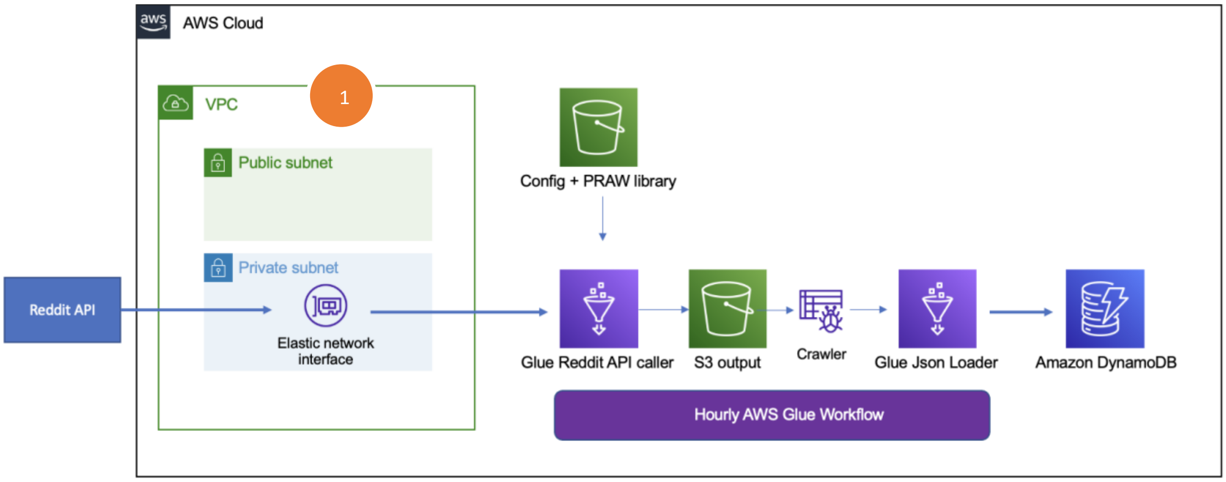
JEROME GADITANO
Interview Coding Challenge with Pelago

Architecture Diagram



Flow of the Architecture:

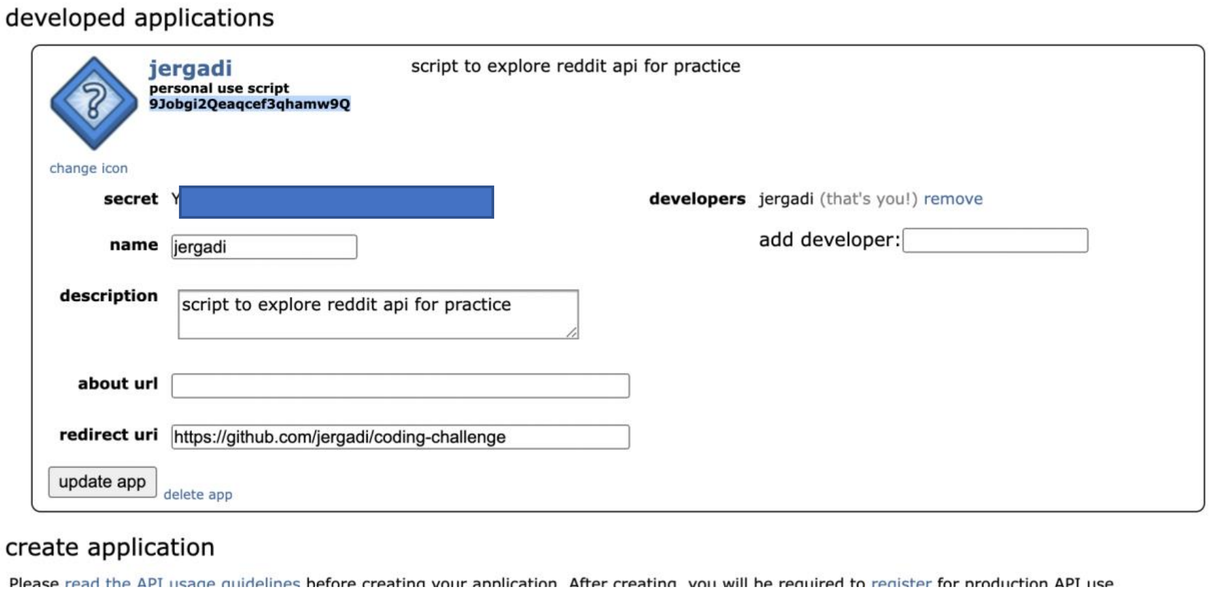
1. We access the Reddit API via AWS VPC that has ENI in the Private subnet, so that any AWS resource that use this VPC would have access outside the public internet, but the public won't have access over our resources.
2. We created an AWS Glue job, I named it as 'Reddit API Caller' that connects over the VPC and accesses the Reddit API.
3. The problem is that PRAW library is not native in AWS Glue. So then this library is manually uploaded in an S3 bucket, along with a configuration file that contains all our access identities. The library and the configuration file are then used by the Reddit API Caller on runtime. This is designed this way so that all IDs and Accesses are not exposed on the scripts.
4. Reddit API Caller then outputs the data into S3, in JSON format. This format is chosen because the fields we use from the response are semi-structured, so it will be problematic when handled in a traditional row-column (CSV) format.
5. The JSON output is then *crawled* by an AWS Glue crawler which results to a Hive table that is accessible by any AWS resource.
6. This hive table is then read by another AWS Glue job that loads the data into Dynamodb.
7. The whole workflow is managed by AWS Glue Workflow. Each job is triggered whenever the task before it succeeded. This is scheduled to run every hour.

