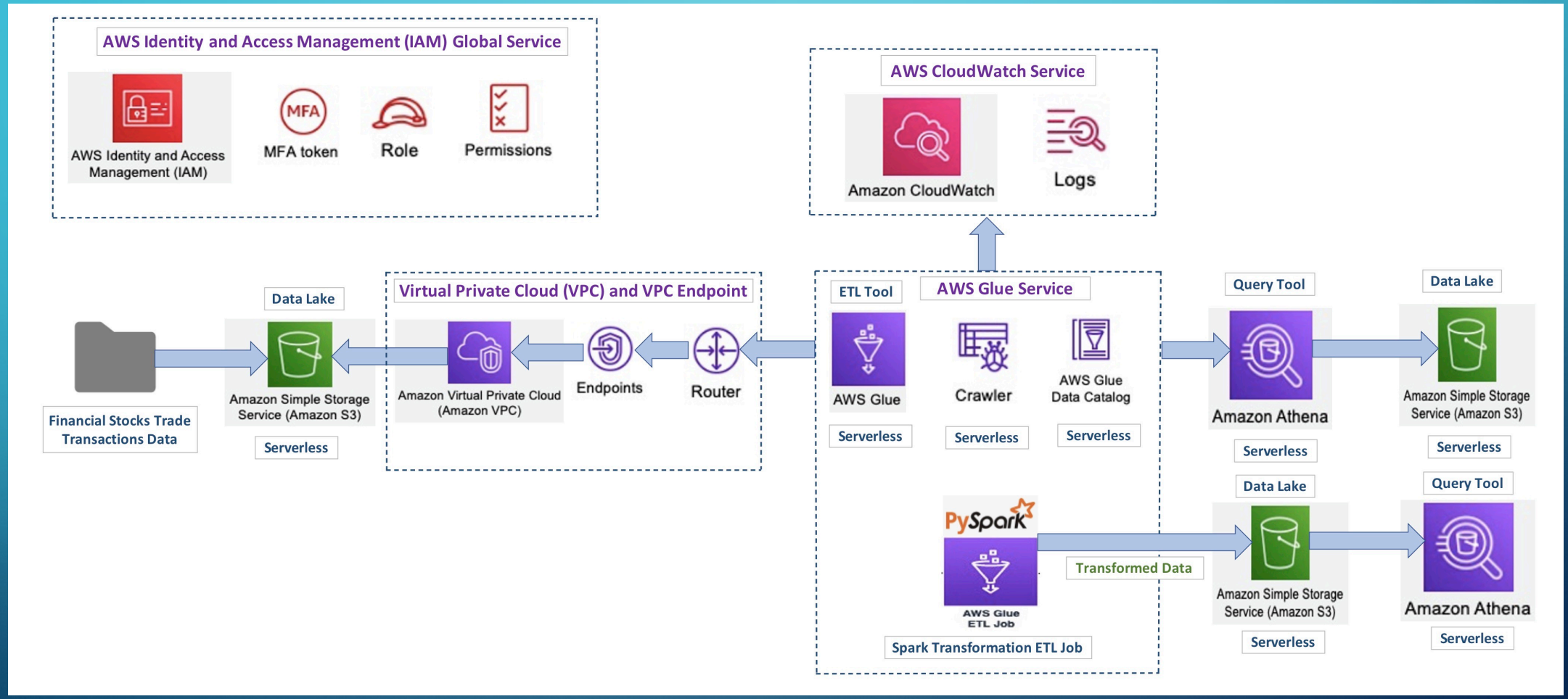
BUILDING A FINANCIAL DATA PIPELINE WITH SPARK TRANSFORMATION USING PYTHON AND AWS

FINANCIAL STOCKS TRADE TRANSACTIONS DATA

PROJECT ARCHITECTURE



TECHNOLOGY STACK

Programming/Scripting Language:

- Python

Query Language:

- SQL

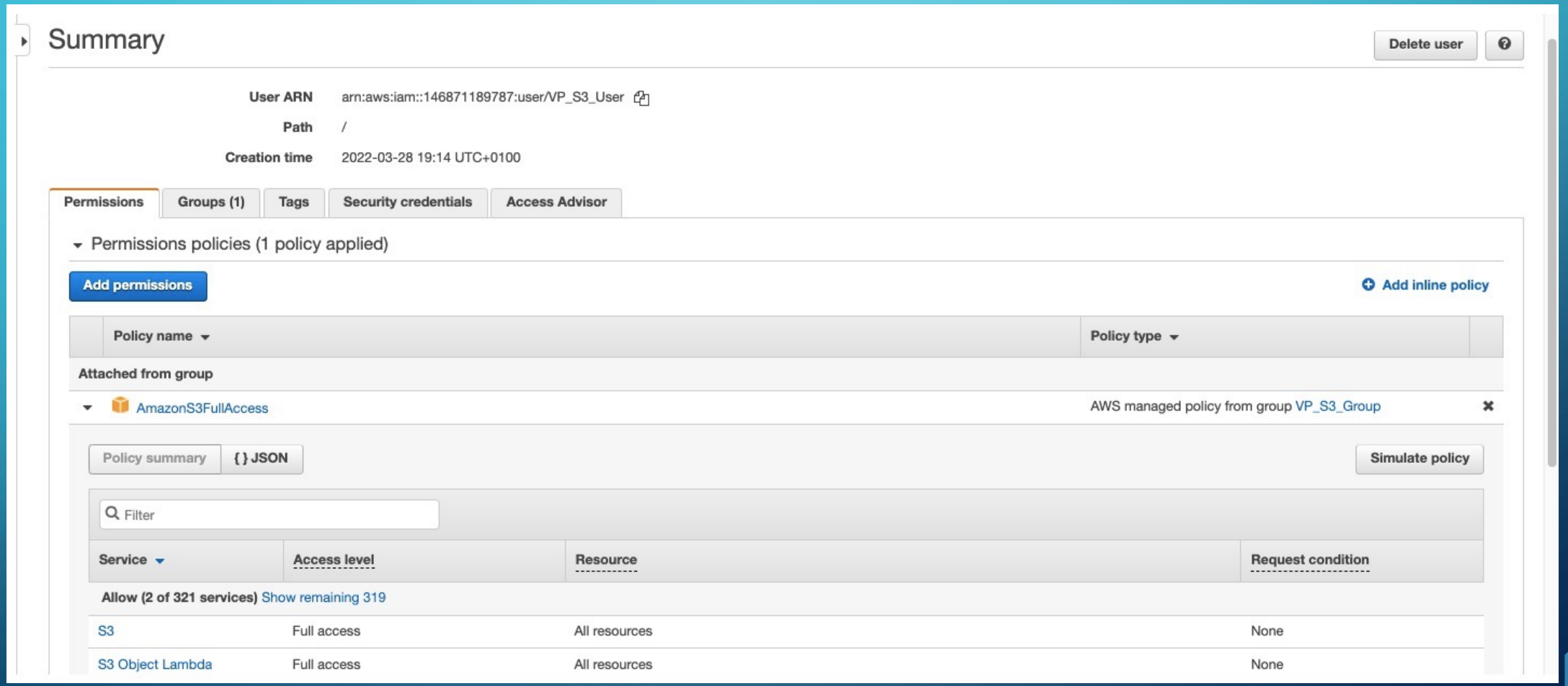
Command Line Interface (CLI):

- AWS CLI

AWS Services:

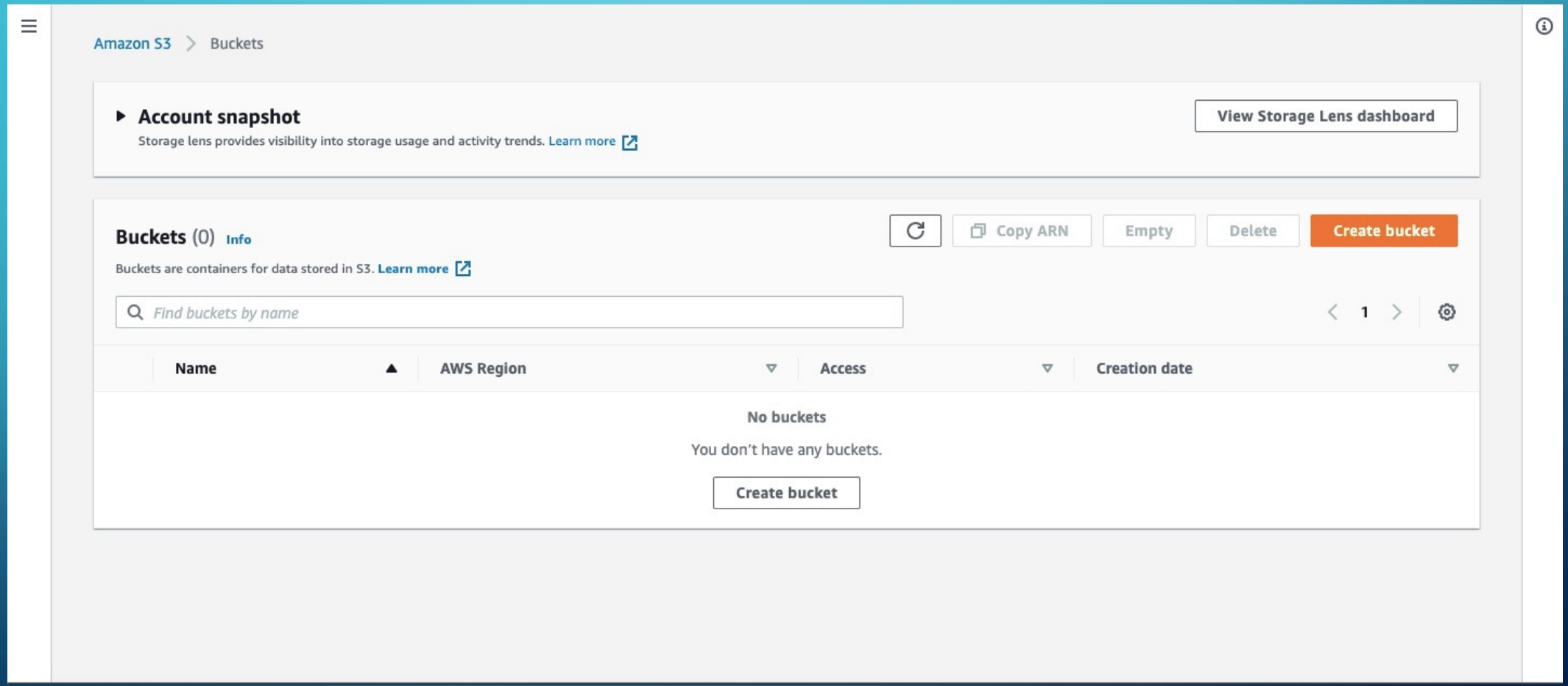
- Identity and Access Management (IAM)
- Virtual Private Cloud (VPC) and VPC Endpoint
- Single Storage Service (S3)
- Glue (Glue Crawler, Glue Catalog, Glue Database & Tables, ETL Job, Spark Transformation)
- CloudWatch
- Athena
- Spark

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



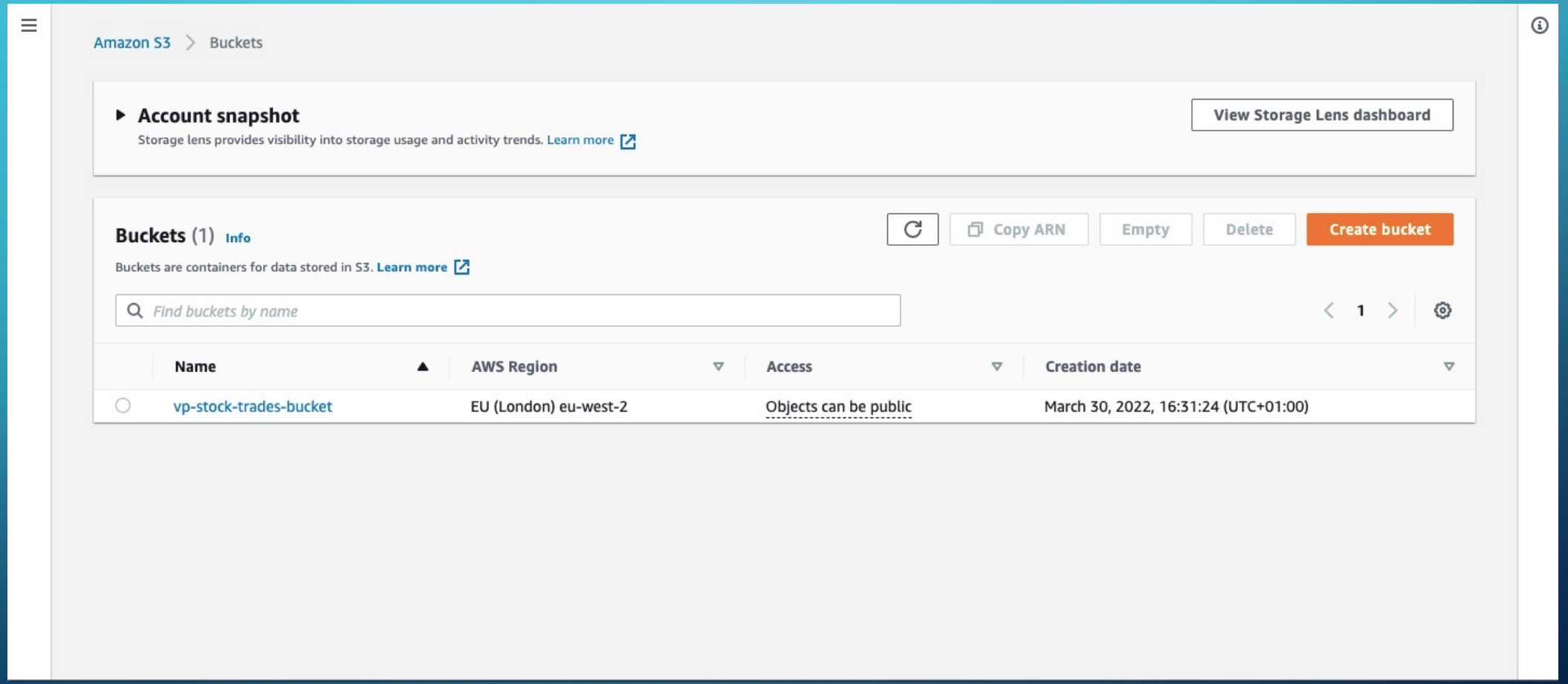
2. CREATION OF AWS S3 BUCKET *(CONTINUED)*

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

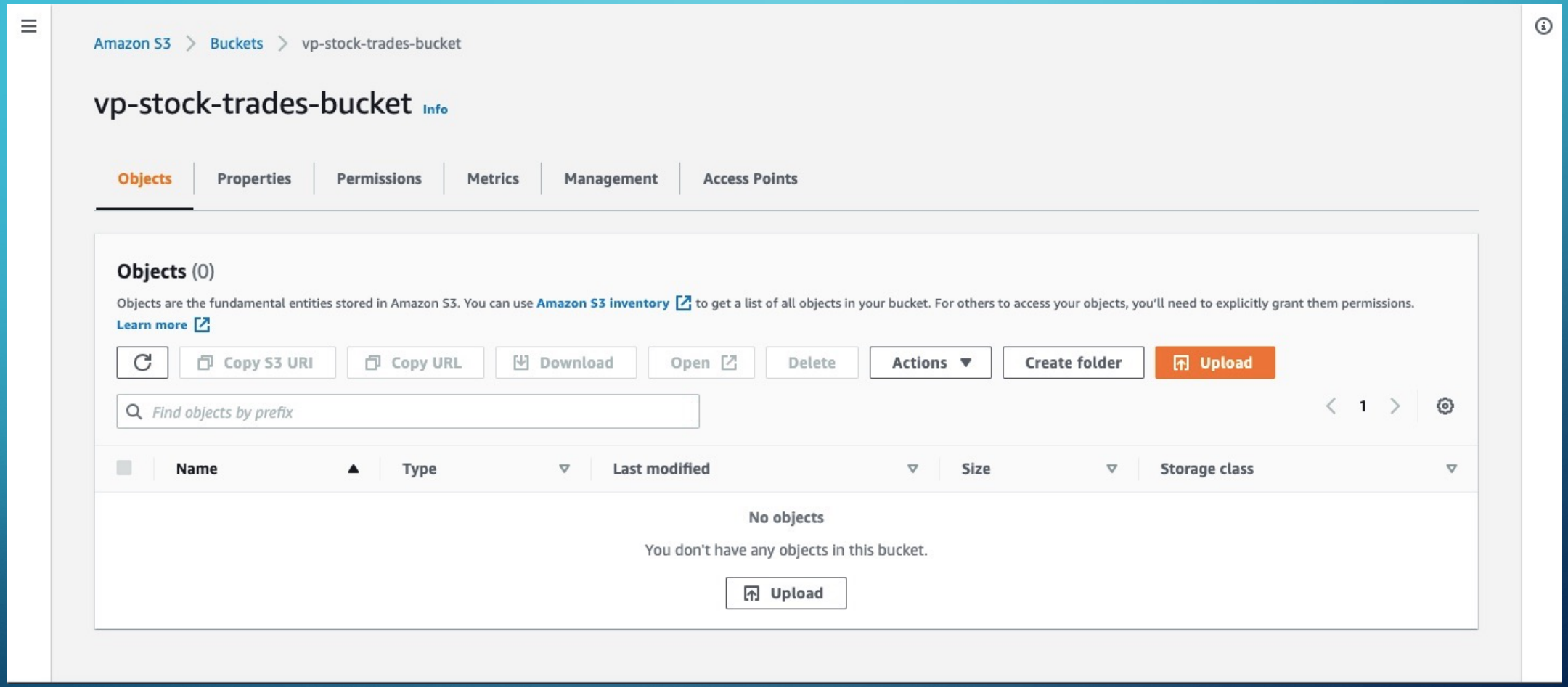


2. CREATION OF AWS S3 BUCKET *(CONTINUED)*

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)*

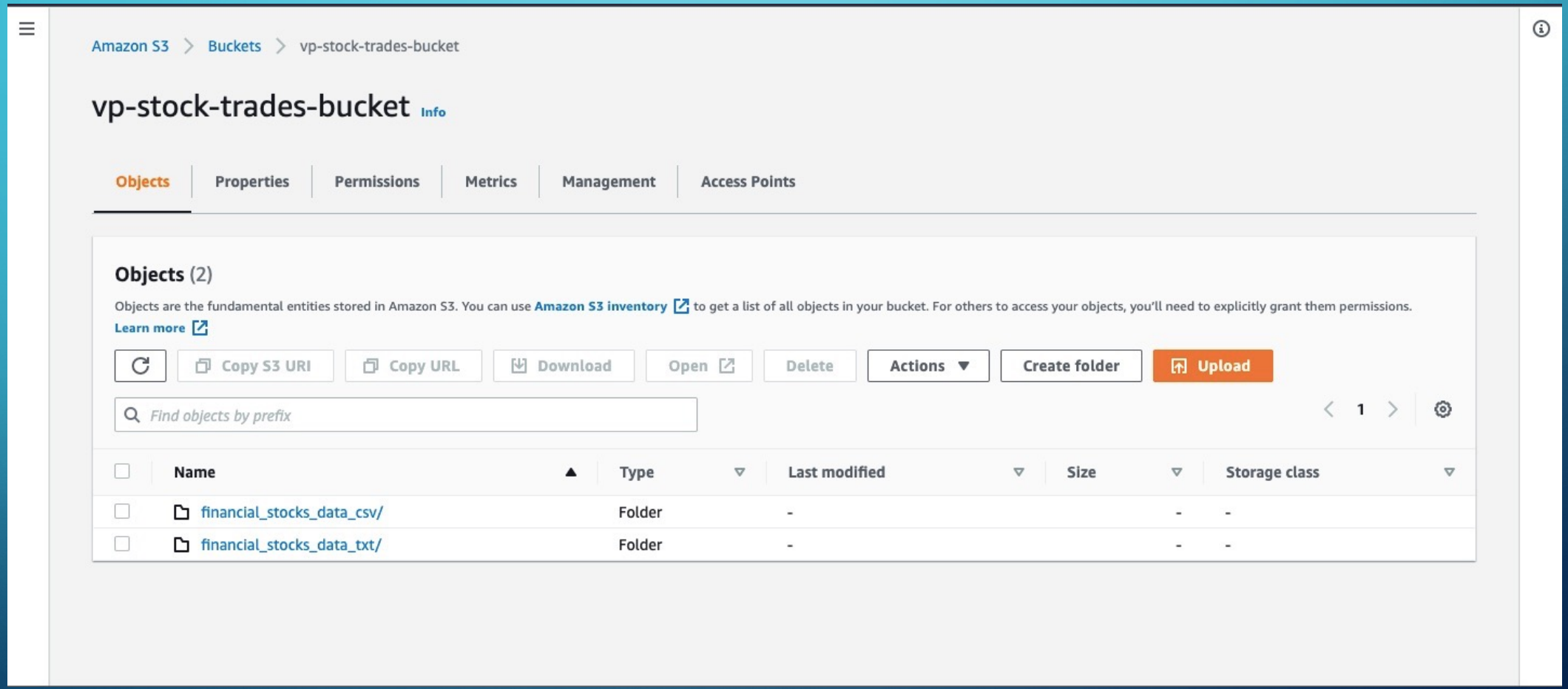
main.py

```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)

```

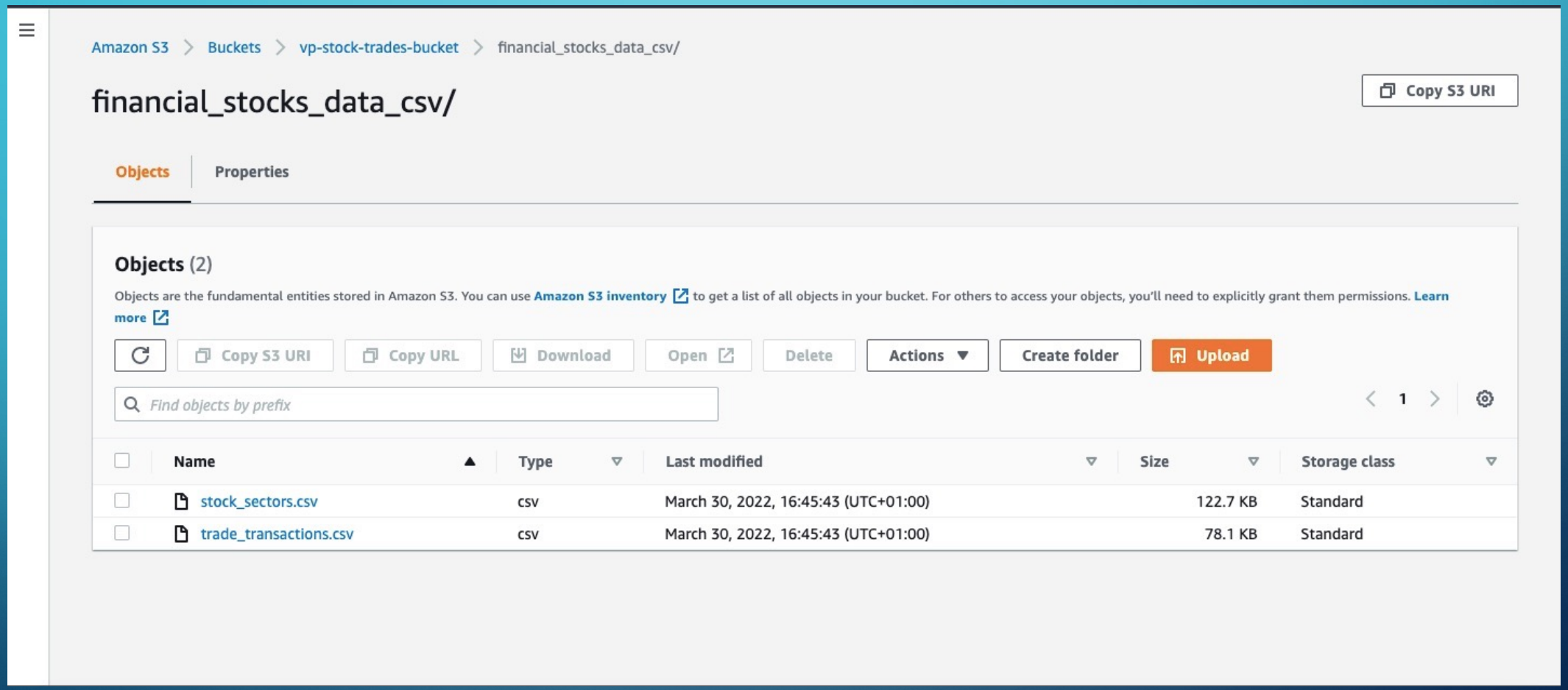
Terminal: Local

```
(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$
```

