Lambda job: To load the data from S3 Bucket (pelago-reddit) to RDS (pelago_db) thru S3 event notification wherever the object is created in the bucket.

   Why lambda as ETL?
   Lambda is a serverless compute service which is cheaper . The above jobs (extraction and loading) run less than a minute. Hence Lambda is preferred over Glue.

b. **S3:**

   Three AWS S3 buckets are created in this project.
   pelago-reddit: This is used as a data lake to store the raw data (top 100 hot posts) extracted from the reddit.
   pelago-key: Use Secure Socket Layer (SSL) from reddit to encrypt a connection to a DB. When connecting using SSL, the client should choose to verify the certificate. If the connection parameters specify sslmode=verify-full, then the client app requires the RDS CA certificates to be referenced in the connection URL. The certificate .pem file is stored in the bucket.
   Datalake-codefiles: The code files are stored in this bucket when deploying the code via cloud formation stack.

   Why S3 as Data Lake?
   The raw data is stored in the S3 bucket (Standard storage class) before loading to the Database. S3 is a cheaper storage service wherein the data can be stored in different file formats. In addition, the historical files can be transferred to Glacier storage class (storage is very cheap) after a certain time period using lifecycle configuration.

c. **Secret Manager:**

   The credentials like (client-id and client-secret) used to access reddit praw API and Database credentials are stored in Secret Manager.

d. **Cloud Watch:**

   CloudWatch Logs : All the logs in extraction and loading lambda jobs are logged in here.
   CloudWatch Events: The extraction lambda job is scheduled hourly thru cron expression.

e. RDS:

The RDS Postgres instance is used as a Database. As the data is in structured format with the defined number of columns , a relational database is used. A fier tier for 750 hours is available , hence RDS Postgres is used as a Database.

Note : In the assignment it is specified to store the data in Database. But the best way would be to store the data in S3 , create metadata(table schema) via AWS Glue and access it via Athena for reports and other dashboards.

f. EC2:

A free tier EC2 instance Amazon Linux t2.micro instance type is used as Bastion host. The RDS DB instance in a private subnet is connected via Bastion host.

It is not secure to expose the database to the public. Hence the DB instance is placed in a private subnet and accessed via a bastion host in a public subnet.

g. Cloudformation template:

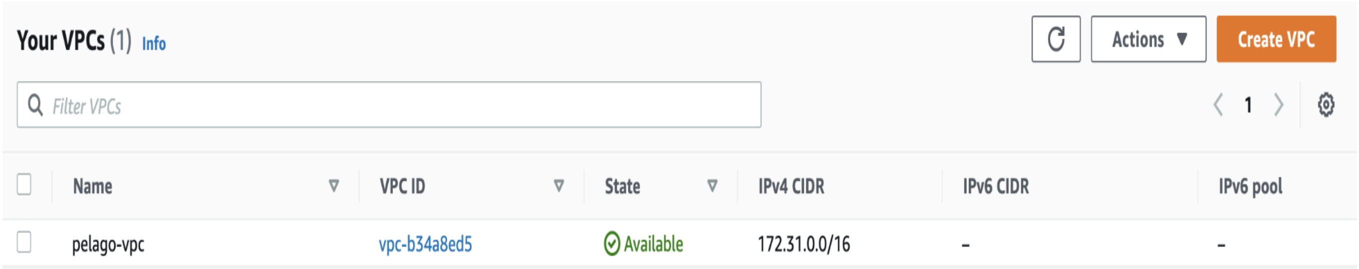
The Lambda and other associated AWS resources are created and managed using cloud formation stack.

- A CloudFormation template is created using the YAML format.
- The code files are saved in S3 bucket (datalake-codefiles).
- The bash script is used to deploy the cloudFormation stack that physically creates the stack resources.

The above way of deployment is used in this project . But the best option would be to use Codepipeline.