Factors in Designing the Architecture

- Serverless
    - I went for a serverless design. Putting this in EC2 server and bootstrapping it is one way to do it, but it would cap the performance acc to the size of the server. And once the server has an issue, the whole process stops. So designing it serverless-ly means microservice and scalability.
- Scalability
    - AWS Glue in PySpark is scalable. Taking advantage of the distributed parallel processing power and its elasticity is beneficial to this project i.e. we can size up or down the DPU of each job anytime we want.
- AWS Cloud Cost Optimization
    - Only paying for when you would run the job, and only upscaling when you need to, and archiving old data in S3, are all ways to make the project cost efficient.
- Semi-structured Data Compatibility
    - I chose JSON and Dynamodb because the response we get are semi-structured. We need the flexibility and ease of use in the future downstream processes so I opted to use Dynamodb.

GITHUB REPO: https://github.com/jergadi/coding-challenge

REDDIT dev account:



developed applications

jergadi
personal use script
9Jobgi2Qeaqcef3qhamw9Q

script to explore reddit api for practice

secret [redacted]

developers jergadi (that's you!) remove

name jergadi

add developer:

description script to explore reddit api for practice

about url

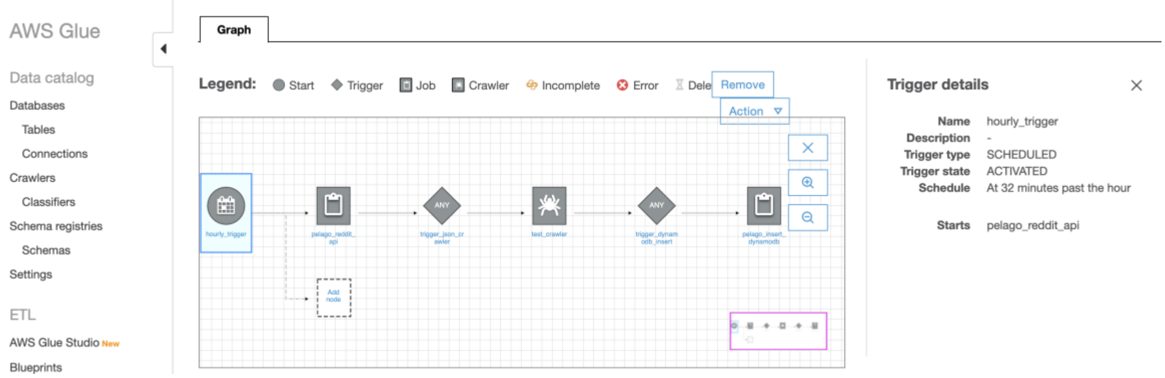
redirect uri https://github.com/jergadi/coding-challenge

create application

Please read the API usage guidelines before creating your application. After creating, you will be required to register for production API use.

