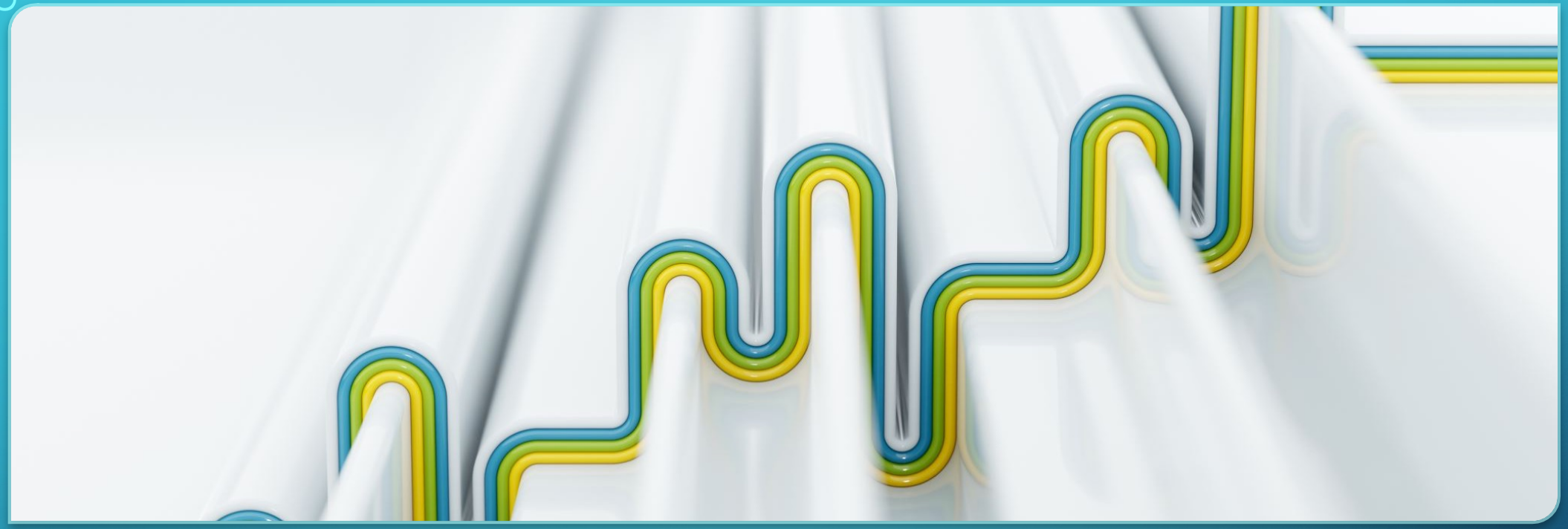
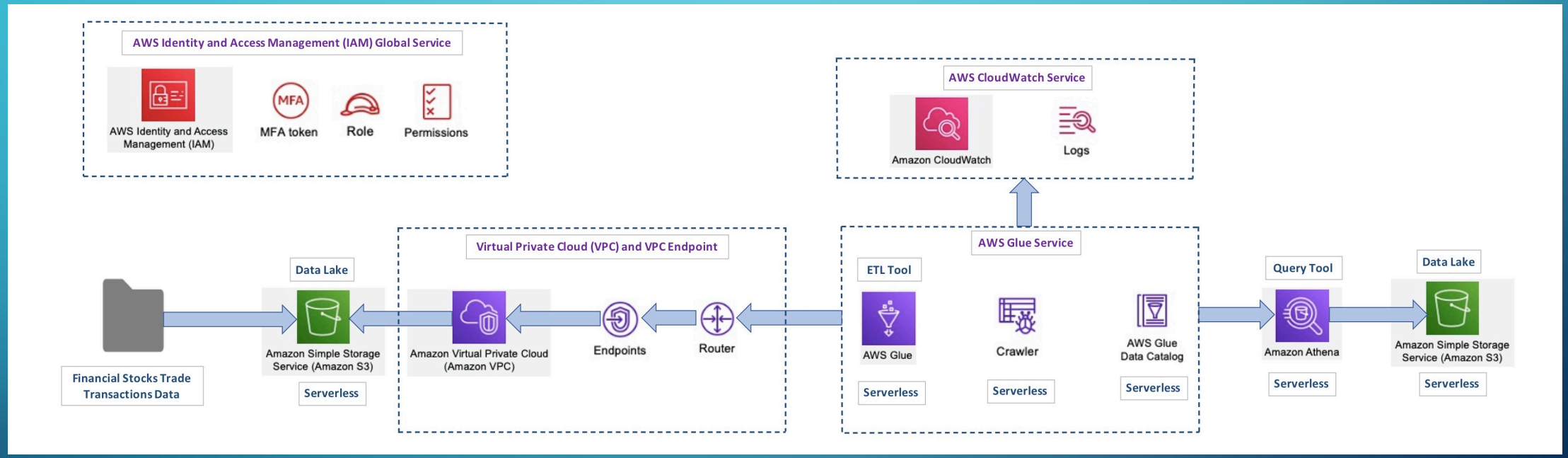


BUILDING A FINANCIAL DATA PIPELINE USING PYTHON AND AWS

FINANCIAL STOCKS TRADE TRANSACTIONS DATA

PROJECT ARCHITECTURE



TECHNOLOGY STACK

Programming/Scripting Language:

- Python

Query Language:

- SQL

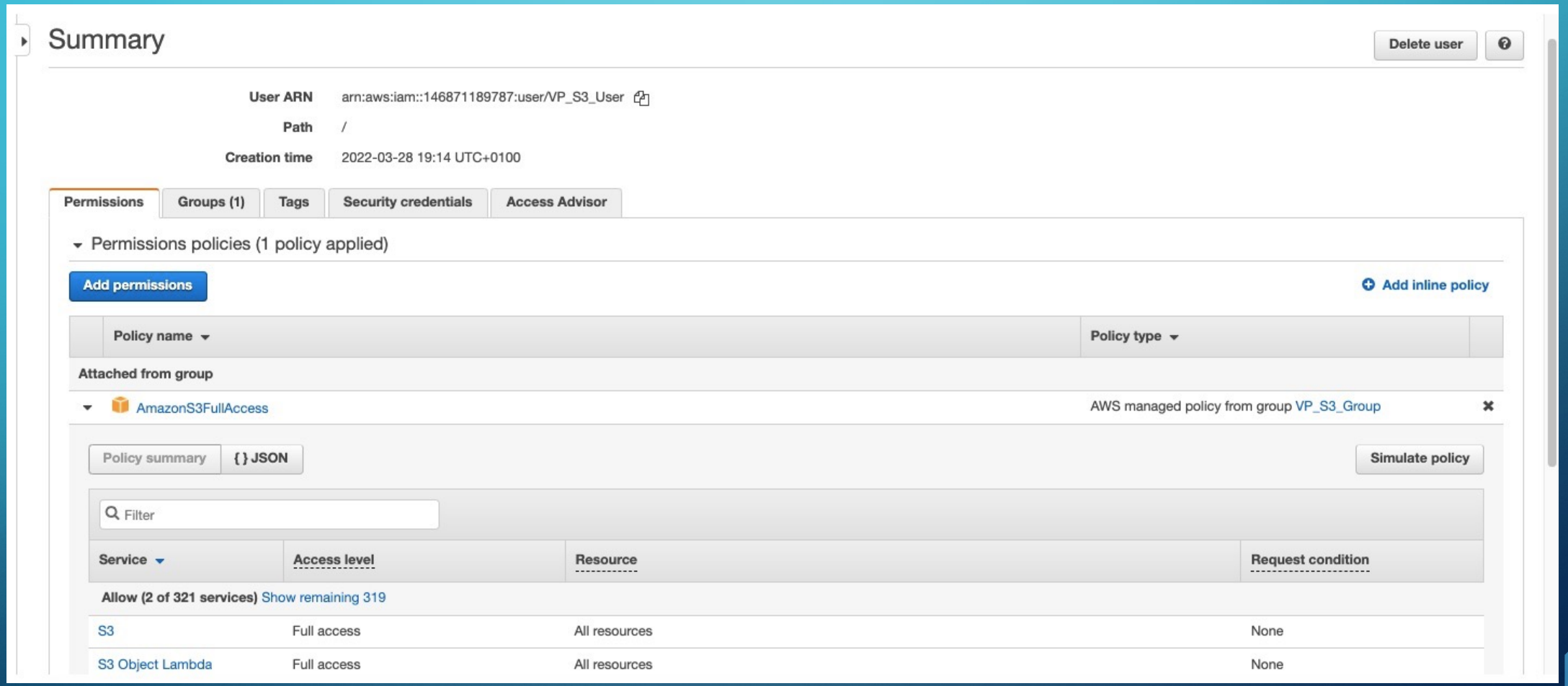
Command Line Interface (CLI):

- AWS CLI

AWS Services:

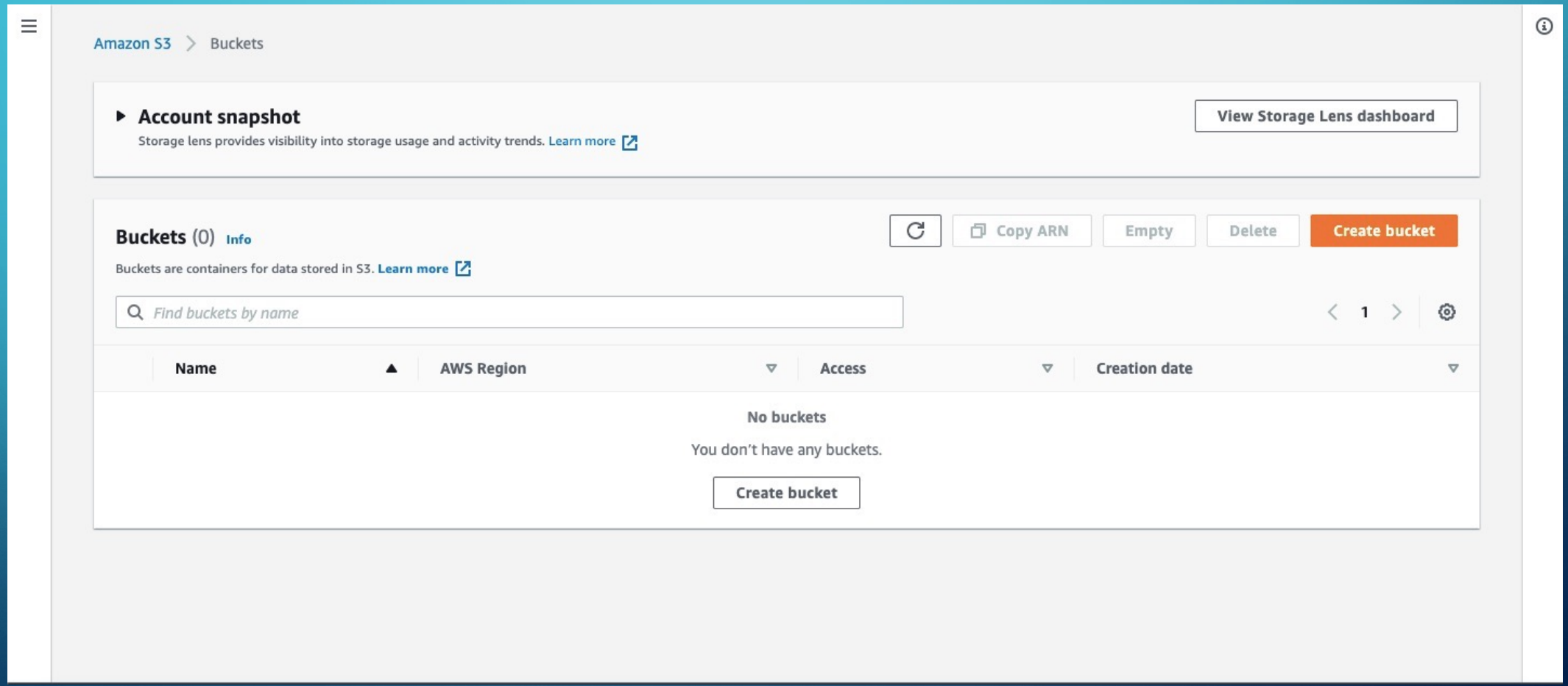
- Identity and Access Management (IAM)
- Virtual Private Cloud (VPC) and VPC Endpoint
- Single Storage Service (S3)
- Glue (Glue Crawler, Glue Catalog, Glue Database & Tables)
- CloudWatch
- Athena

1. CREATION OF AWS IAM USER WITH FULL ACCESS TO AWS S3 BUCKET



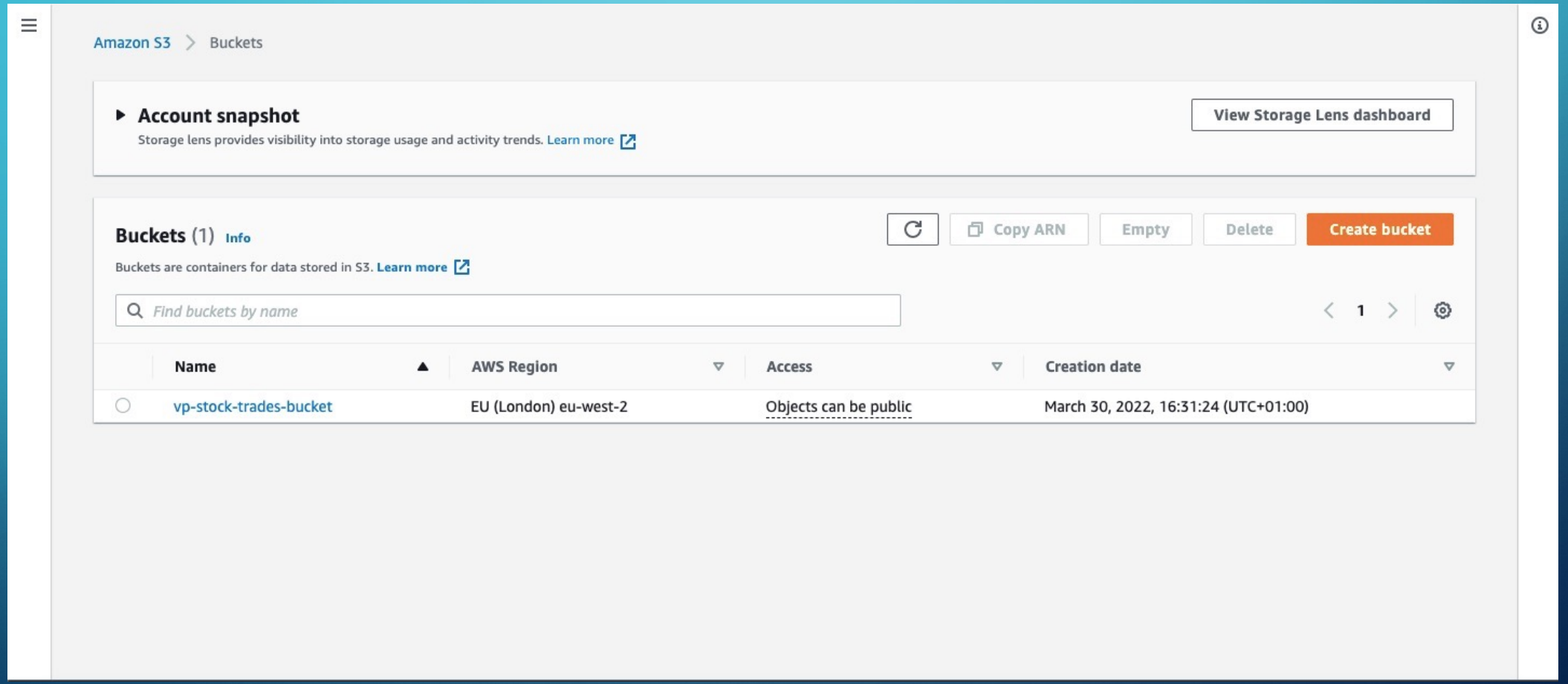
2. CREATION OF AWS S3 BUCKET *(CONTINUED)*

AWS S3 Bucket before executing the python scripts to create a bucket

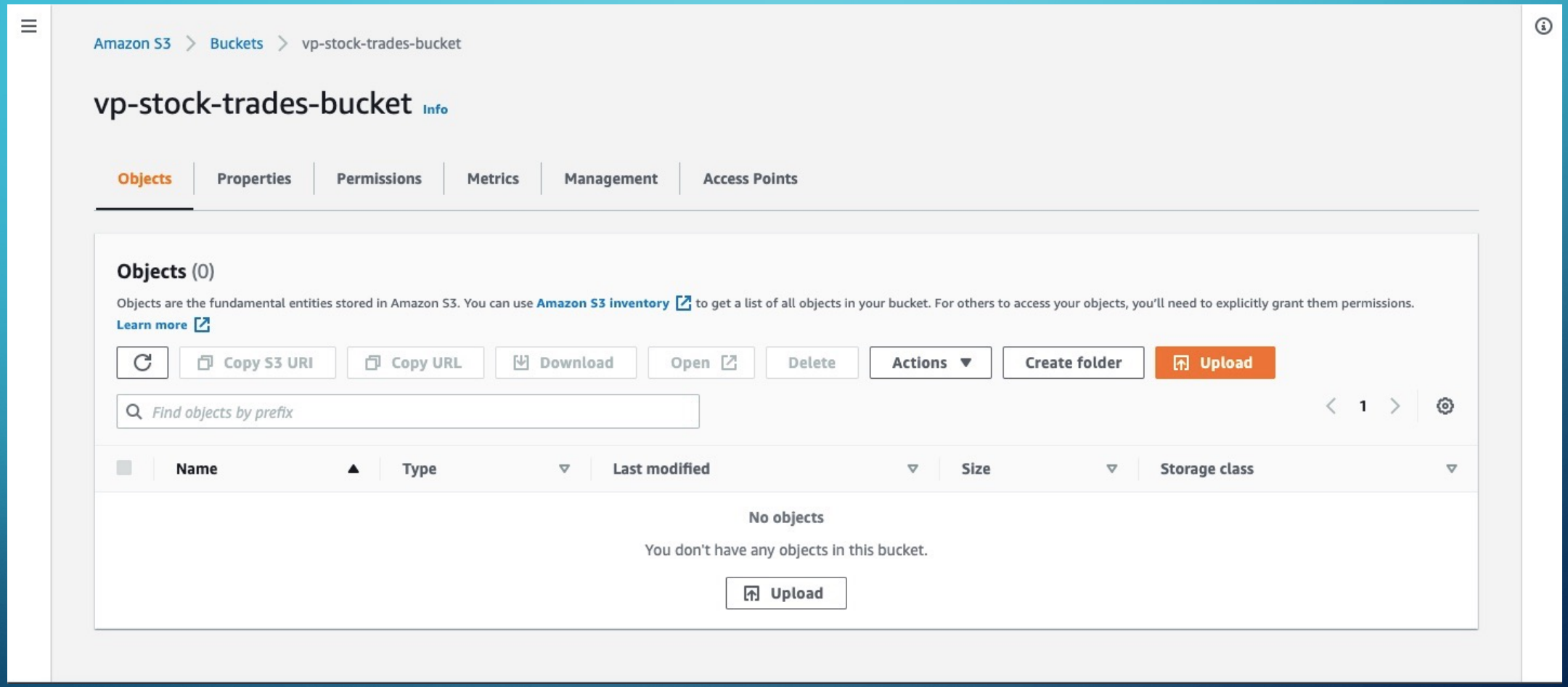


2. CREATION OF AWS S3 BUCKET *(CONTINUED)*

AWS S3 Bucket after executing the python scripts to create a bucket



2. CREATION OF AWS S3 BUCKET



3. INGESTING SOURCE DATA FILES IN TO AWS S3 BUCKET *(CONTINUED)*

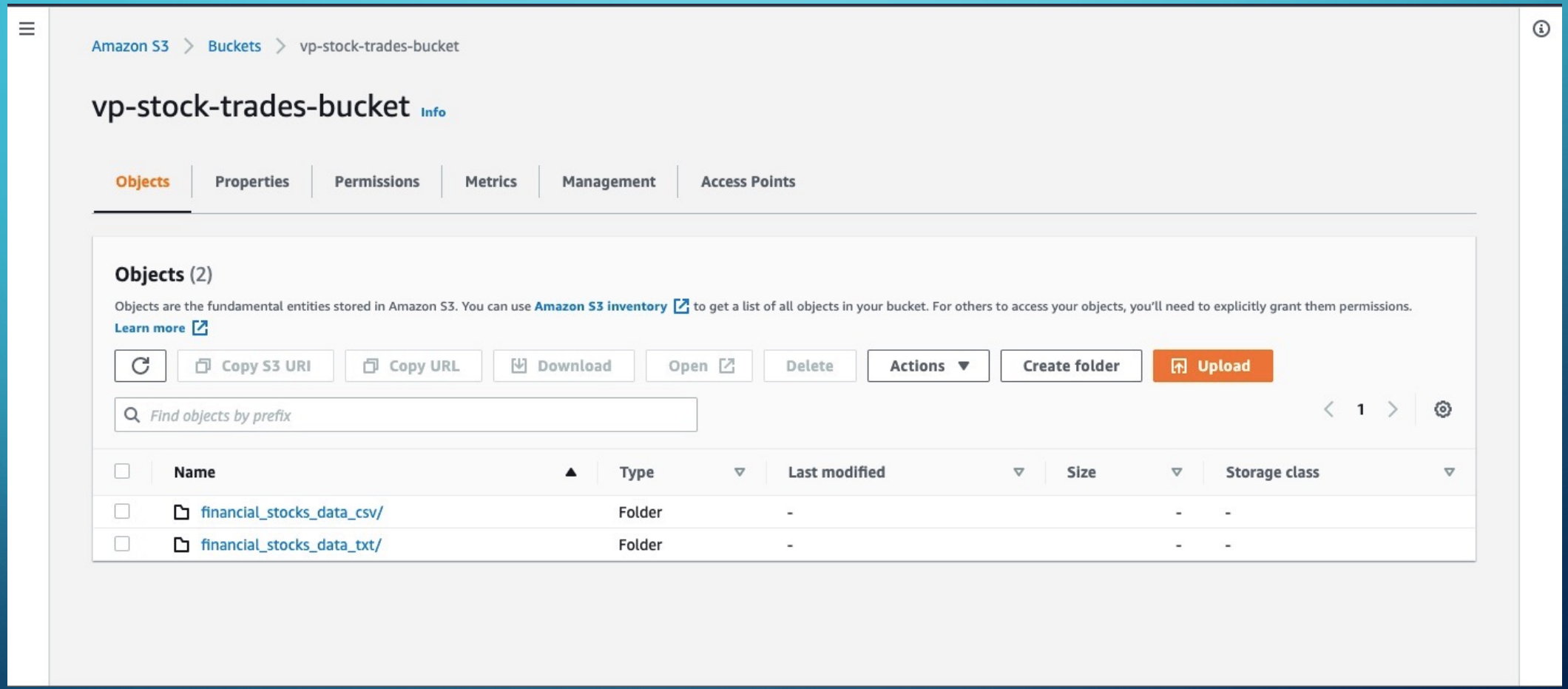
main.py

```python
# Stock Trade Transactions Data - ETL Data Pipeline using Python, AWS (CLI, S3, Glue, Athena, QuickSight)
# Main Program

from security import access_key_id, secret_access_key
from config import bucket_name, bucket_location, s3_bucket_metadata_file_path, source_data_files_path
from s3_bucket_creation import s3_create_bucket
from s3_bucket_metadata_retrieval import s3_response_metadata
from s3_bucket_data_ingestion import s3_bucket_data_ingest

# Creating a Bucket in AWS S3
s3_response = s3_create_bucket(access_key_id, secret_access_key, bucket_name, bucket_location)

# Writing the S3 Bucket Metadata Details to a Text File
s3_response_metadata(s3_response, s3_bucket_metadata_file_path)

# Ingesting Source Data Files into AWS S3 Bucket
s3_bucket_data_ingest(access_key_id, secret_access_key, bucket_name, source_data_files_path)