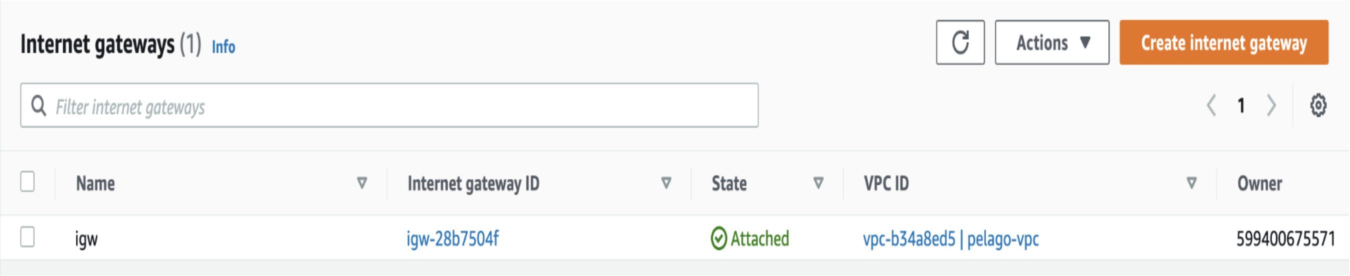
3. Network configurations :

1. Create VPC (default vpc is used) .



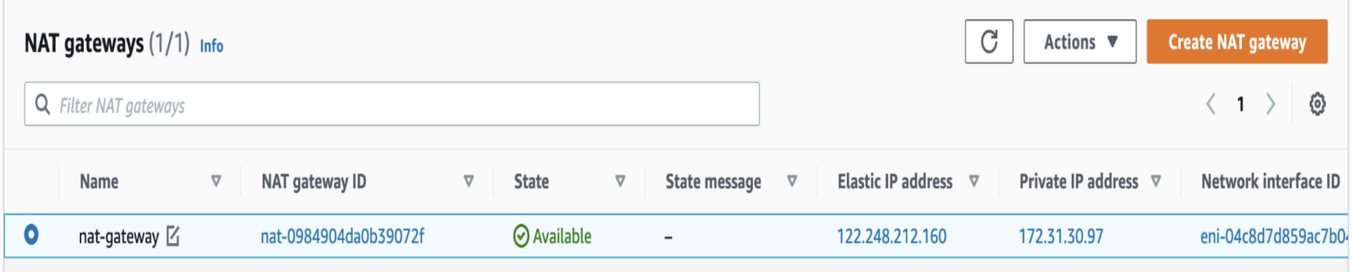
□	Name	VPC ID	State	IPv4 CIDR	IPv6 CIDR	IPv6 pool
□	pelago-vpc	vpc-b34a8ed5	✓ Available	172.31.0.0/16	-	-

2. Create Internet Gateway .



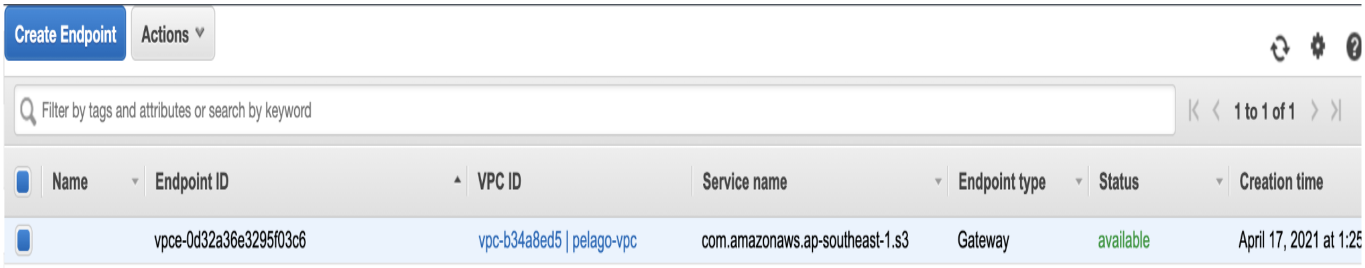
□	Name	Internet gateway ID	State	VPC ID	Owner
□	igw	igw-28b7504f	✓ Attached	vpc-b34a8ed5 | pelago-vpc	599400675571