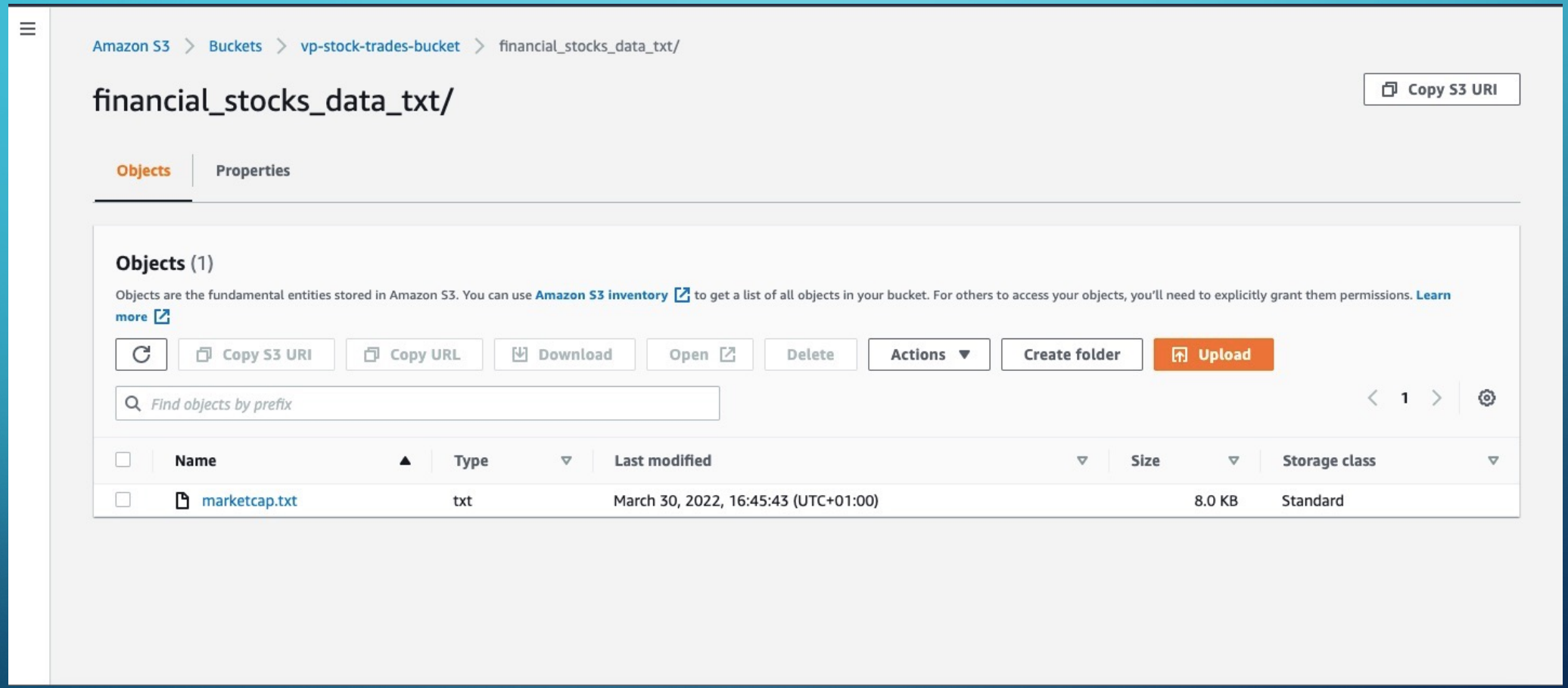
3. INGESTING SOURCE DATA FILES IN TO AWS S3 BUCKET *(CONTINUED)*



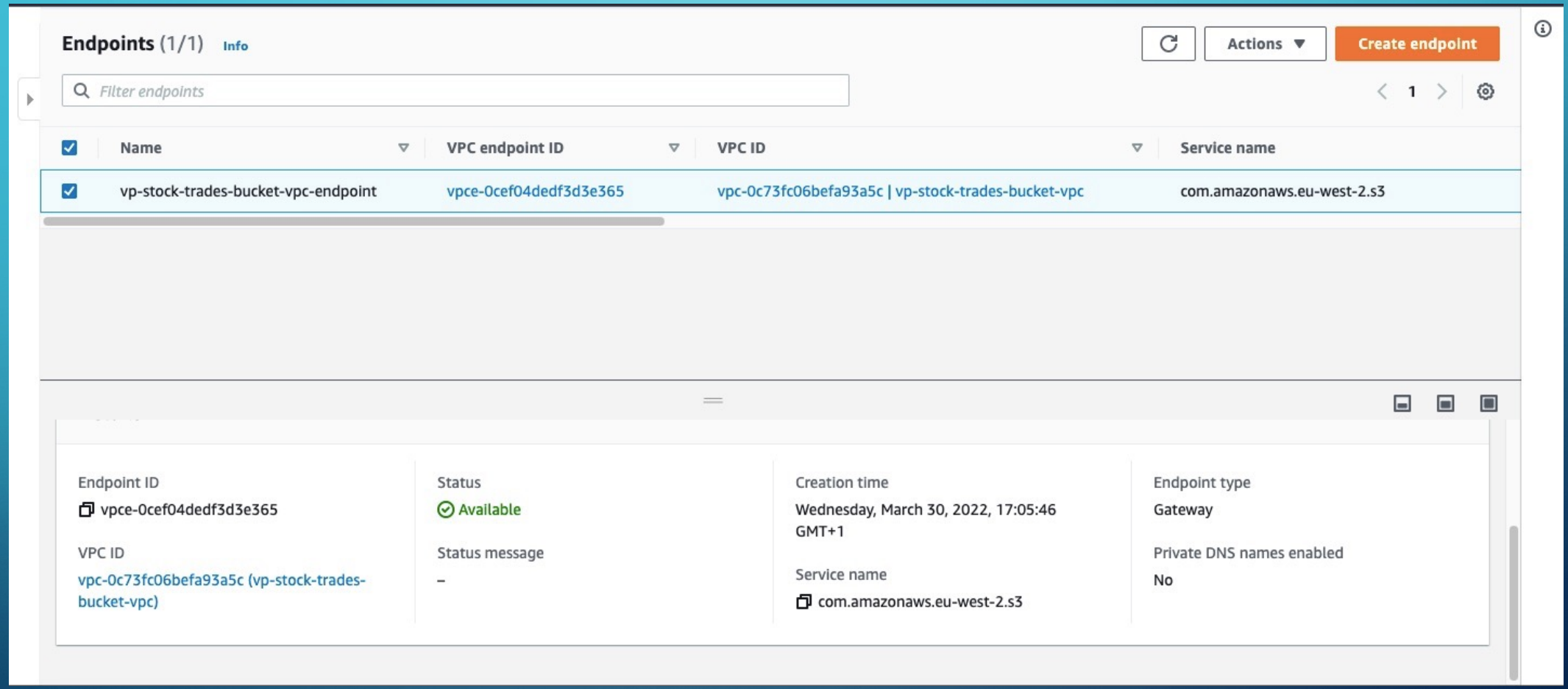
3. INGESTING SOURCE DATA FILES IN TO AWS S3 BUCKET *(CONTINUED)*



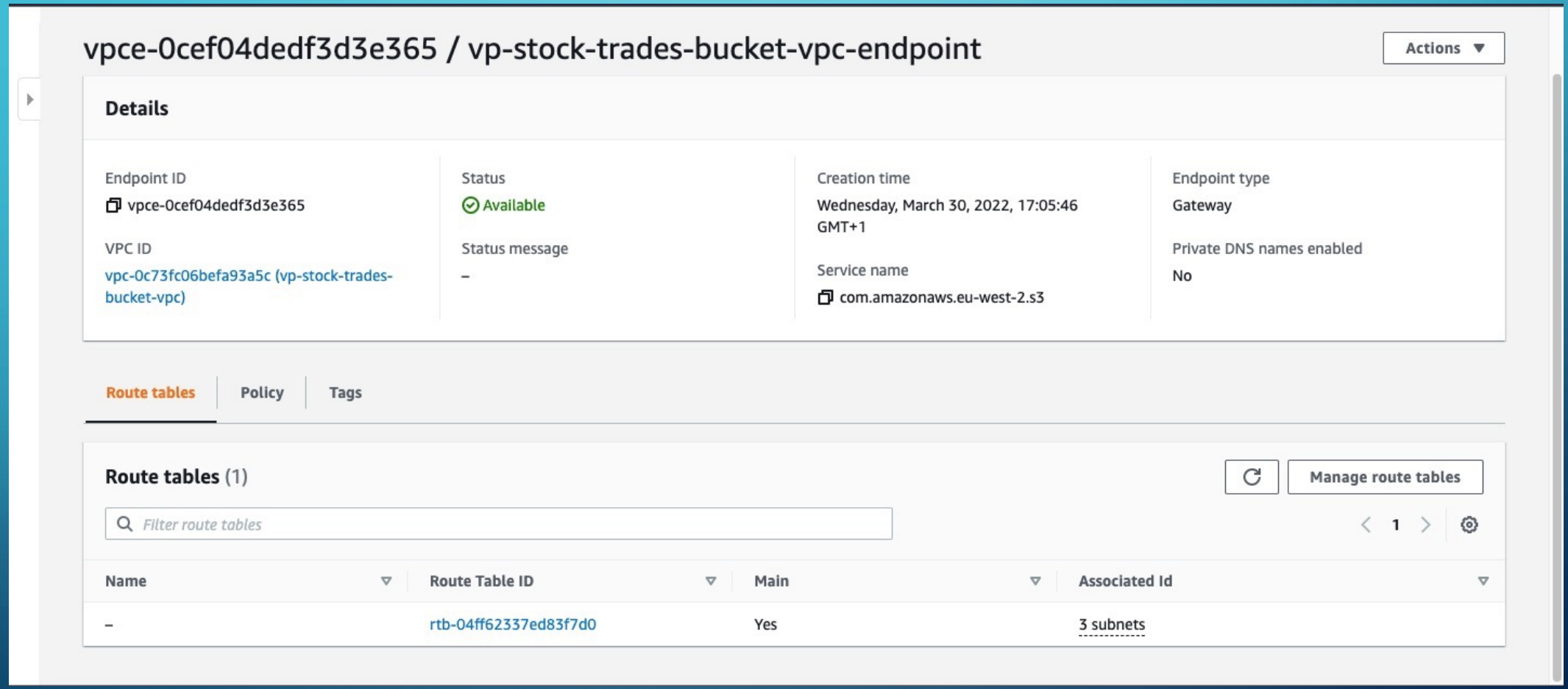
3. INGESTING SOURCE DATA FILES IN TO AWS S3 BUCKET



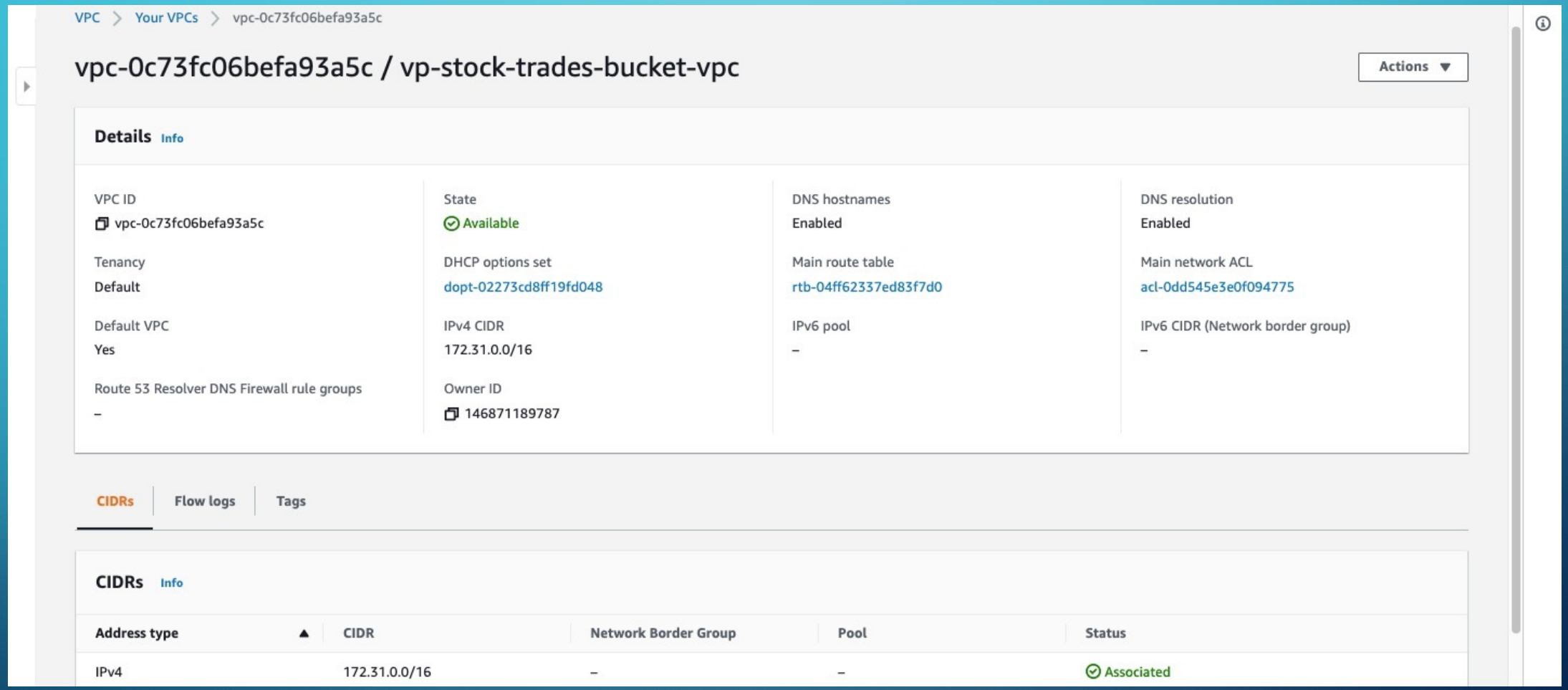
4. CREATION OF A VPC ENDPOINT INTERFACE TO ENABLE AWS GLUE TO ACCESS AWS S3 *(CONTINUED)*



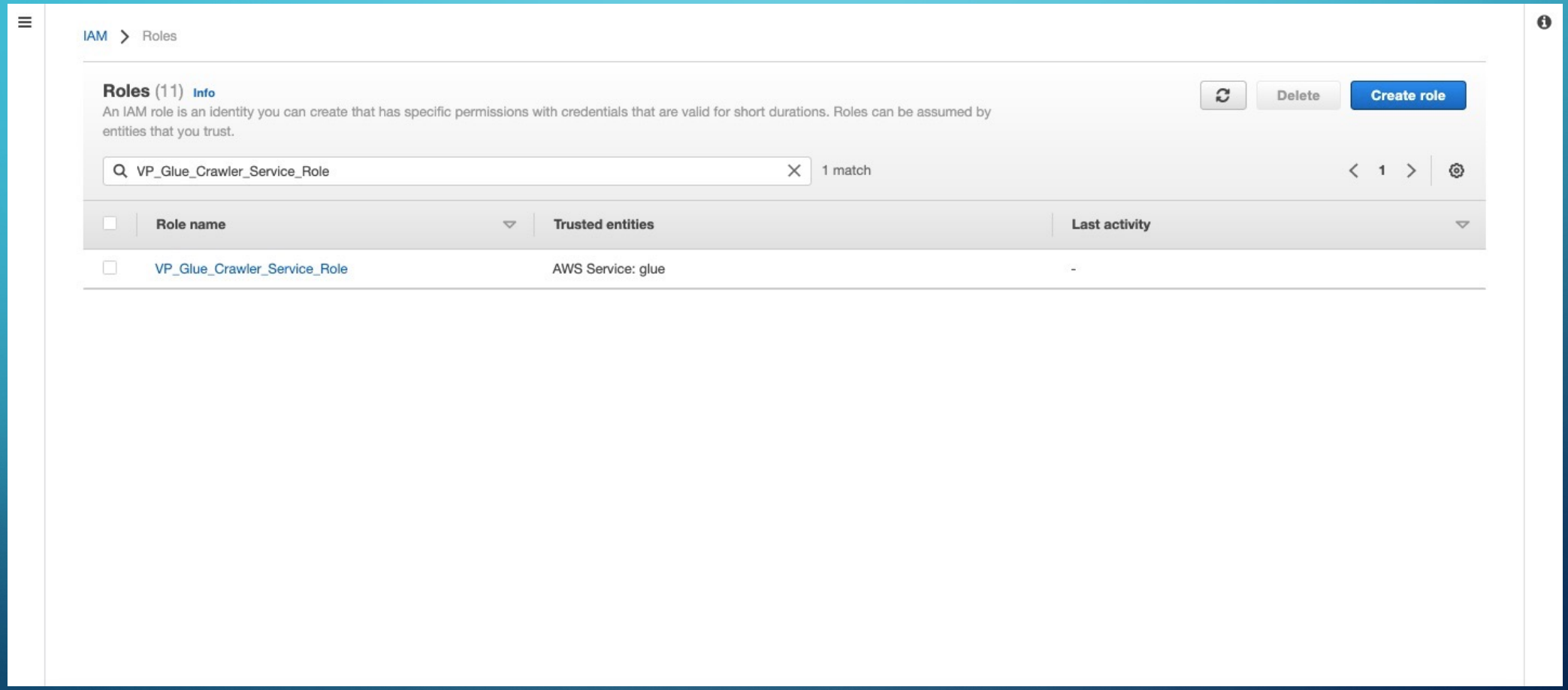
4. CREATION OF A VPC ENDPOINT INTERFACE TO ENABLE AWS GLUE TO ACCESS AWS S3 *(CONTINUED)*



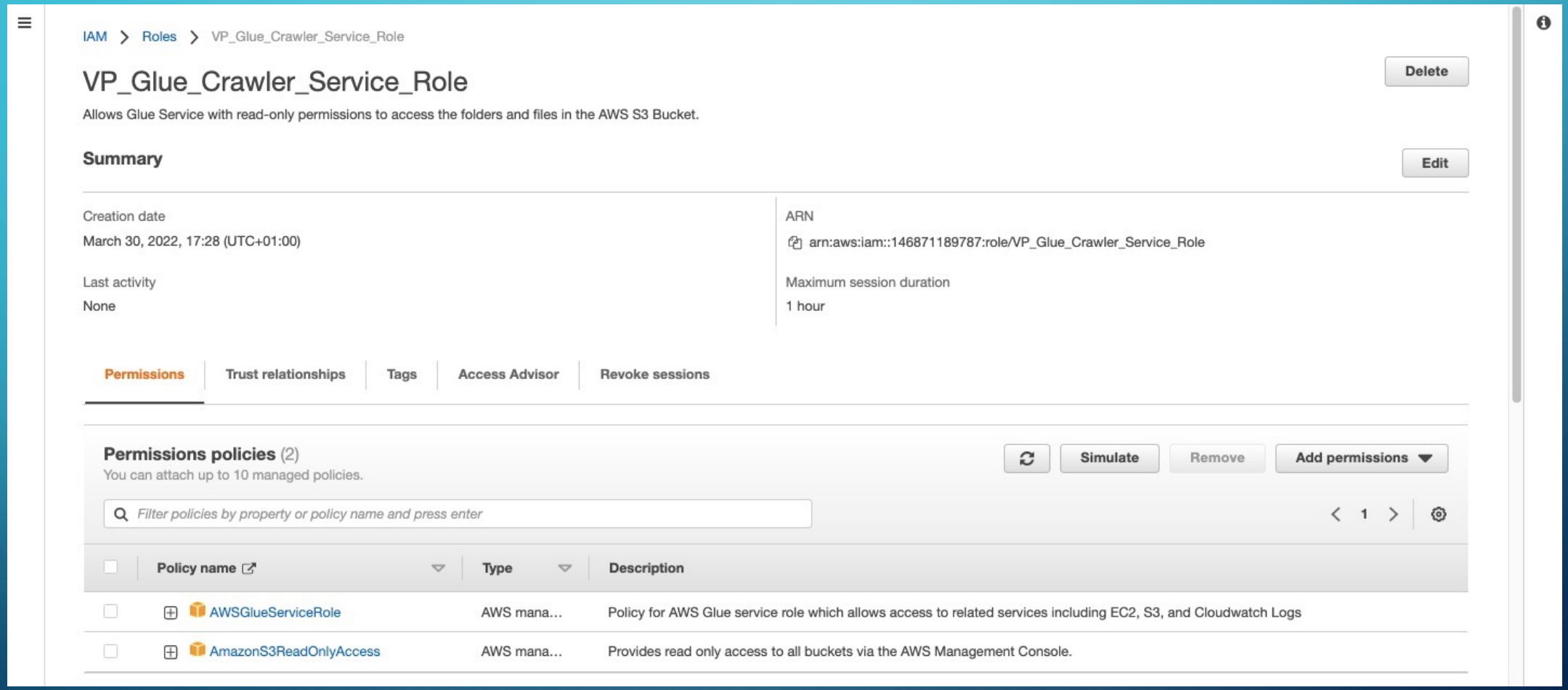
4. CREATION OF A VPC ENDPOINT INTERFACE TO ENABLE AWS GLUE TO ACCESS AWS S3



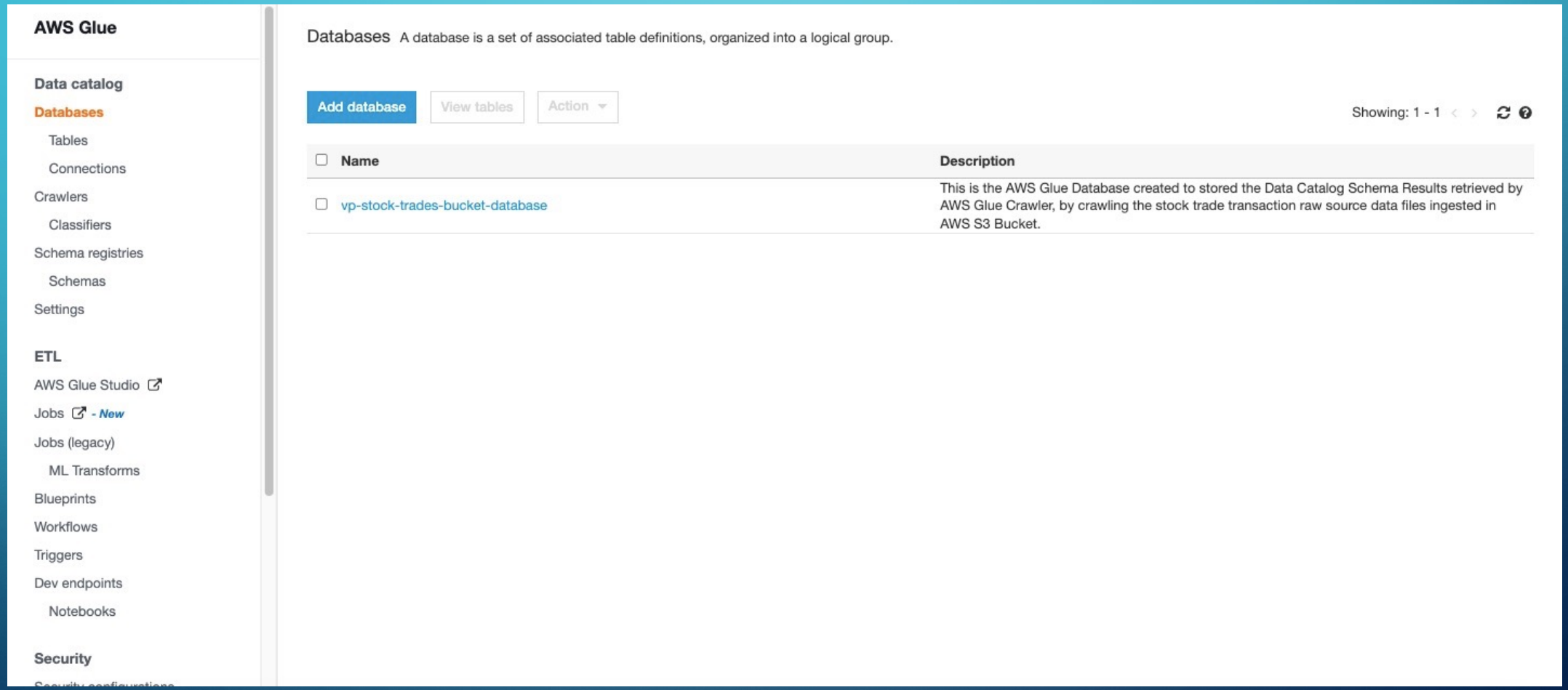
5. CREATION OF AN IAM ROLE TO GRANT PERMISSIONS TO AWS GLUE CRAWLER TO ACCESS AWS S3 *(CONTINUED)*



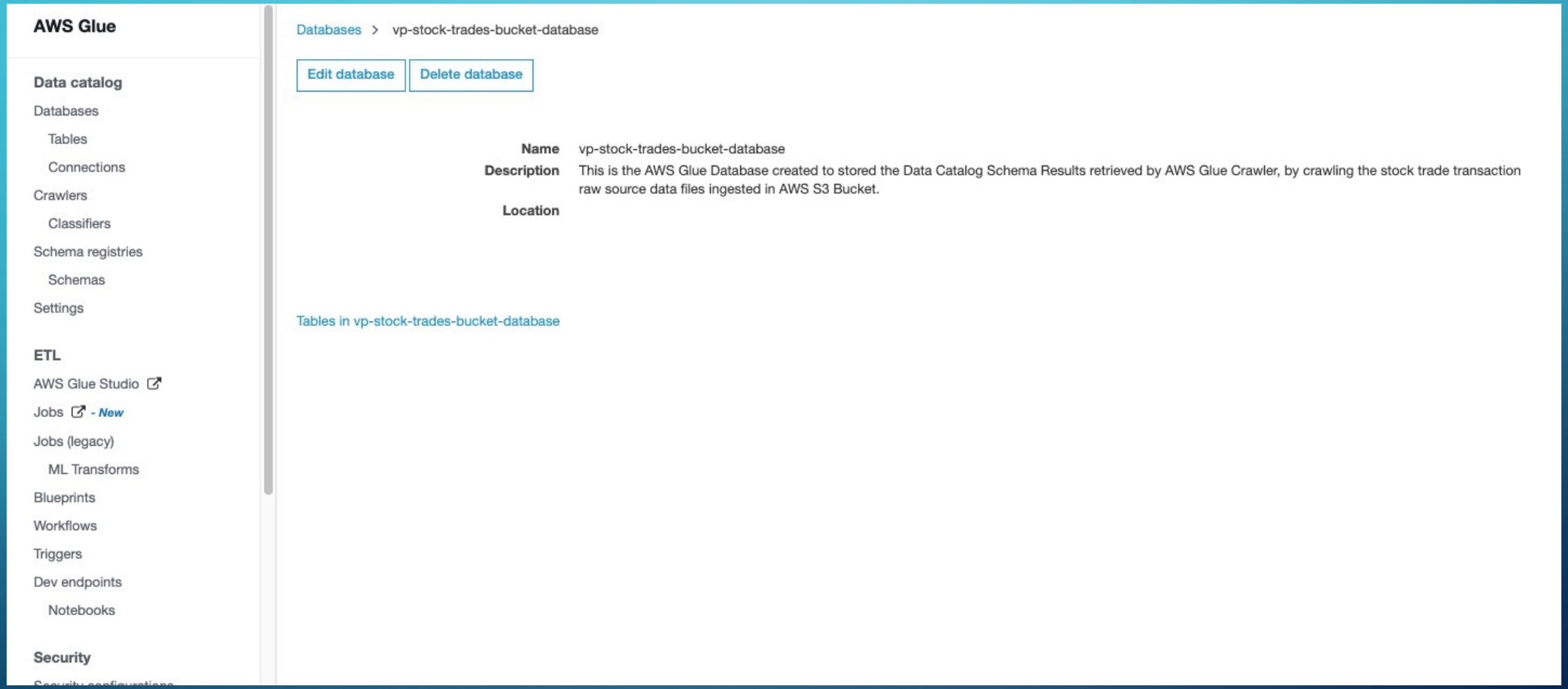
5. CREATION OF AN IAM ROLE TO GRANT PERMISSIONS TO AWS GLUE CRAWLER TO ACCESS AWS S3



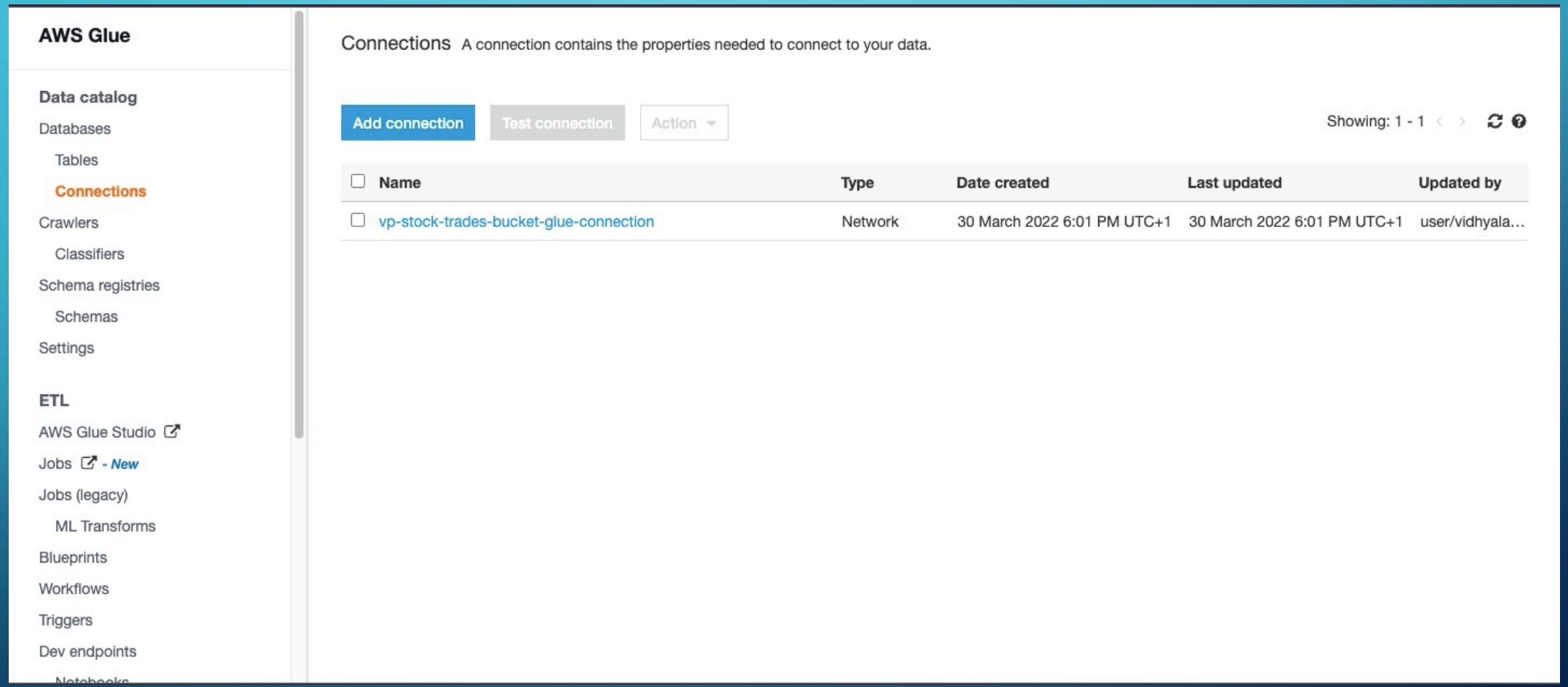
6. CREATION OF A DATABASE IN THE AWS GLUE TO STORE THE GLUE CRAWLER DATA CATALOG SCHEMA RESULTS *(CONTINUED)*