```

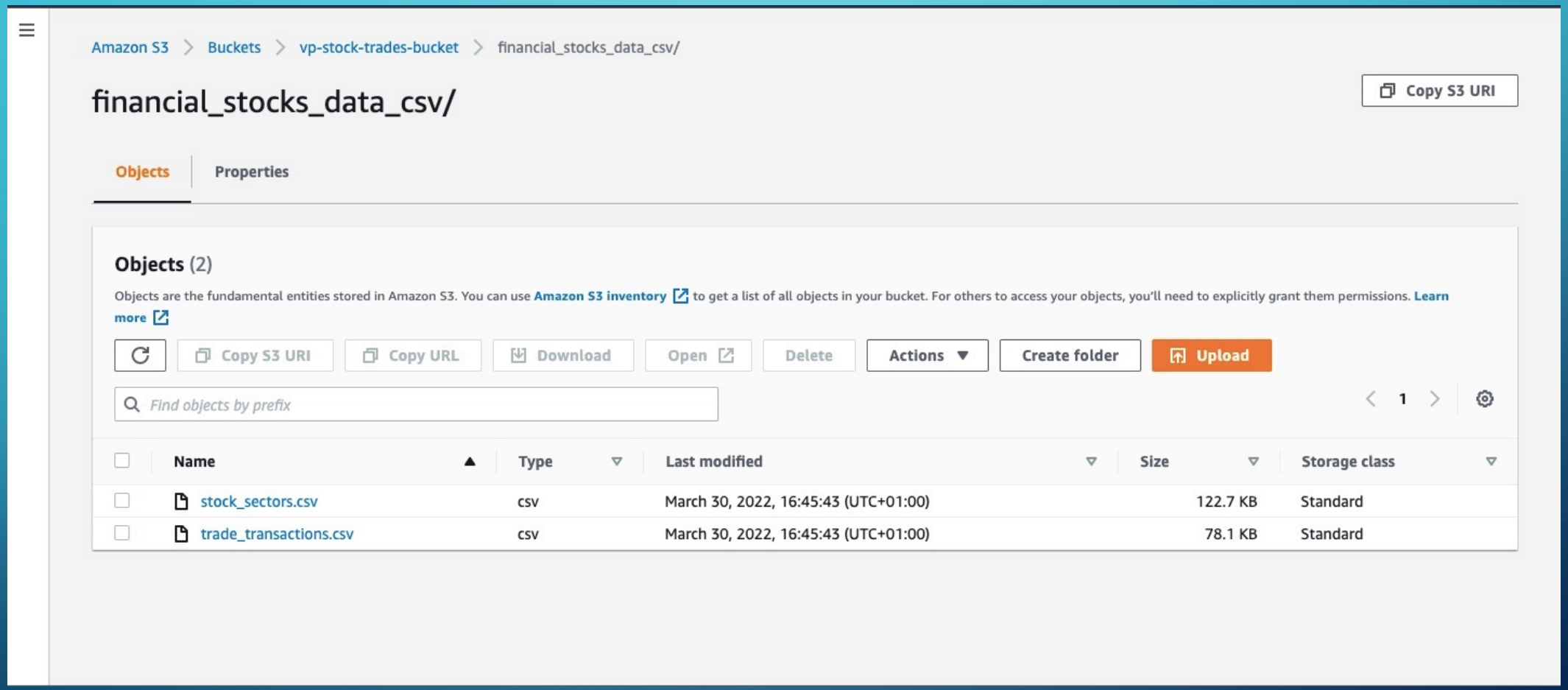
Terminal: Local

```
(base) VidhyalshmisAir:source_code vidhyalakshmiparthasarathy$ clear
(base) VidhyalshmisAir:source_code vidhyalakshmiparthasarathy$ python3 main.py
Unrecognised Source File Format: dummy.json
Unrecognised Source File Format: test.xml
Total Source Files Ingested Successfully into AWS S3: 3
Total Unrecognised Source Files Not Ingested into AWS S3: 2
(base) VidhyalshmisAir:source_code vidhyalakshmiparthasarathy$
```

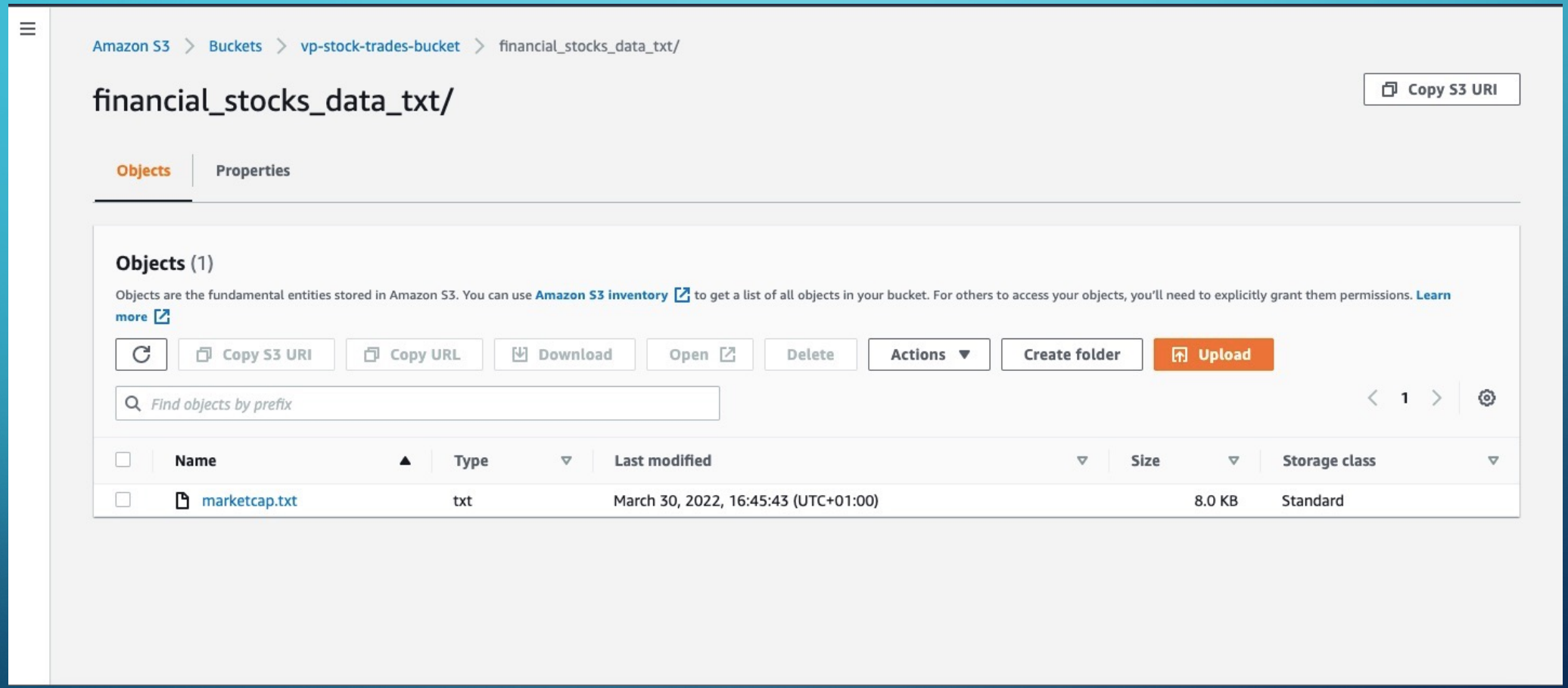
3. INGESTING SOURCE DATA FILES IN TO AWS S3 BUCKET *(CONTINUED)*



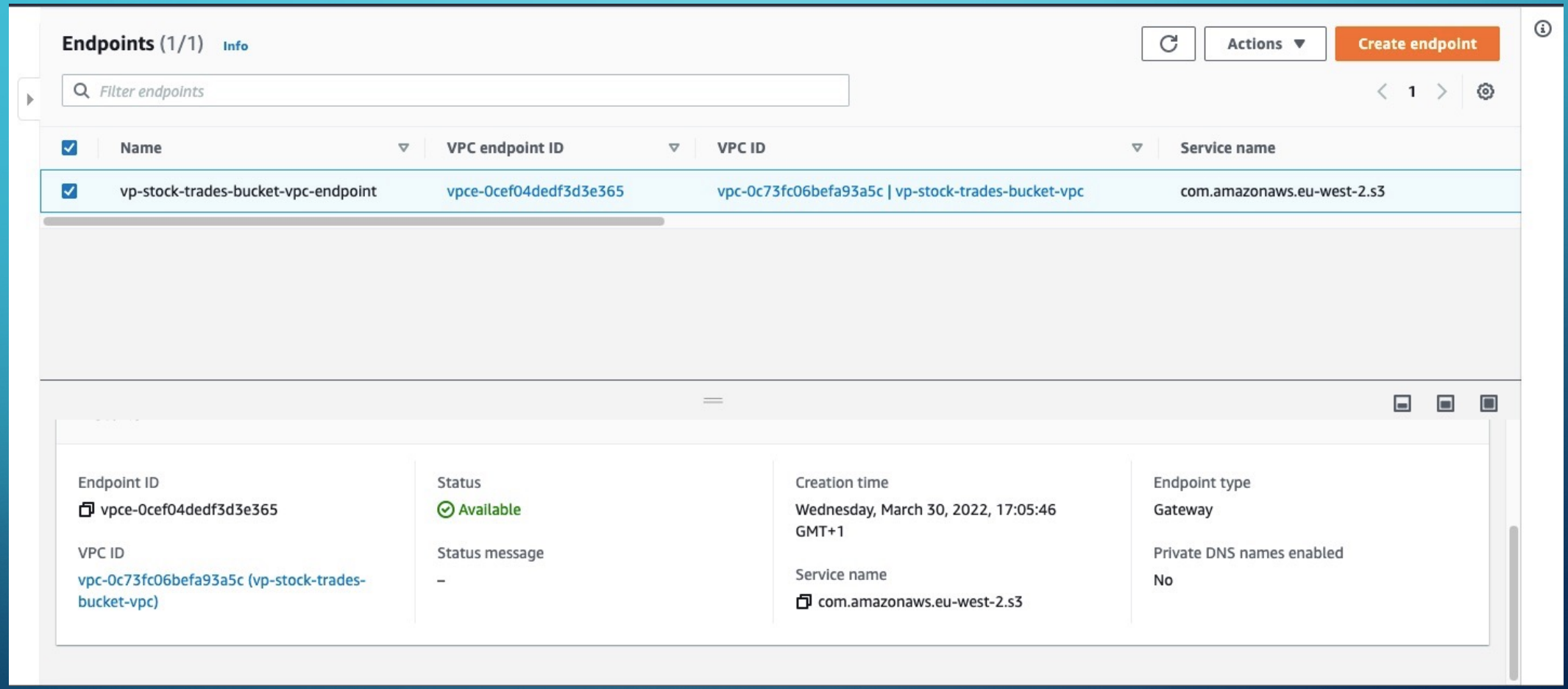
3. INGESTING SOURCE DATA FILES IN TO AWS S3 BUCKET *(CONTINUED)*



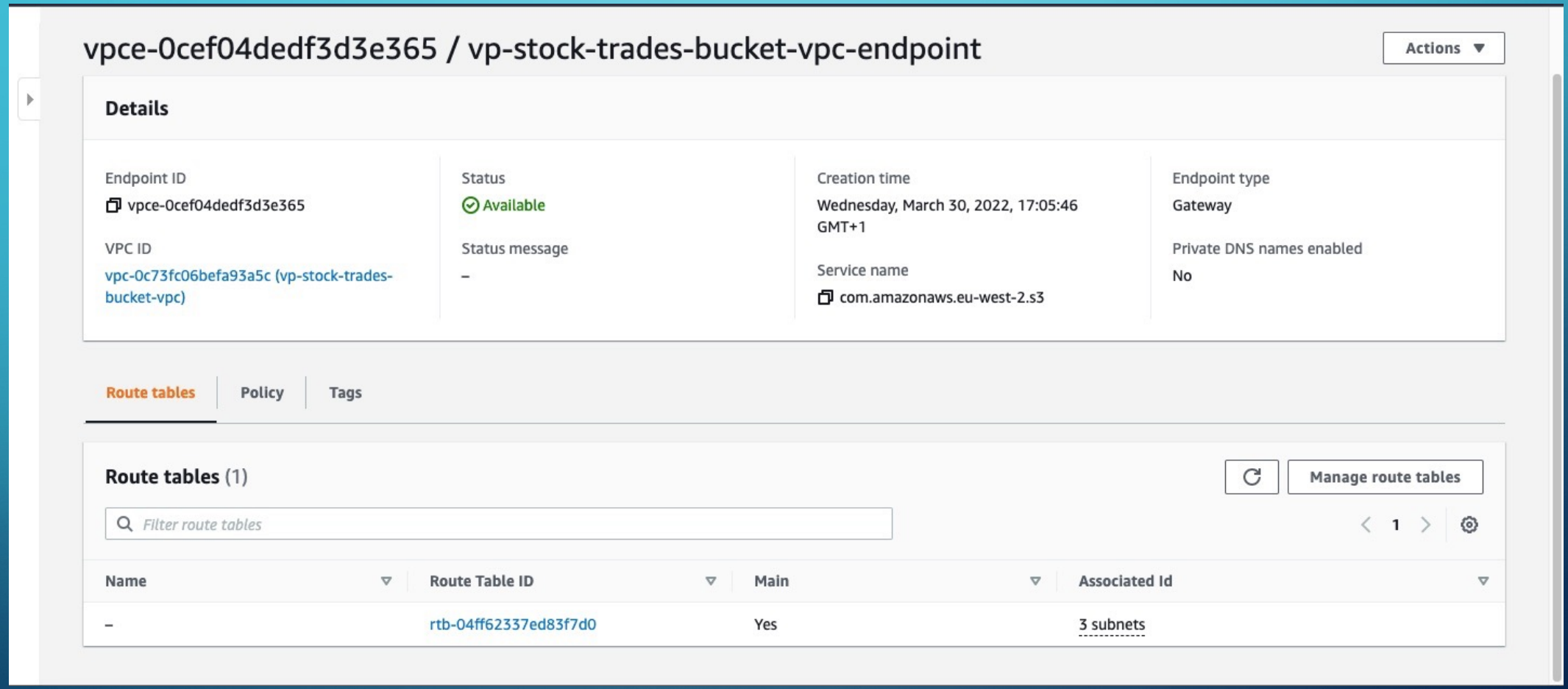
3. INGESTING SOURCE DATA FILES IN TO AWS S3 BUCKET



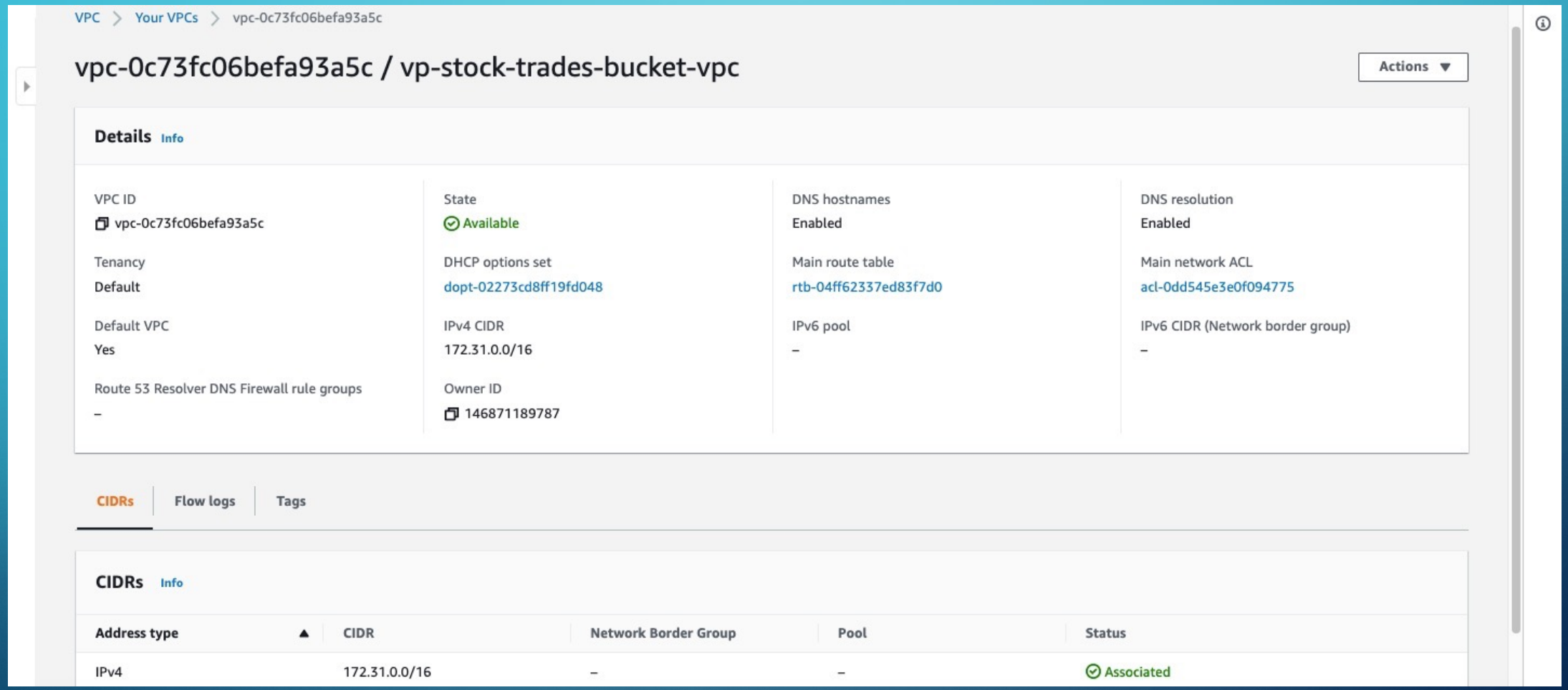
4. CREATION OF A VPC ENDPOINT INTERFACE TO ENABLE AWS GLUE TO ACCESS AWS S3 *(CONTINUED)*