3. Create NAT Gateway .



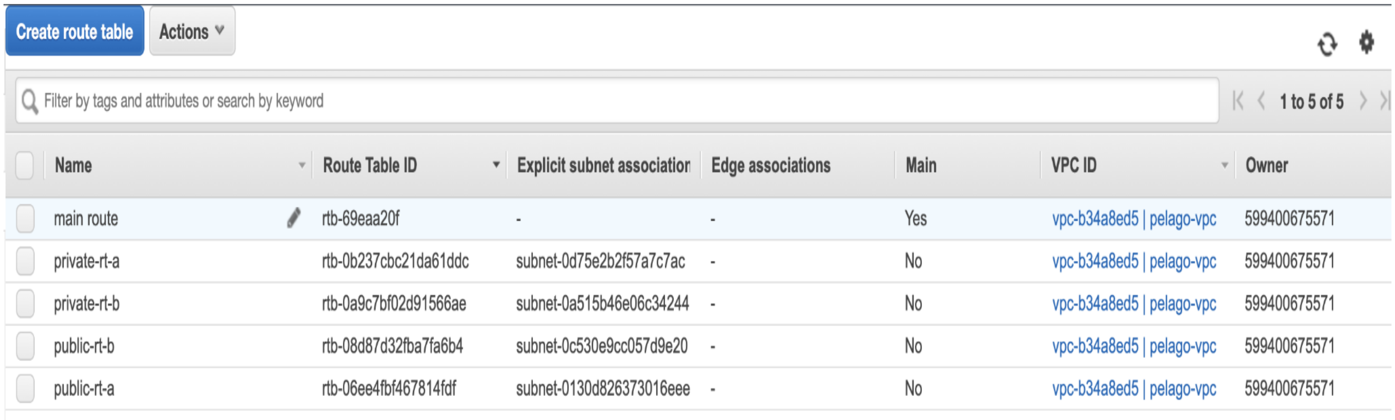
	Name	NAT gateway ID	State	State message	Elastic IP address	Private IP address	Network interface ID
●	nat-gateway	nat-0984904da0b39072f	✓ Available	–	122.248.212.160	172.31.30.97	eni-04c8d7d859ac7b0

4. Create Endpoint (Gateway) for S3.
   To communicate with other AWS resources within the vpc network .



☑	Name	Endpoint ID	VPC ID	Service name	Endpoint type	Status	Creation time
☑		vpce-0d32a36e3295f03c6	vpc-b34a8ed5 | pelago-vpc	com.amazonaws.ap-southeast-1.s3	Gateway	available	April 17, 2021 at 1:25

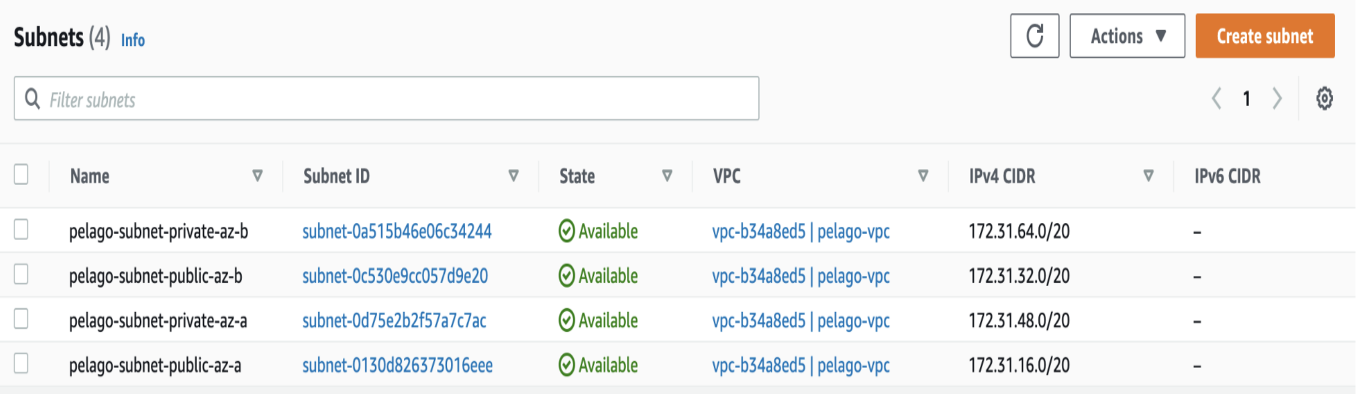
5. Create private and public route tables.