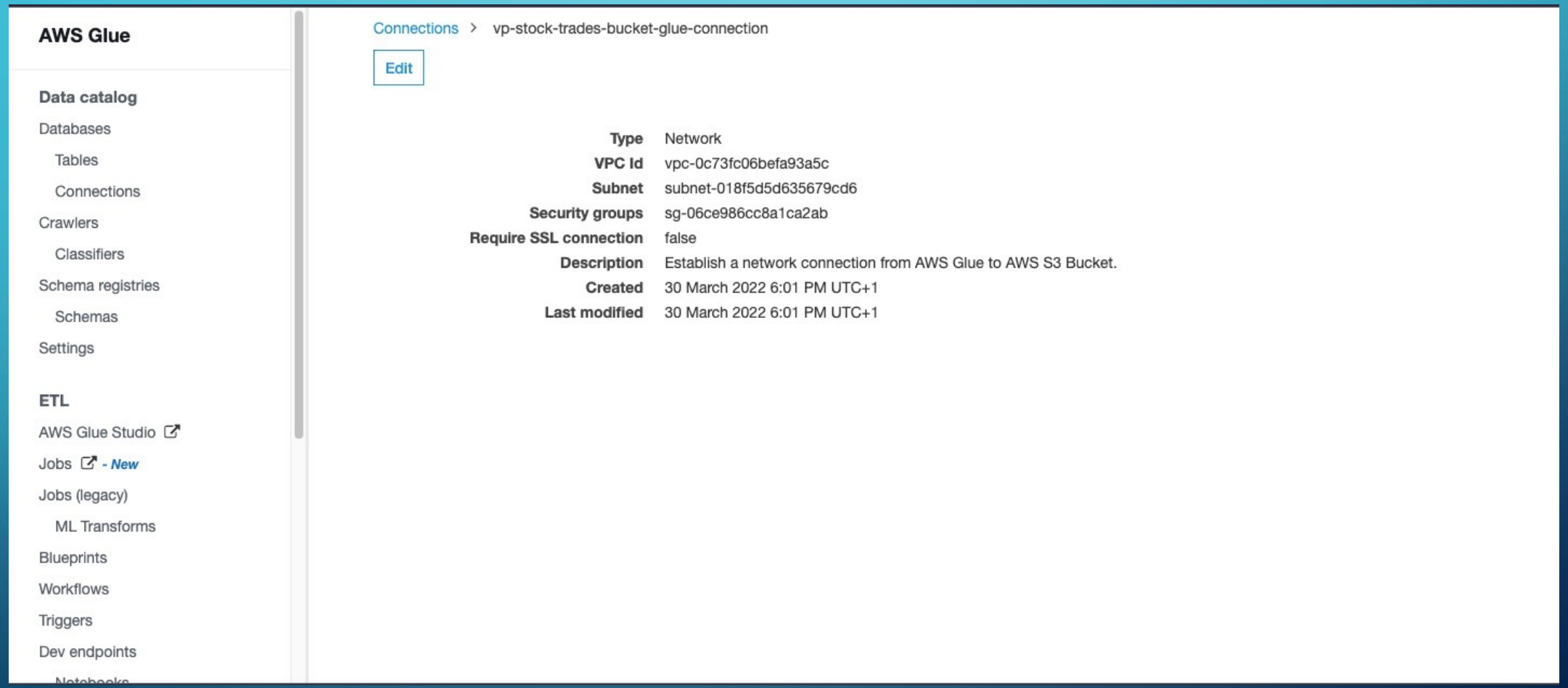
6. CREATION OF A DATABASE IN THE AWS GLUE TO STORE THE GLUE CRAWLER DATA CATALOG SCHEMA RESULTS



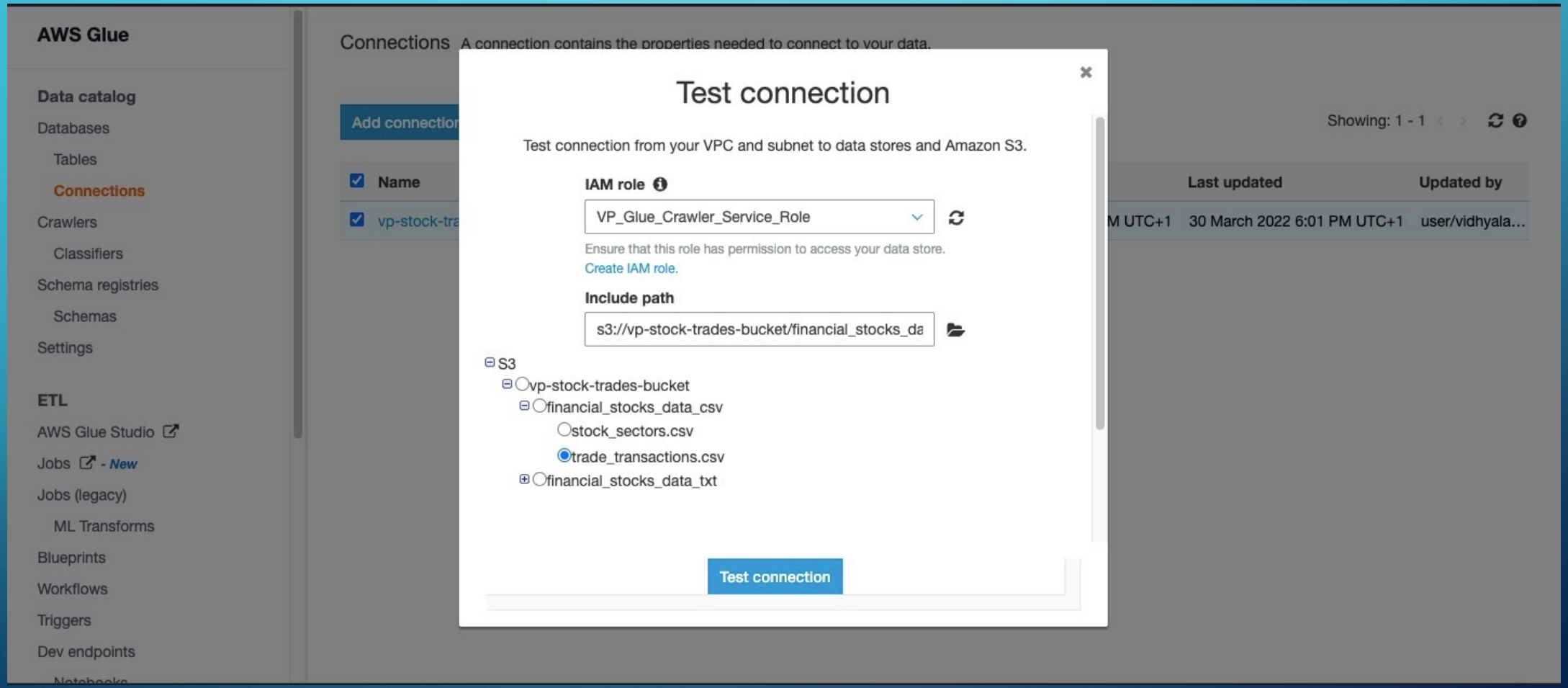
7. ESTABLISH A NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET *(CONTINUED)*



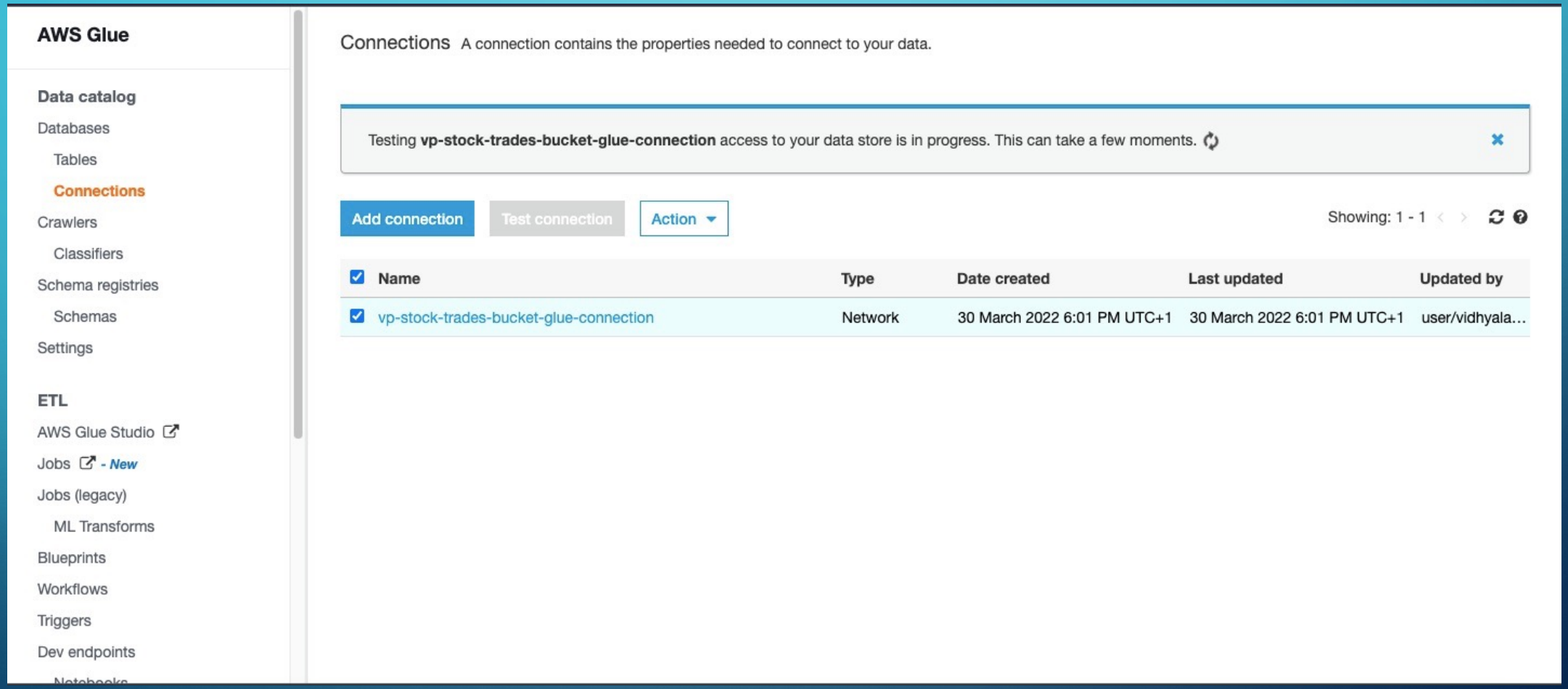
7. ESTABLISH A NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET



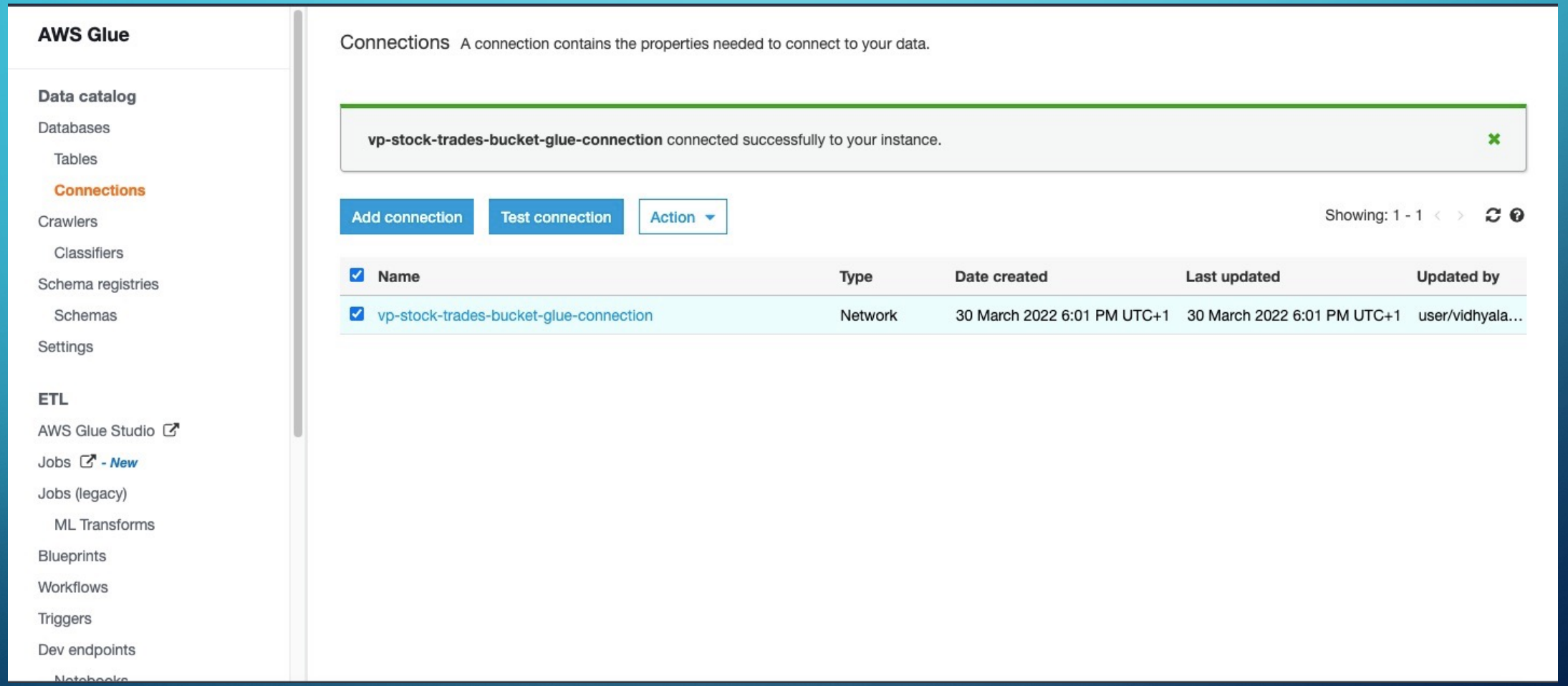
8. VERIFY/TEST THE NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET *(CONTINUED)*



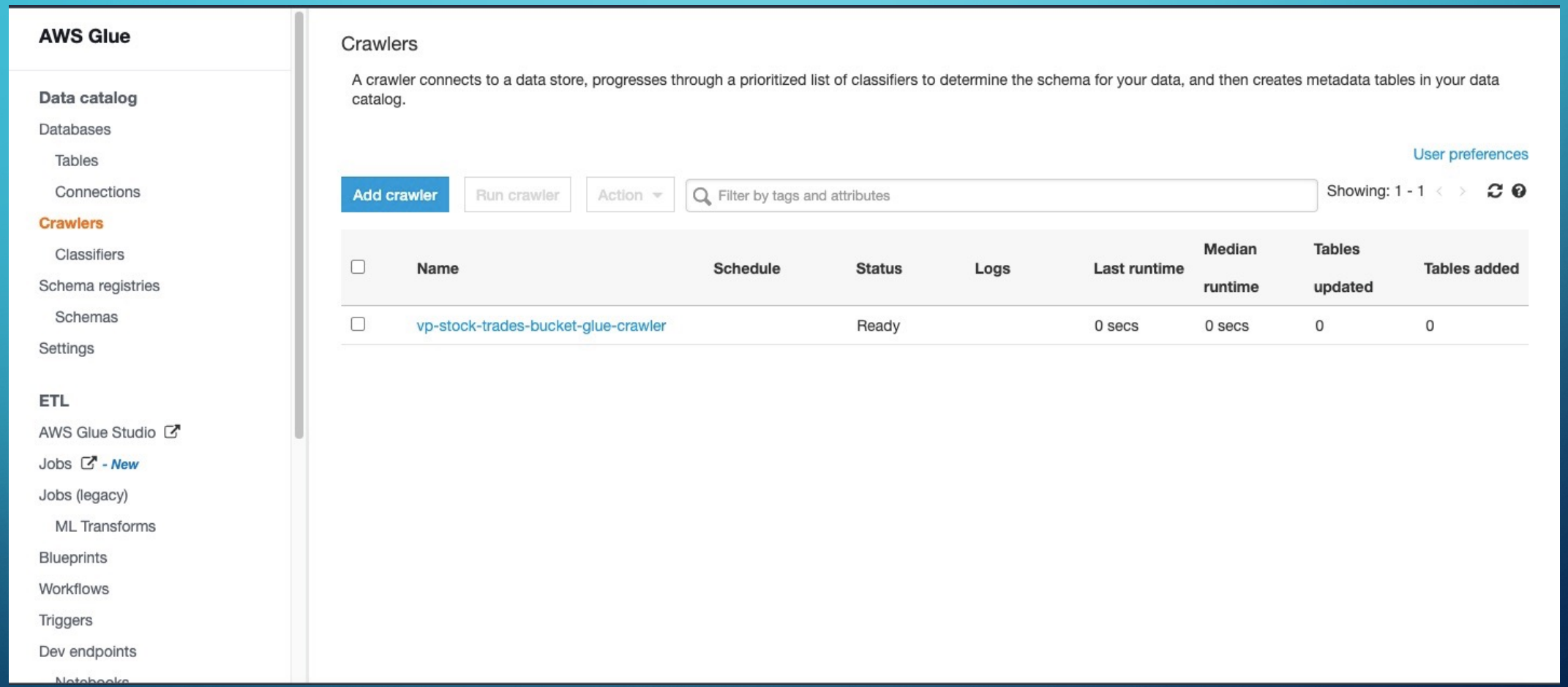
8. VERIFY/TEST THE NETWORK CONNECTION FROM AWS GLUE CRAWLER THROUGH VPC ENDPOINT TO AWS S3 BUCKET *(CONTINUED)*



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

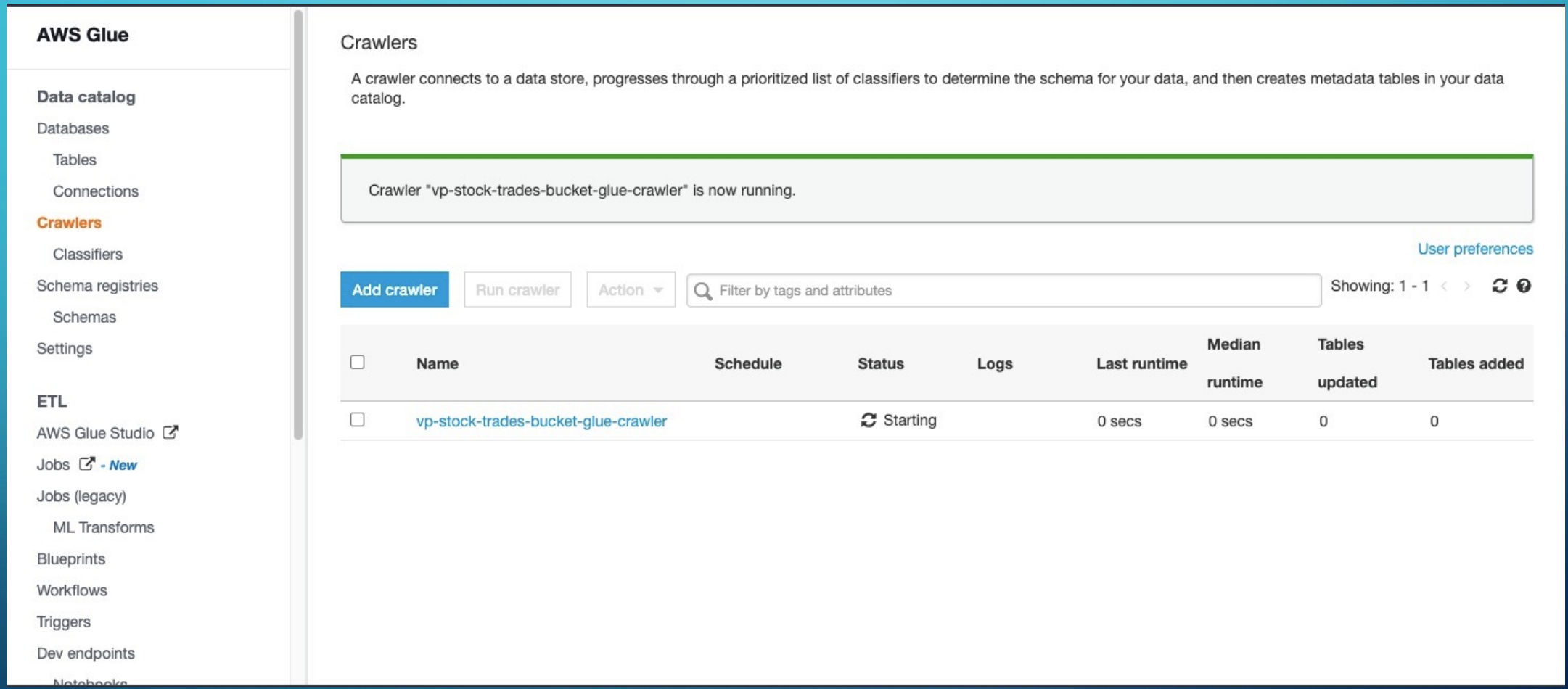


9. CREATION OF AWS GLUE CRAWLER TO RETRIEVE DATA CATALOG SCHEMA RESULTS OF THE SPECIFIED SOURCE FILES IN AWS S3 BUCKET *(CONTINUED)*

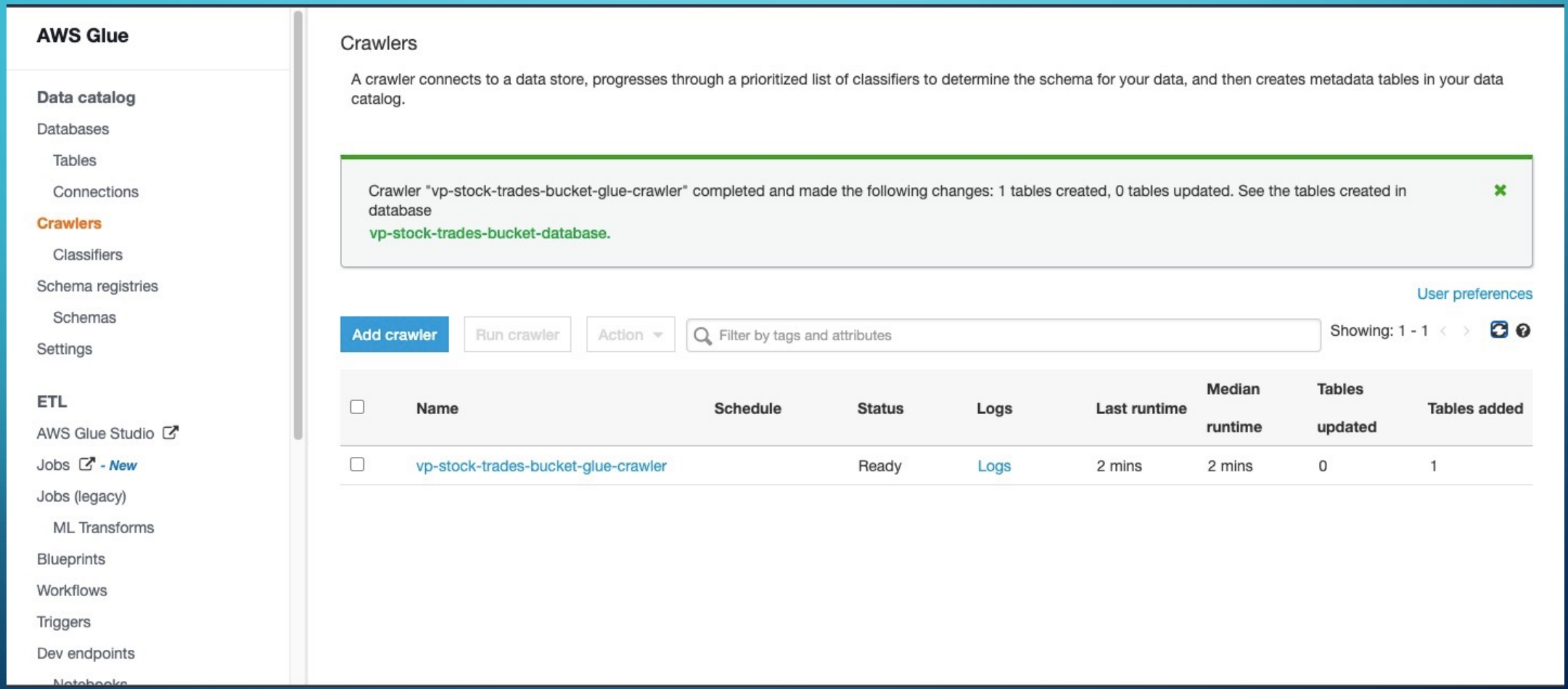


9. CREATION OF AWS GLUE CRAWLER TO RETRIEVE DATA CATALOG SCHEMA RESULTS OF THE SPECIFIED SOURCE FILES IN AWS S3 BUCKET

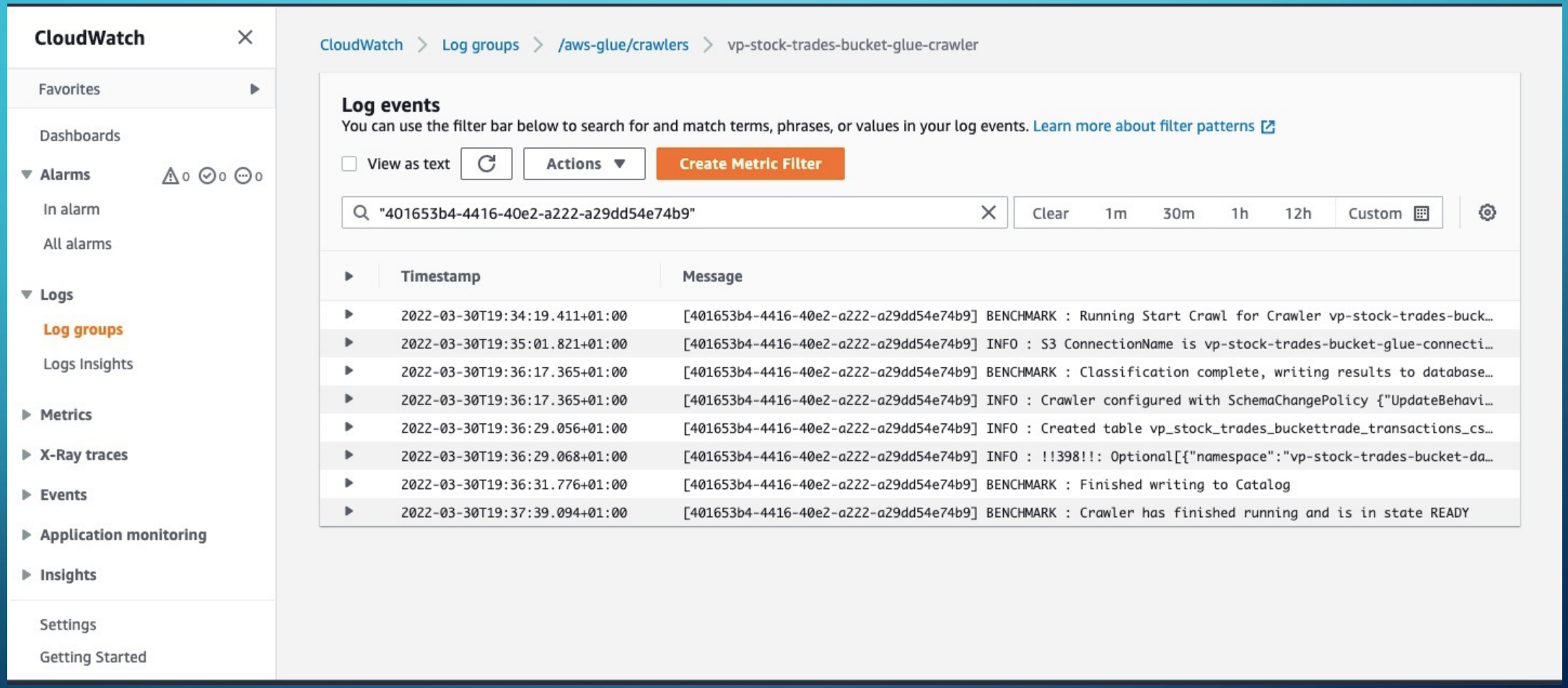
10. RUNNING THE AWS GLUE CRAWLER TO RETRIEVE DATA CATALOG SCHEMA RESULTS OF THE SPECIFIED SOURCE FILES IN AWS S3 BUCKET *(CONTINUED)*