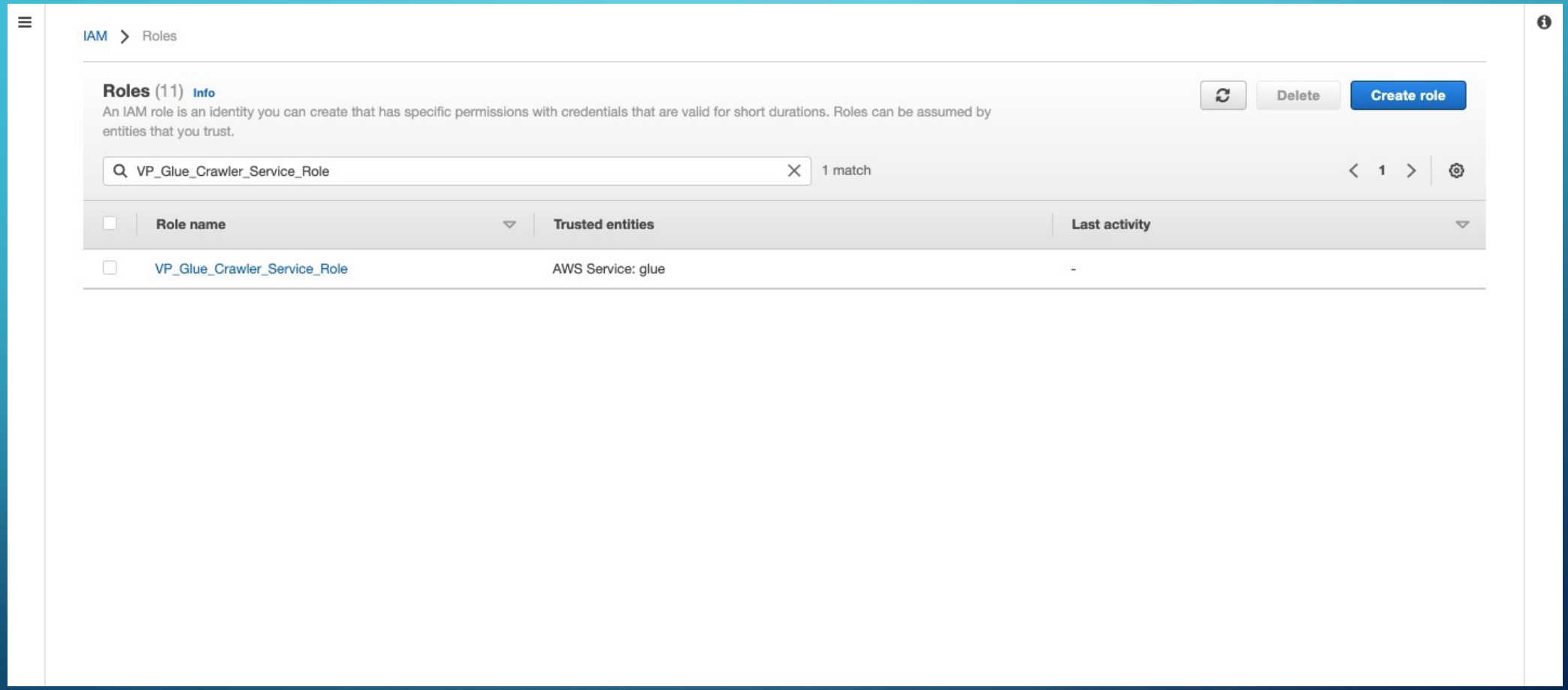
4. CREATION OF A VPC ENDPOINT INTERFACE TO ENABLE AWS GLUE TO ACCESS AWS S3 *(CONTINUED)*



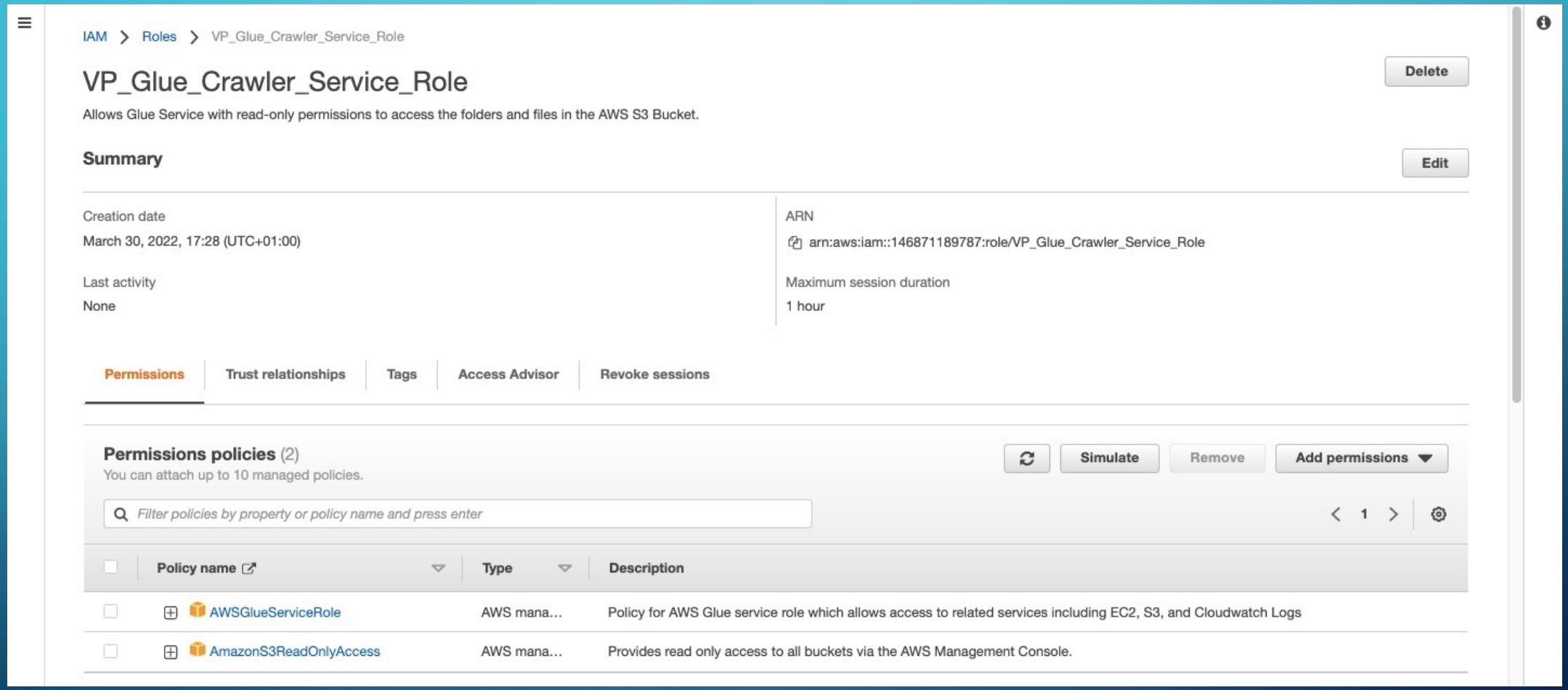
4. CREATION OF A VPC ENDPOINT INTERFACE TO ENABLE AWS GLUE TO ACCESS AWS S3



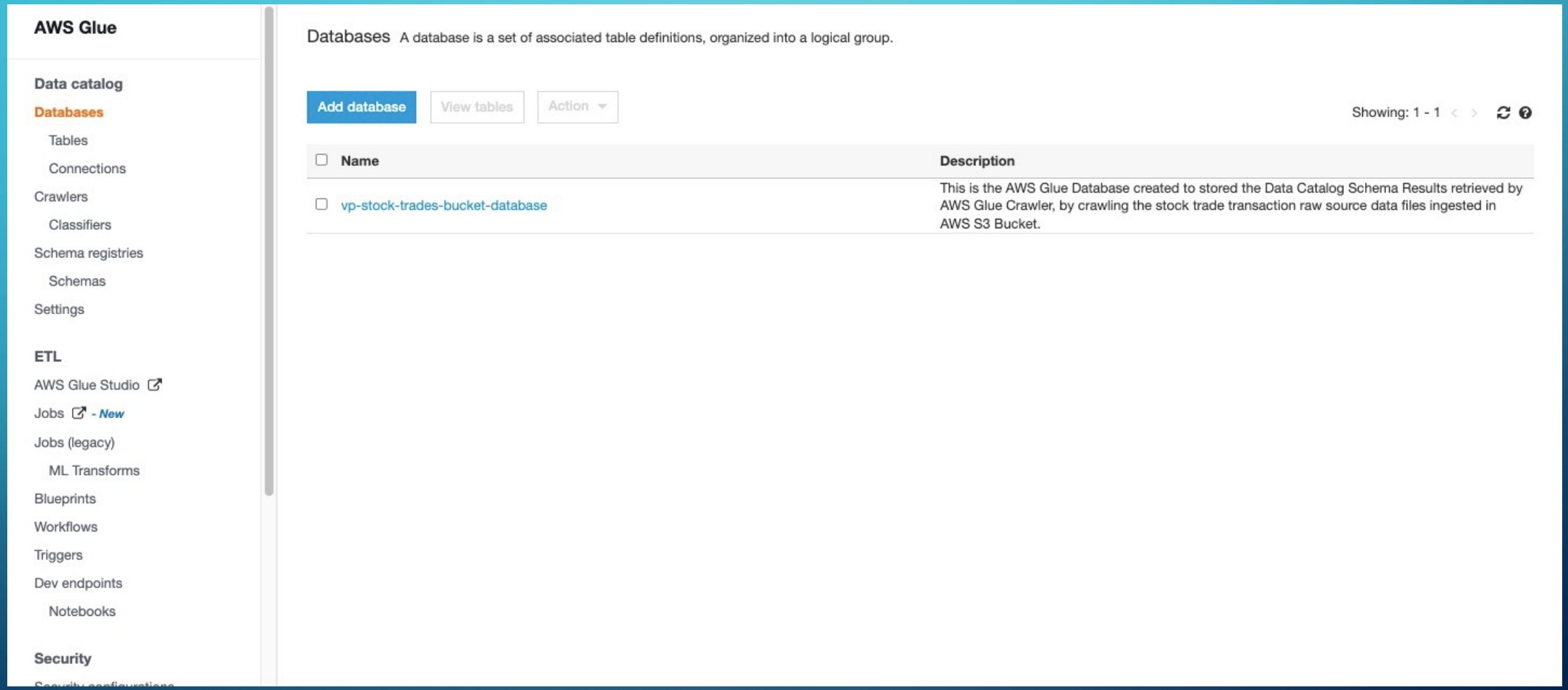
5. CREATION OF AN IAM ROLE TO GRANT PERMISSIONS TO AWS GLUE CRAWLER TO ACCESS AWS S3 *(CONTINUED)*



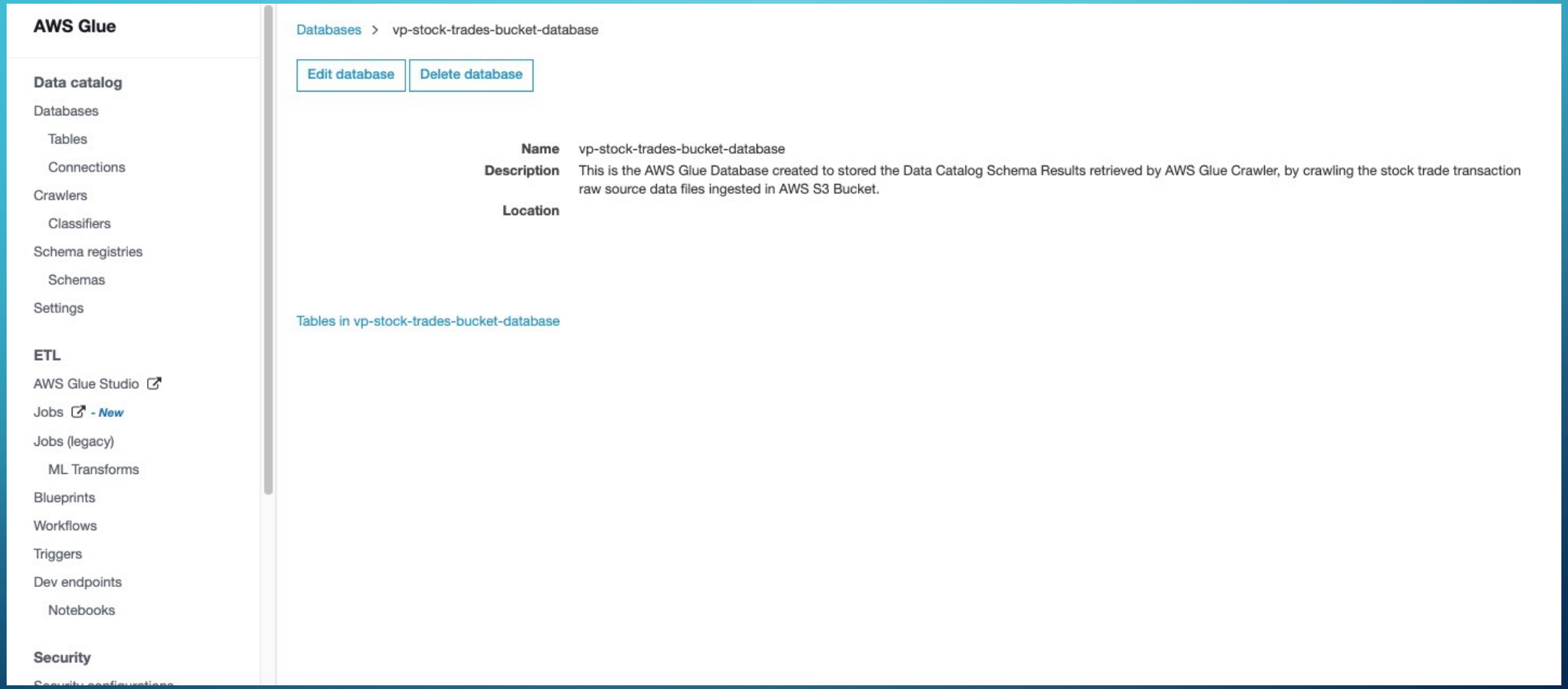
5. CREATION OF AN IAM ROLE TO GRANT PERMISSIONS TO AWS GLUE CRAWLER TO ACCESS AWS S3



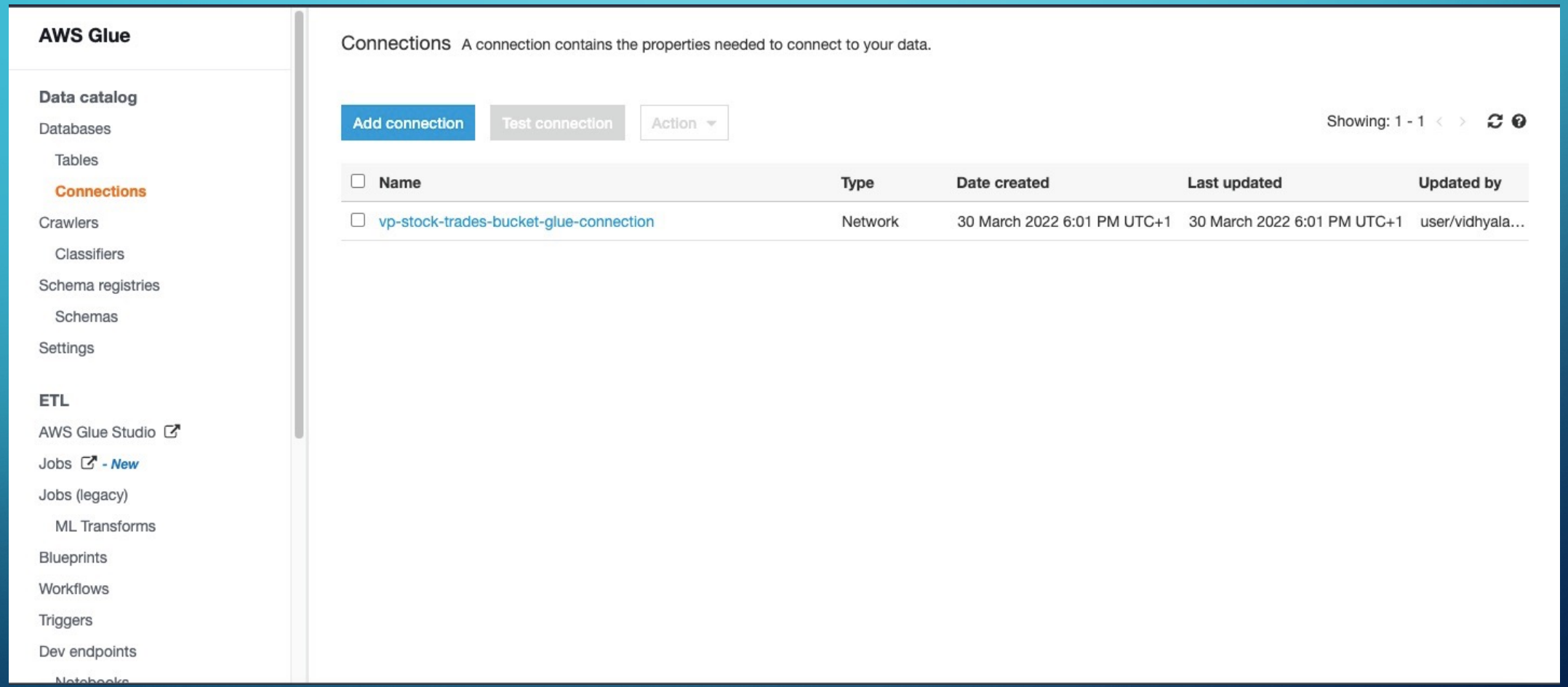
6. CREATION OF A DATABASE IN THE AWS GLUE TO STORE THE GLUE CRAWLER DATA CATALOG SCHEMA RESULTS *(CONTINUED)*



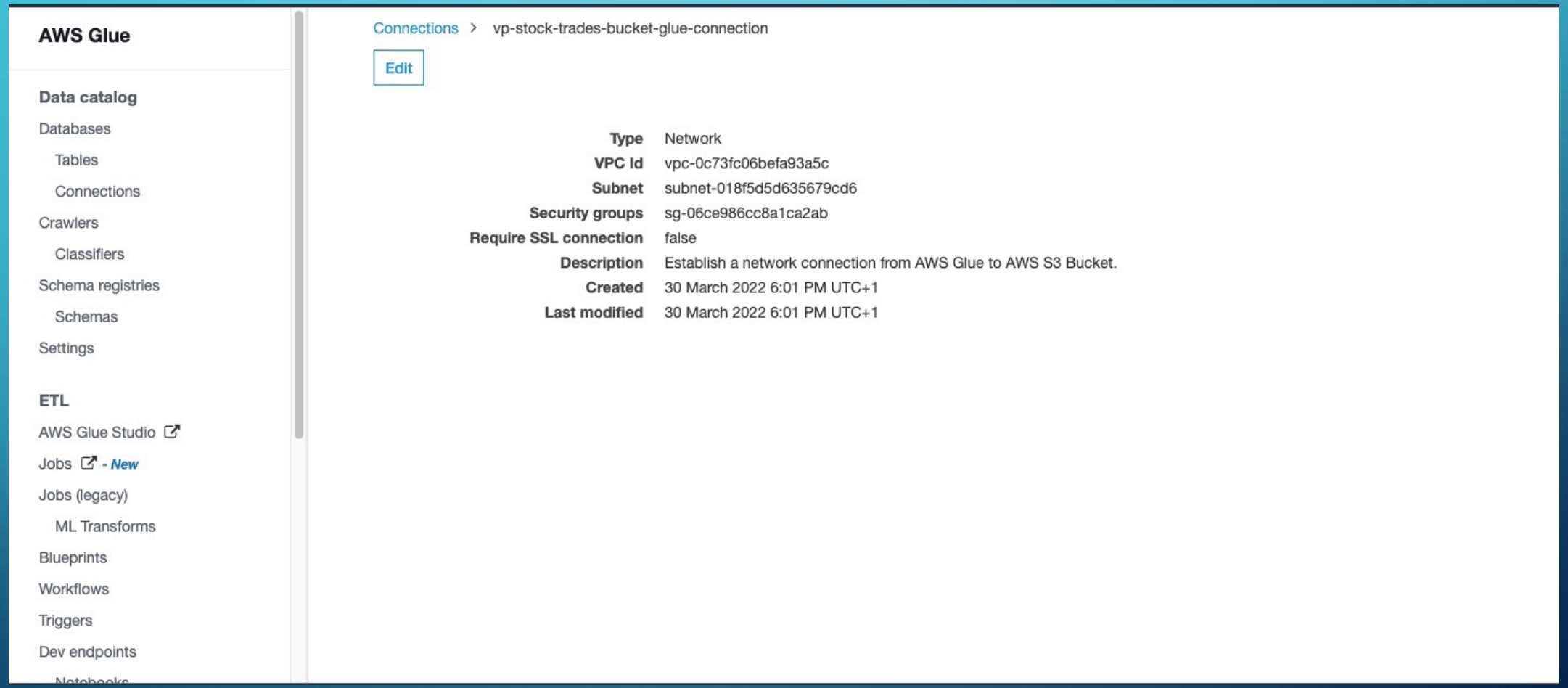
6. CREATION OF A DATABASE IN THE AWS GLUE TO STORE THE GLUE CRAWLER DATA CATALOG SCHEMA RESULTS