□	Name	Route Table ID	Explicit subnet association	Edge associations	Main	VPC ID	Owner
□	main route	rtb-69eaa20f	-	-	Yes	vpc-b34a8ed5 | pelago-vpc	599400675571
□	private-rt-a	rtb-0b237cbc21da61ddc	subnet-0d75e2b2f57a7c7ac	-	No	vpc-b34a8ed5 | pelago-vpc	599400675571
□	private-rt-b	rtb-0a9c7bf02d91566ae	subnet-0a515b46e06c34244	-	No	vpc-b34a8ed5 | pelago-vpc	599400675571
□	public-rt-b	rtb-08d87d32fba7fa6b4	subnet-0c530e9cc057d9e20	-	No	vpc-b34a8ed5 | pelago-vpc	599400675571
□	public-rt-a	rtb-06ee4fbf467814fdf	subnet-0130d826373016eee	-	No	vpc-b34a8ed5 | pelago-vpc	599400675571

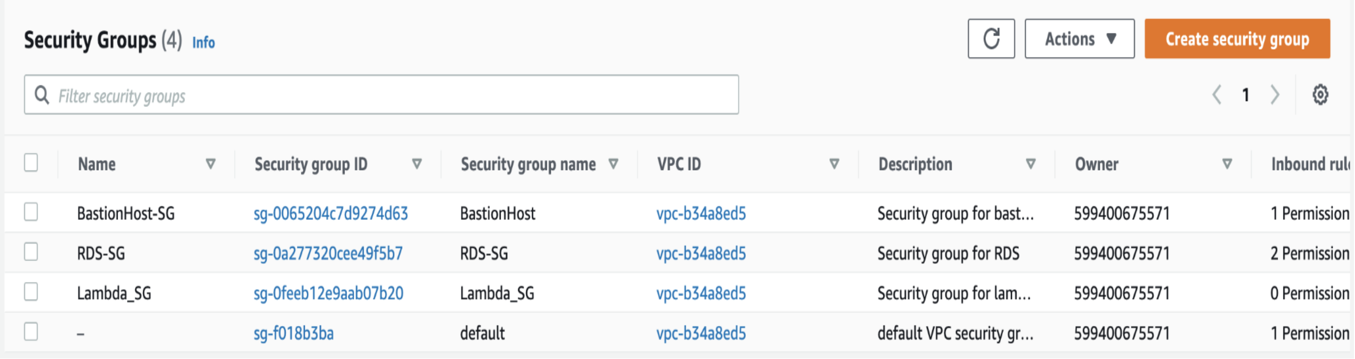
6. Add Internet Gateway to the public route table and NAT Gateway to the private route table.

7. Create Private and Public Subnets in availability zone A and B , attach to VPC.



□	Name	Subnet ID	State	VPC	IPv4 CIDR	IPv6 CIDR
□	pelago-subnet-private-az-b	subnet-0a515b46e06c34244	✓ Available	vpc-b34a8ed5 | pelago-vpc	172.31.64.0/20	–
□	pelago-subnet-public-az-b	subnet-0c530e9cc057d9e20	✓ Available	vpc-b34a8ed5 | pelago-vpc	172.31.32.0/20	–
□	pelago-subnet-private-az-a	subnet-0d75e2b2f57a7c7ac	✓ Available	vpc-b34a8ed5 | pelago-vpc	172.31.48.0/20	–
□	pelago-subnet-public-az-a	subnet-0130d826373016eee	✓ Available	vpc-b34a8ed5 | pelago-vpc	172.31.16.0/20	–