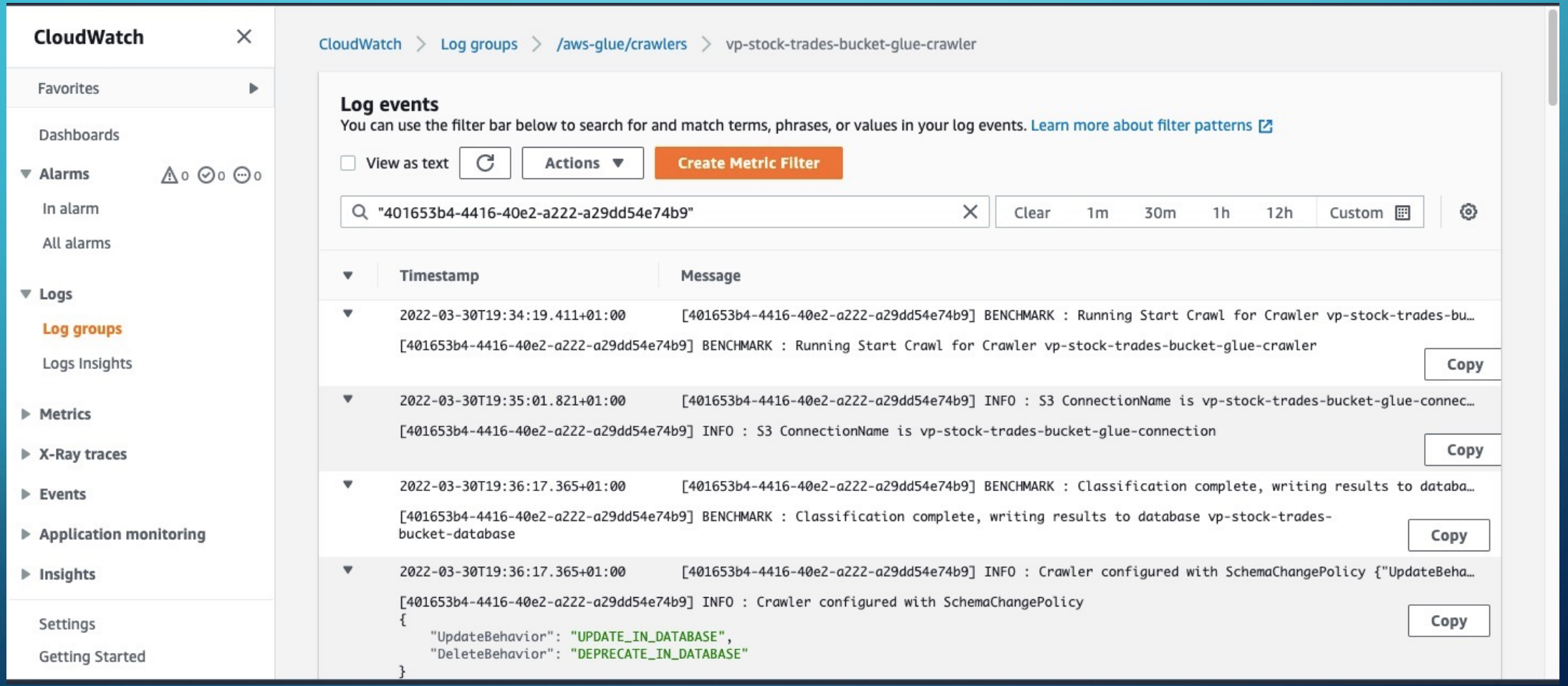
10. RUNNING THE AWS GLUE CRAWLER TO RETRIEVE DATA CATALOG SCHEMA RESULTS OF THE SPECIFIED SOURCE FILES IN AWS S3 BUCKET



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

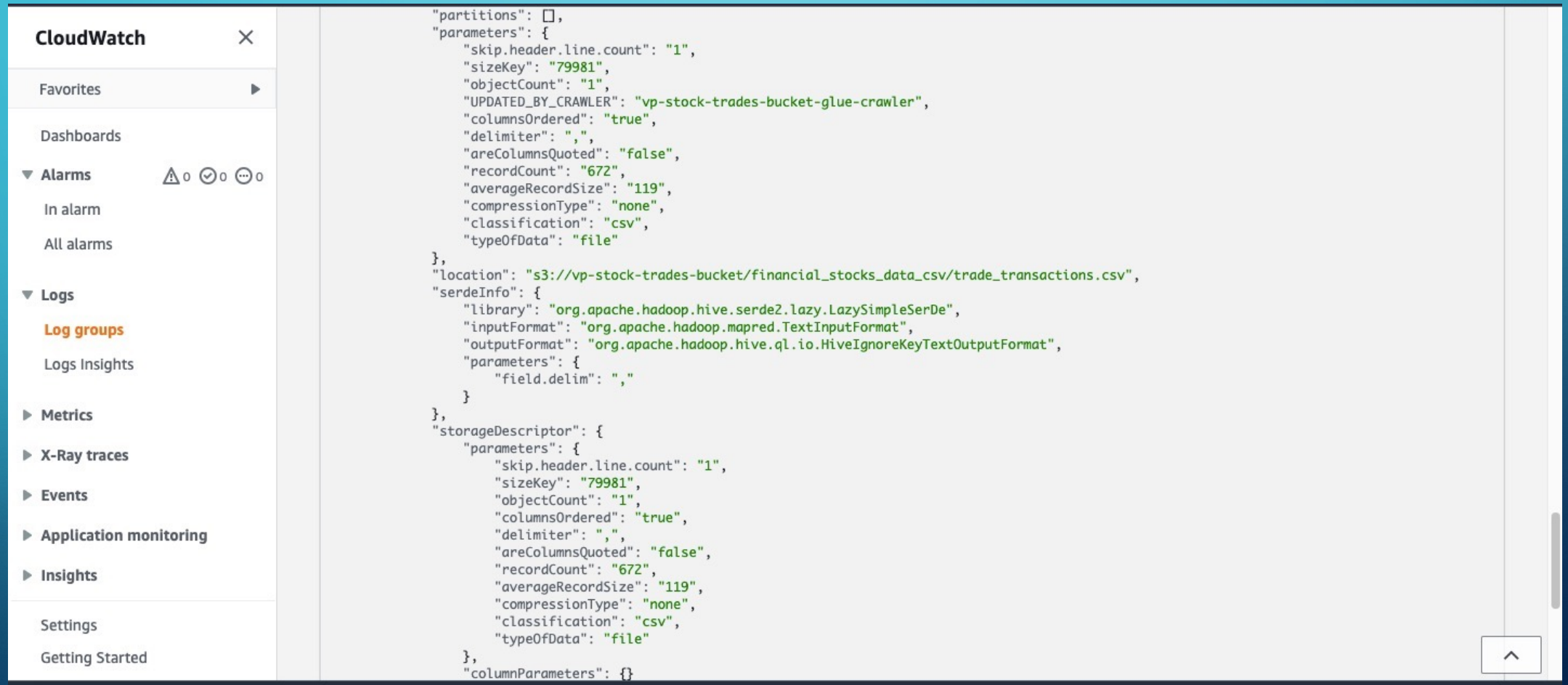


11. VIEWING THE AWS GLUE CRAWLER EXECUTION LOGS IN AWS CLOUDWATCH *(CONTINUED)*

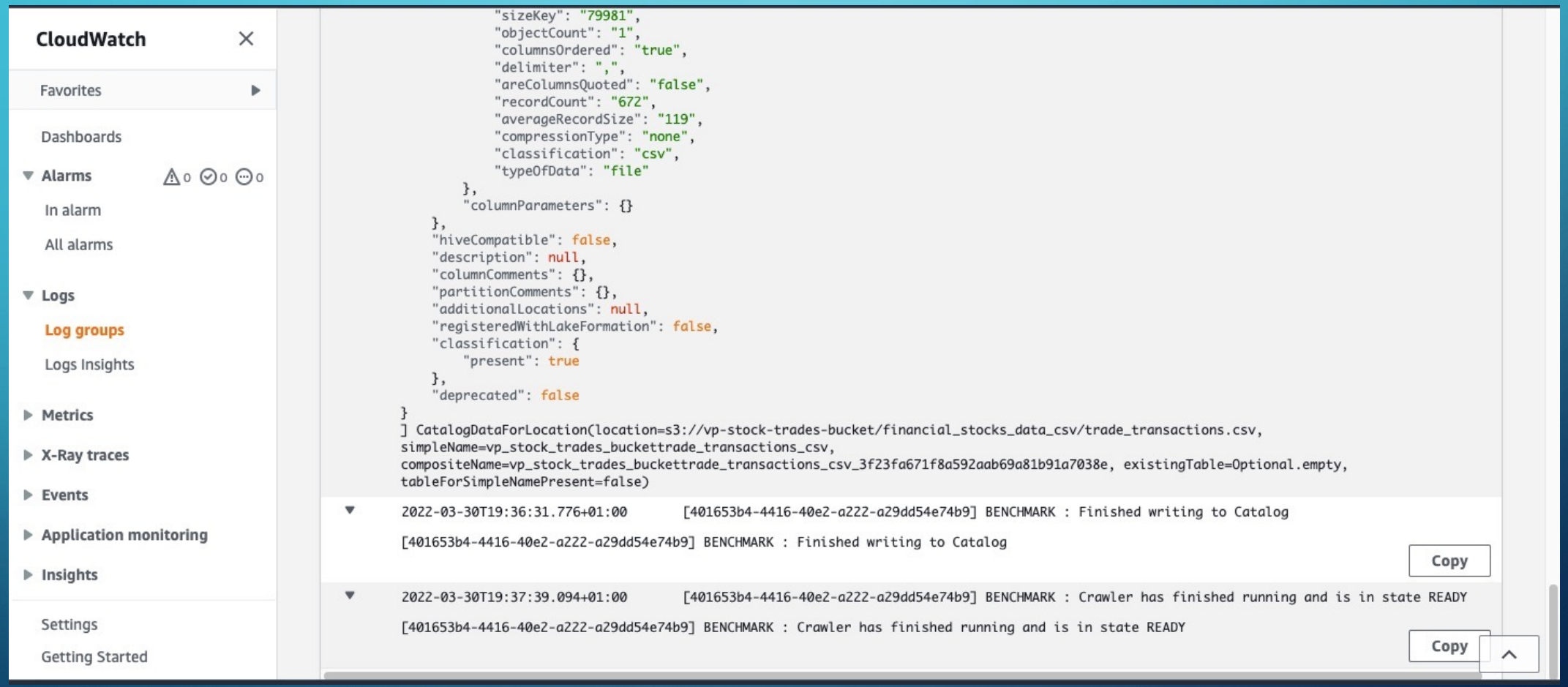


11. VIEWING THE AWS GLUE CRAWLER EXECUTION LOGS IN AWS CLOUDWATCH *(CONTINUED)*

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

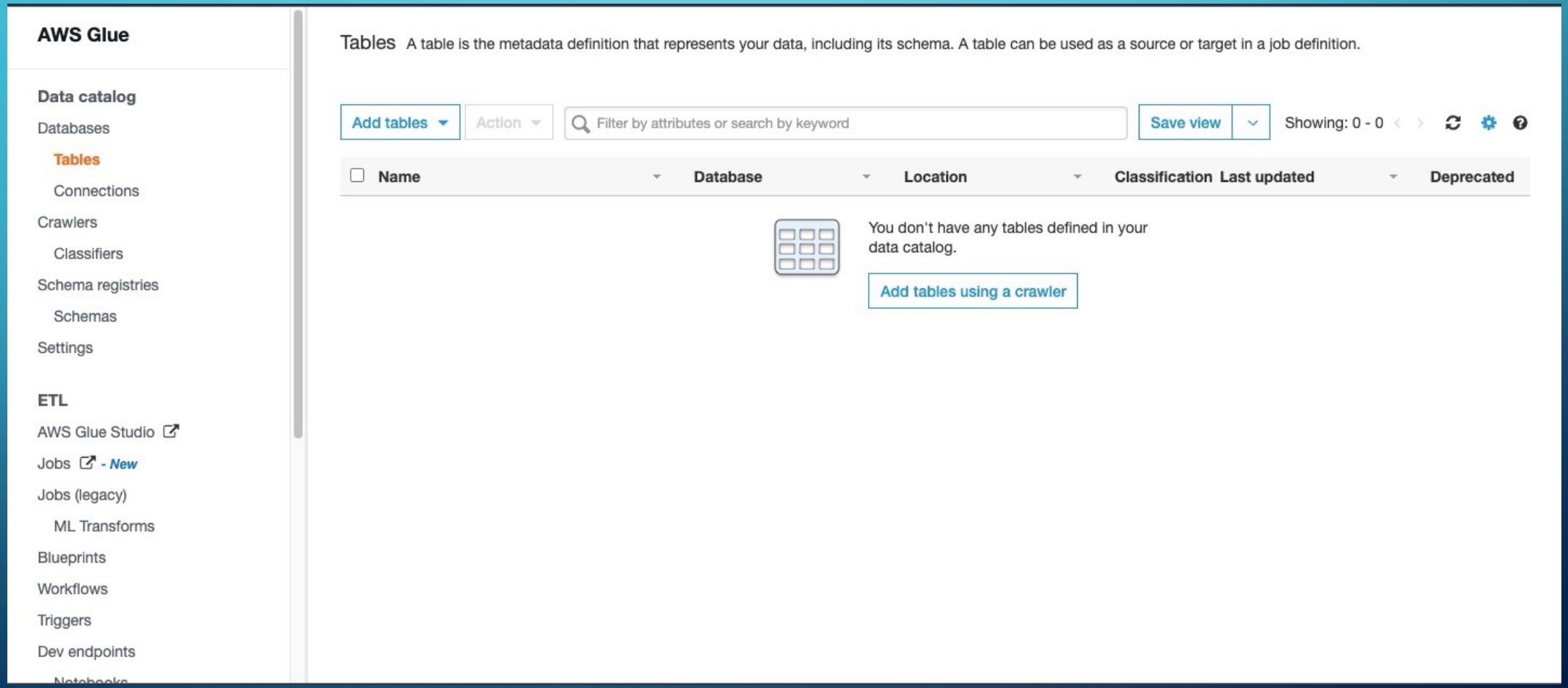


11. VIEWING THE AWS GLUE CRAWLER EXECUTION LOGS IN AWS CLOUDWATCH



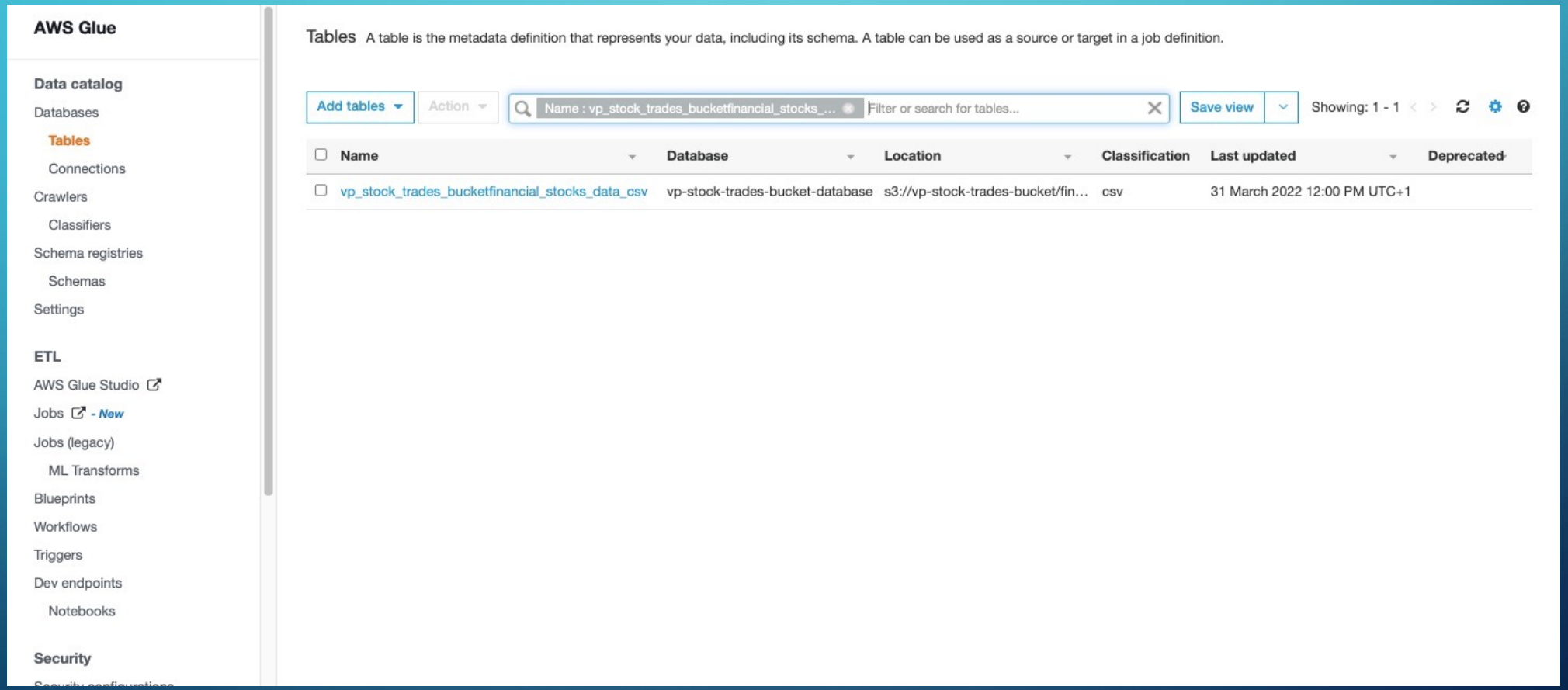
12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE *(CONTINUED)*

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

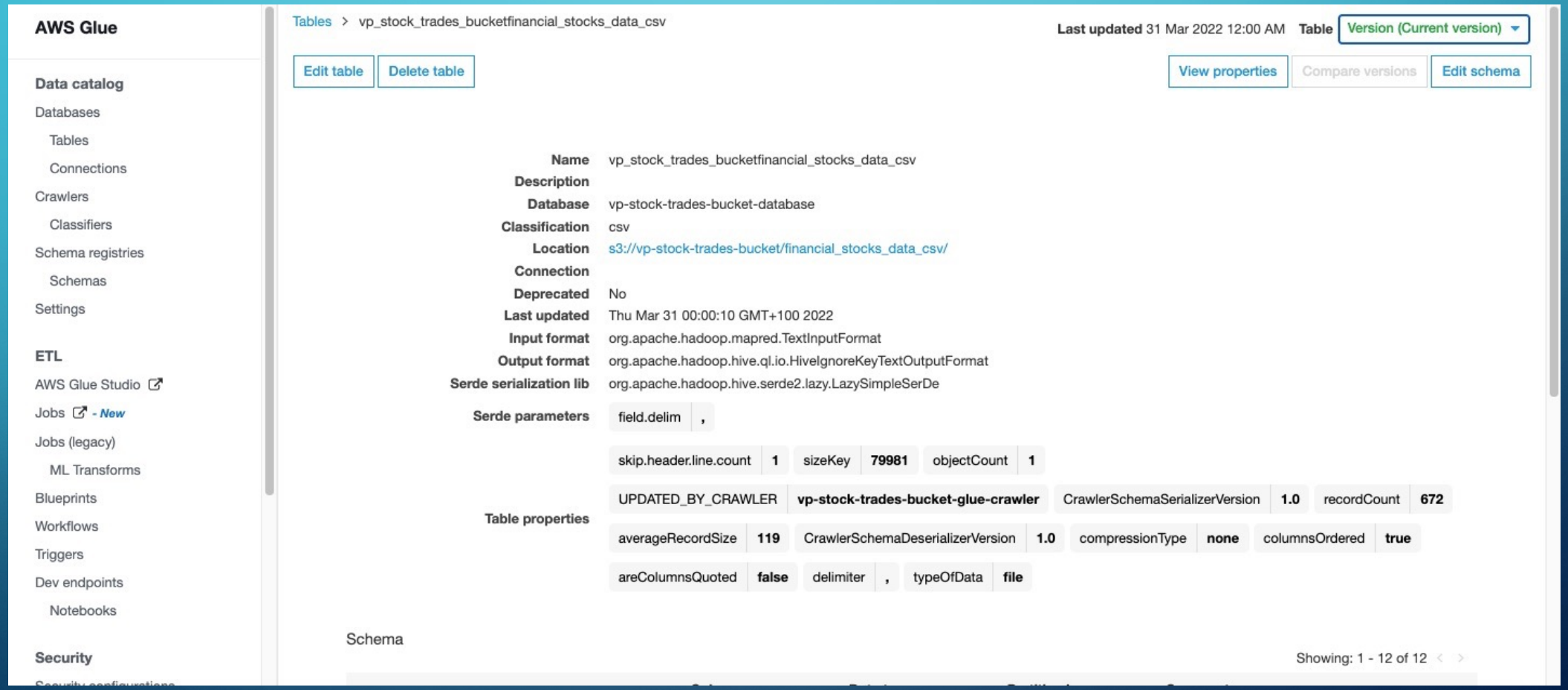


12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE *(CONTINUED)*

AWS Glue Database Table after executing the AWS Glue Crawler

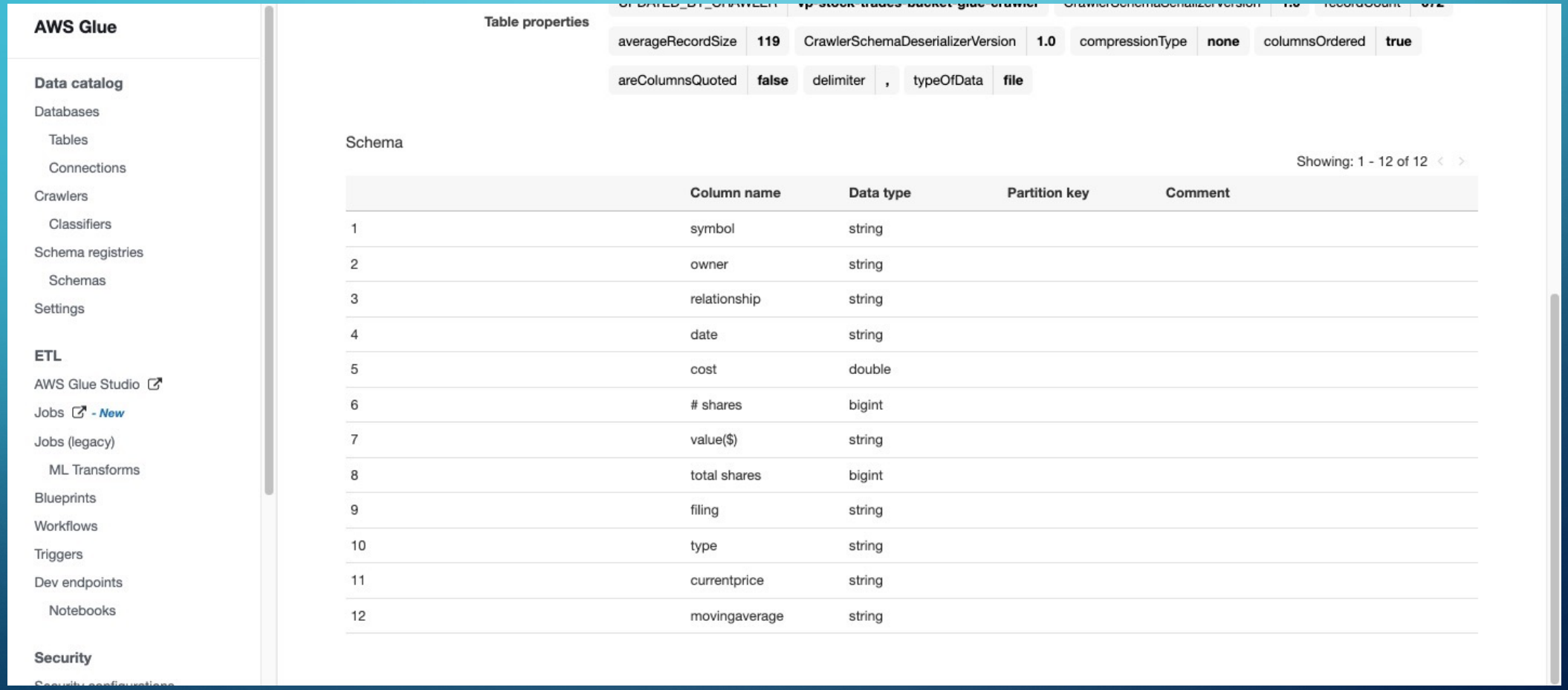
12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE *(CONTINUED)*