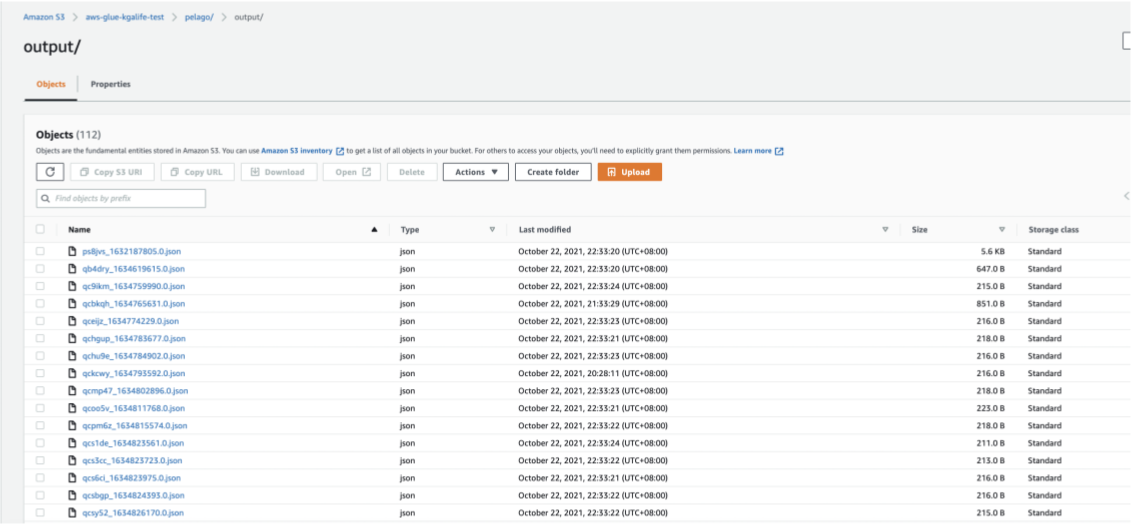
AWS RESOURCES:

1. AWS Glue Workflow – hourly scheduled.



Name	hourly_trigger
Description	-
Trigger type	SCHEDULED
Trigger state	ACTIVATED
Schedule	At 32 minutes past the hour
Starts	pelago_reddit_api

2. S3 bucket containing the JSON output records



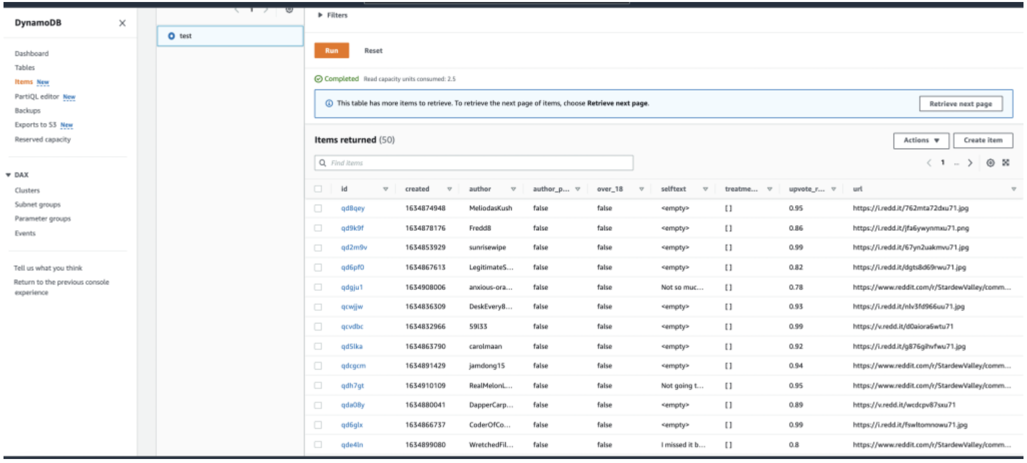
output/

Objects (112)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
ps8jvs_1632187805.0.json	json	October 22, 2021, 22:33:20 (UTC+08:00)	5.6 KB	Standard
qb4dry_1634619615.0.json	json	October 22, 2021, 22:33:20 (UTC+08:00)	647.0 B	Standard
qc9ikm_1634759990.0.json	json	October 22, 2021, 22:33:24 (UTC+08:00)	215.0 B	Standard
qcbkqh_1634765631.0.json	json	October 22, 2021, 21:33:29 (UTC+08:00)	851.0 B	Standard
qceijz_1634774229.0.json	json	October 22, 2021, 22:33:23 (UTC+08:00)	216.0 B	Standard
qchgup_1634783677.0.json	json	October 22, 2021, 22:33:21 (UTC+08:00)	218.0 B	Standard
qchu9e_1634784902.0.json	json	October 22, 2021, 22:33:23 (UTC+08:00)	216.0 B	Standard
qckcwy_1634793592.0.json	json	October 22, 2021, 20:28:11 (UTC+08:00)	216.0 B	Standard
qcmp47_1634802896.0.json	json	October 22, 2021, 22:33:23 (UTC+08:00)	218.0 B	Standard
qcoo5v_1634811768.0.json	json	October 22, 2021, 22:33:21 (UTC+08:00)	223.0 B	Standard
qcpm6z_1634815574.0.json	json	October 22, 2021, 22:33:22 (UTC+08:00)	218.0 B	Standard
qcs1de_1634823561.0.json	json	October 22, 2021, 22:33:24 (UTC+08:00)	211.0 B	Standard
qcs3cc_1634823723.0.json	json	October 22, 2021, 22:33:22 (UTC+08:00)	213.0 B	Standard
qcs6ci_1634823975.0.json	json	October 22, 2021, 22:33:21 (UTC+08:00)	216.0 B	Standard
qcsbgp_1634824393.0.json	json	October 22, 2021, 22:33:22 (UTC+08:00)	216.0 B	Standard
qcsy52_1634826170.0.json	json	October 22, 2021, 22:33:22 (UTC+08:00)	215.0 B	Standard