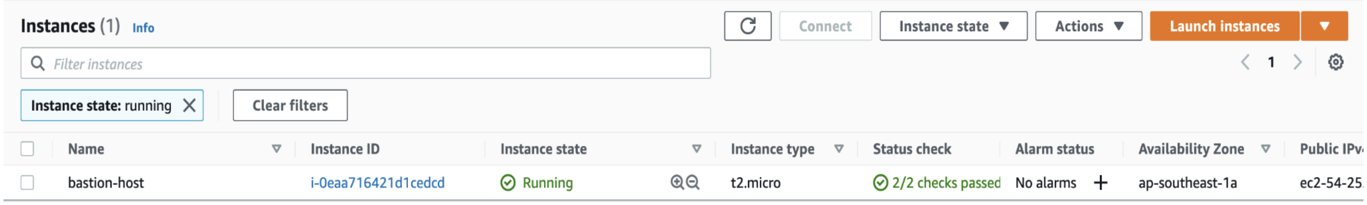
8. Attach the private subnet to private route tables and public subnet to public route table.
9. Create a security group for Lambda,EC2 and RDS.



□	Name	Security group ID	Security group name	VPC ID	Description	Owner	Inbound rul
□	BastionHost-SG	sg-0065204c7d9274d63	BastionHost	vpc-b34a8ed5	Security group for bast...	599400675571	1 Permission
□	RDS-SG	sg-0a277320cee49f5b7	RDS-SG	vpc-b34a8ed5	Security group for RDS	599400675571	2 Permission
□	Lambda_SG	sg-0feeb12e9aab07b20	Lambda_SG	vpc-b34a8ed5	Security group for lam...	599400675571	0 Permission
□	–	sg-f018b3ba	default	vpc-b34a8ed5	default VPC security gr...	599400675571	1 Permission

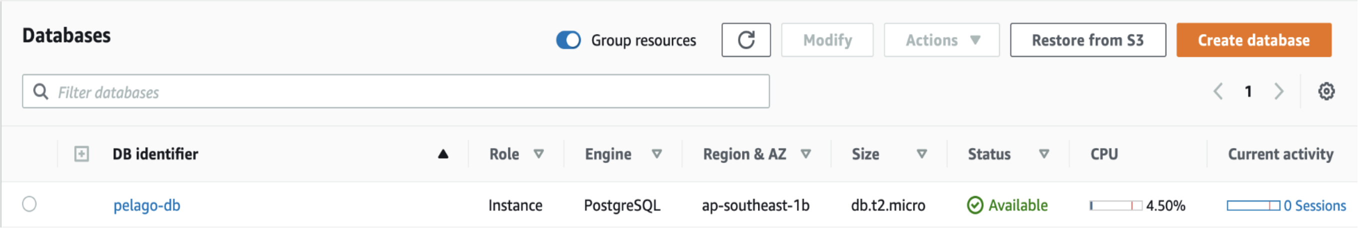
4. Bastion host :

1. Create a EC2 Instance (Bastion host) and add a security group .



□	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv
□	bastion-host	i-0eaa716421d1cedcd	✓ Running	t2.micro	✓ 2/2 checks passed	No alarms +	ap-southeast-1a	ec2-54-25

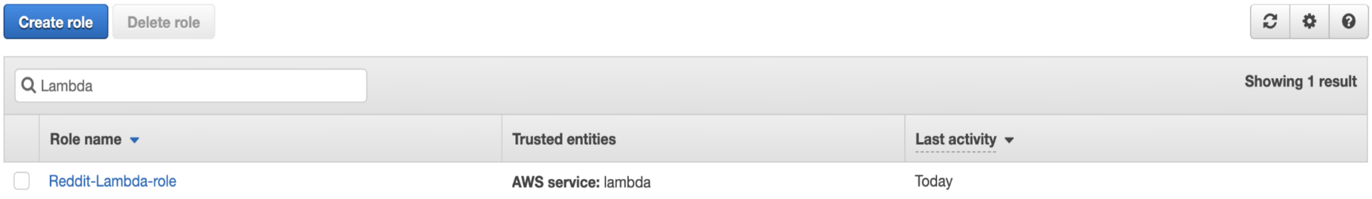
2. Create RDS Postgres Database Instance.