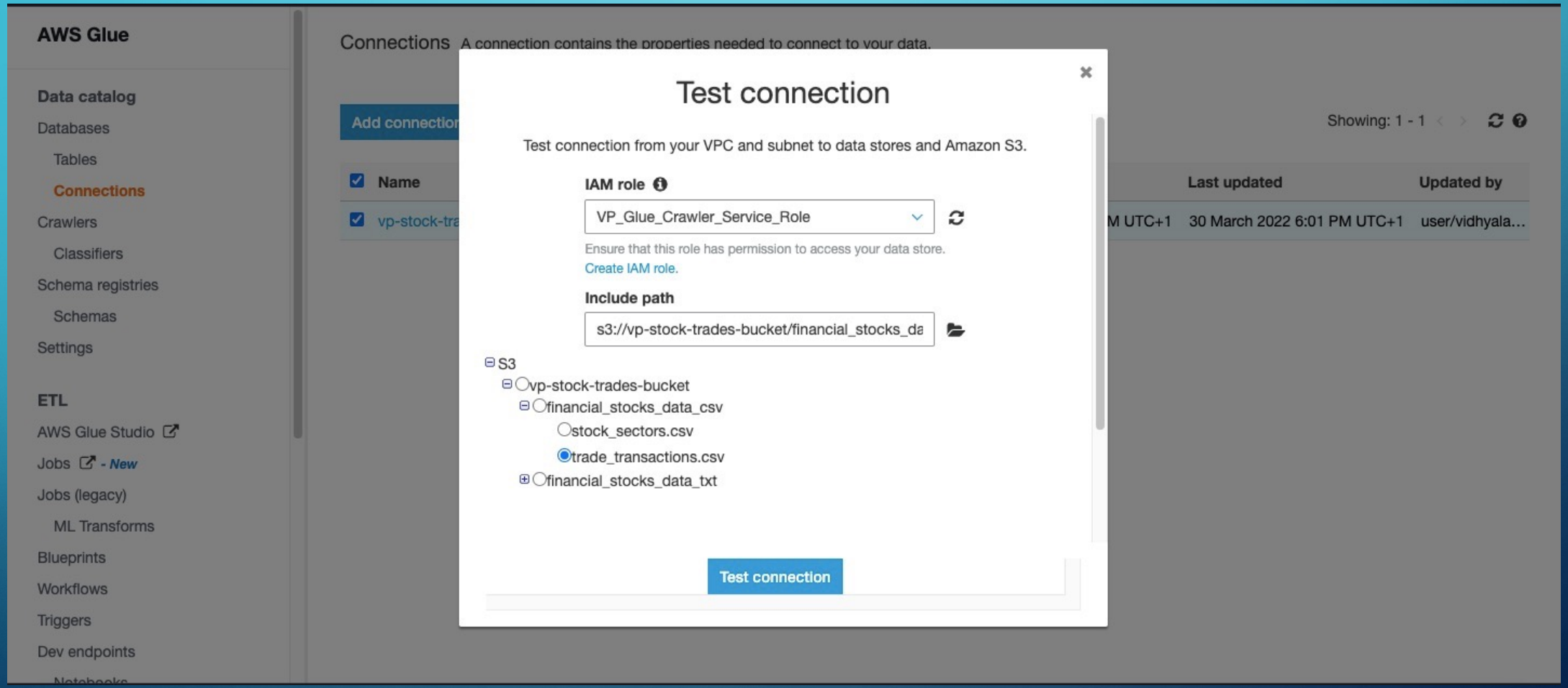
7. ESTABLISH A NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET *(CONTINUED)*



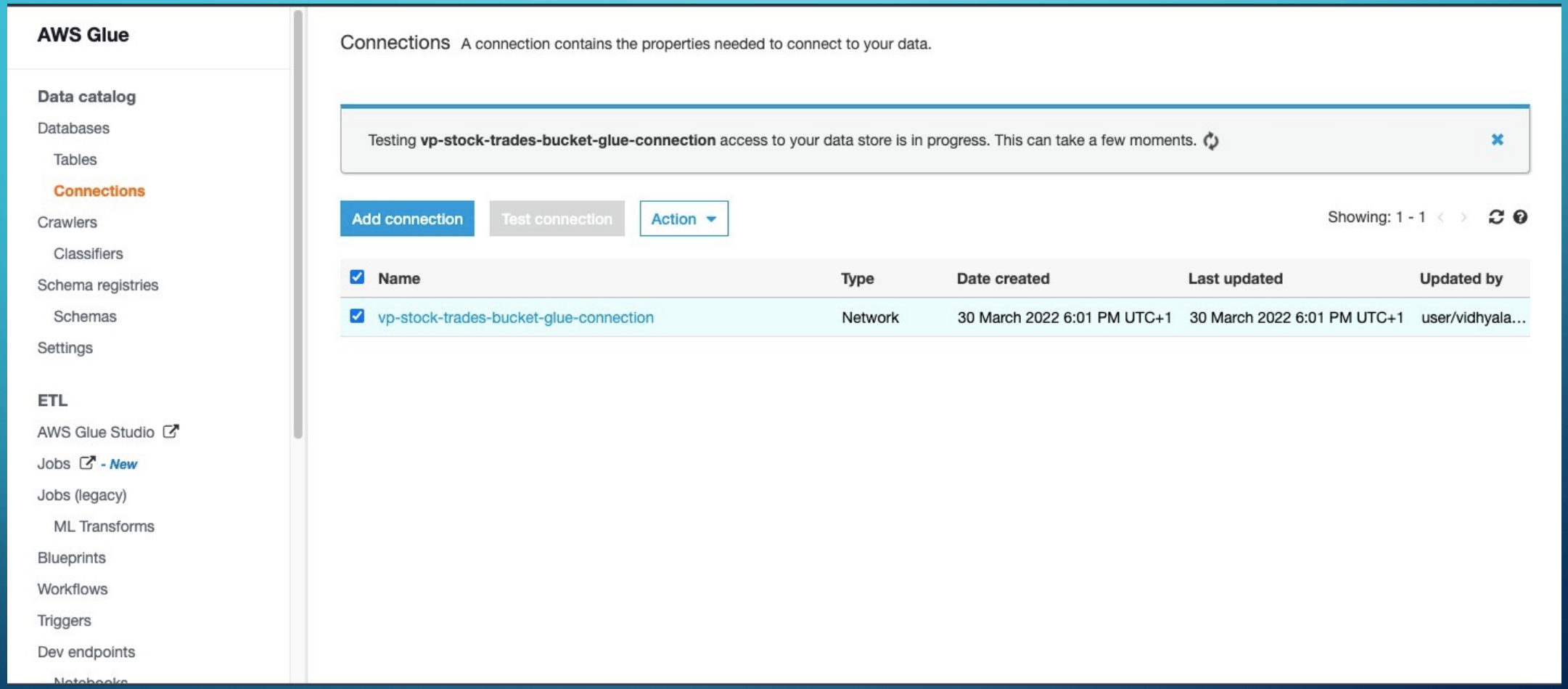
7. ESTABLISH A NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET



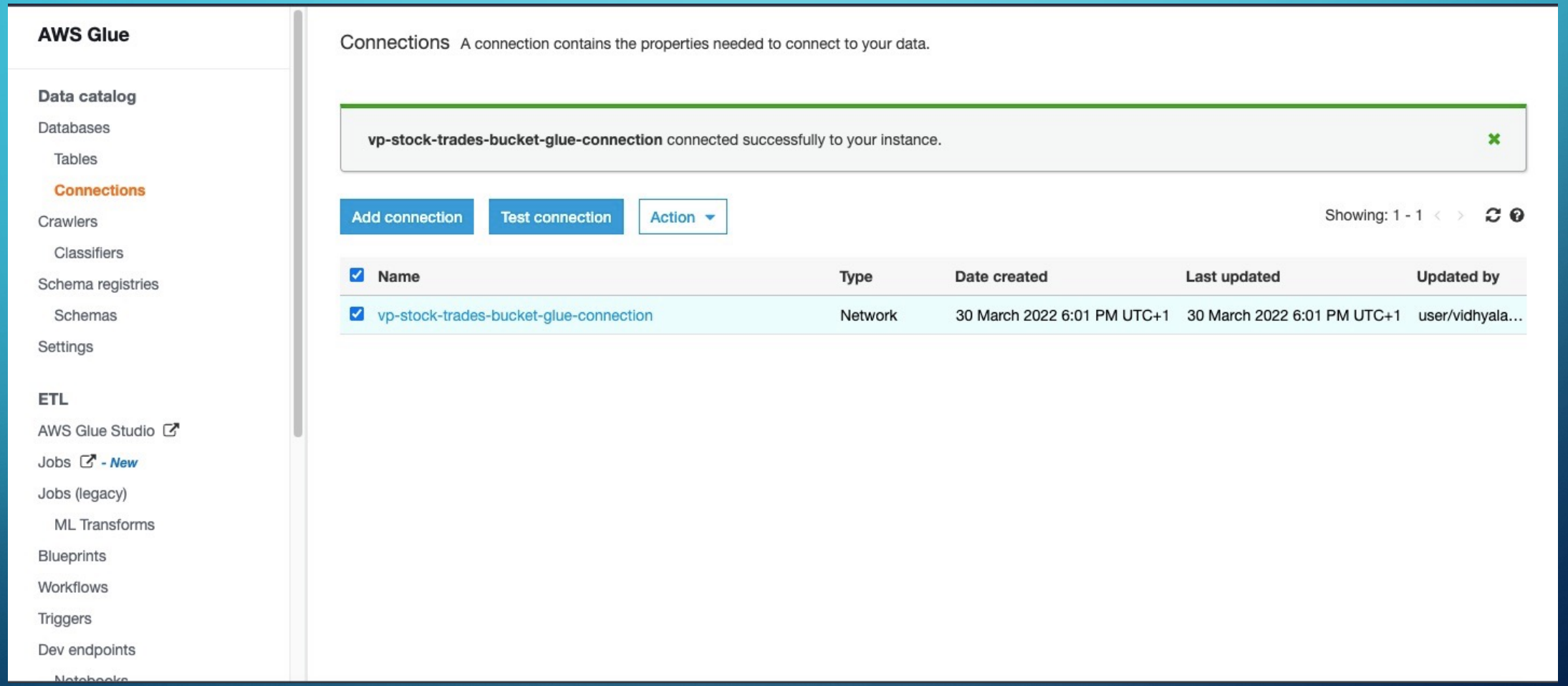
8. VERIFY/TEST THE NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET *(CONTINUED)*



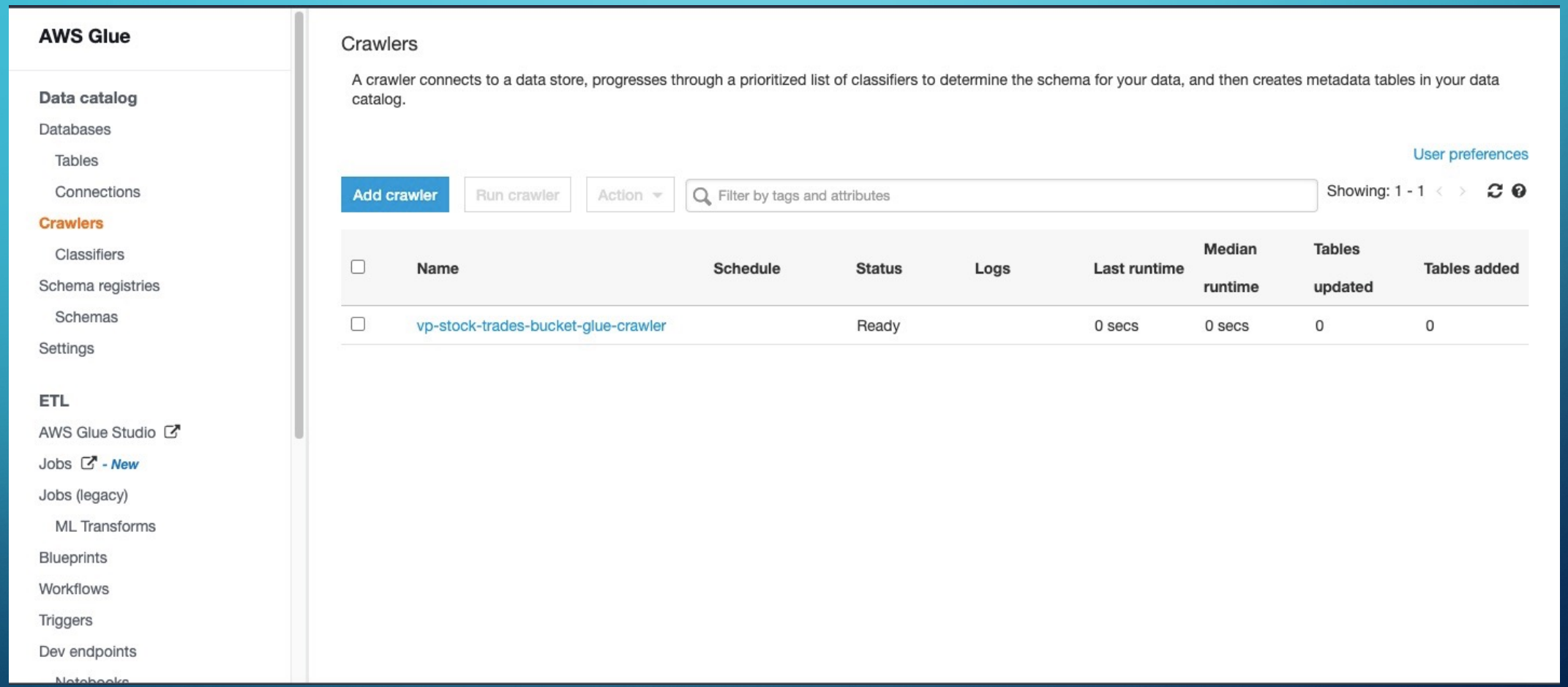
8. VERIFY/TEST THE NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET *(CONTINUED)*



8. VERIFY/TEST THE NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET

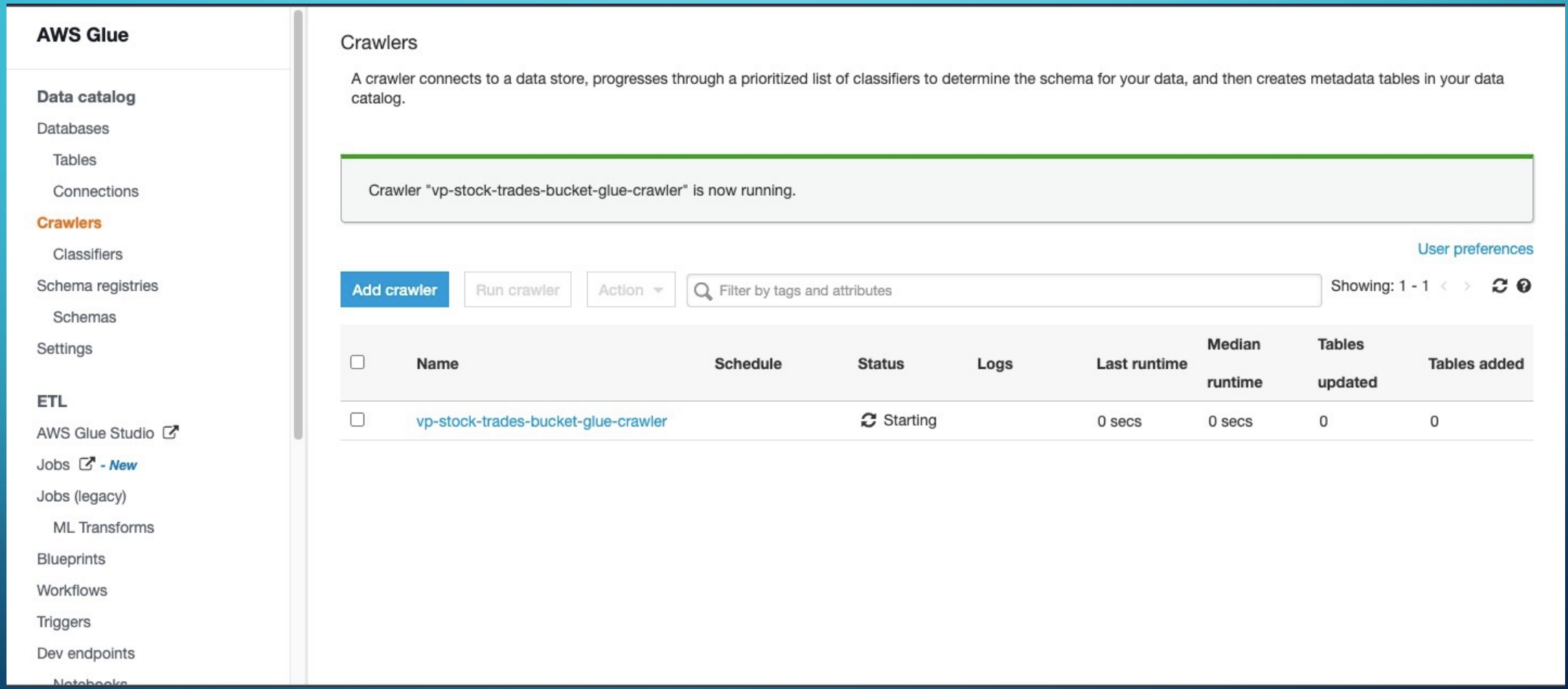


9. CREATION OF AWS GLUE CRAWLER TO RETRIEVE DATA CATALOG SCHEMA RESULTS OF THE SPECIFIED SOURCE FILES IN AWS S3 BUCKET *(CONTINUED)*

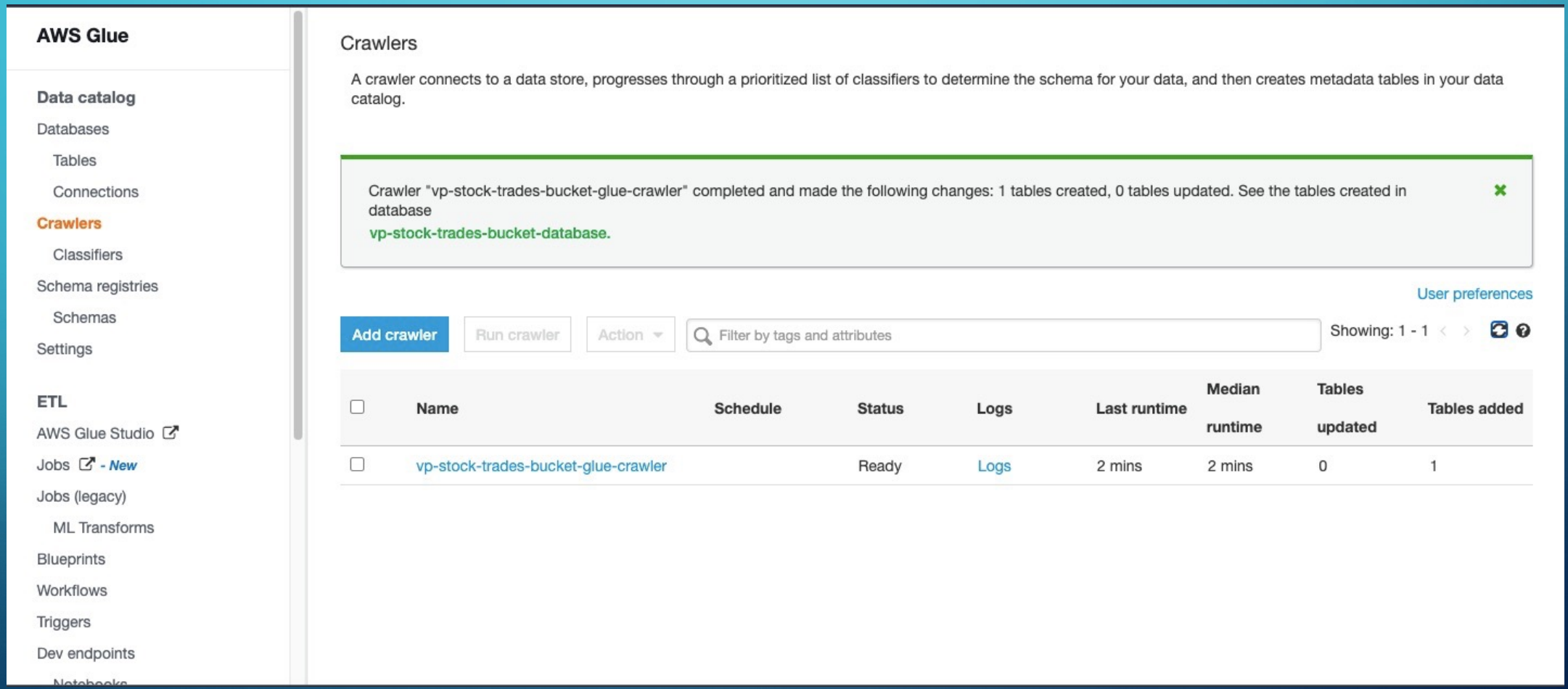


9. CREATION OF AWS GLUE CRAWLER TO RETRIEVE DATA CATALOG SCHEMA RESULTS OF THE SPECIFIED SOURCE FILES IN AWS S3 BUCKET

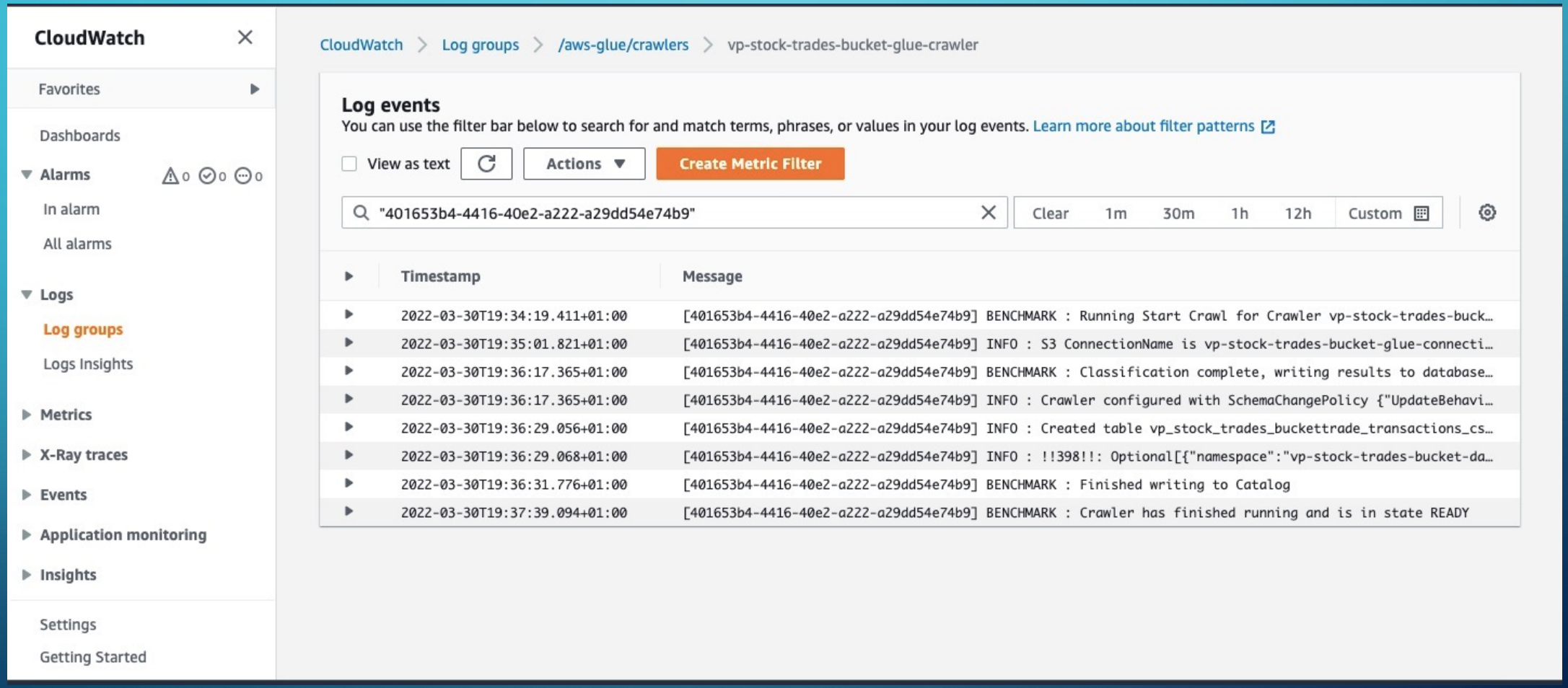
10. RUNNING THE AWS GLUE CRAWLER TO RETRIEVE DATA CATALOG SCHEMA RESULTS OF THE SPECIFIED SOURCE FILES IN AWS S3 BUCKET *(CONTINUED)*