AWS Glue Database Table after executing the AWS Glue Crawler

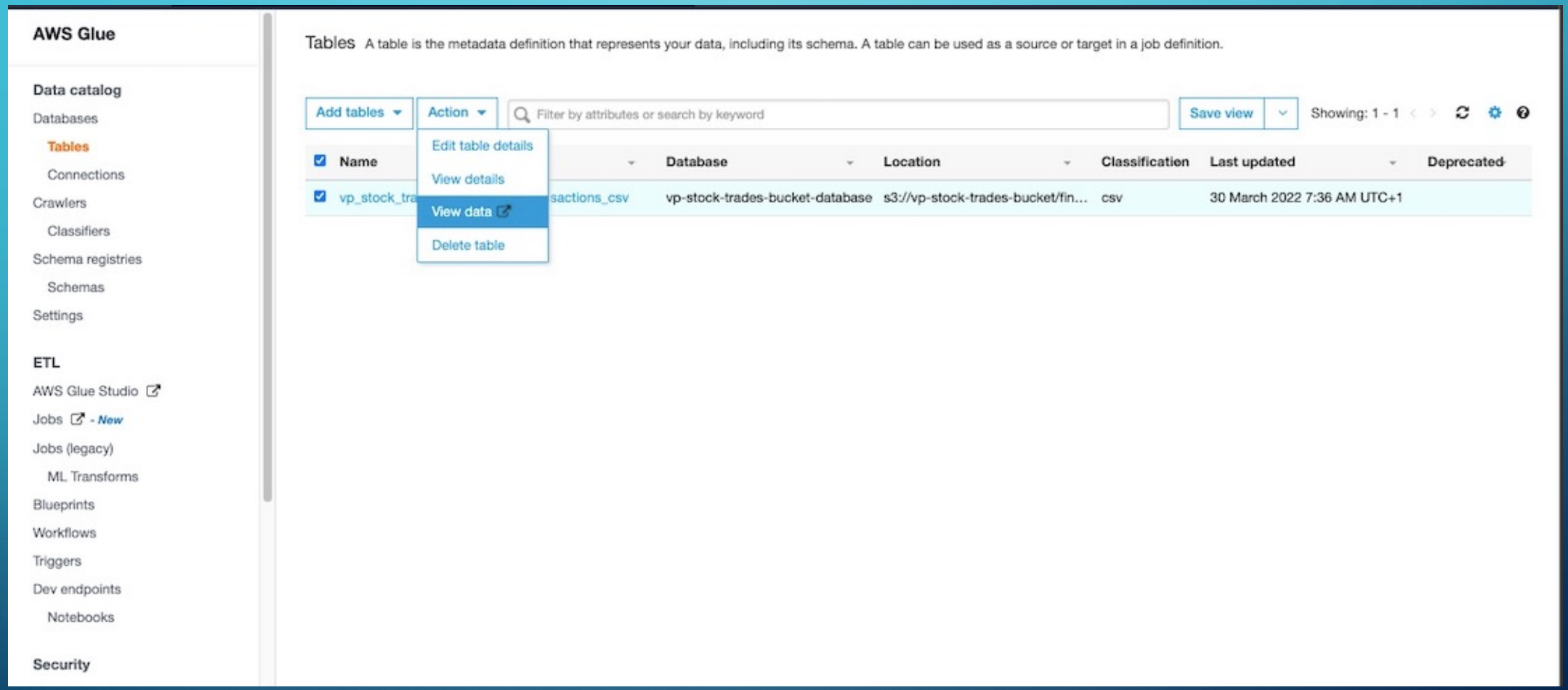


12. VERIFYING THE DATA CATALOG SCHEMA RESULTS STORED IN THE AWS GLUE DATABASE TABLE

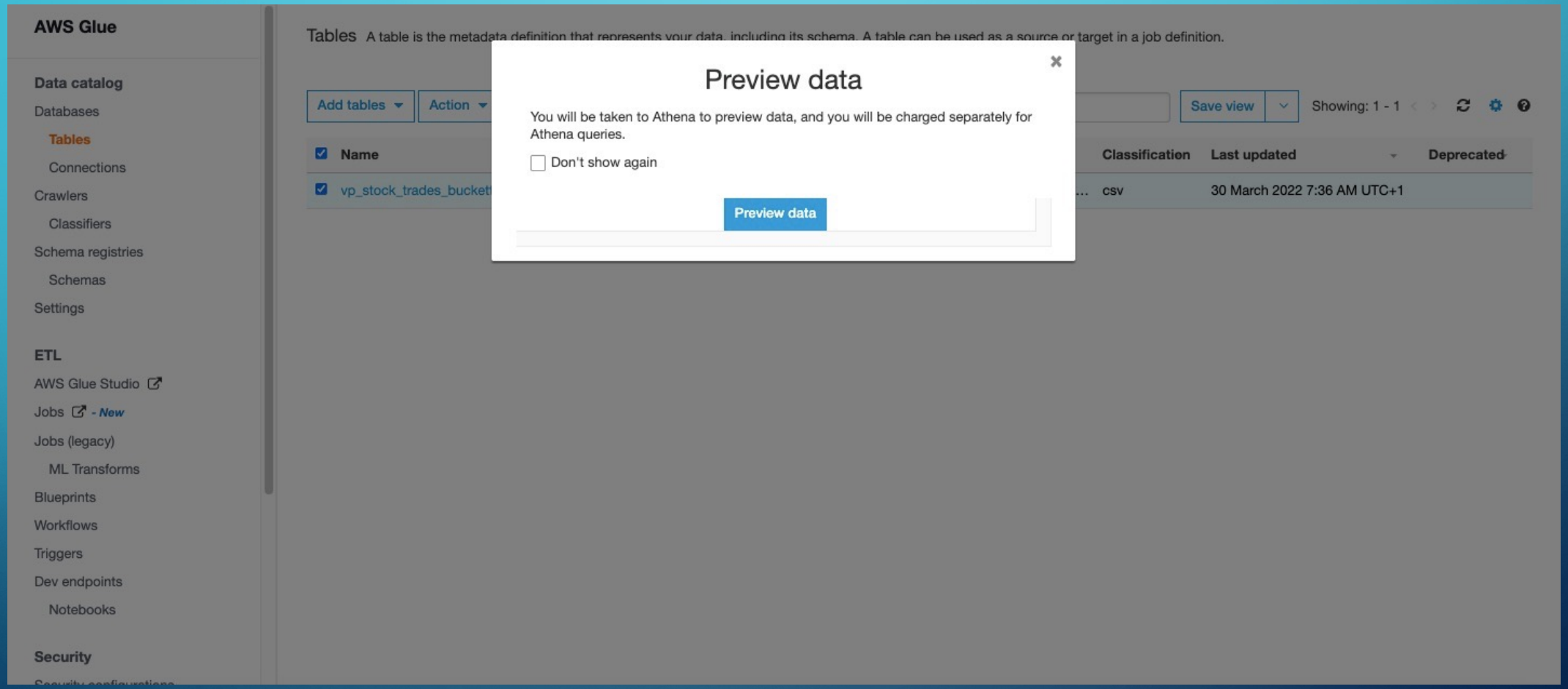
AWS Glue Database Table after executing the AWS Glue Crawler



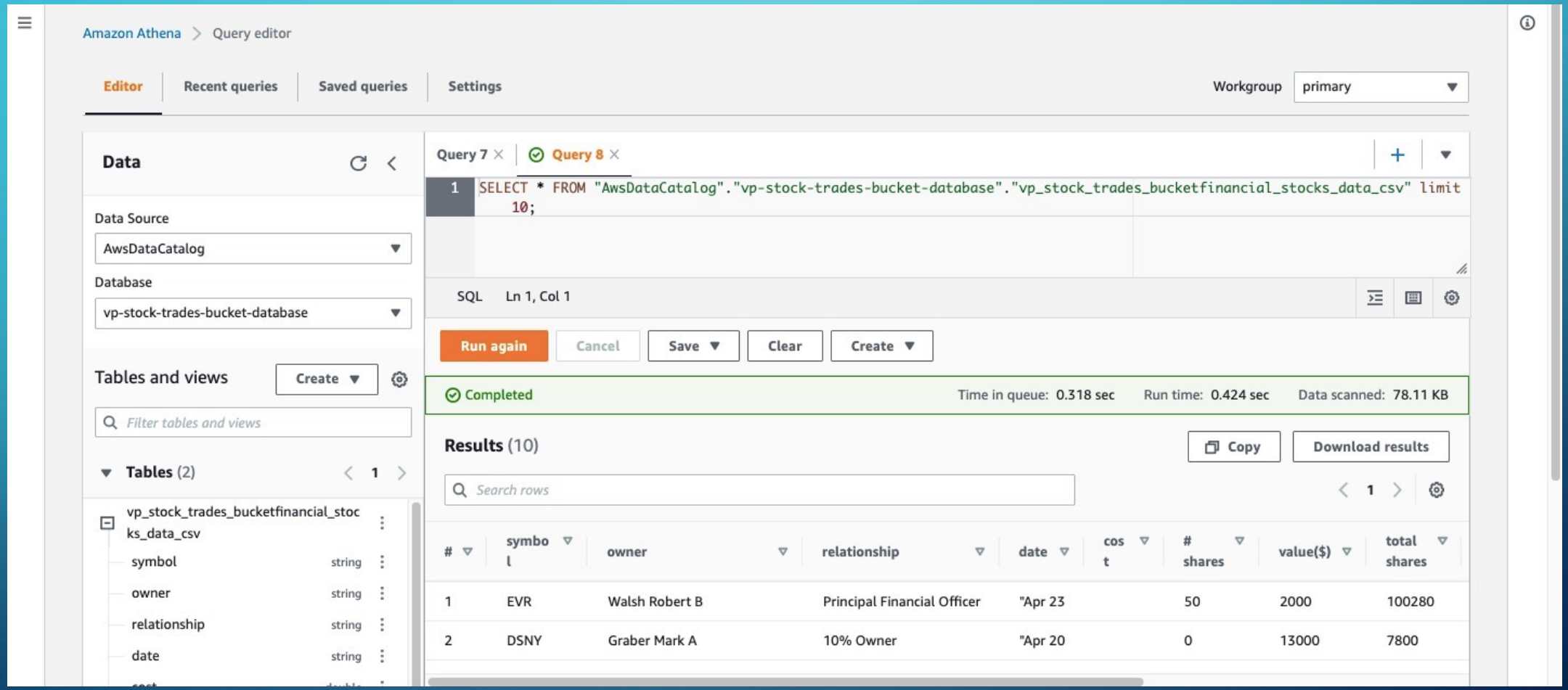
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*



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

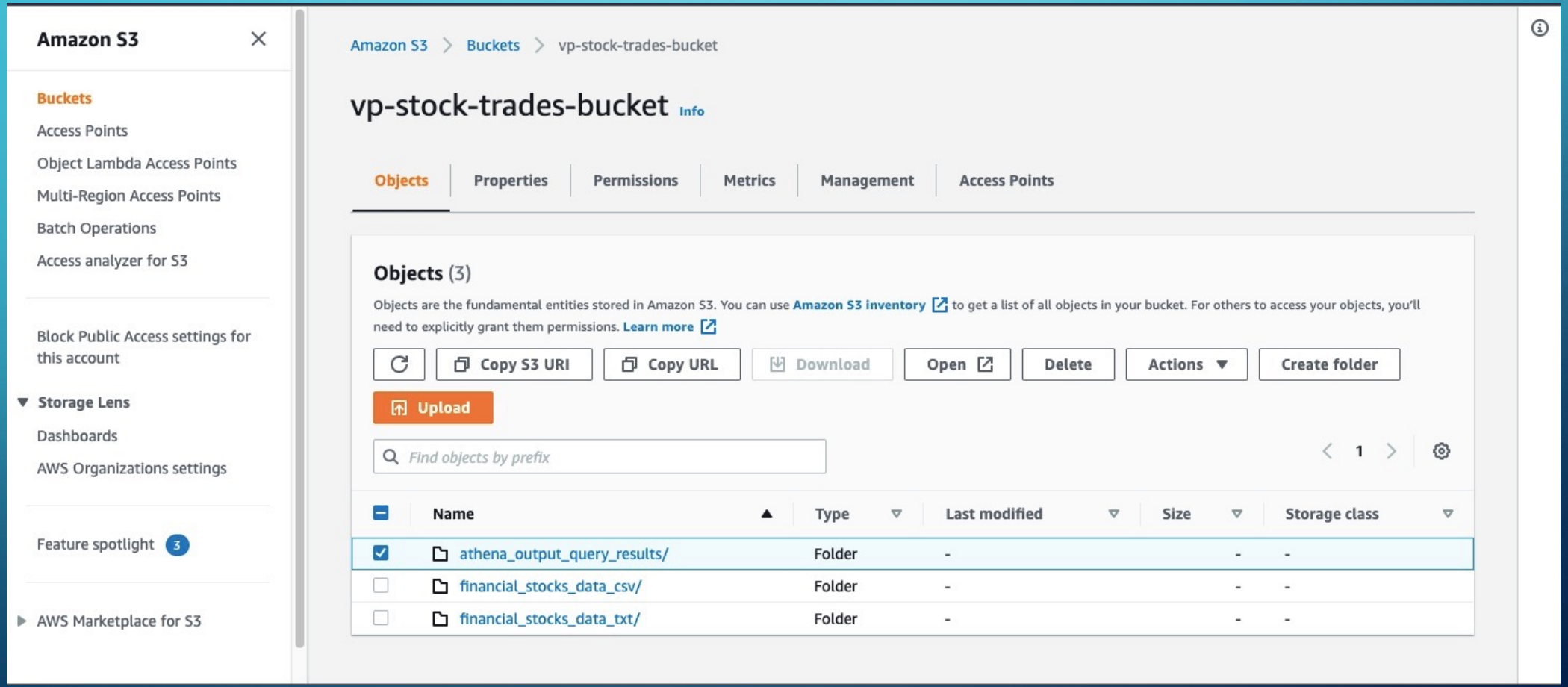


13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*



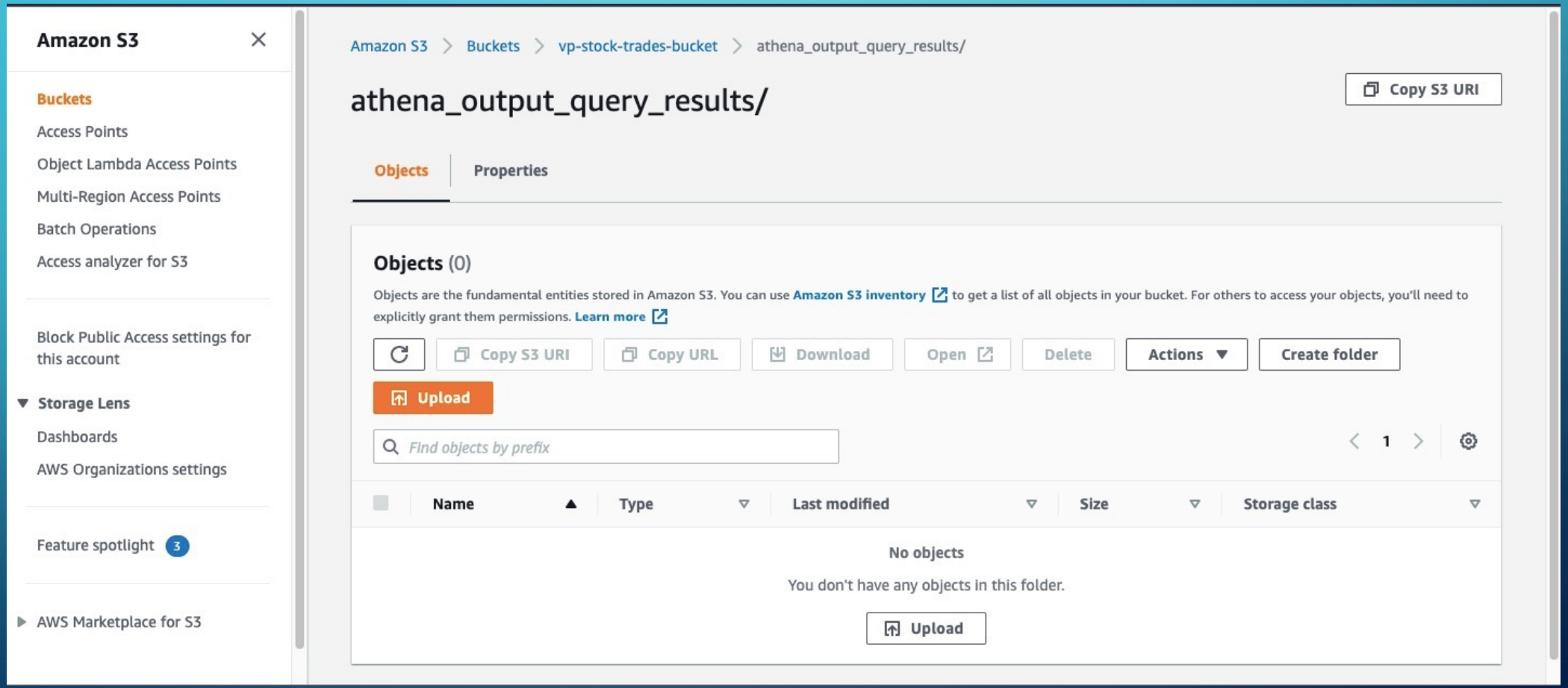
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

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



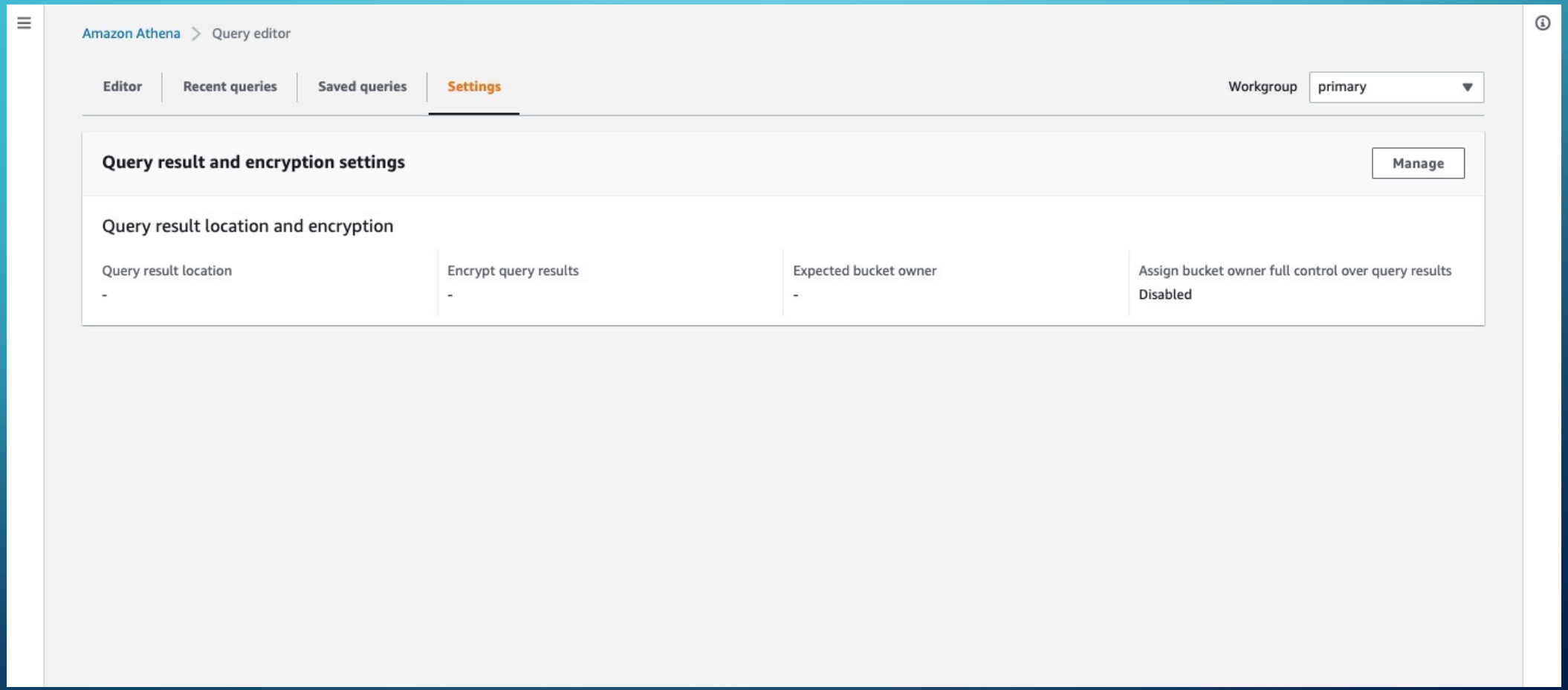
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

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



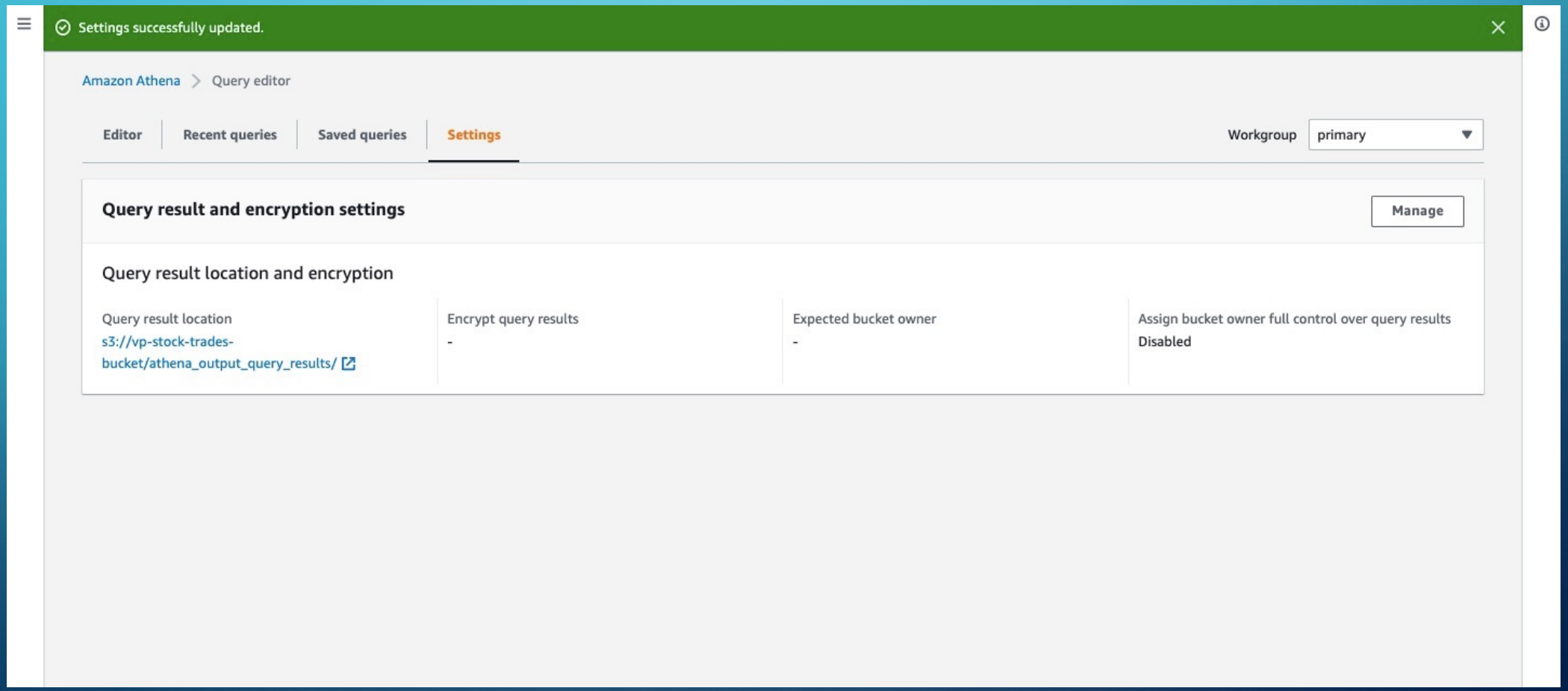
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

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



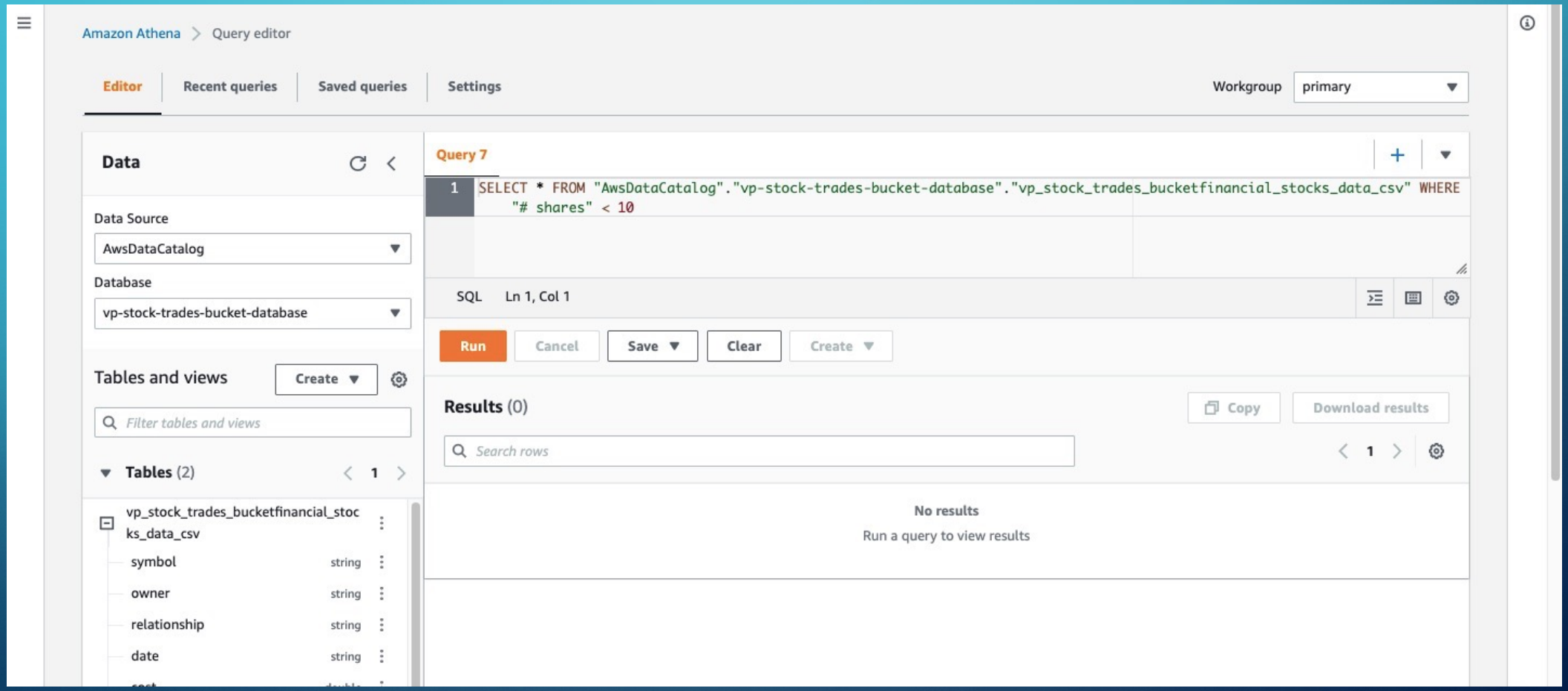
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