3. Dynamodb containing the inserted Records



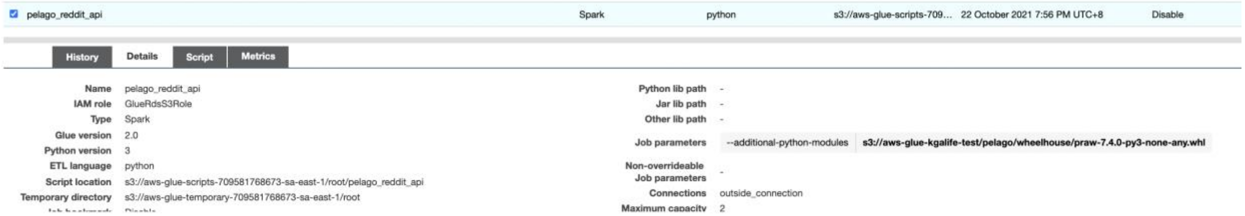
Completed Read capacity units consumed: 2.5

This table has more items to retrieve. To retrieve the next page of items, choose **Retrieve next page**.

Items returned (50)

id	created	author	author_p...	over_18	selftext	treatme...	upvote_r...	url
qd8qey	1634874948	MeliodasKush	false	false	<empty>	[]	0.95	https://i.redd.it/762mta72dxu71.jpg
qd9k9f	1634878176	Fredd8	false	false	<empty>	[]	0.86	https://i.redd.it/jfa6ywynmxu71.png
qd2m9v	1634853929	sunrisewipe	false	false	<empty>	[]	0.99	https://i.redd.it/67yn2uakmvu71.jpg
qd6pf0	1634867613	LegitimateS...	false	false	<empty>	[]	0.82	https://i.redd.it/dgts8d69rwu71.jpg
qdgju1	1634908006	anxious-ora...	false	false	Not so muc...	[]	0.78	https://www.reddit.com/r/StardewValley/comm...
qcwjjw	1634836309	DeskEvery8...	false	false	<empty>	[]	0.93	https://i.redd.it/nlv3fd966uu71.jpg
qcvdbc	1634832966	59l33	false	false	<empty>	[]	0.99	https://v.redd.it/d0aiora6wtu71
qd5lka	1634863790	carolmaan	false	false	<empty>	[]	0.92	https://i.redd.it/g876gihvfwu71.jpg
qdcgcm	1634891429	jamdong15	false	false	<empty>	[]	0.94	https://www.reddit.com/r/StardewValley/comm...
qdh7gt	1634910109	RealMelonL...	false	false	Not going t...	[]	0.95	https://www.reddit.com/r/StardewValley/comm...
qda08y	1634880041	DapperCarp...	false	false	<empty>	[]	0.89	https://v.redd.it/wcdcpv87sxu71
qd6glx	1634866737	CoderOfCo...	false	false	<empty>	[]	0.99	https://i.redd.it/fswltomnowu71.jpg
qde4in	1634899080	WretchedFil...	false	false	I missed it b...	[]	0.8	https://www.reddit.com/r/StardewValley/comm...

4. AWS Glue job Reddit API Caller settings

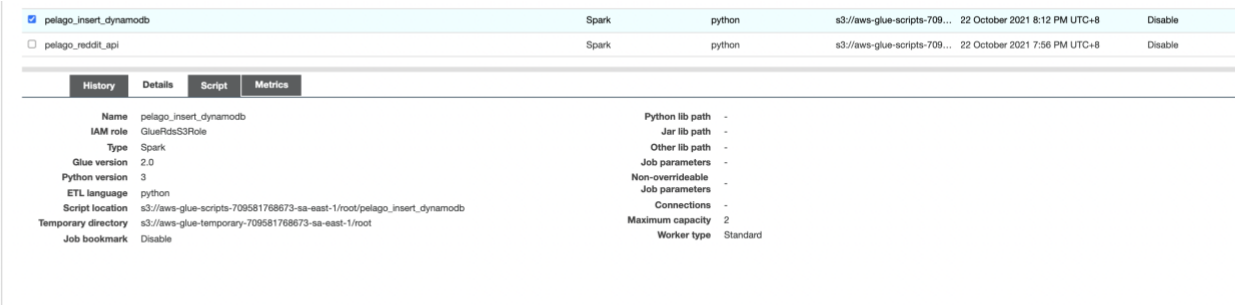


☑	pelago_reddit_api	Spark	python	s3://aws-glue-scripts-709…	22 October 2021 7:56 PM UTC+8	Disable