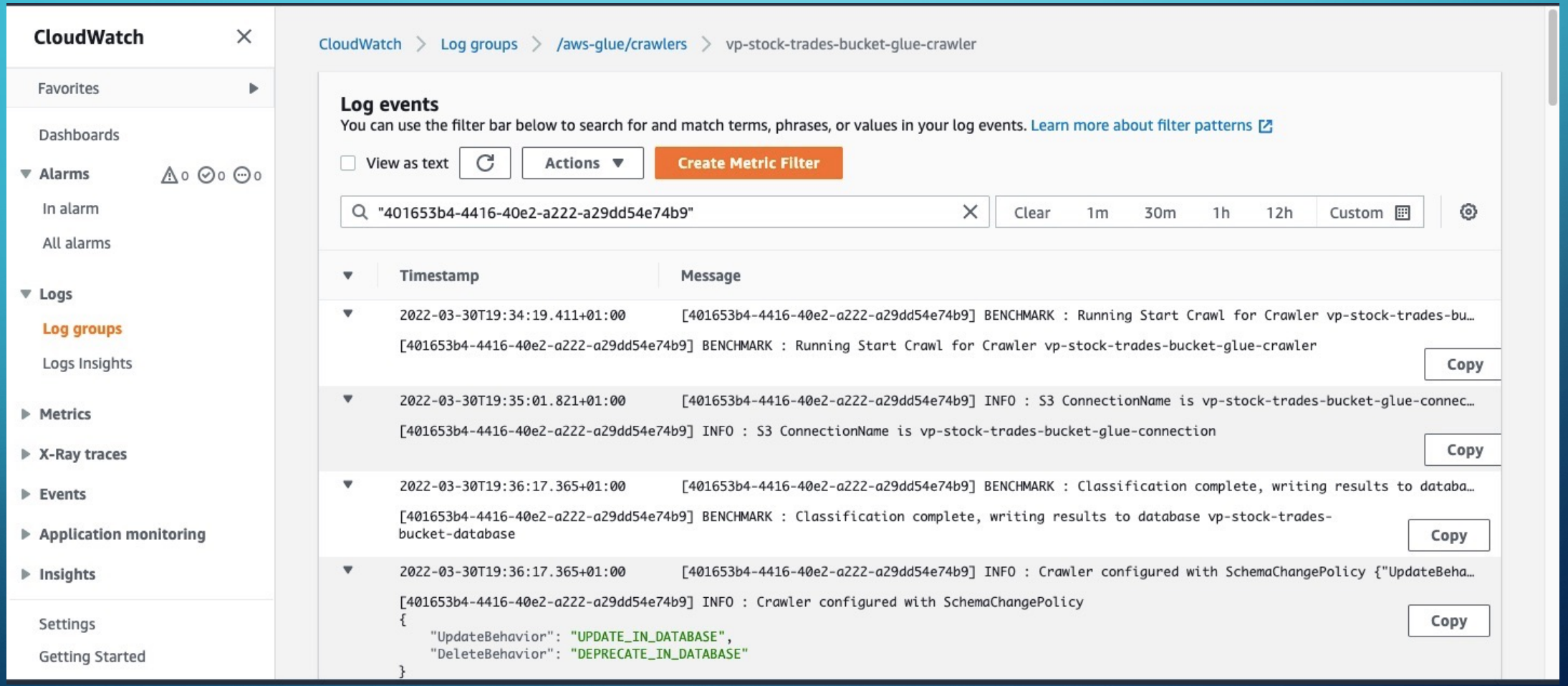
10. RUNNING THE AWS GLUE CRAWLER TO RETRIEVE DATA CATALOG SCHEMA RESULTS OF THE SPECIFIED SOURCE FILES IN AWS S3 BUCKET



11. VIEWING THE AWS GLUE CRAWLER EXECUTION LOGS IN AWS CLOUDWATCH *(CONTINUED)*

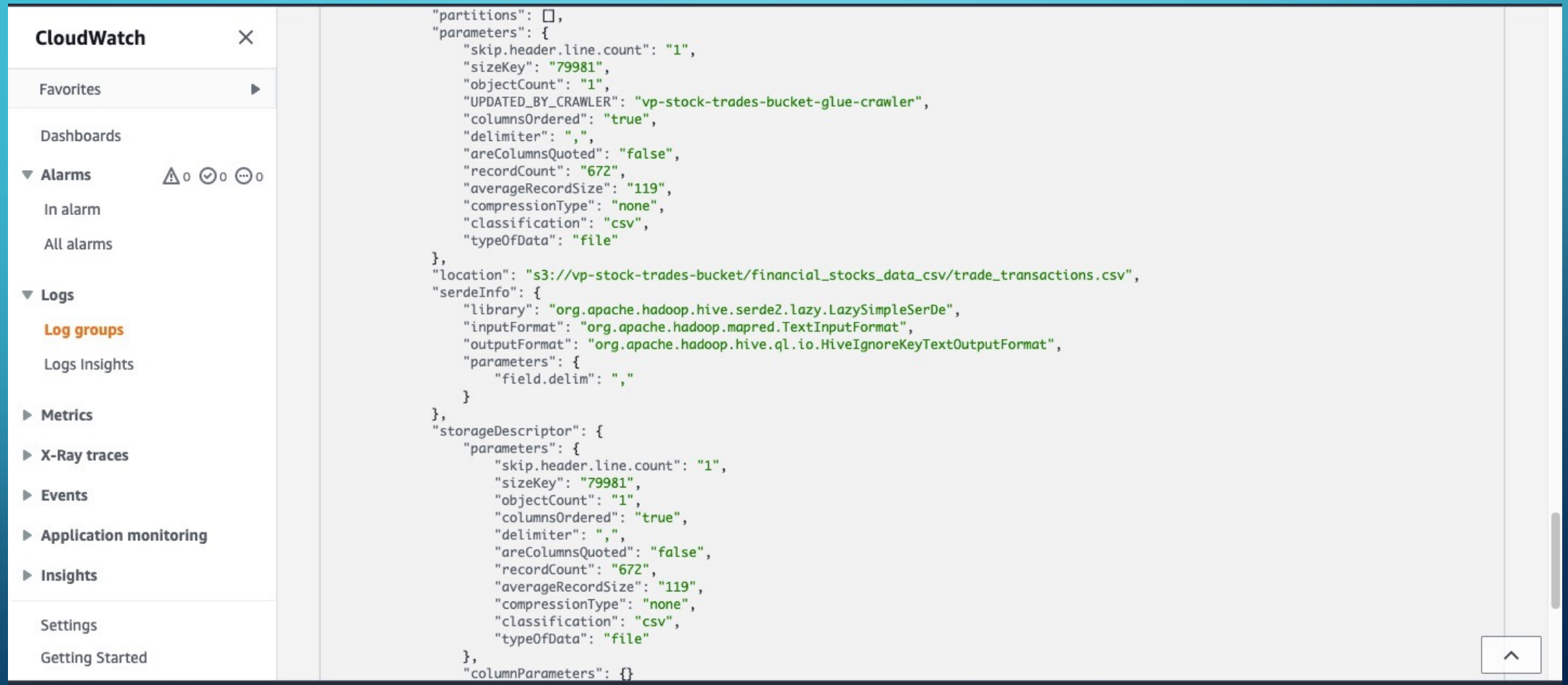


11. VIEWING THE AWS GLUE CRAWLER EXECUTION LOGS IN AWS CLOUDWATCH *(CONTINUED)*

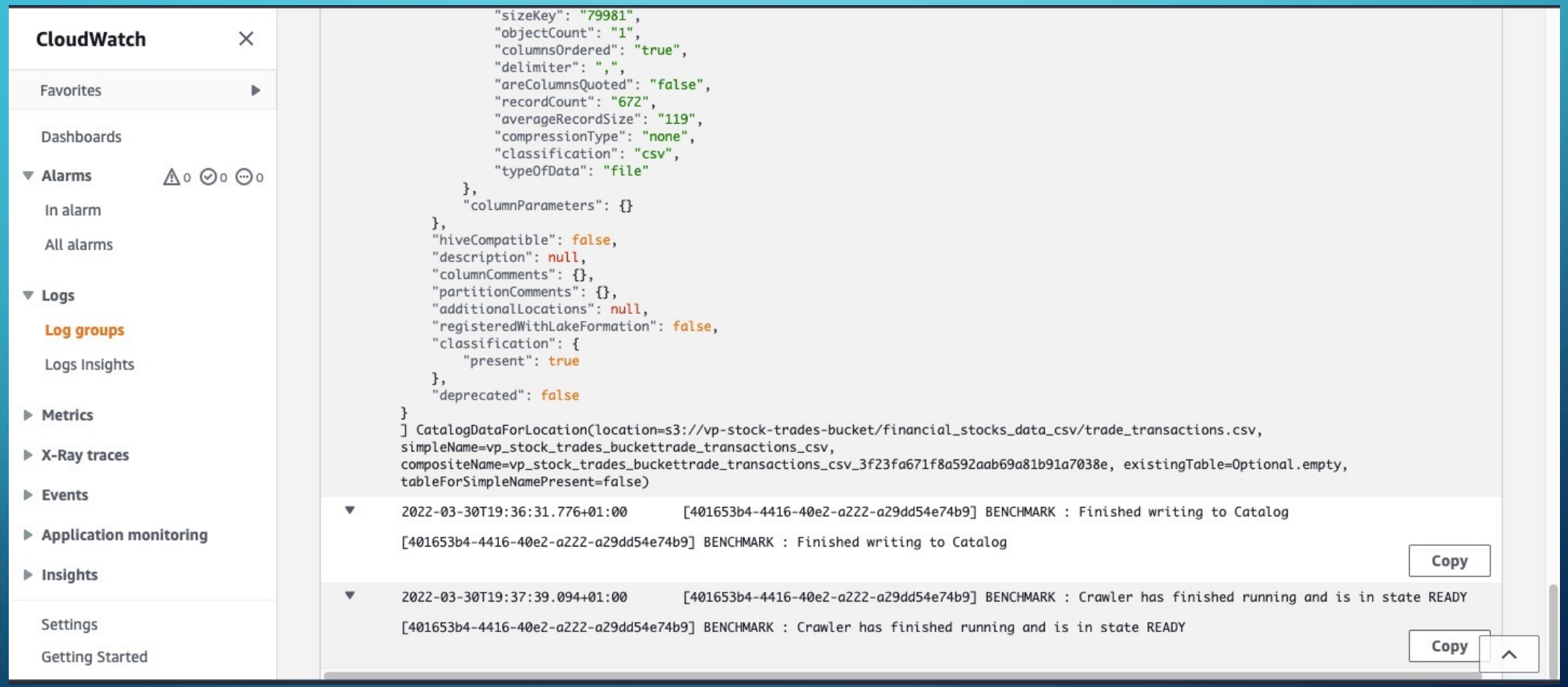


11. VIEWING THE AWS GLUE CRAWLER EXECUTION LOGS IN AWS CLOUDWATCH *(CONTINUED)*

11. VIEWING THE AWS GLUE CRAWLER EXECUTION LOGS IN AWS CLOUDWATCH *(CONTINUED)*

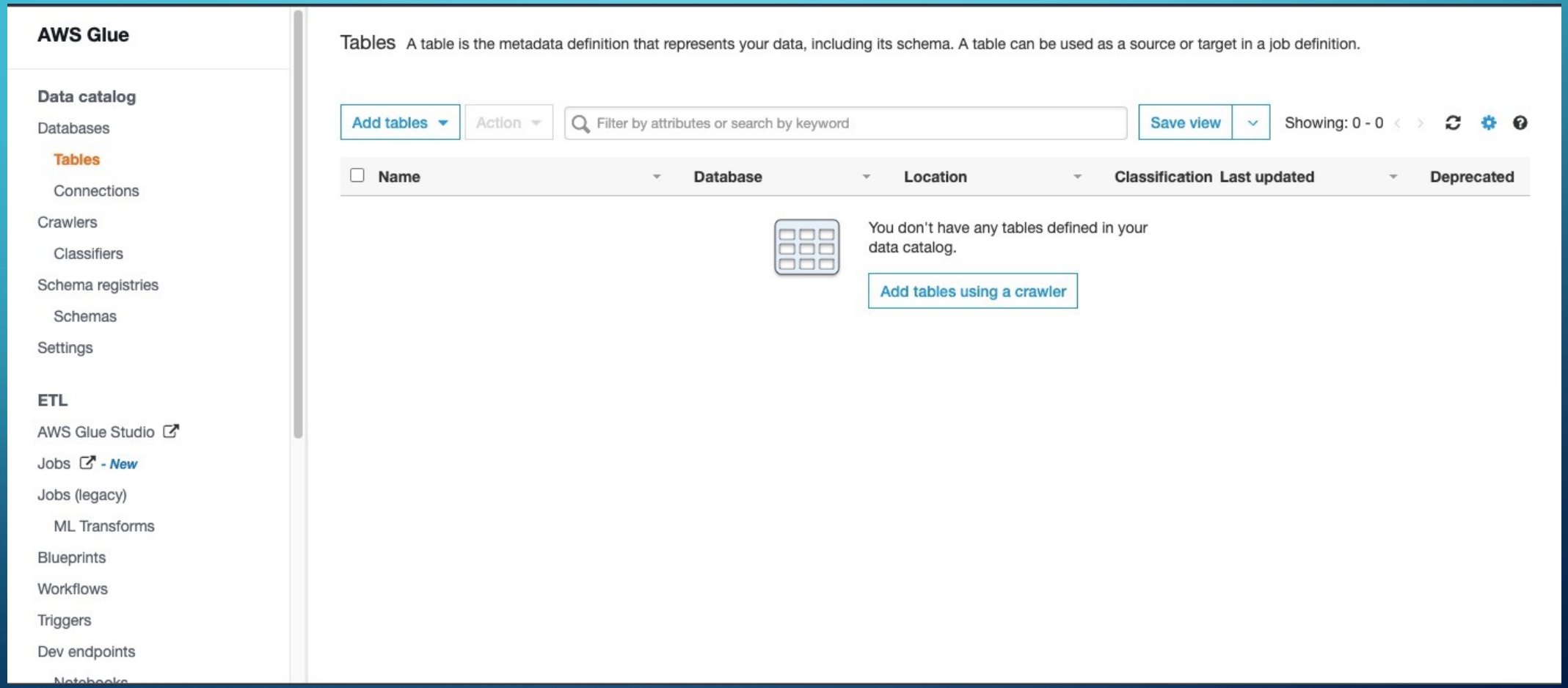


11. VIEWING THE AWS GLUE CRAWLER EXECUTION LOGS IN AWS CLOUDWATCH



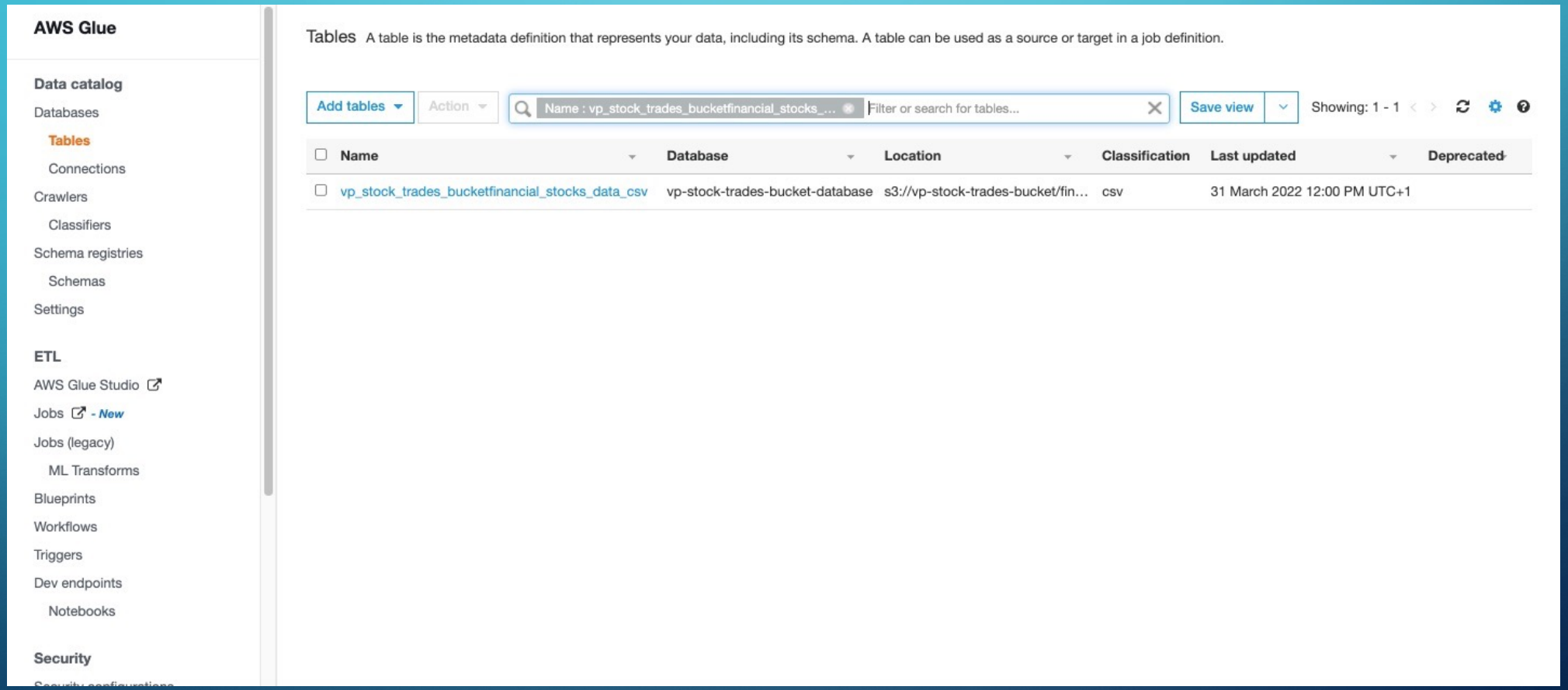
12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE *(CONTINUED)*