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



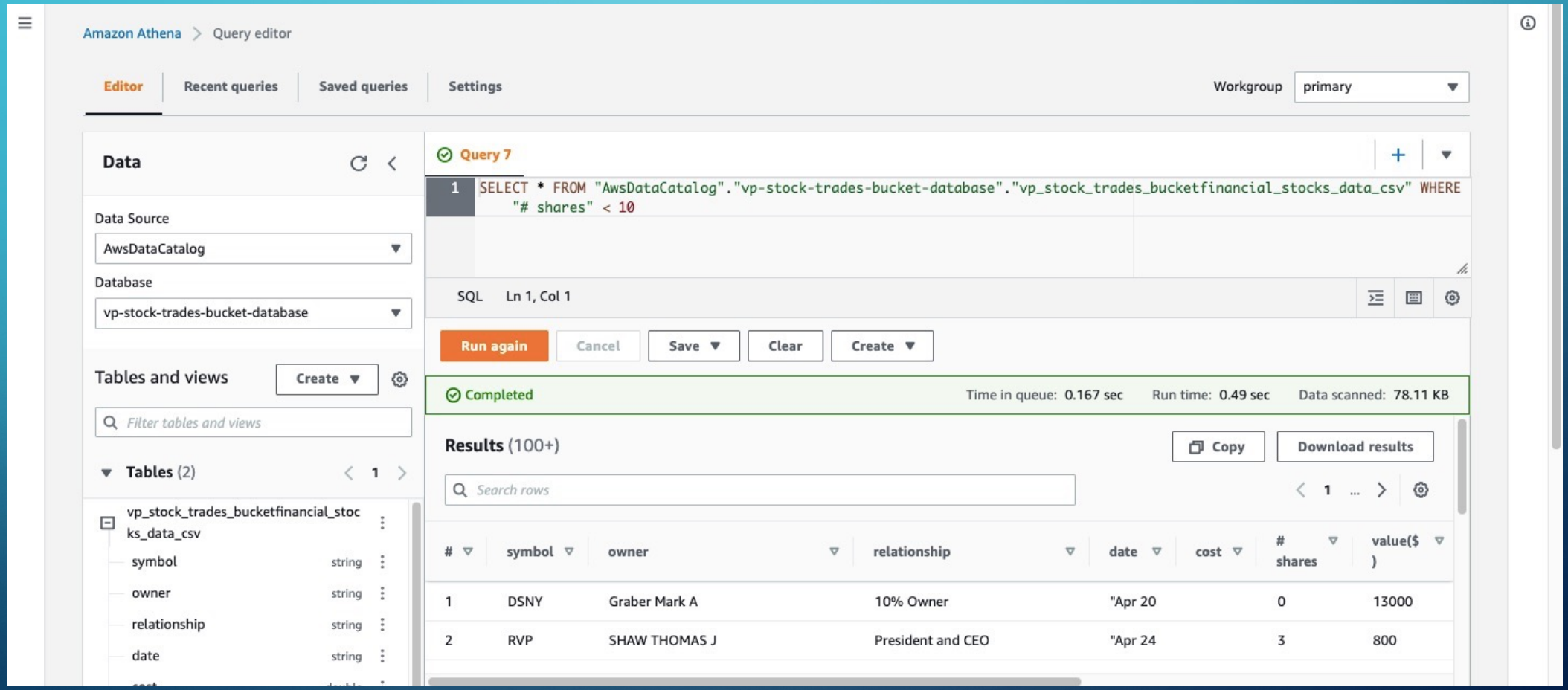
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA *(CONTINUED)*

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

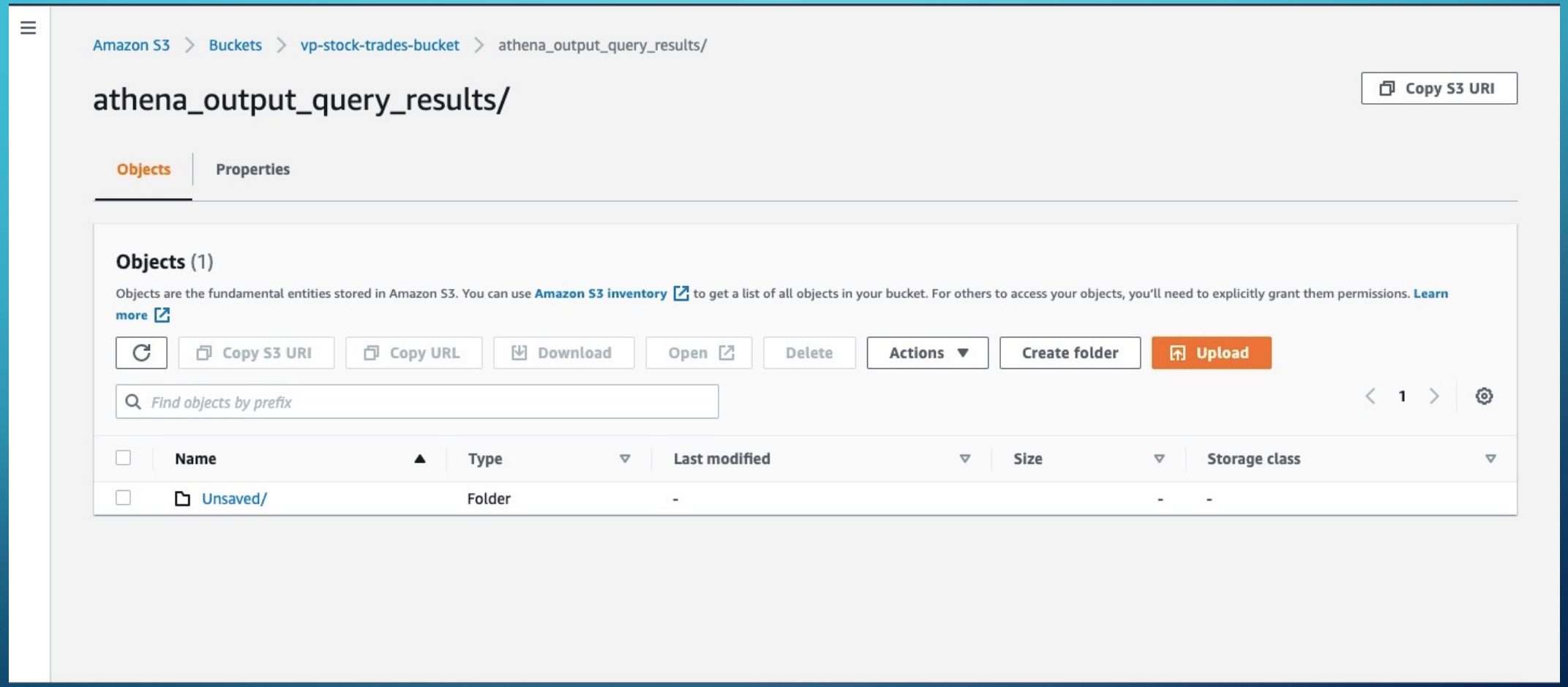


13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA

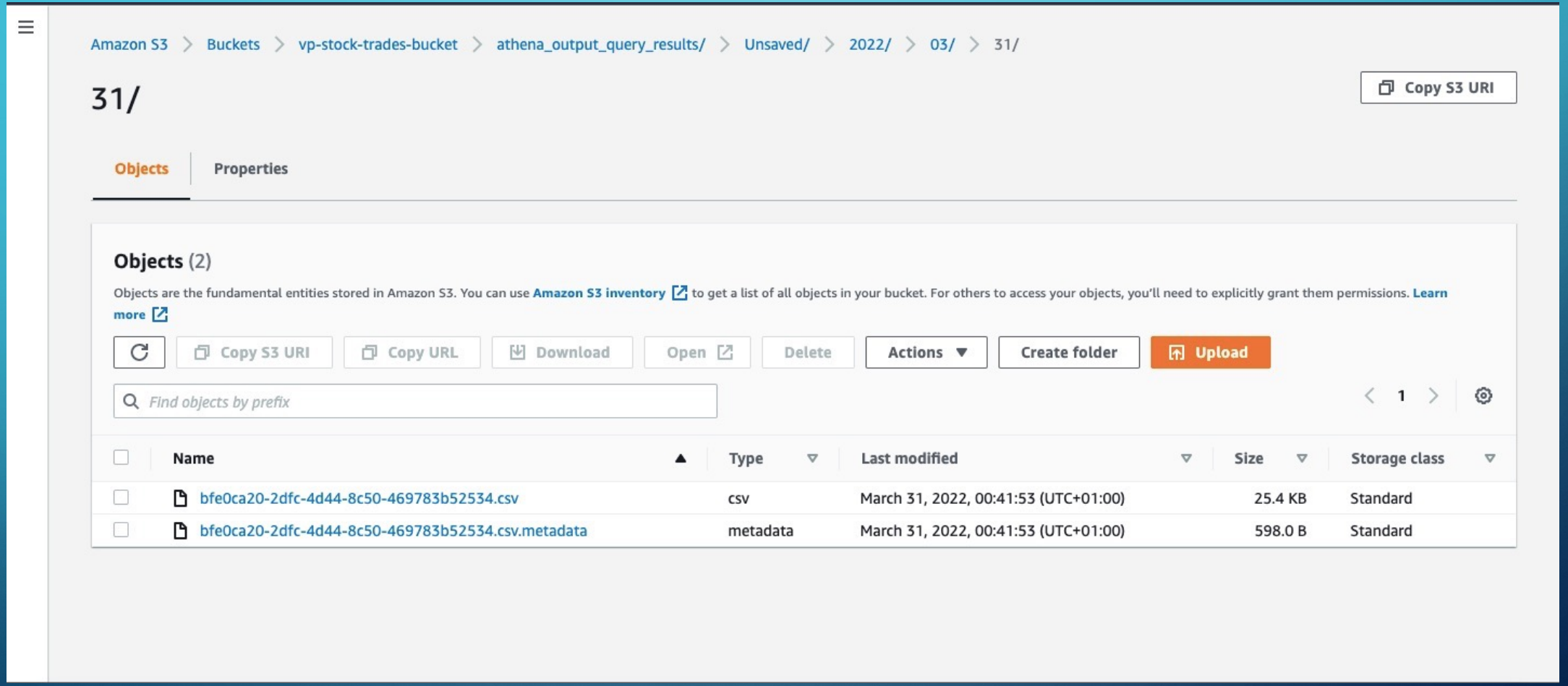
SQL Query Execution to query and filter the trade transactions for which number of shares processed is less than 10



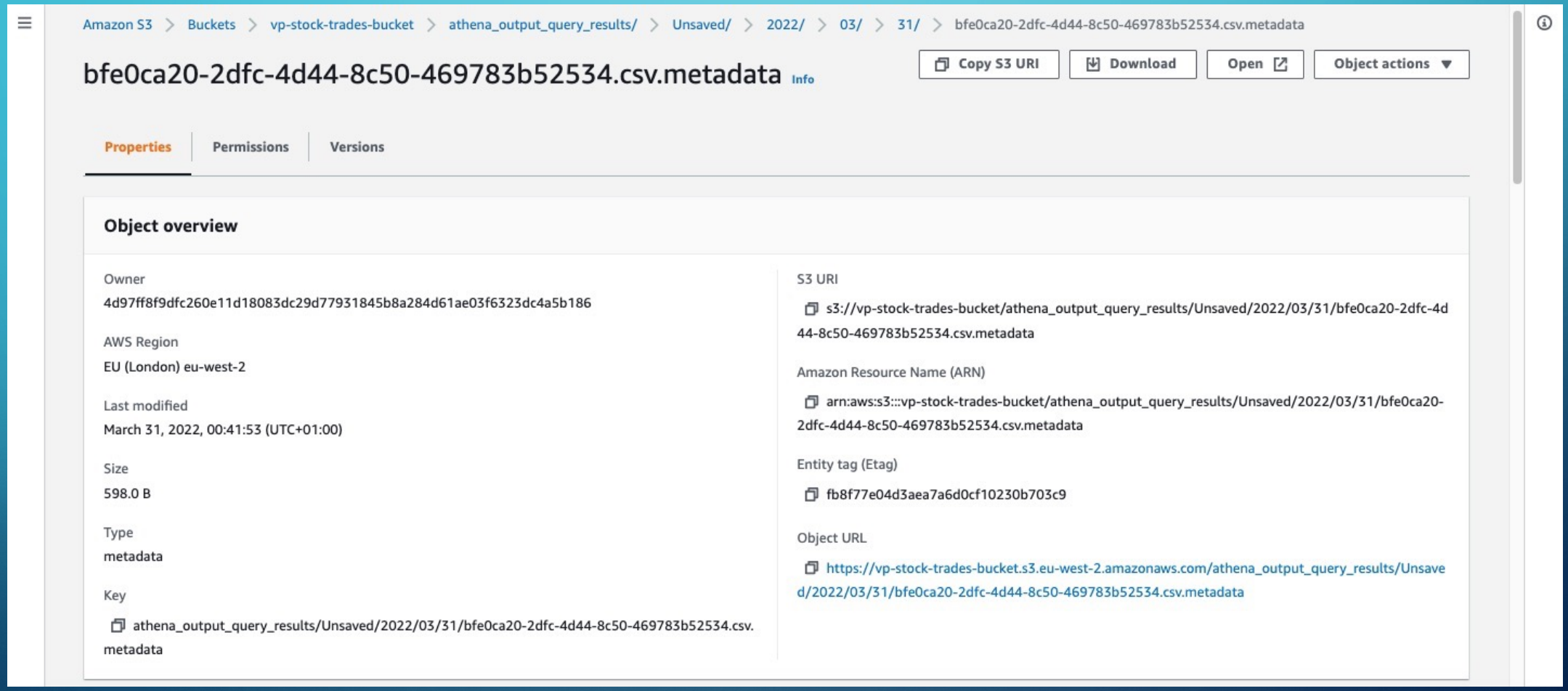
14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



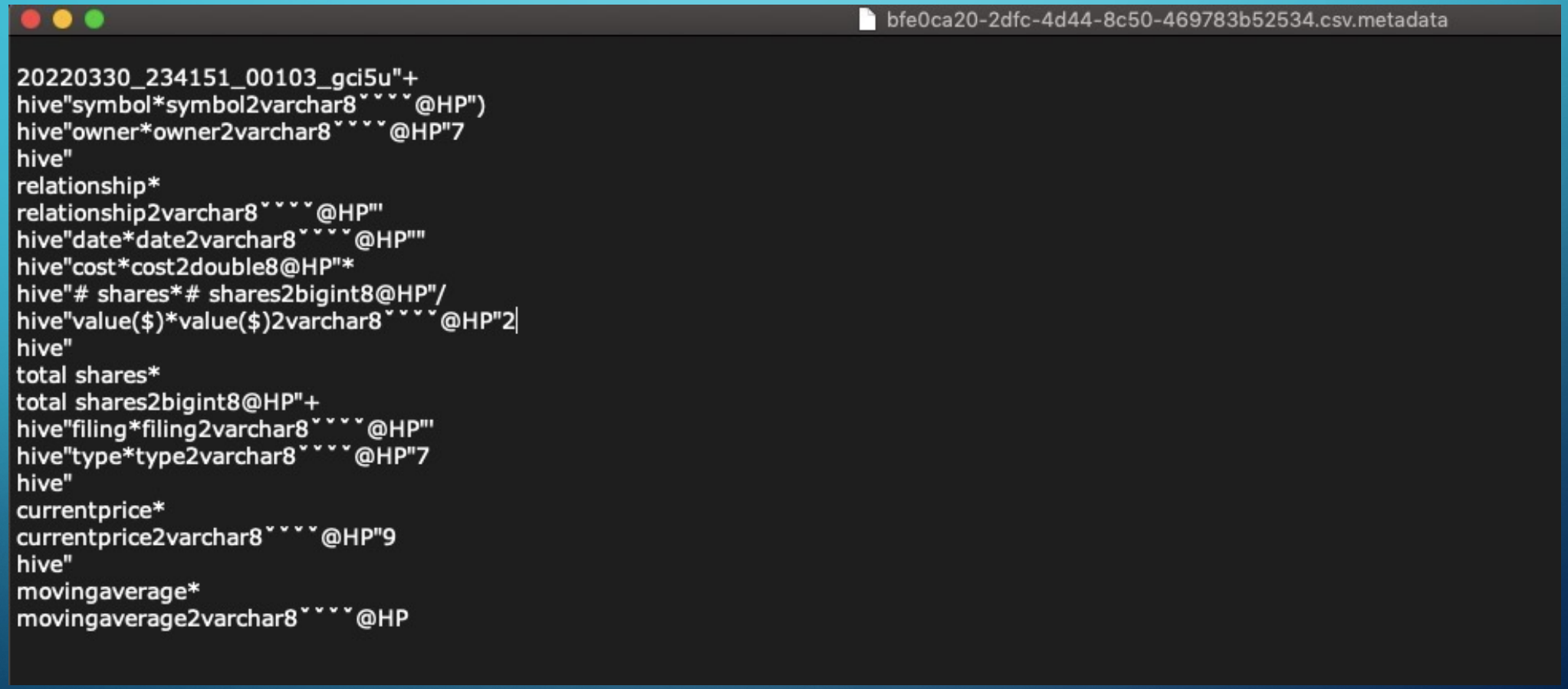
14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



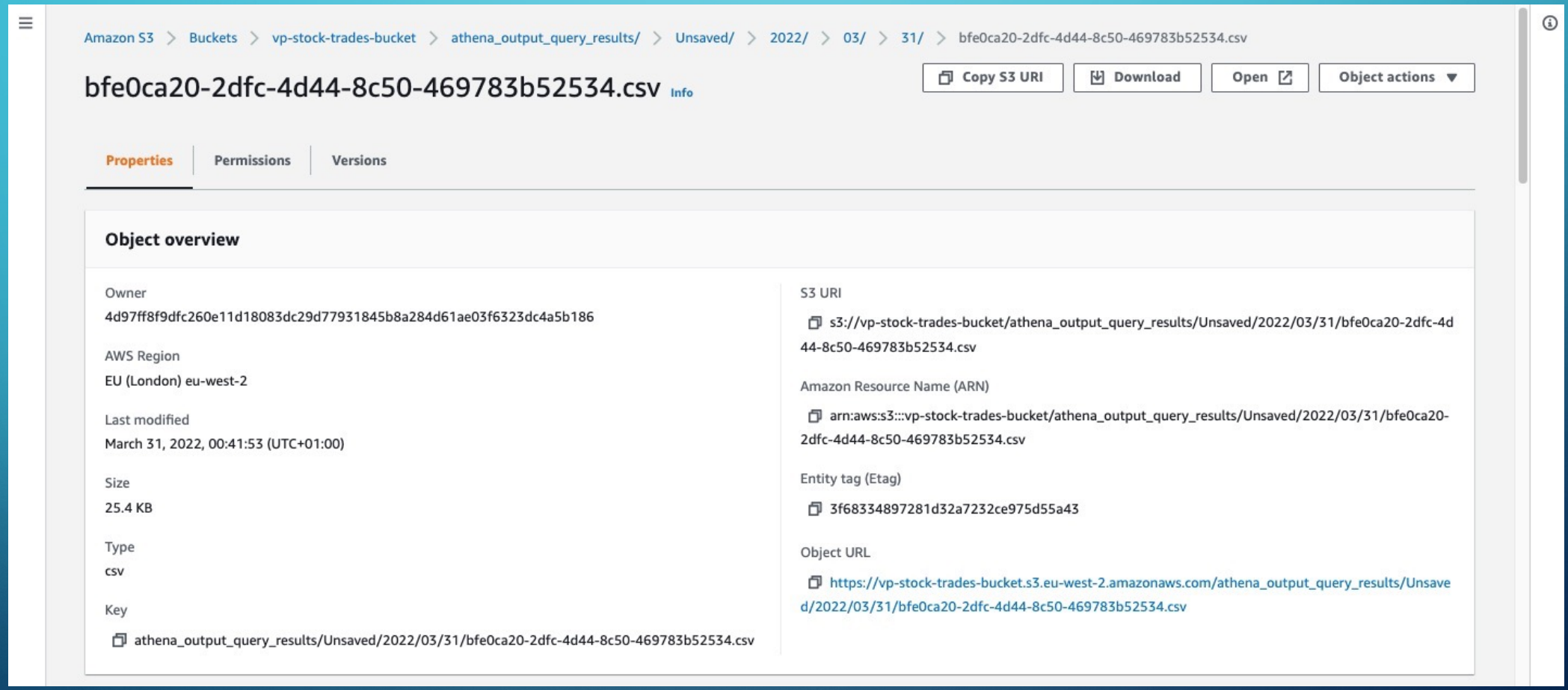
14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



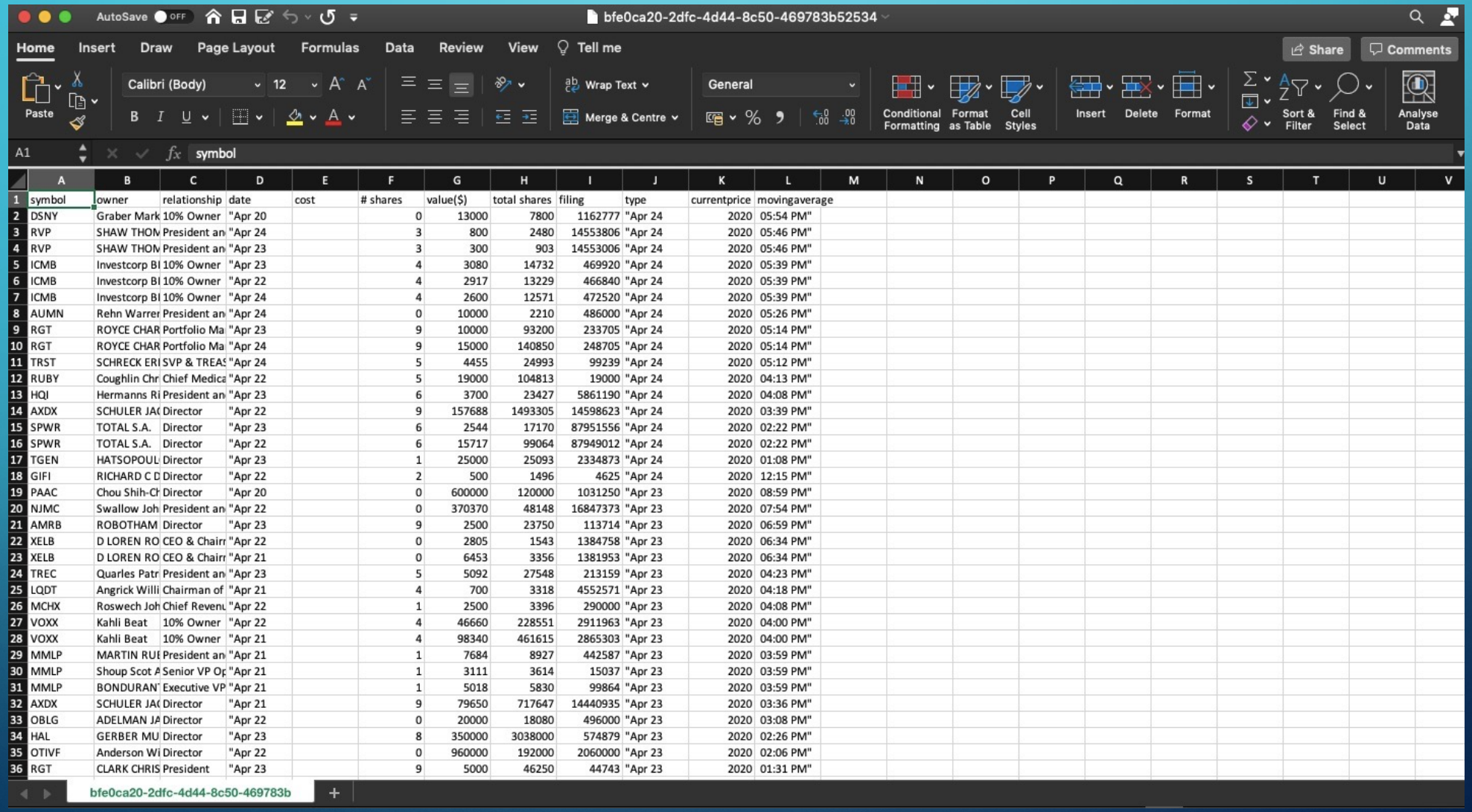
14. VERIFYING THE AWS ATHENA QUERY OUTPUT RESULTS STORED IN TO AWS S3 BUCKET *(CONTINUED)*



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



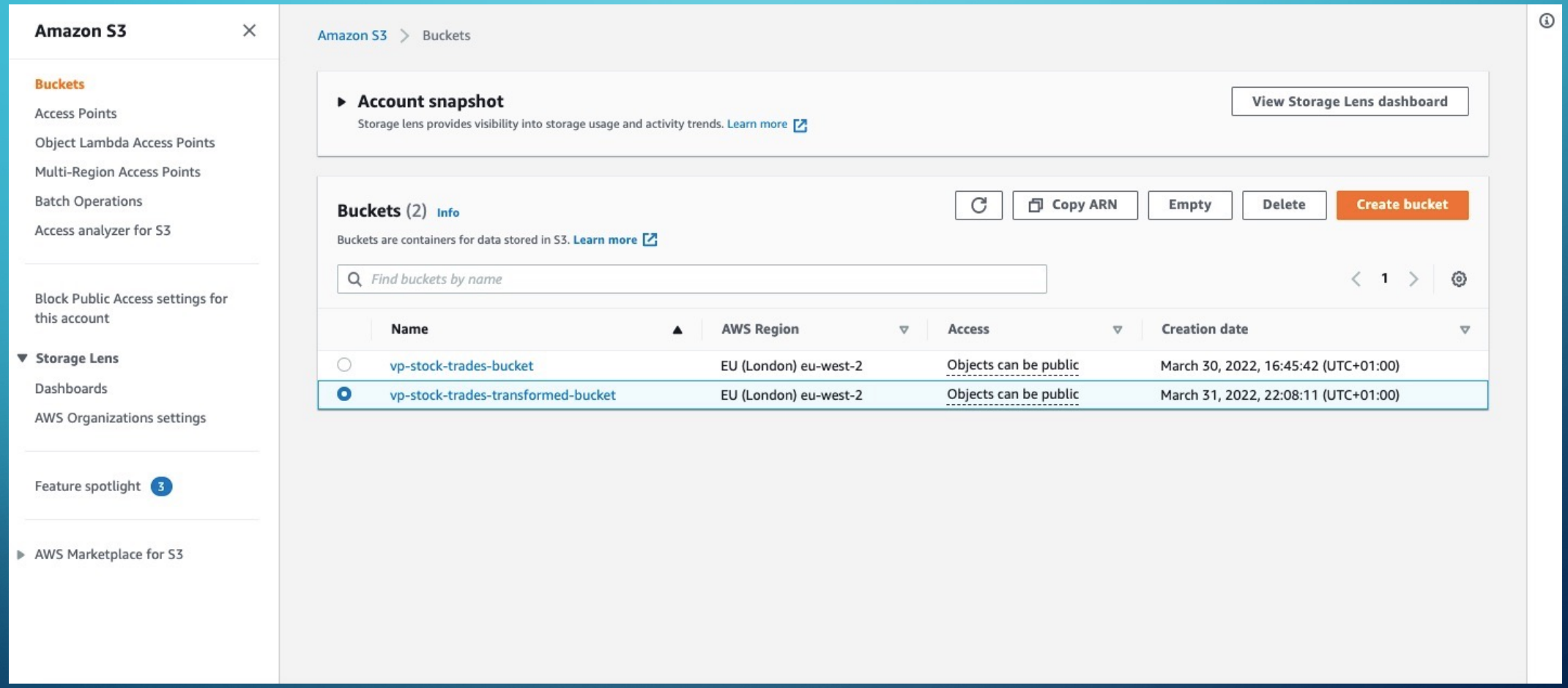
bfe0ca20-2dfc-4d44-8c50-469783b52534

	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"