	DB identifier ▲	Role	Engine	Region & AZ	Size	Status	CPU	Current activity
○	pelago-db	Instance	PostgreSQL	ap-southeast-1b	db.t2.micro	✓ Available	4.50%	0 Sessions

5. IAM Configuration :

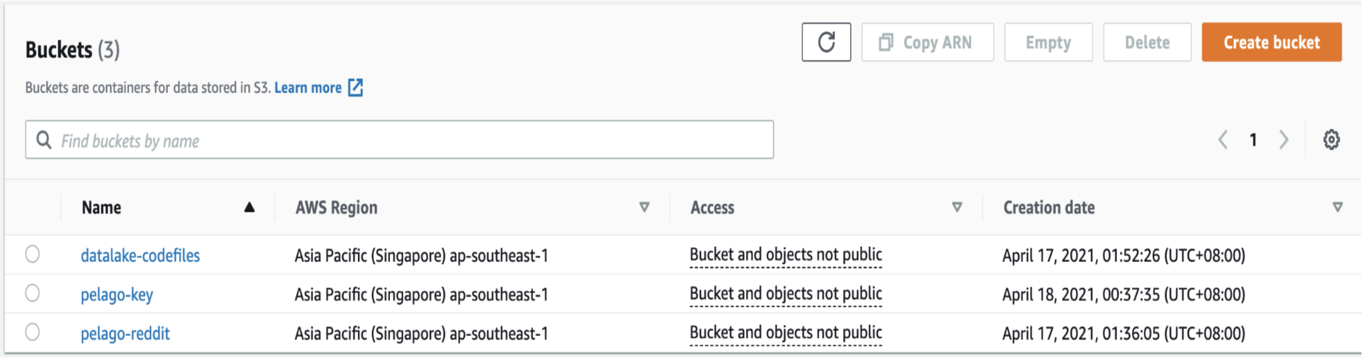
1. Create an IAM role for Lambda and attach policy.



Role name	Trusted entities	Last activity
Reddit-Lambda-role	**AWS service:** lambda	Today

6. S3 Bucket:

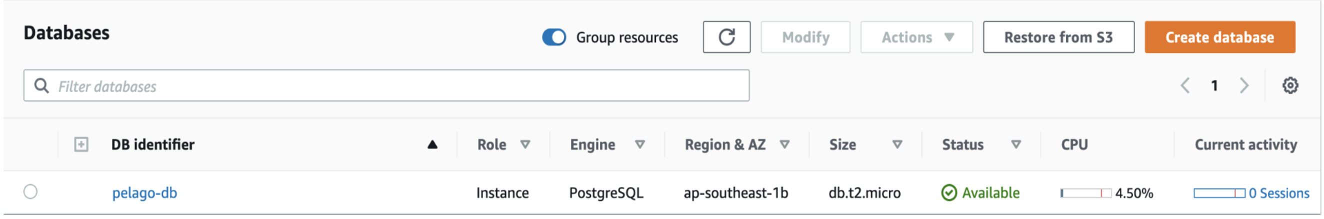
1. Create S3 bucket pelago-reddit (to store the extracted reddit post raw data) , pelago-key (to store the RDS CA certificate) and datalake-codefiles(to store the cloudFormation stack code files).
2. Create bucket policy to manage the access at bucket level.(Refer github for bucket policy).



Name	AWS Region	Access	Creation date
datalake-codefiles	Asia Pacific (Singapore) ap-southeast-1	Bucket and objects not public	April 17, 2021, 01:52:26 (UTC+08:00)
pelago-key	Asia Pacific (Singapore) ap-southeast-1	Bucket and objects not public	April 18, 2021, 00:37:35 (UTC+08:00)
pelago-reddit	Asia Pacific (Singapore) ap-southeast-1	Bucket and objects not public	April 17, 2021, 01:36:05 (UTC+08:00)