AWS Glue Database Table before executing the AWS Glue Crawler



12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE *(CONTINUED)*

AWS Glue Database Table after executing the AWS Glue Crawler

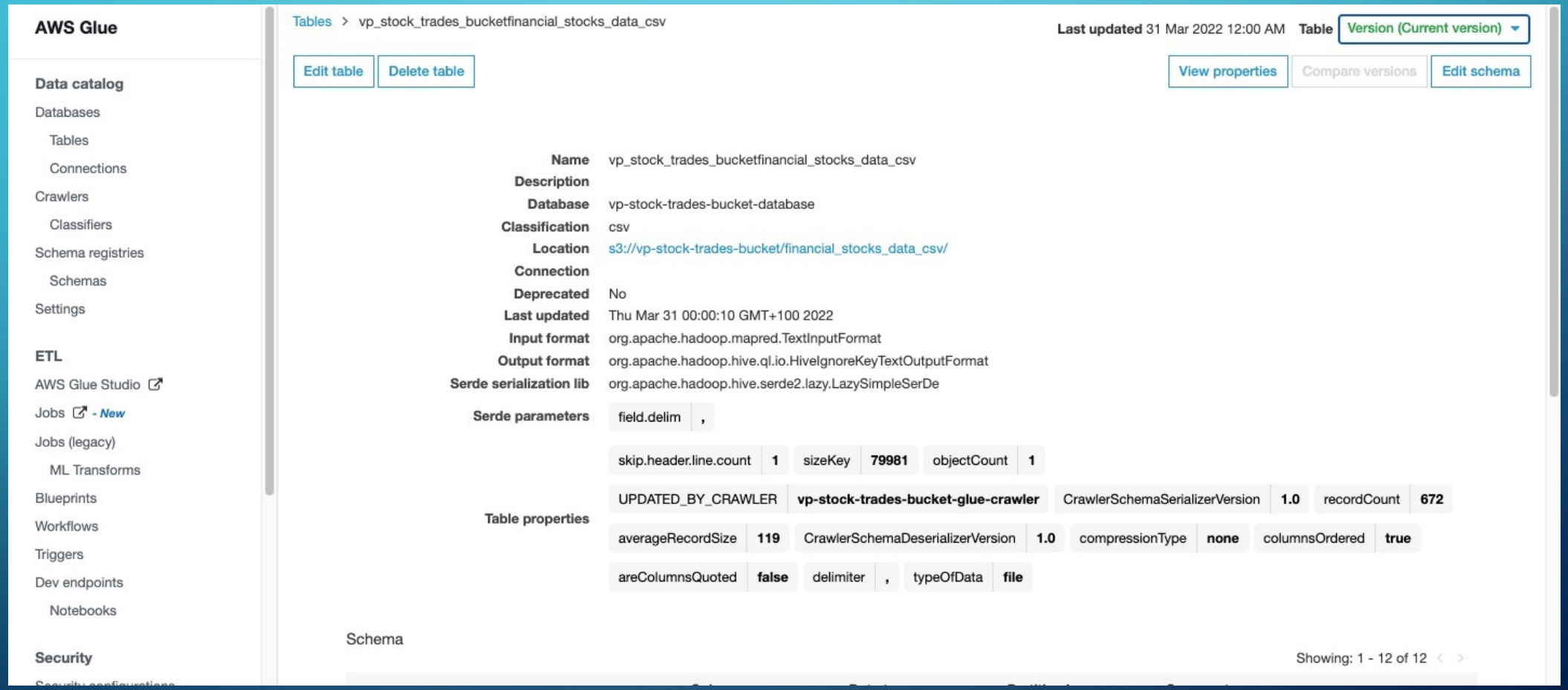


12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE *(CONTINUED)*

AWS Glue Database Table after executing the AWS Glue Crawler

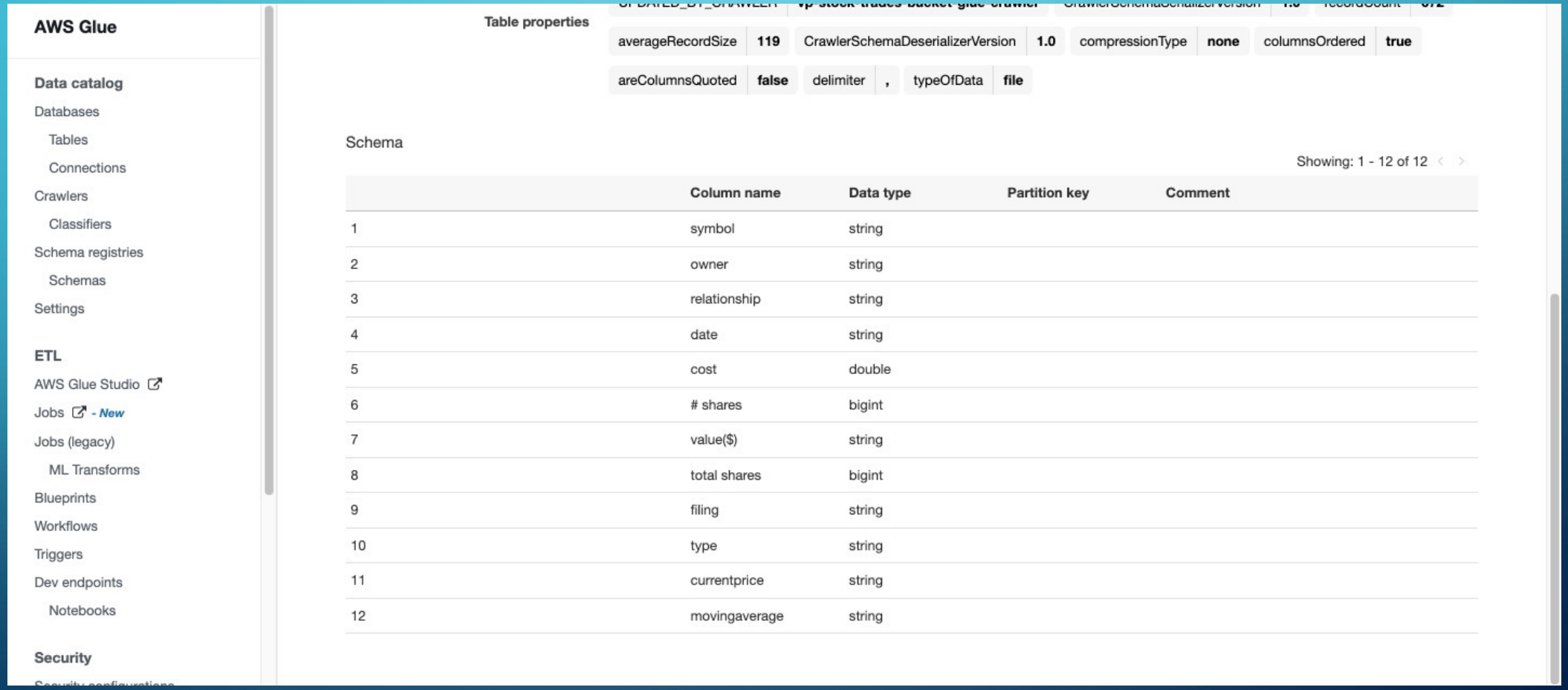
12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE *(CONTINUED)*

AWS Glue Database Table after executing the AWS Glue Crawler

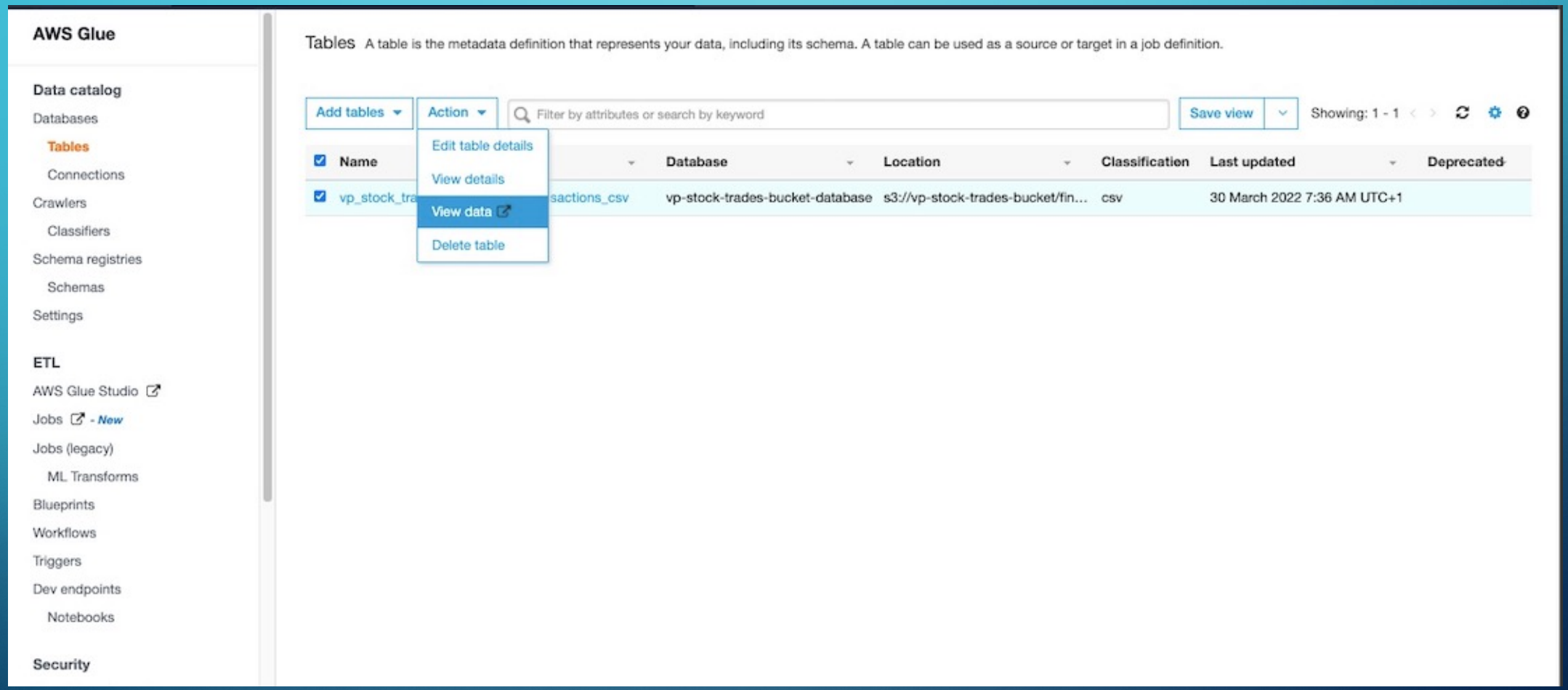


12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE

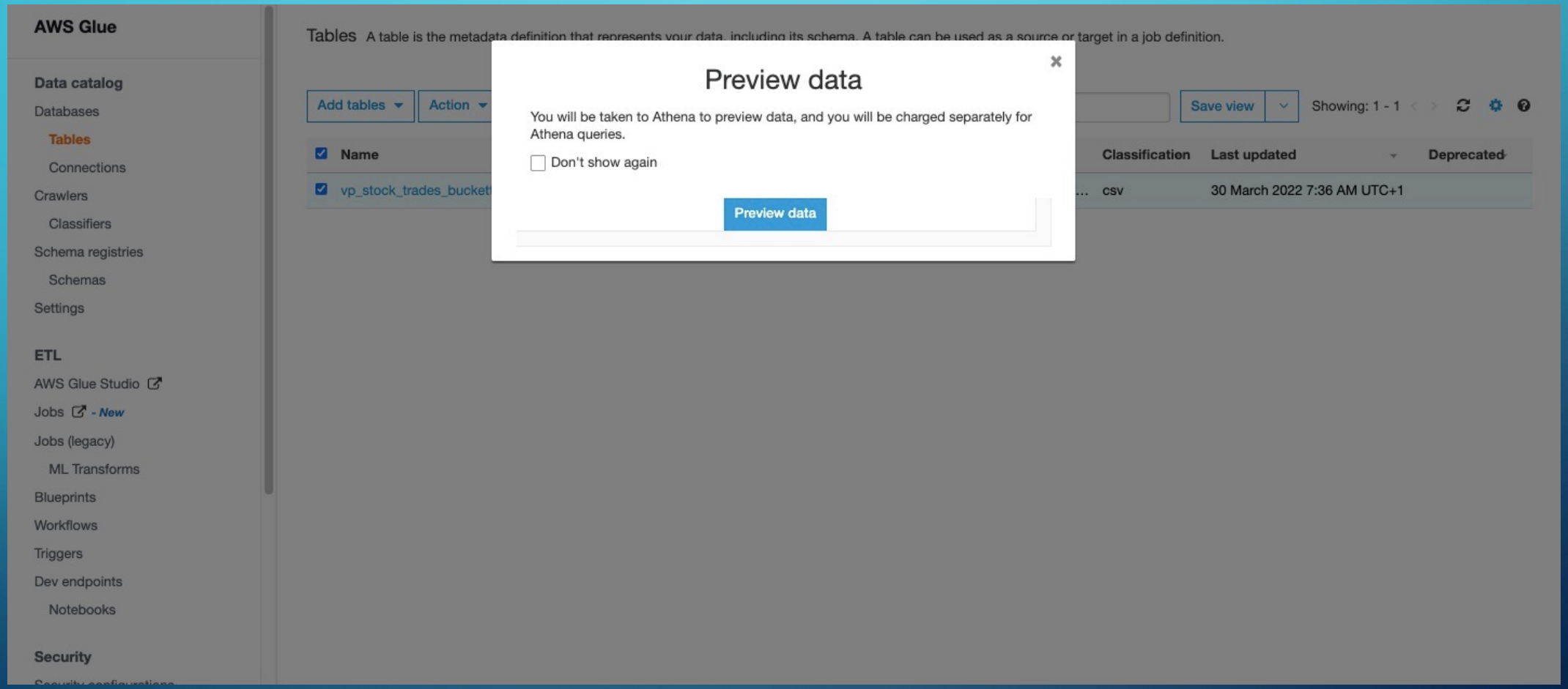
AWS Glue Database Table after executing the AWS Glue Crawler



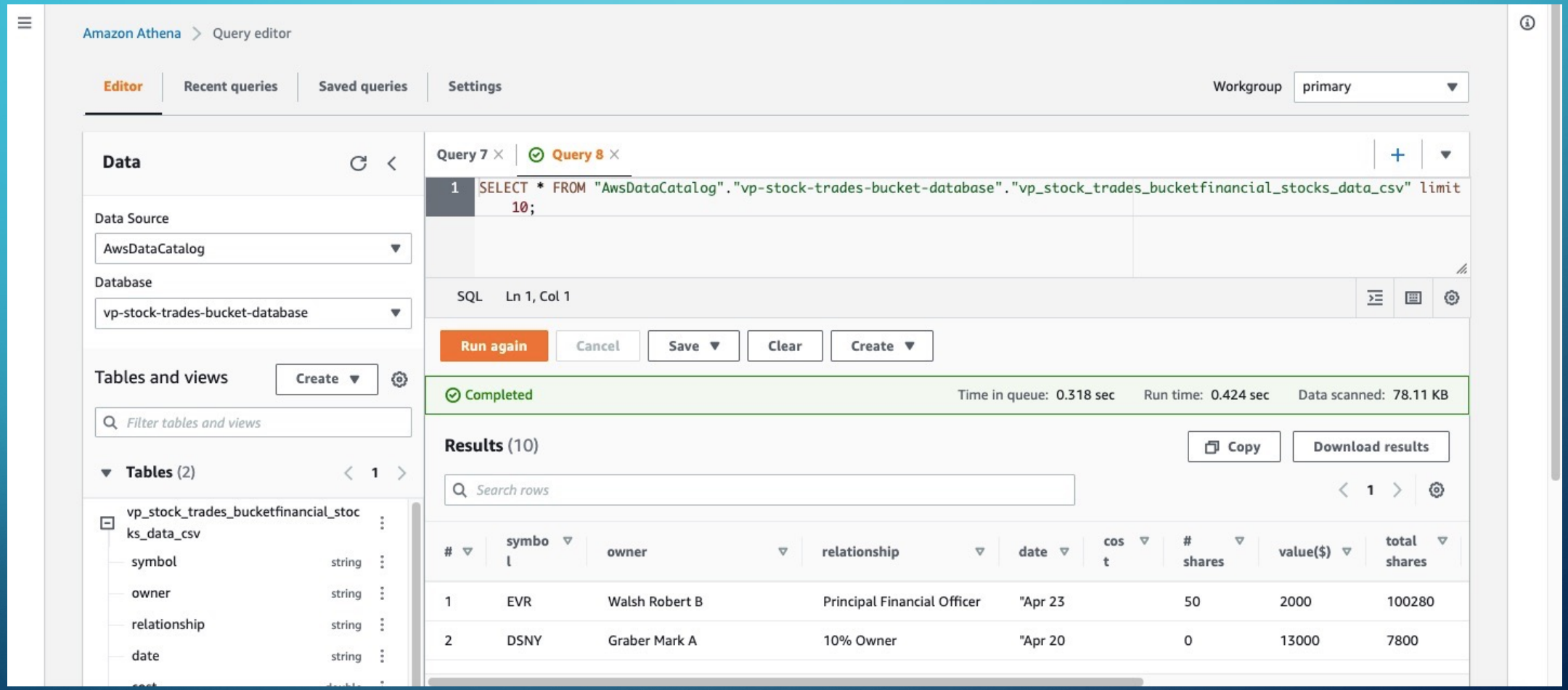
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*



13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

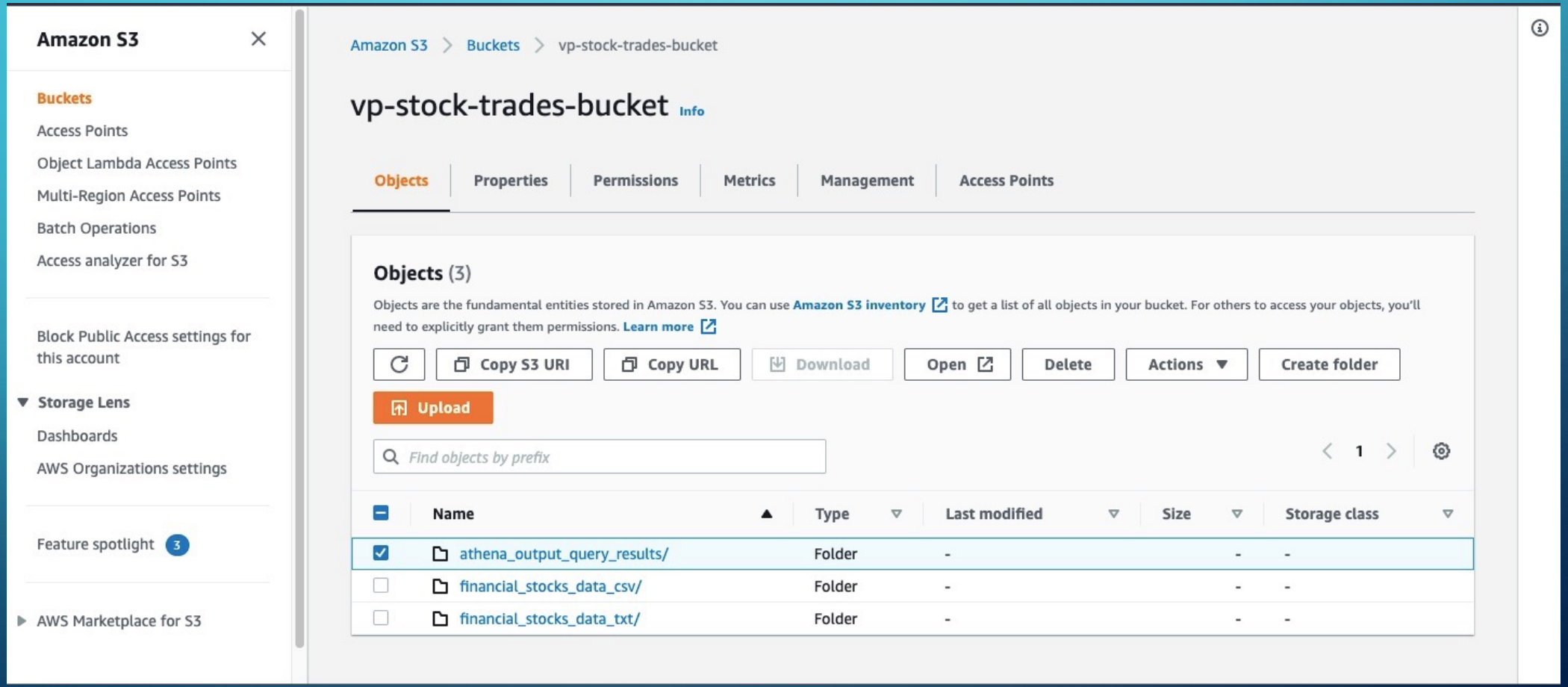


13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*



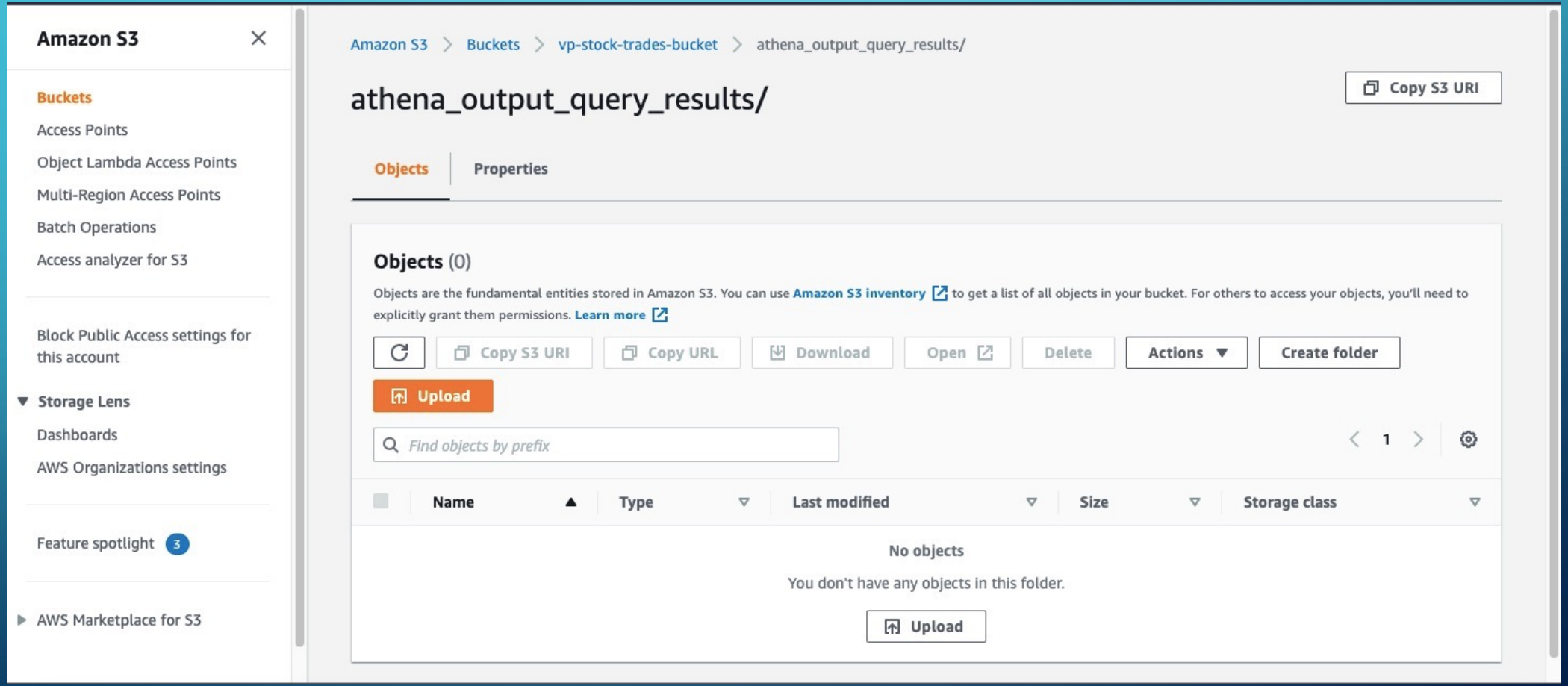
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