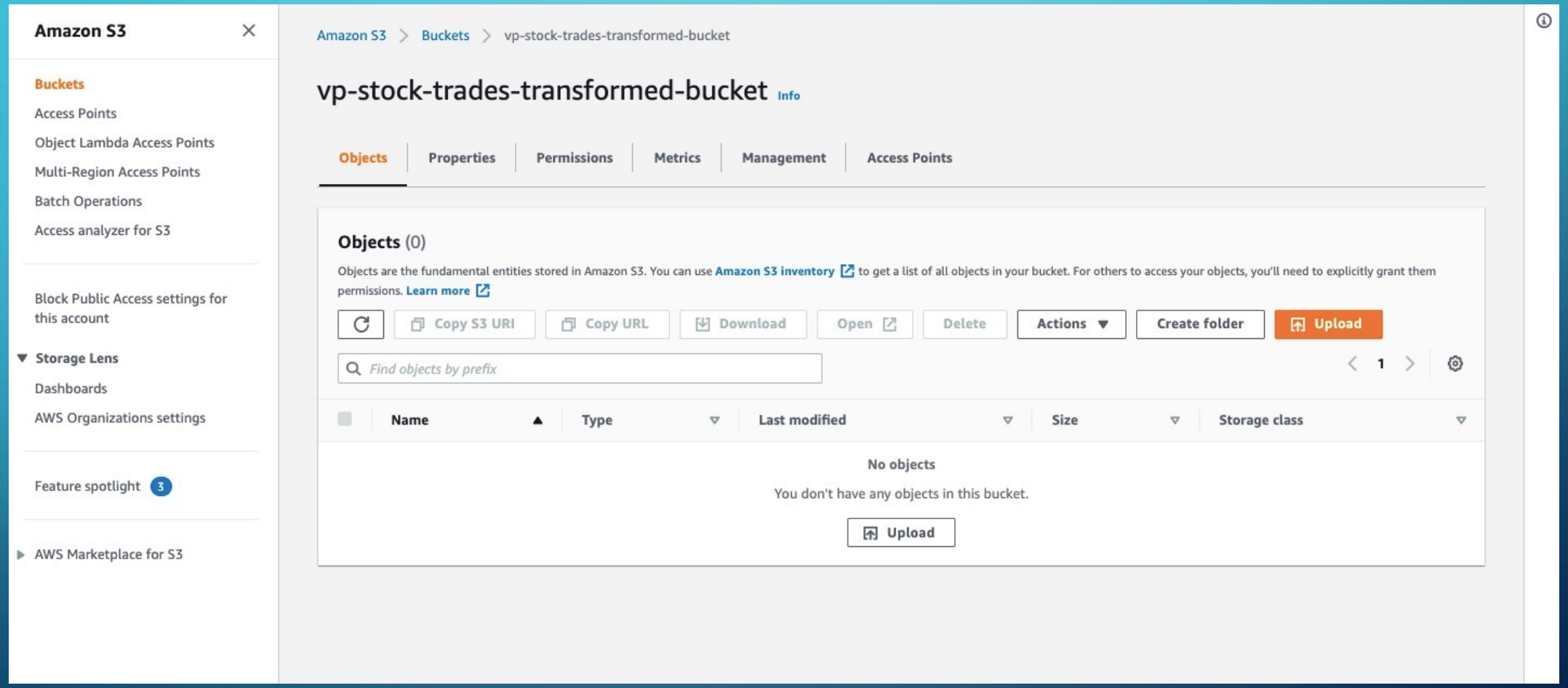
bfe0ca20-2dfc-4d44-8c50-469783b

AWS GLUE ETL JOB WITH SPARK TRANSFORMATION

15. CREATING A NEW AWS S3 BUCKET TO STORE THE TRANSFORMED DATA *(CONTINUED)*

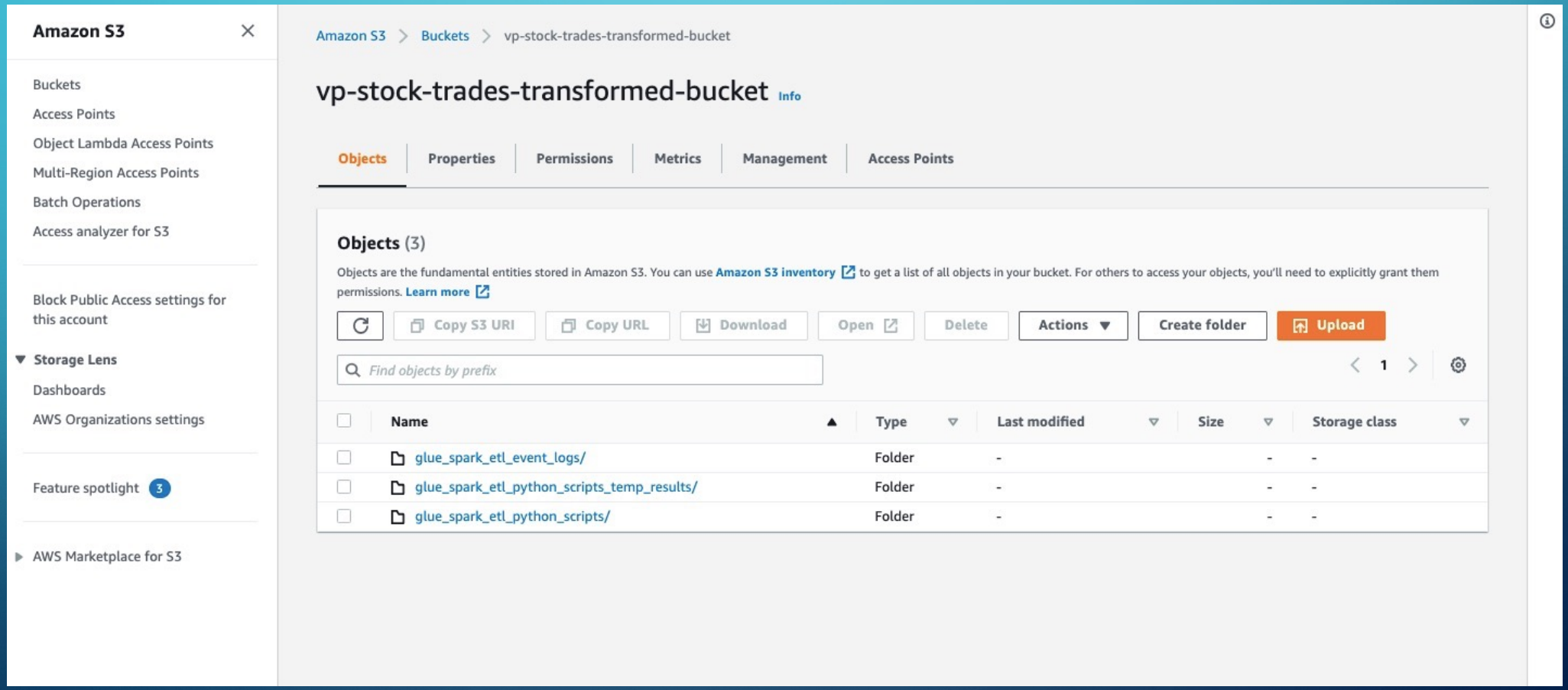


15. CREATING A NEW AWS S3 BUCKET TO STORE THE TRANSFORMED DATA



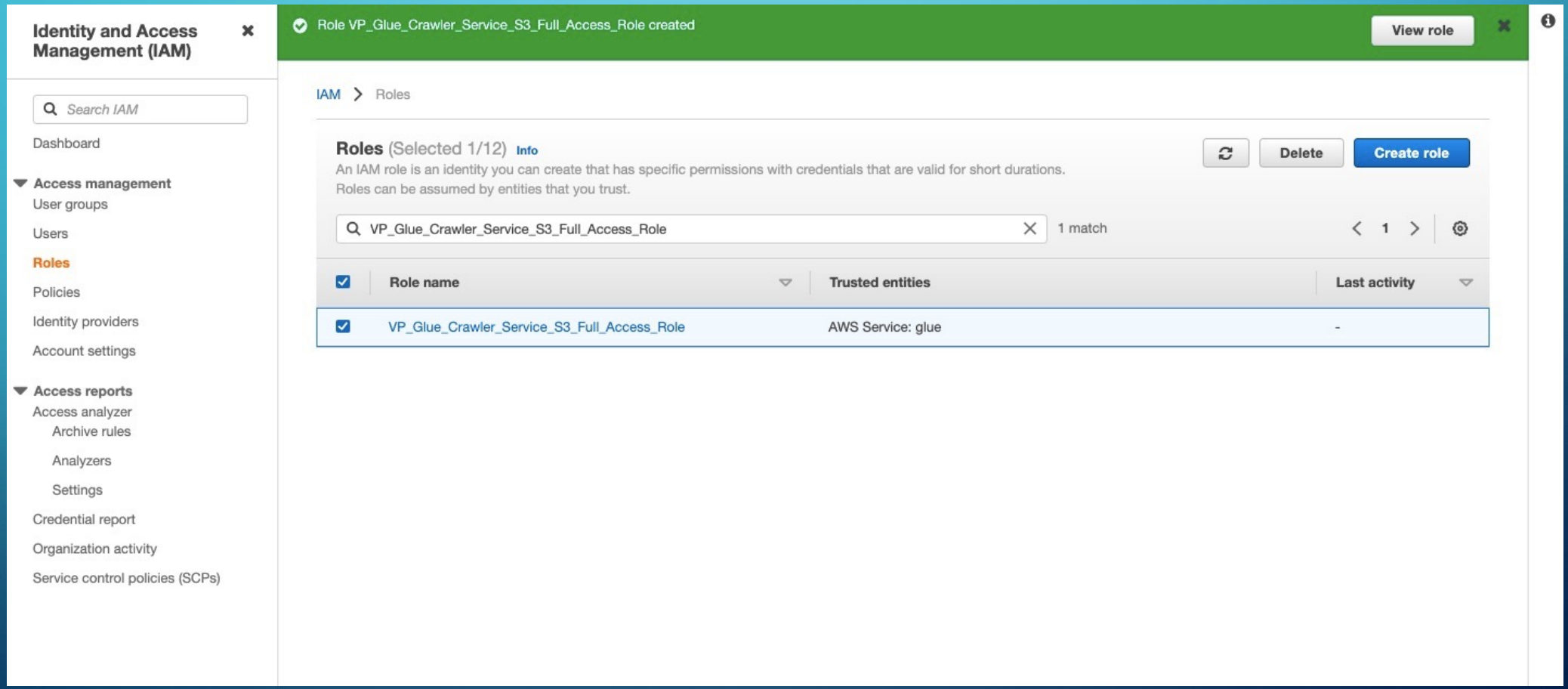
16. CREATION OF FOLDERS IN THE AWS S3 NEW BUCKET (TRANSFORMATION DATA BUCKET)

Three directories to store the AWS Glue Spark ETL Job Python Script, to store the temporary intermediate results and to store the spark job event logs respectively



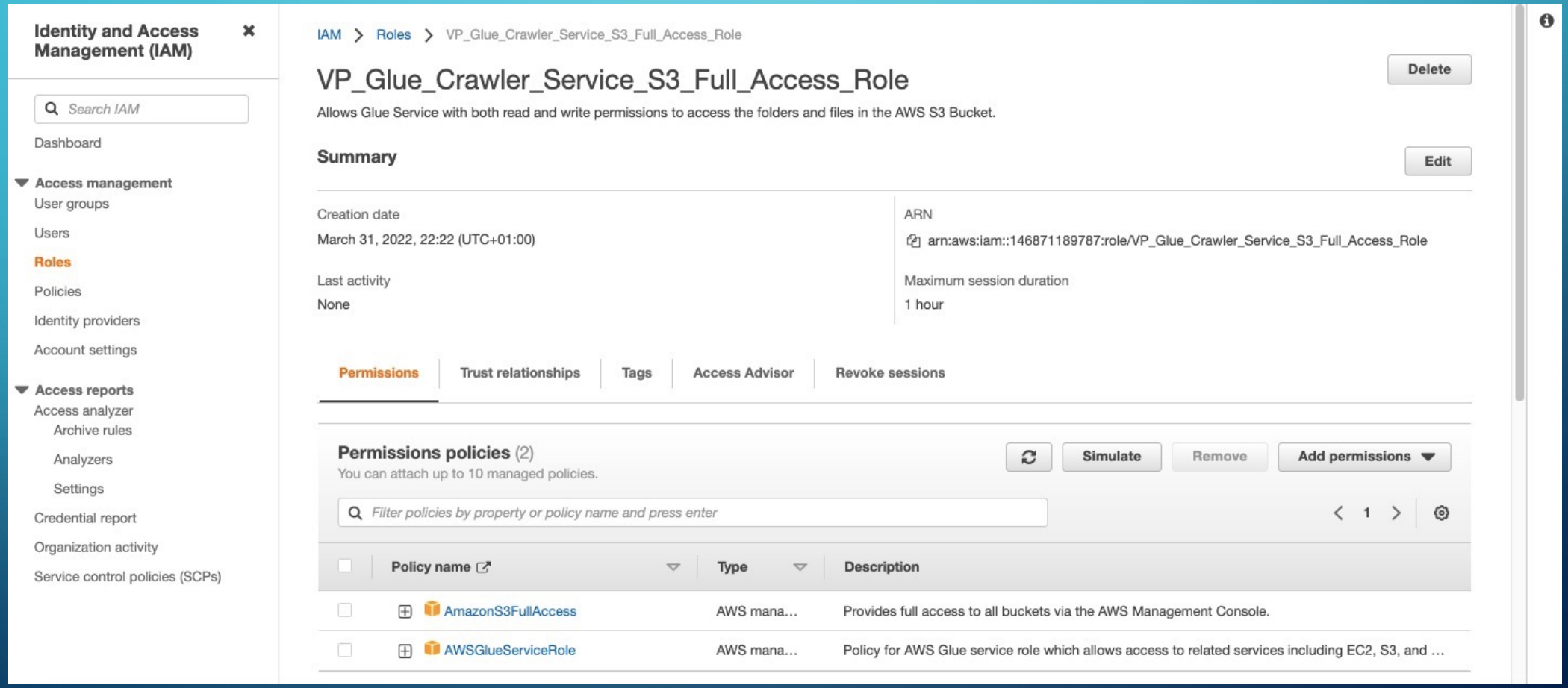
17. CREATION OF AN IAM ROLE TO GRANT PERMISSIONS TO AWS GLUE CRAWLER TO ACCESS AWS S3 *(CONTINUED)*

IAM Role with Read and Write permissions to access AWS S3 Bucket



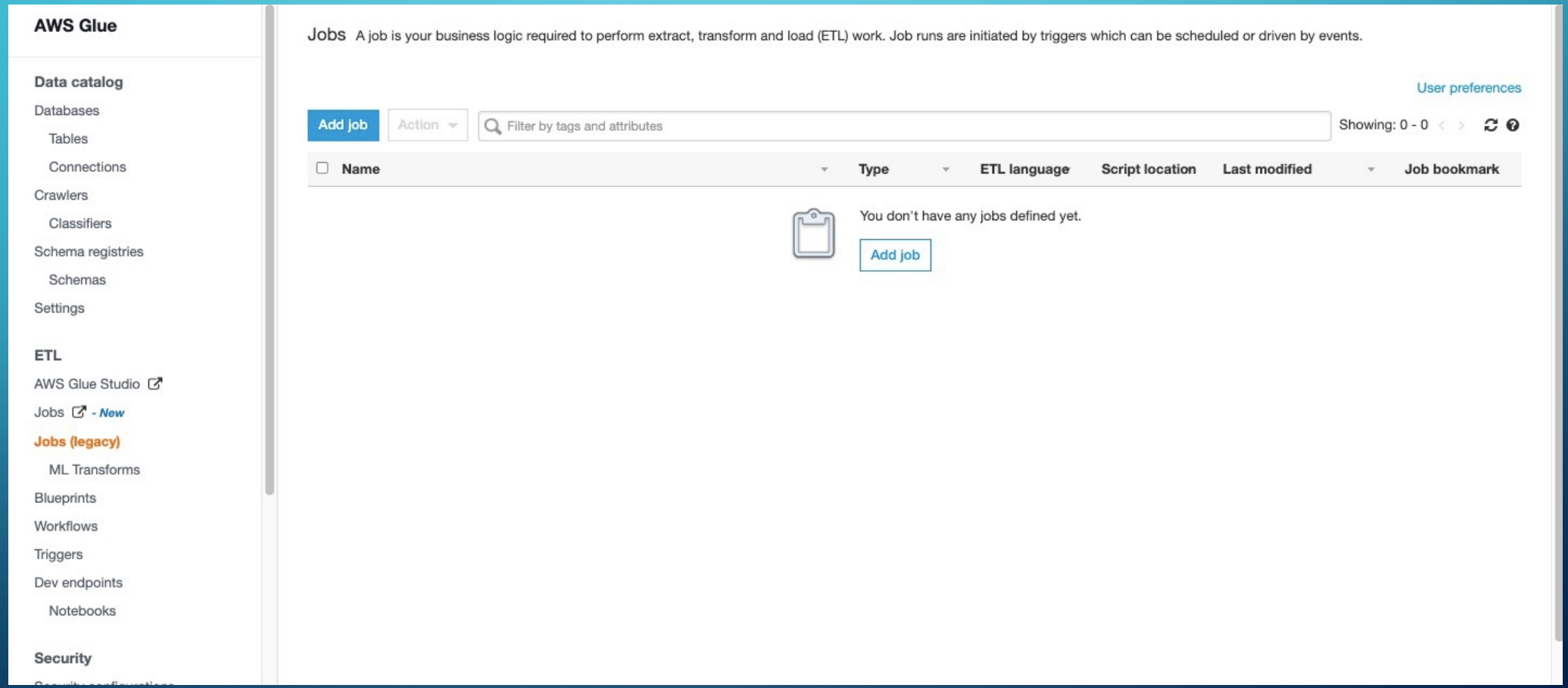
17. CREATION OF AN IAM ROLE TO GRANT PERMISSIONS TO AWS GLUE CRAWLER TO ACCESS AWS S3

IAM Role with Read and Write permissions to access AWS S3 Bucket



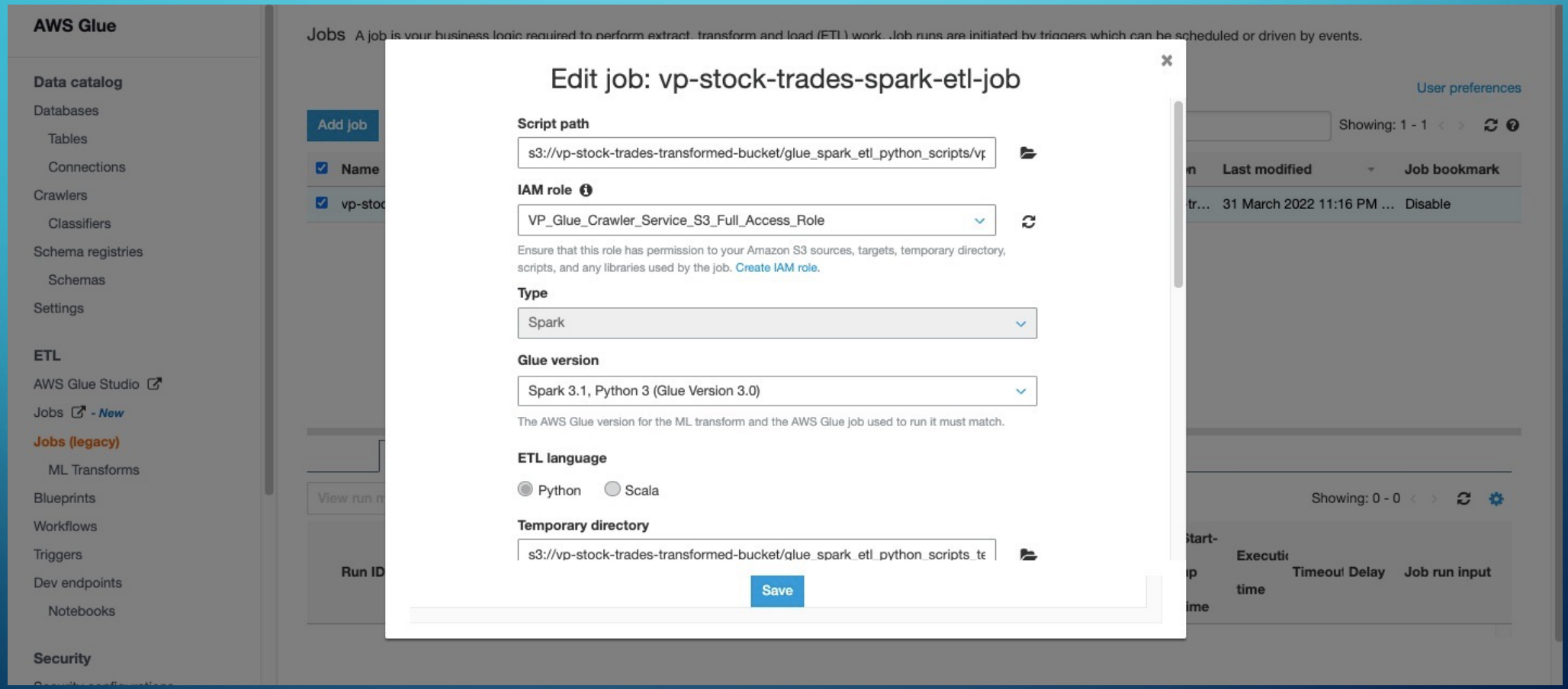
18. CREATION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



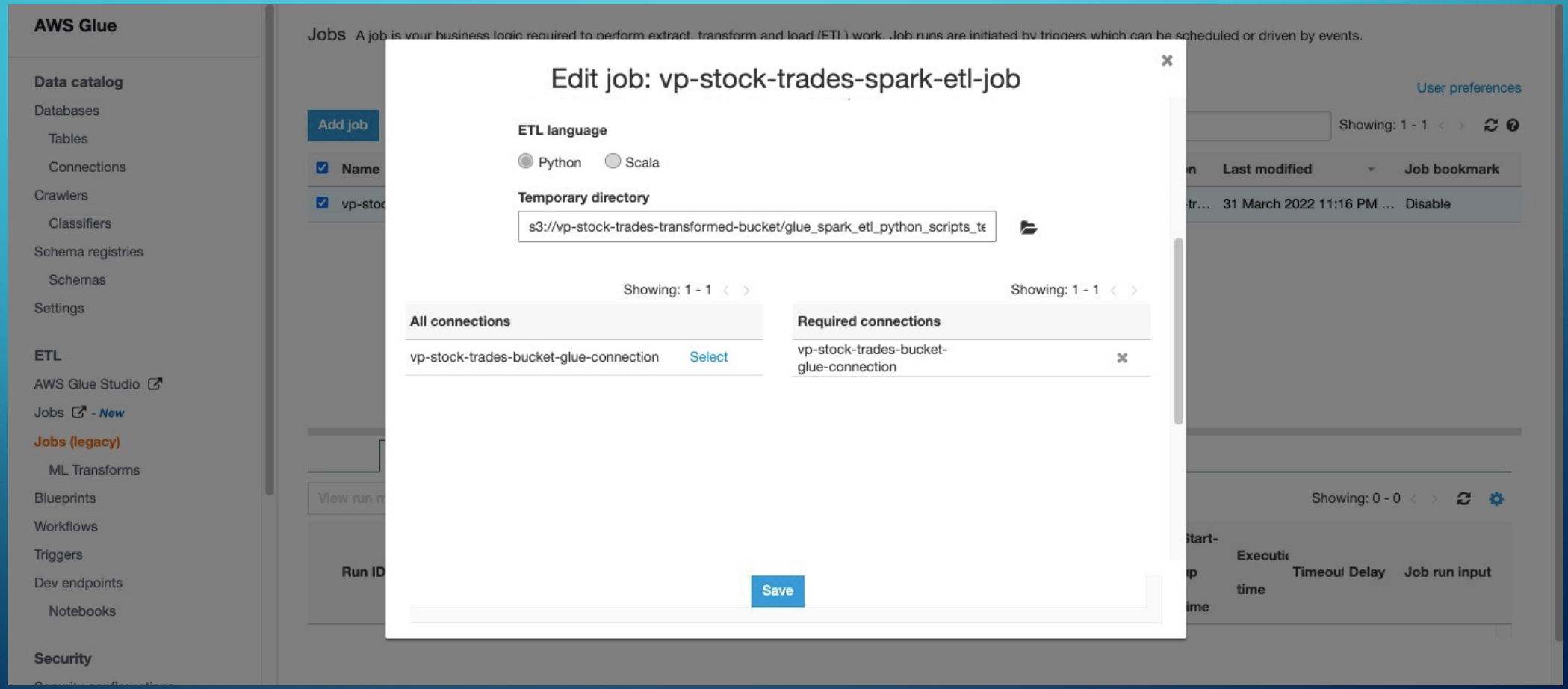
18. CREATION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



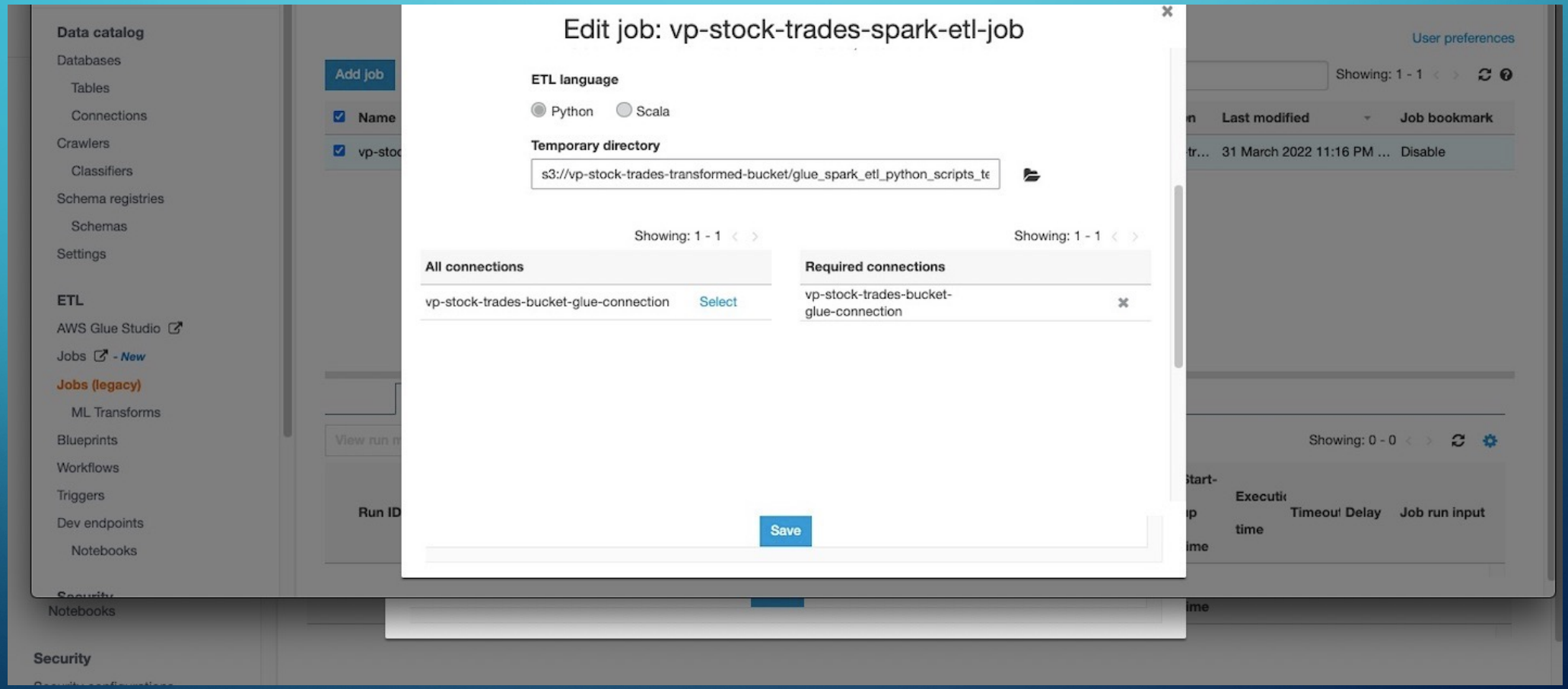
18. CREATION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