7. RDS:

1. Create the RDS Postgres DB Instance.
2. Attach the RDS Security group while creating the instance.

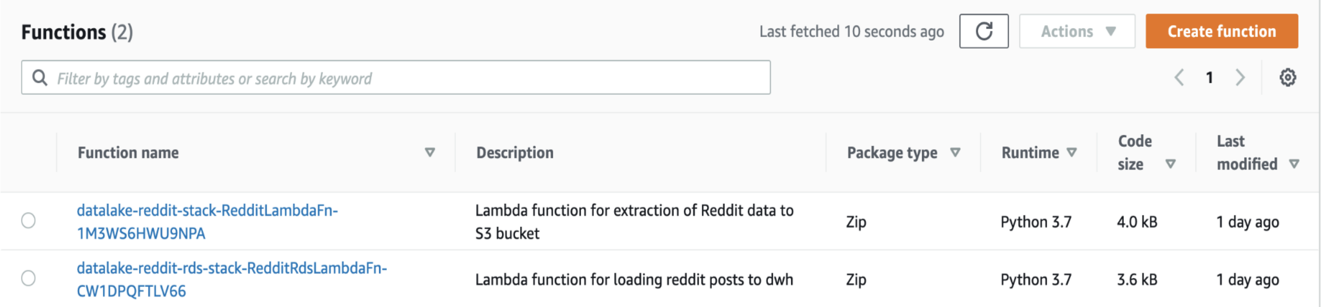


DB identifier	Role	Engine	Region & AZ	Size	Status	CPU	Current activity
pelago-db	Instance	PostgreSQL	ap-southeast-1b	db.t2.micro	Available	4.50%	0 Sessions

8. Lambda:

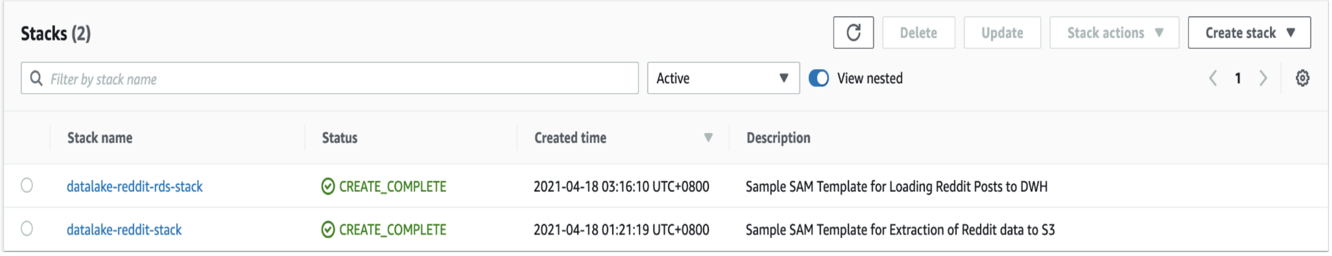
1. Extraction lambda job.
2. Loading lambda job.

Prerequisite : Create an IAM role for lambda to access S3, RDS,CloudWatch and SecretManager.


Functions (2)
Last fetched 10 seconds ago
Actions ▼
Create function
Filter by tags and attributes or search by keyword
1

Function name	Description	Package type	Runtime	Code size	Last modified
datalake-reddit-stack-RedditLambdaFn-1M3WS6HWU9NPA	Lambda function for extraction of Reddit data to S3 bucket	Zip	Python 3.7	4.0 kB	1 day ago
datalake-reddit-rds-stack-RedditRdsLambdaFn-CW1DPQFTLV66	Lambda function for loading reddit posts to dwh	Zip	Python 3.7	3.6 kB	1 day ago

9. CloudFormation stack:



Stack name	Status	Created time	Description
datalake-reddit-rds-stack	✓ CREATE_COMPLETE	2021-04-18 03:16:10 UTC+0800	Sample SAM Template for Loading Reddit Posts to DWH
datalake-reddit-stack	✓ CREATE_COMPLETE	2021-04-18 01:21:19 UTC+0800	Sample SAM Template for Extraction of Reddit data to S3

References:

What is uploaded to github-

1. ***Lambda job (python scripts)***
2. ***Cloud formation yaml file***
3. ***Bash script to automate the creation of resources (lambda) physically thru cloud formation stack.***
4. ***Sql files***
5. ***Bucket policy***