Creating a folder in the current AWS S3 Bucket to store the AWS Athena Query Output Results



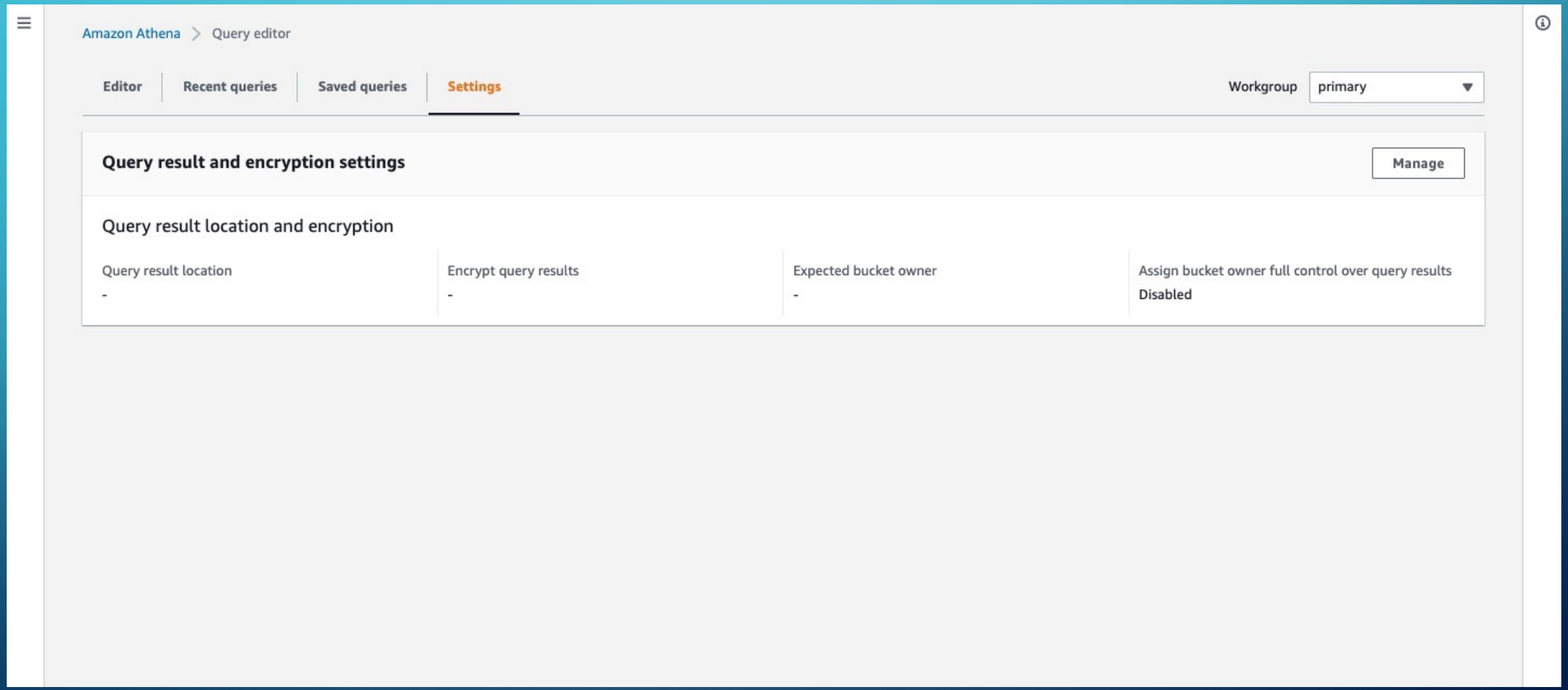
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

Creating a folder in the current AWS S3 Bucket to store the AWS Athena Query Output Results



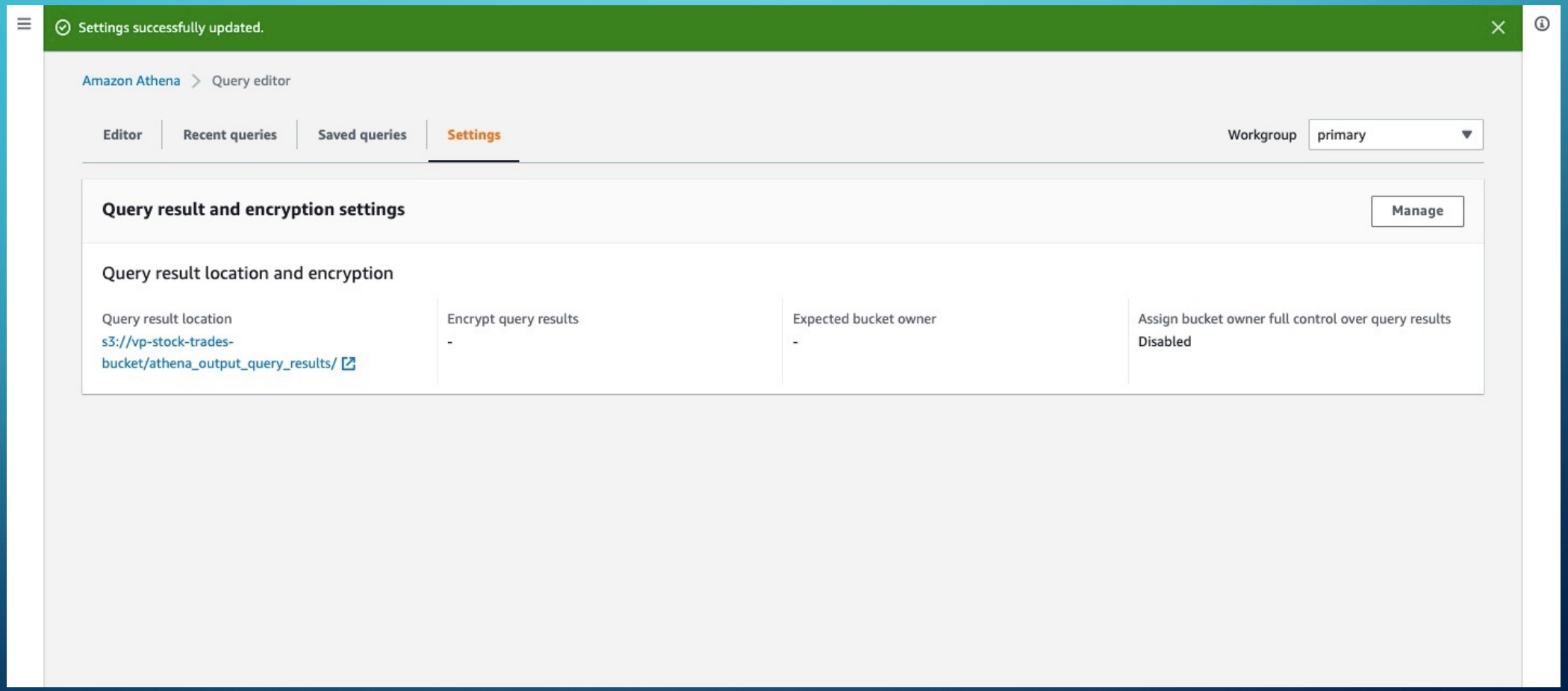
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

Setting-up the AWS S3 Bucket location to store the AWS Athena Query Output Results in the AWS Athena Settings



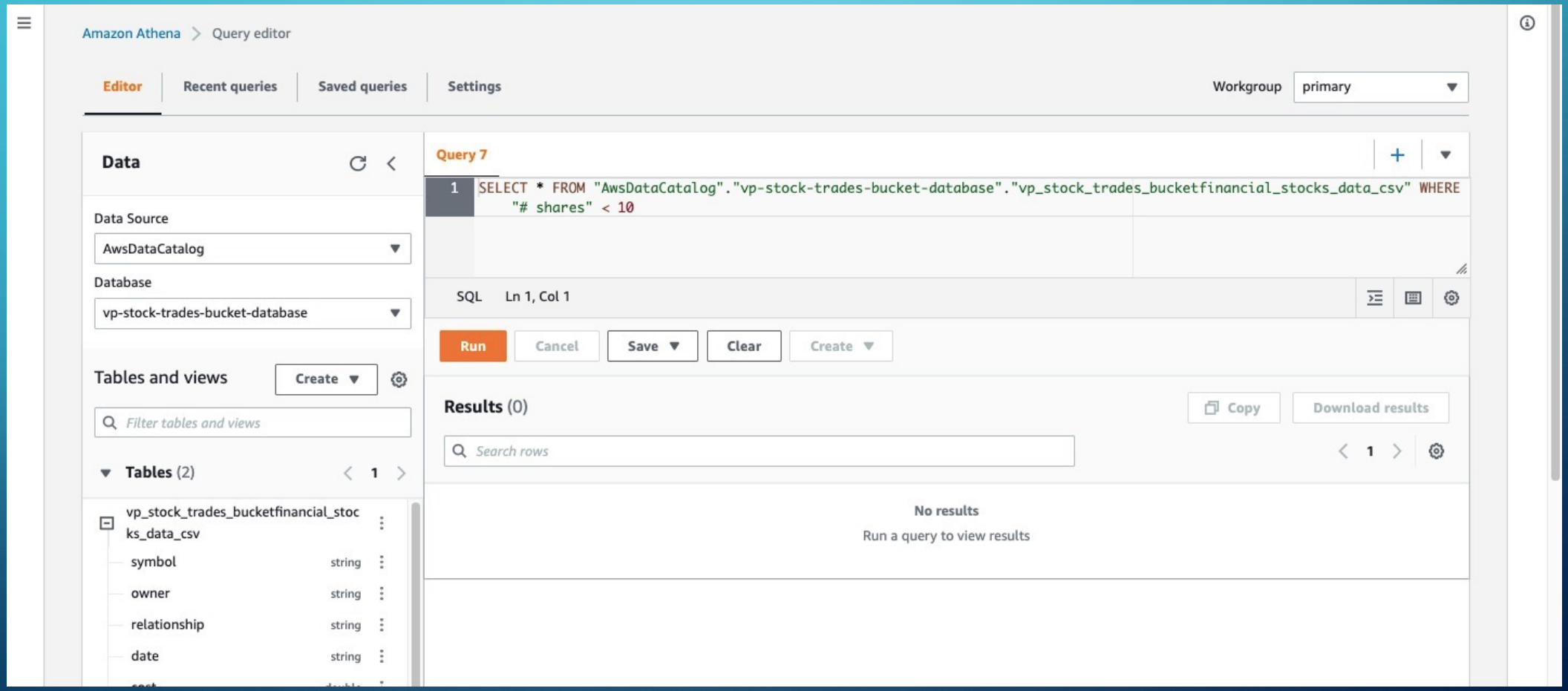
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

Setting-up the AWS S3 Bucket location to store the AWS Athena Query Output Results in the AWS Athena Settings



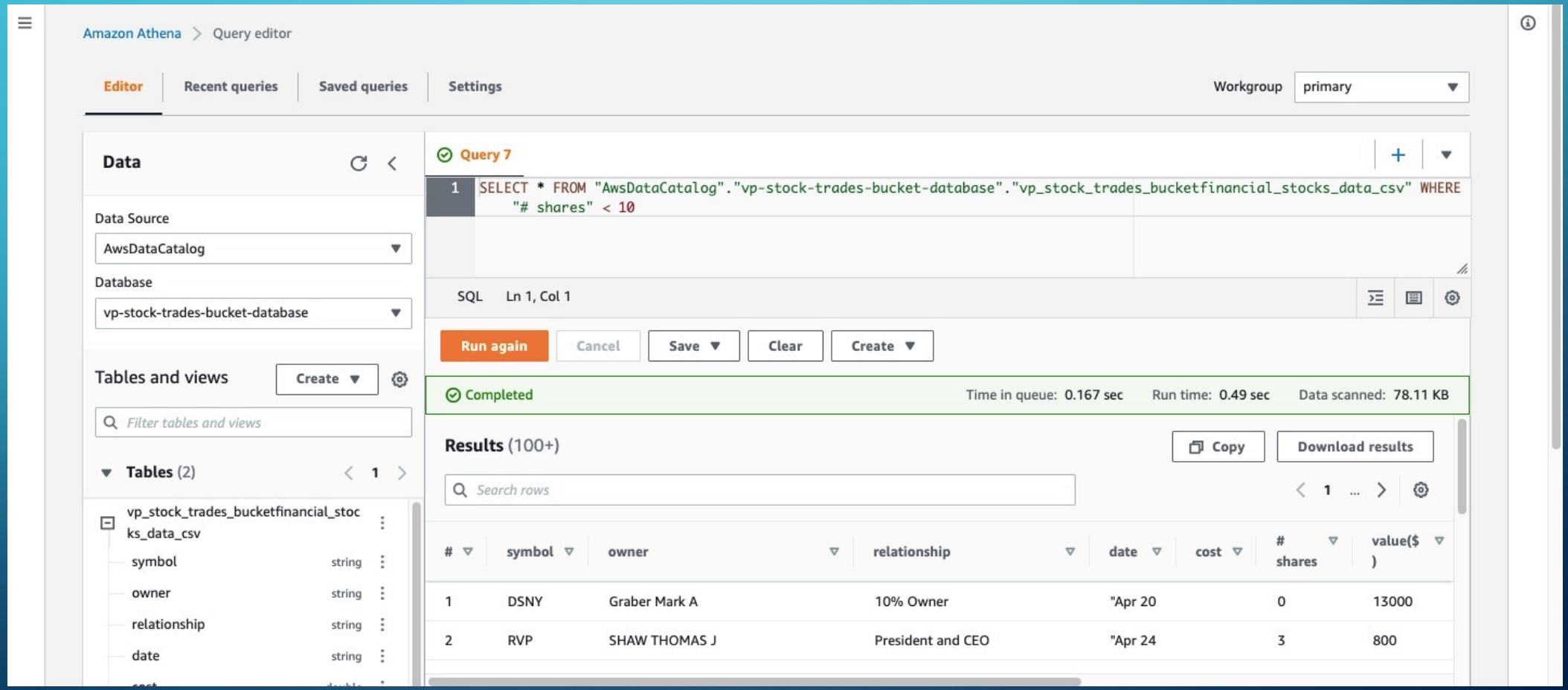
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

SQL Query Execution to query and filter the trade transactions for which number of shares processed is less than 10

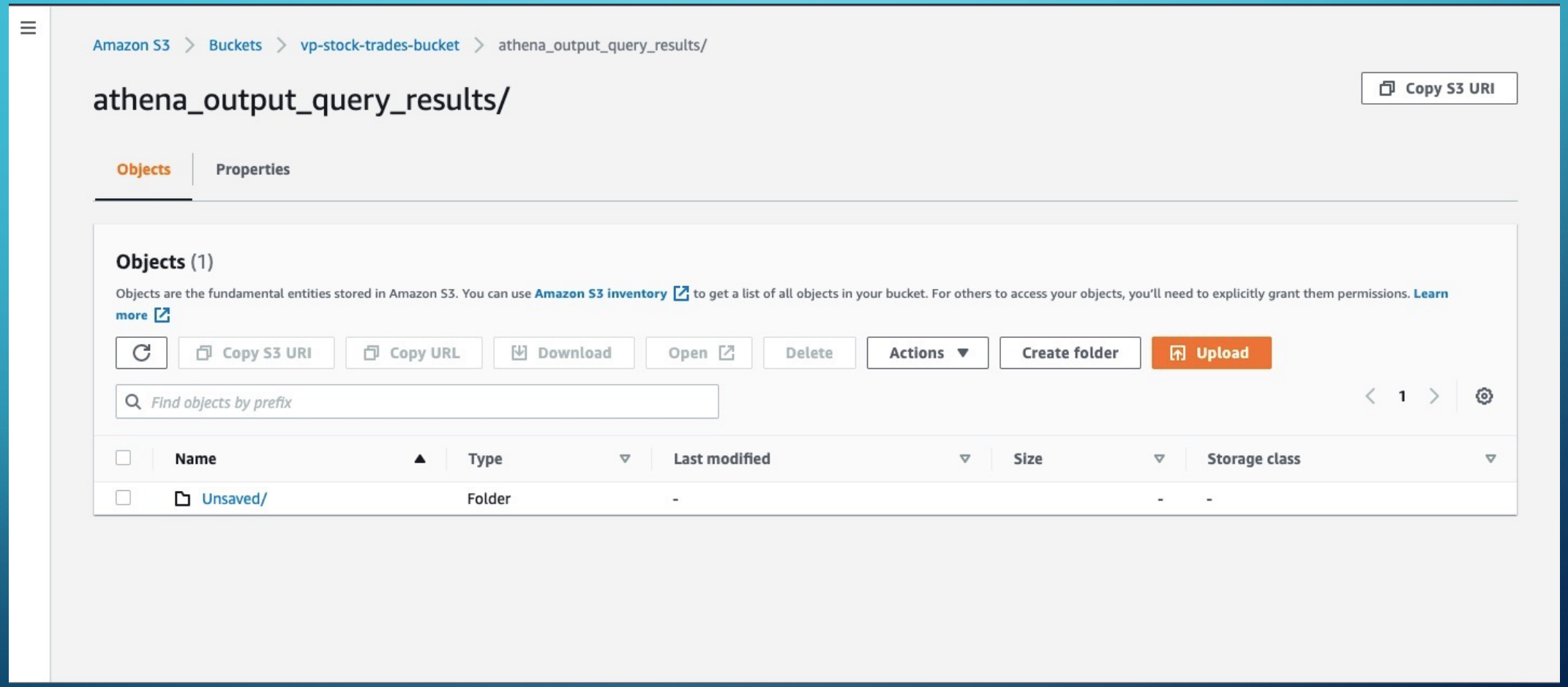


13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA

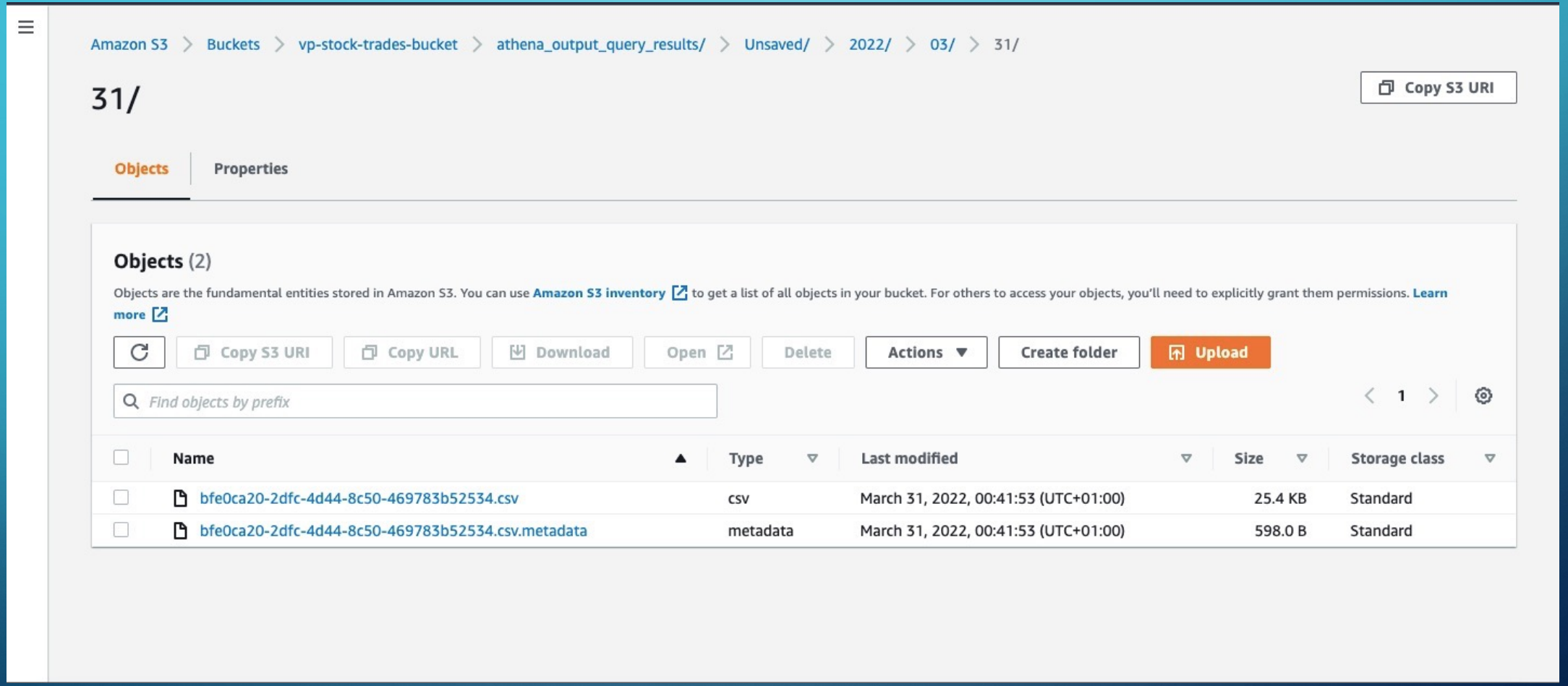
SQL Query Execution to query and filter the trade transactions for which number of shares processed is less than 10