Name pelago_reddit_api
IAM role GlueRdsS3Role
Type Spark
Glue version 2.0
Python version 3
ETL language python
Script location s3://aws-glue-scripts-709581768673-sa-east-1/root/pelago_reddit_api
Temporary directory s3://aws-glue-temporary-709581768673-sa-east-1/root
Job bookmark Disable

Python lib path -
Jar lib path -
Other lib path -
Job parameters --additional-python-modules **s3://aws-glue-kgalife-test/pelago/wheelhouse/praw-7.4.0-py3-none-any.whl**
Non-overrideable Job parameters -
Connections outside_connection
Maximum capacity 2

5. Aws Glue job JSON loader settings

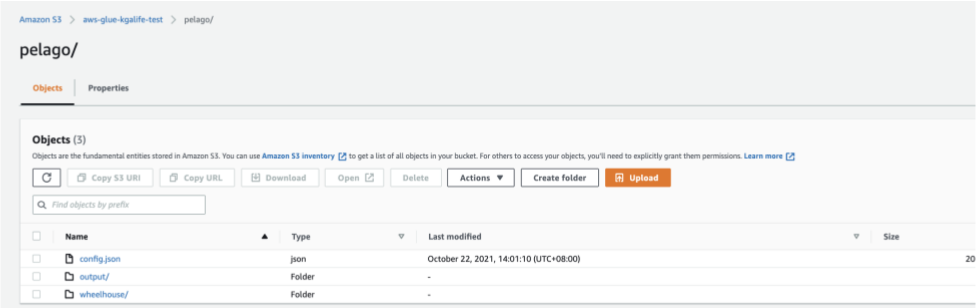


☑	pelago_insert_dynamodb	Spark	python	s3://aws-glue-scripts-709…	22 October 2021 8:12 PM UTC+8	Disable
□	pelago_reddit_api	Spark	python	s3://aws-glue-scripts-709…	22 October 2021 7:56 PM UTC+8	Disable

Name pelago_insert_dynamodb
IAM role GlueRdsS3Role
Type Spark
Glue version 2.0
Python version 3
ETL language python
Script location s3://aws-glue-scripts-709581768673-sa-east-1/root/pelago_insert_dynamodb
Temporary directory s3://aws-glue-temporary-709581768673-sa-east-1/root
Job bookmark Disable

Python lib path -
Jar lib path -
Other lib path -
Job parameters -
Non-overrideable Job parameters -
Connections -
Maximum capacity 2
Worker type Standard

6. S3 Bucket containg the config file and PRAW library

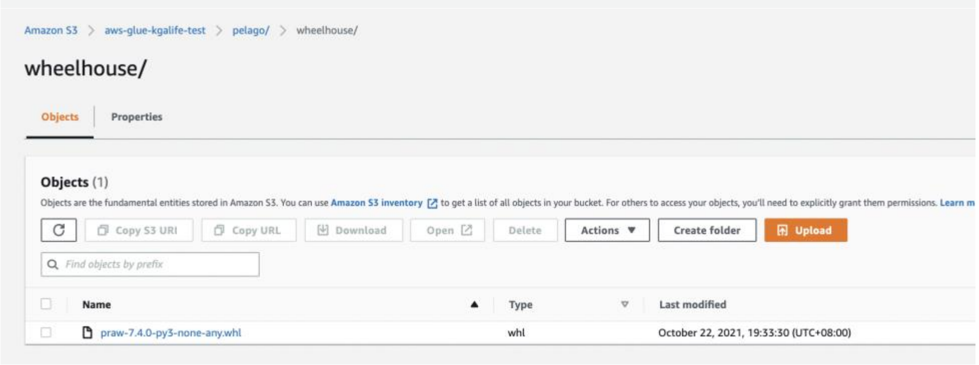


pelago/

Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size
□	config.json	json	October 22, 2021, 14:01:10 (UTC+08:00)	20
□	output/	Folder	-	
□	wheelhouse/	Folder	-	



wheelhouse/

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn m

□	Name ▲	Type	Last modified
□	praw-7.4.0-py3-none-any.whl	whl	October 22, 2021, 19:33:30 (UTC+08:00)