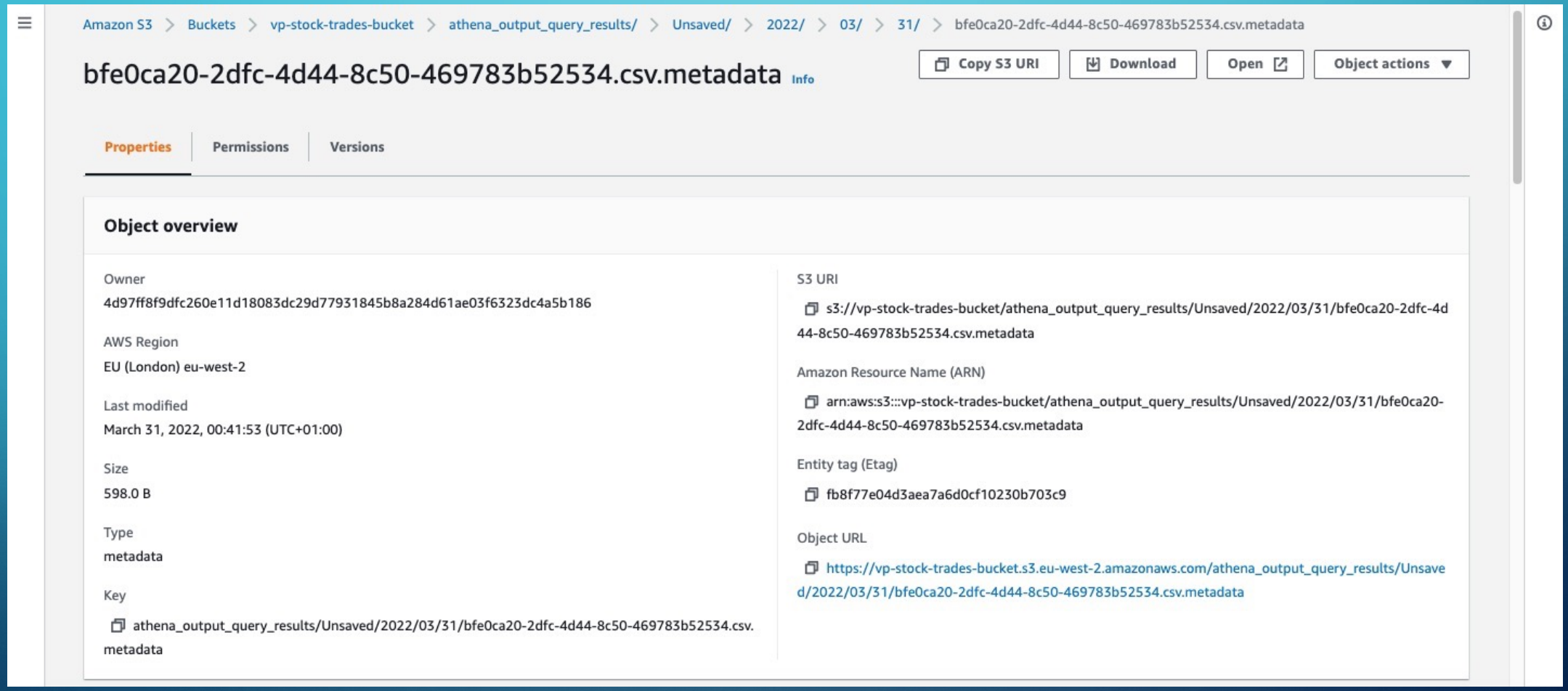
14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



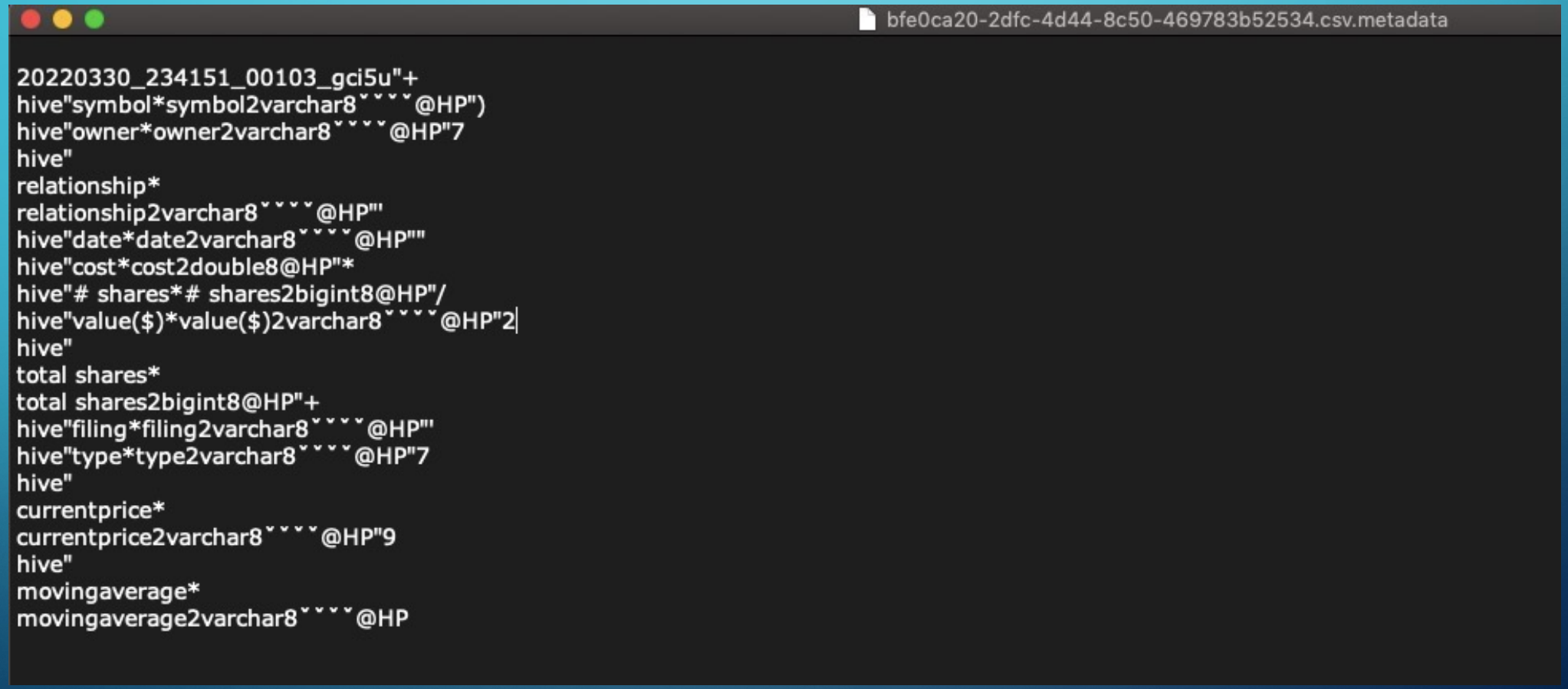
14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



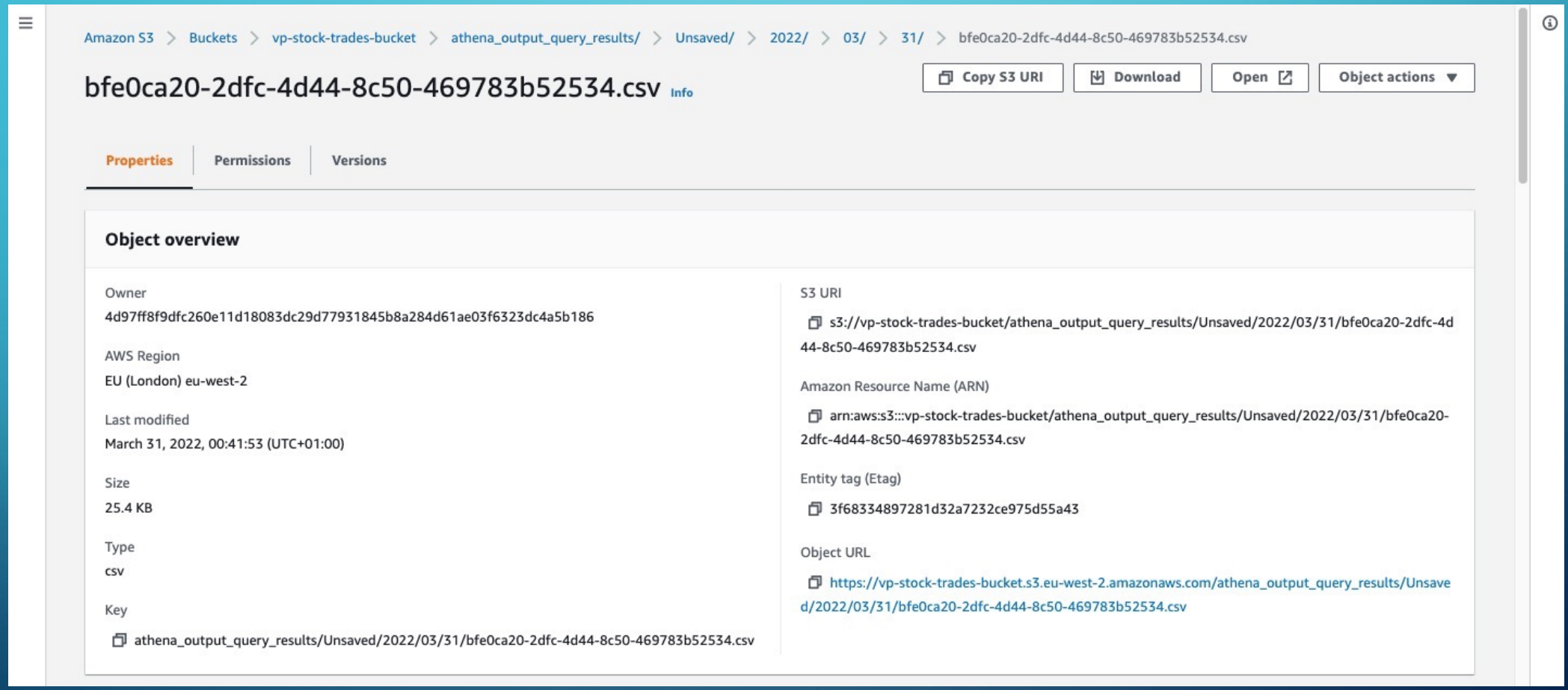
14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



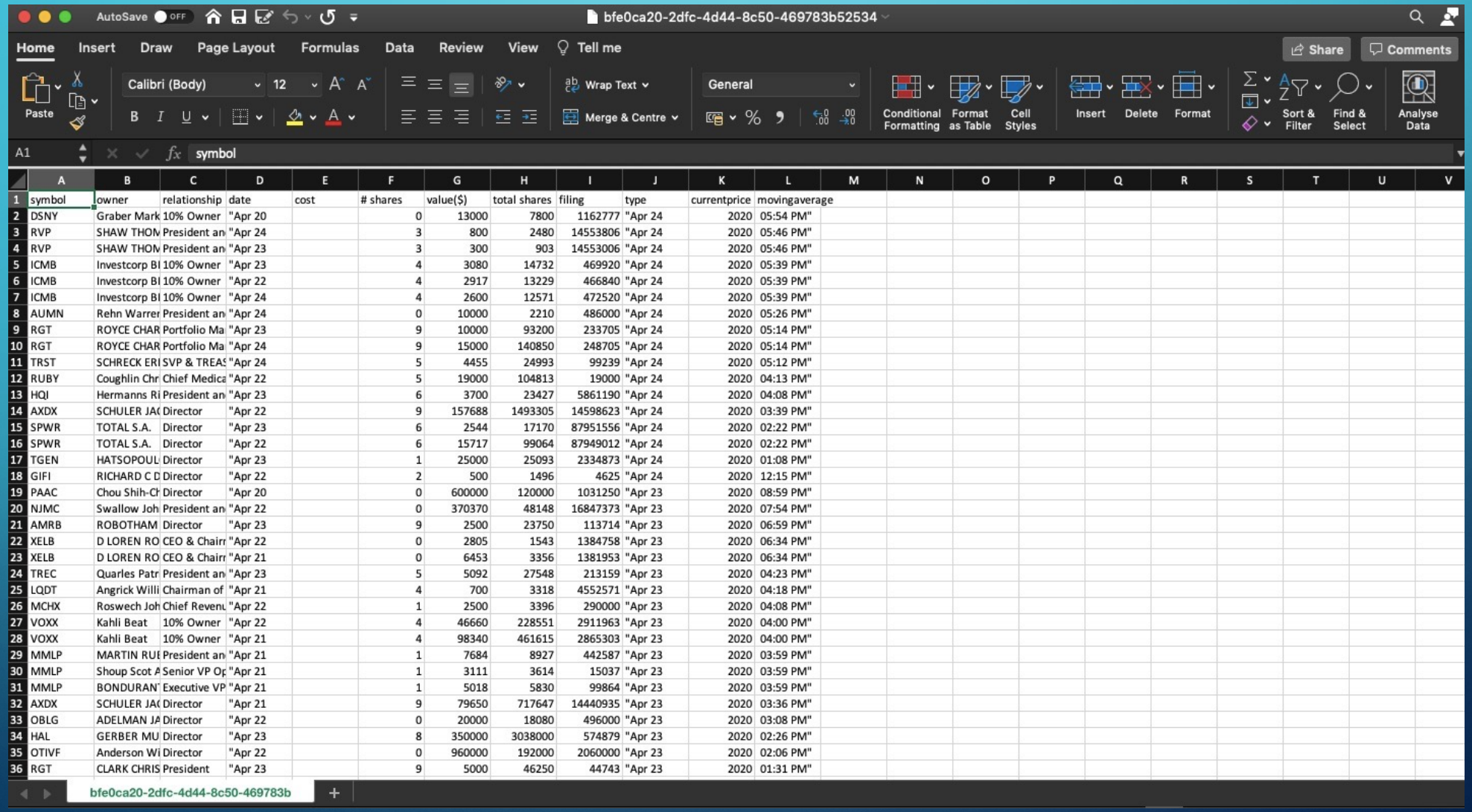
14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET



bfe0ca20-2dfc-4d44-8c50-469783b52534

	symbol	owner	relationship	date	cost	# shares	value($)	total shares	filing	type	currentprice	movingaverage
2	DSNY	Graber Mark	10% Owner	"Apr 20		0	13000	7800	1162777	"Apr 24	2020	05:54 PM"
3	RVP	SHAW THON	President an	"Apr 24		3	800	2480	14553806	"Apr 24	2020	05:46 PM"
4	RVP	SHAW THON	President an	"Apr 23		3	300	903	14553006	"Apr 24	2020	05:46 PM"
5	ICMB	Investcorp Bl	10% Owner	"Apr 23		4	3080	14732	469920	"Apr 24	2020	05:39 PM"
6	ICMB	Investcorp Bl	10% Owner	"Apr 22		4	2917	13229	466840	"Apr 24	2020	05:39 PM"
7	ICMB	Investcorp Bl	10% Owner	"Apr 24		4	2600	12571	472520	"Apr 24	2020	05:39 PM"
8	AUMN	Rehn Warrer	President an	"Apr 24		0	10000	2210	486000	"Apr 24	2020	05:26 PM"
9	RGT	ROYCE CHAR	Portfolio Ma	"Apr 23		9	10000	93200	233705	"Apr 24	2020	05:14 PM"
10	RGT	ROYCE CHAR	Portfolio Ma	"Apr 24		9	15000	140850	248705	"Apr 24	2020	05:14 PM"
11	TRST	SCHRECK ERI	SVP & TREAS	"Apr 24		5	4455	24993	99239	"Apr 24	2020	05:12 PM"
12	RUBY	Coughlin Chr	Chief Medica	"Apr 22		5	19000	104813	19000	"Apr 24	2020	04:13 PM"
13	HQI	Hermanns Ri	President an	"Apr 23		6	3700	23427	5861190	"Apr 24	2020	04:08 PM"
14	AXDX	SCHULER JA(	Director	"Apr 22		9	157688	1493305	14598623	"Apr 24	2020	03:39 PM"
15	SPWR	TOTAL S.A.	Director	"Apr 23		6	2544	17170	87951556	"Apr 24	2020	02:22 PM"
16	SPWR	TOTAL S.A.	Director	"Apr 22		6	15717	99064	87949012	"Apr 24	2020	02:22 PM"
17	TGEN	HATSOPOUL	Director	"Apr 23		1	25000	25093	2334873	"Apr 24	2020	01:08 PM"
18	GIFI	RICHARD C D	Director	"Apr 22		2	500	1496	4625	"Apr 24	2020	12:15 PM"
19	PAAC	Chou Shih-Ch	Director	"Apr 20		0	600000	120000	1031250	"Apr 23	2020	08:59 PM"
20	NJMC	Swallow Joh	President an	"Apr 22		0	370370	48148	16847373	"Apr 23	2020	07:54 PM"
21	AMRB	ROBOTHAM	Director	"Apr 23		9	2500	23750	113714	"Apr 23	2020	06:59 PM"
22	XELB	D LOREN RO	CEO & Chairr	"Apr 22		0	2805	1543	1384758	"Apr 23	2020	06:34 PM"
23	XELB	D LOREN RO	CEO & Chairr	"Apr 21		0	6453	3356	1381953	"Apr 23	2020	06:34 PM"
24	TREC	Quarles Patr	President an	"Apr 23		5	5092	27548	213159	"Apr 23	2020	04:23 PM"
25	LQDT	Angrick Willi	Chairman of	"Apr 21		4	700	3318	4552571	"Apr 23	2020	04:18 PM"
26	MCHX	Roswech Joh	Chief Revenu	"Apr 22		1	2500	3396	290000	"Apr 23	2020	04:08 PM"
27	VOXX	Kahli Beat	10% Owner	"Apr 22		4	46660	228551	2911963	"Apr 23	2020	04:00 PM"
28	VOXX	Kahli Beat	10% Owner	"Apr 21		4	98340	461615	2865303	"Apr 23	2020	04:00 PM"
29	MMLP	MARTIN RUI	President an	"Apr 21		1	7684	8927	442587	"Apr 23	2020	03:59 PM"
30	MMLP	Shoup Scot A	Senior VP Op	"Apr 21		1	3111	3614	15037	"Apr 23	2020	03:59 PM"
31	MMLP	BONDURAN	Executive VP	"Apr 21		1	5018	5830	99864	"Apr 23	2020	03:59 PM"
32	AXDX	SCHULER JA(	Director	"Apr 21		9	79650	717647	14440935	"Apr 23	2020	03:36 PM"
33	OBLG	ADELMAN JA	Director	"Apr 22		0	20000	18080	496000	"Apr 23	2020	03:08 PM"
34	HAL	GERBER MU	Director	"Apr 23		8	350000	3038000	574879	"Apr 23	2020	02:26 PM"
35	OTIVF	Anderson Wi	Director	"Apr 22		0	960000	192000	2060000	"Apr 23	2020	02:06 PM"
36	RGT	CLARK CHRIS	President	"Apr 23		9	5000	46250	44743	"Apr 23	2020	01:31 PM"

bfe0ca20-2dfc-4d44-8c50-469783b

END OF THE PROJECT