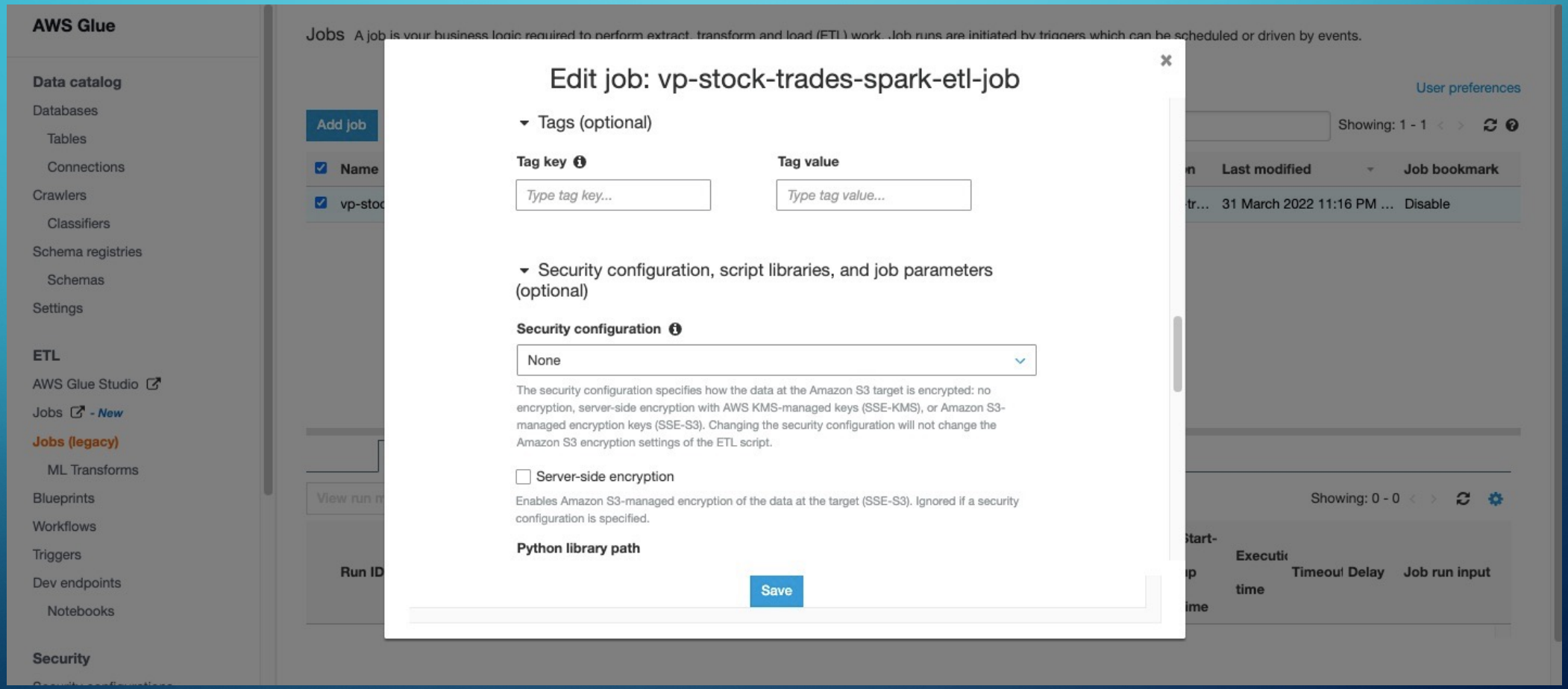
18. CREATION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



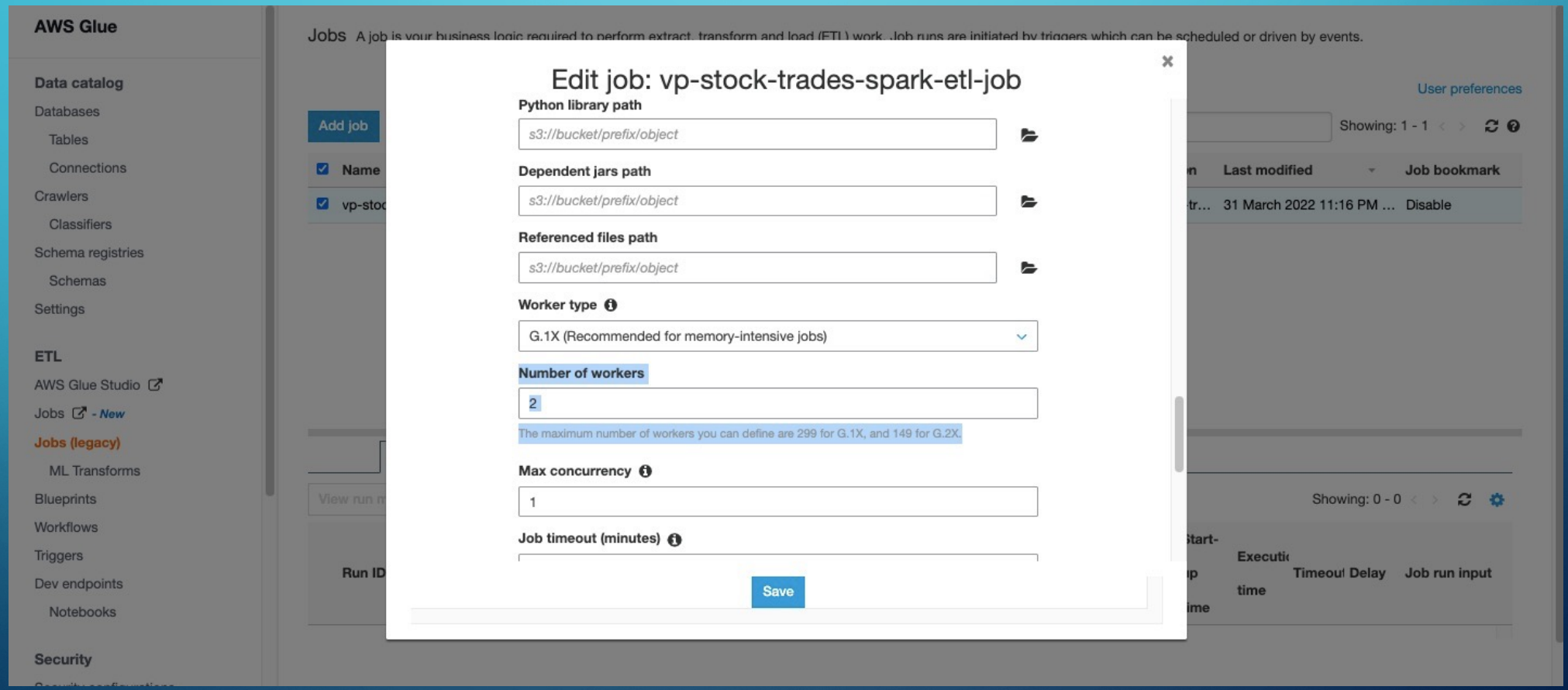
18. CREATION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



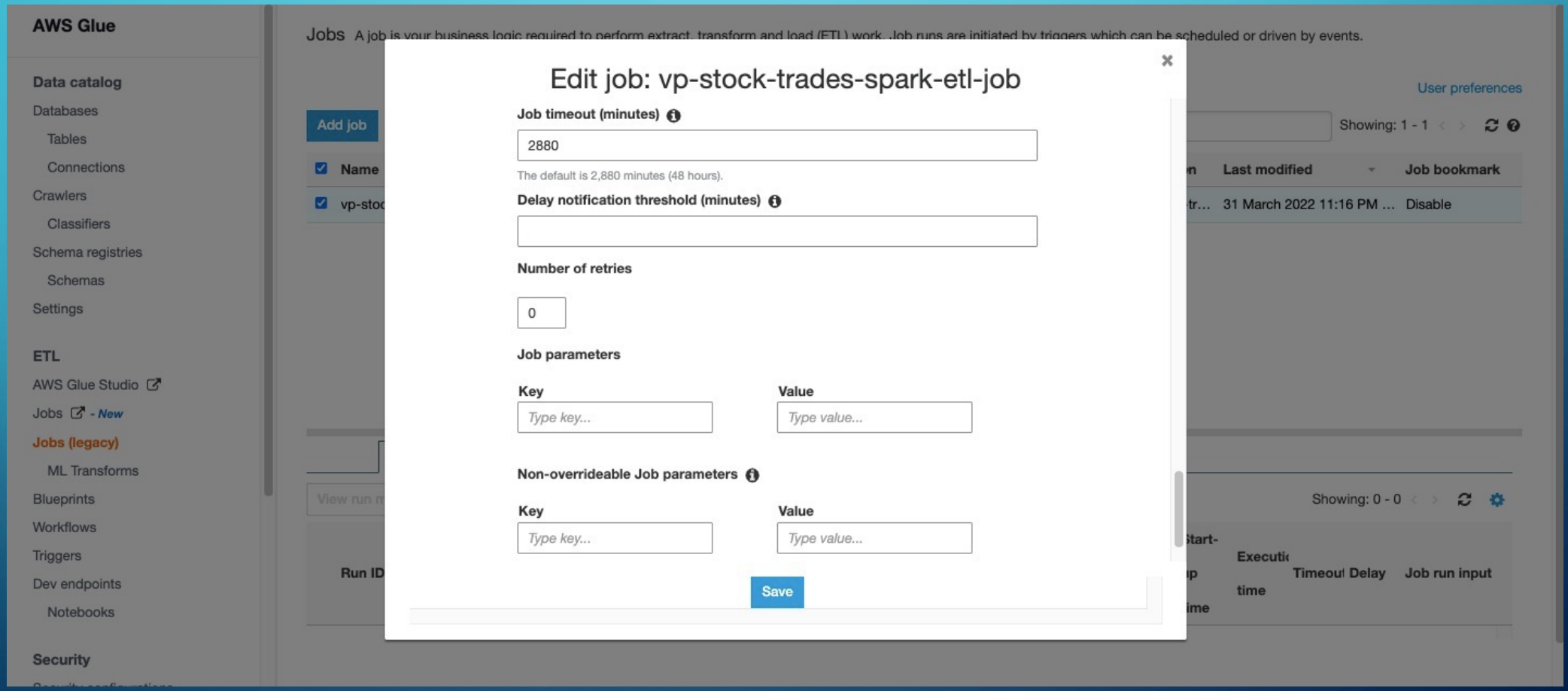
18. CREATION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



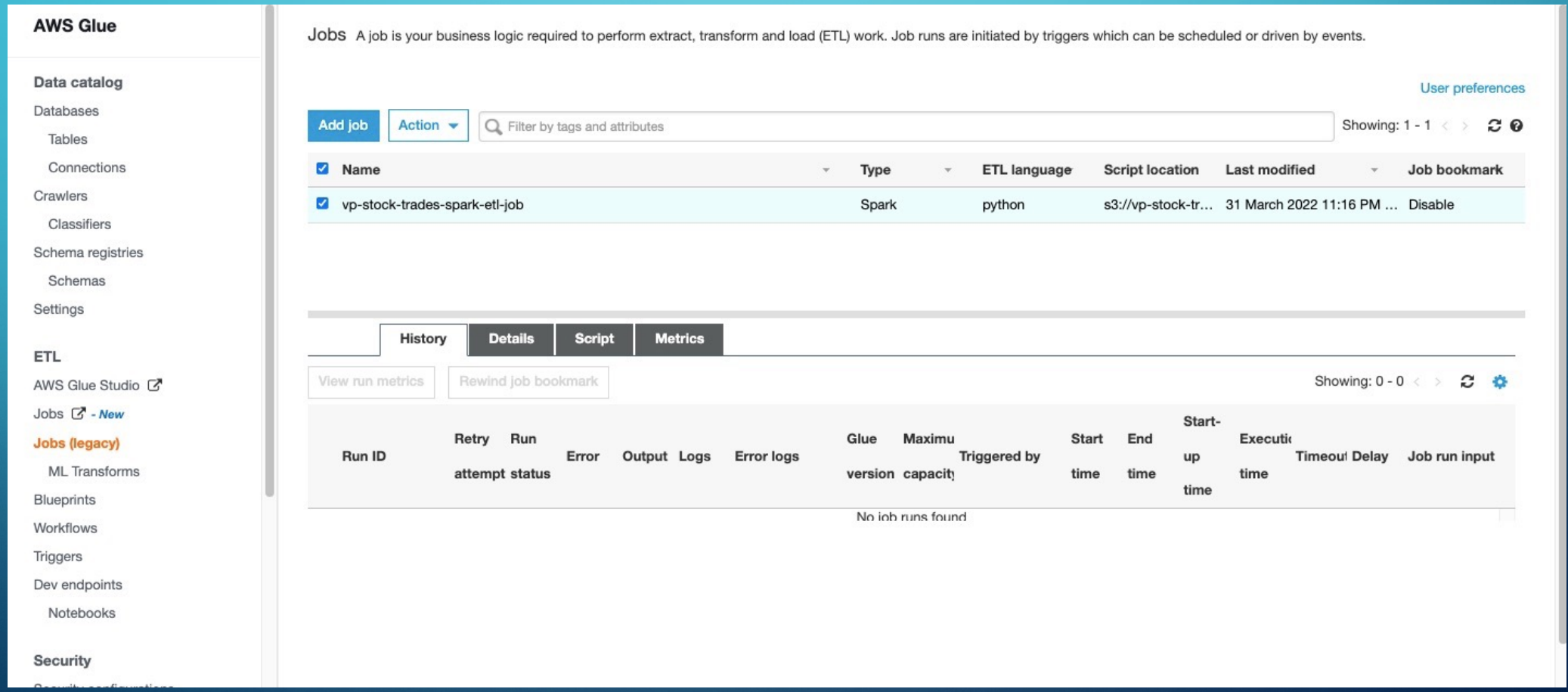
18. CREATION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’

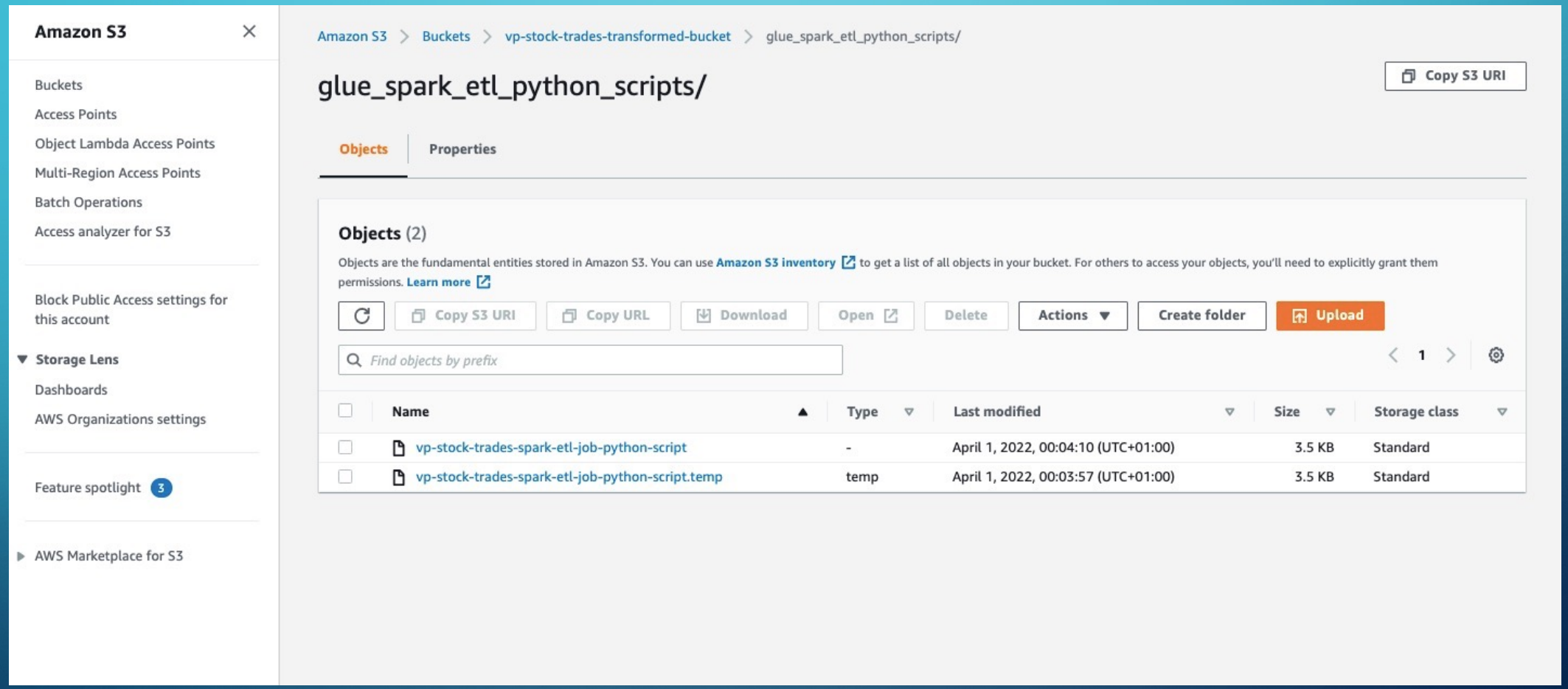


18. CREATION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’

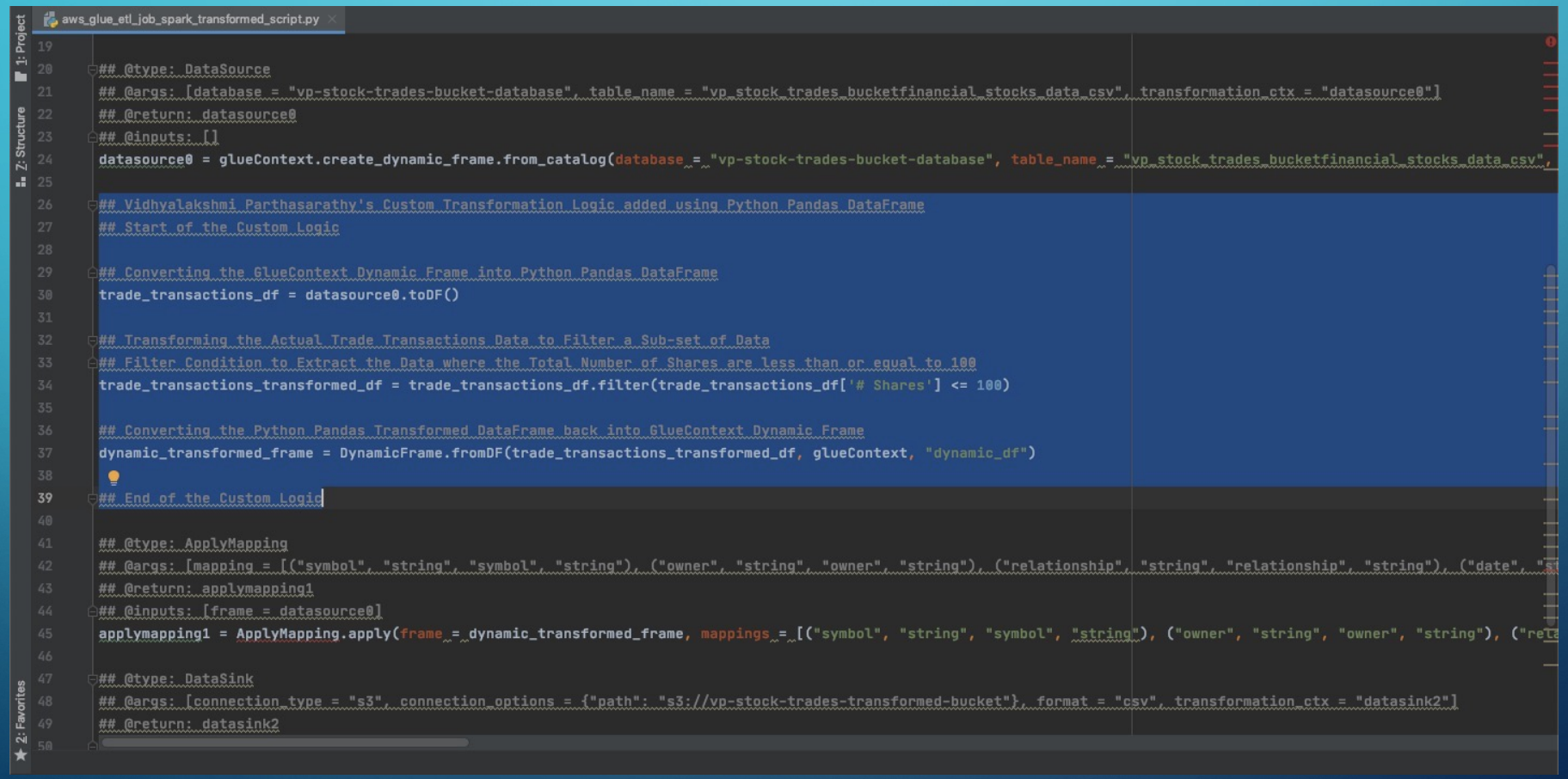


19. A COPY OF AWS GLUE ETL SPARK JOB ETL SCRIPT WITH TRANSFORMATION LOGIC SAVED INTO THE SPECIFIED PATH OF THE S3 BUCKET *(CONTINUED)*



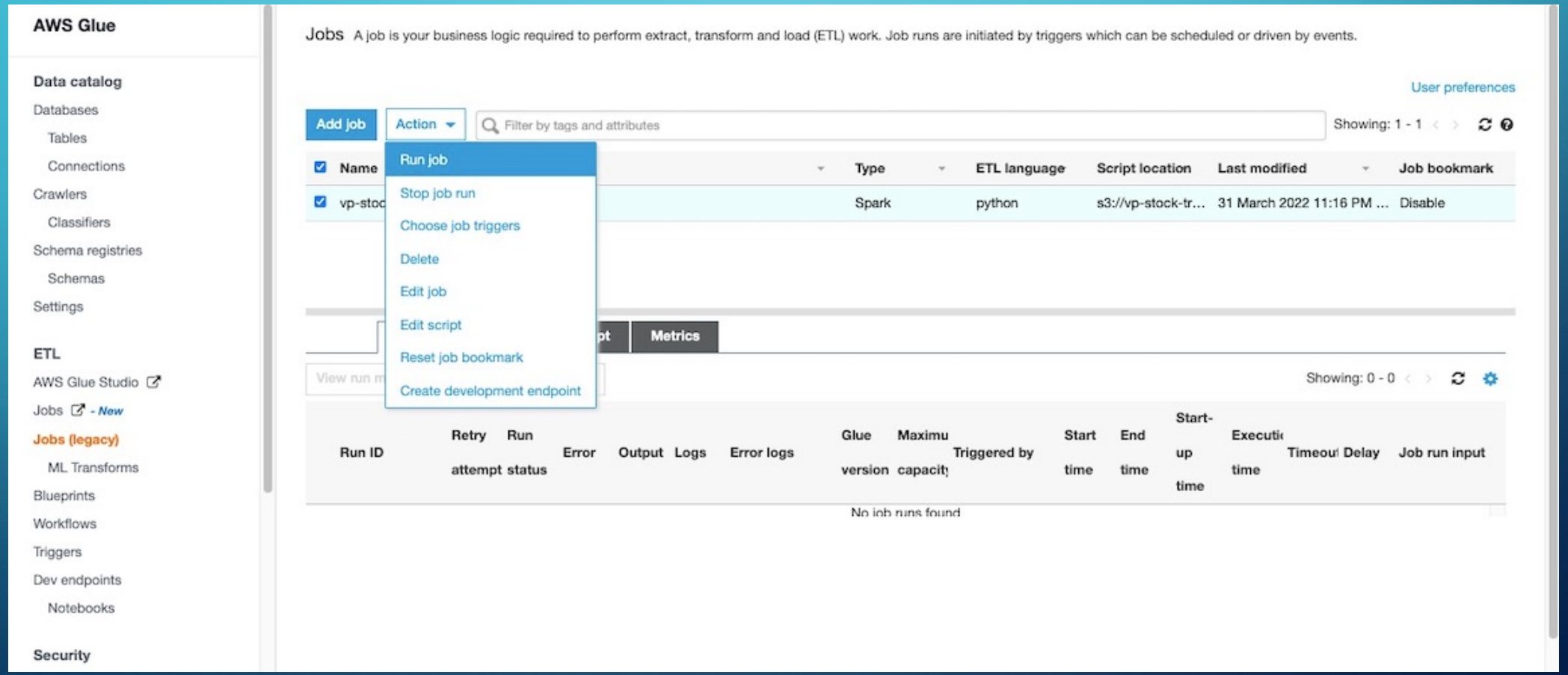
19. A COPY OF AWS GLUE ETL SPARK JOB ETL SCRIPT WITH TRANSFORMATION LOGIC SAVED INTO THE SPECIFIED PATH OF THE S3 BUCKET

Transformation Logic to filter the trade transactions with the total number of shares (‘# Shares’) as less than or equal to 100. Expected Results as 33 Trade Transactions (Rows) in the output transformed data file.



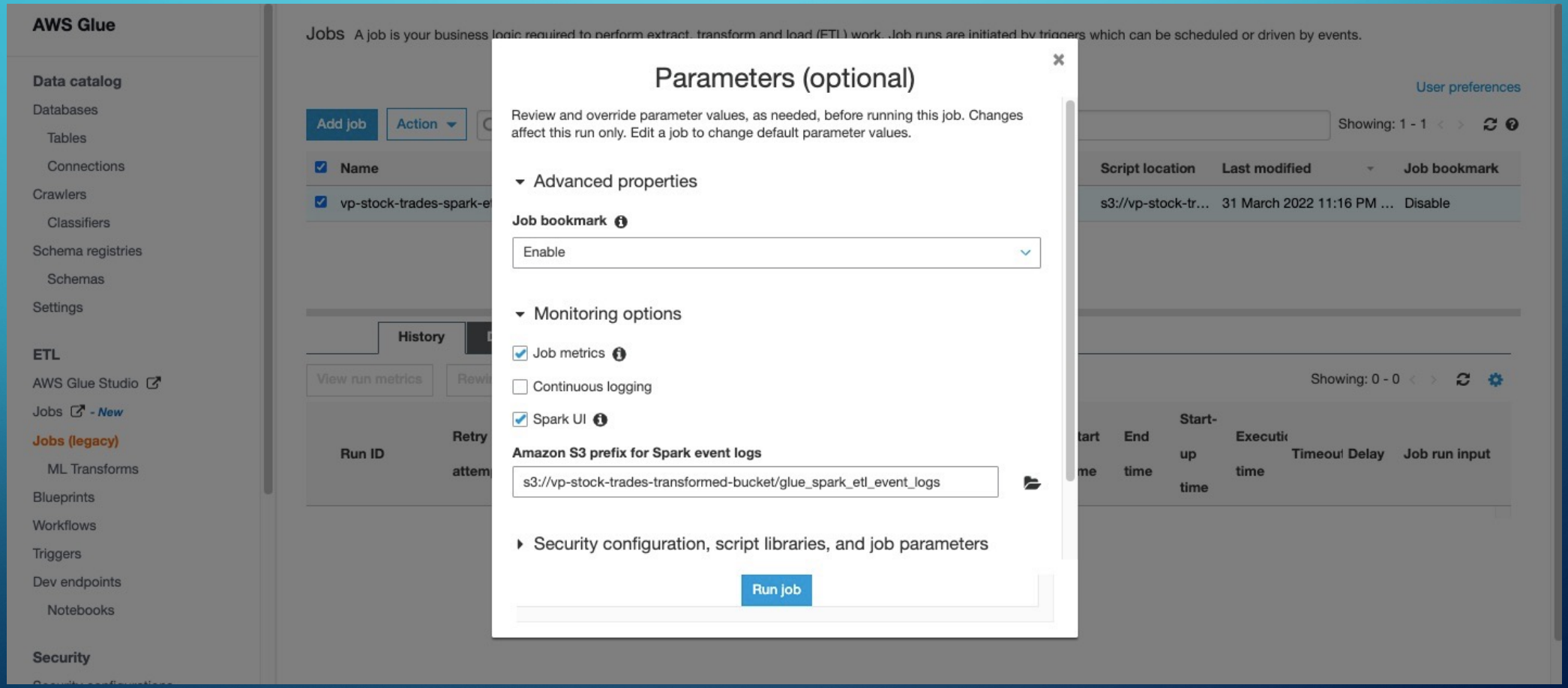
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



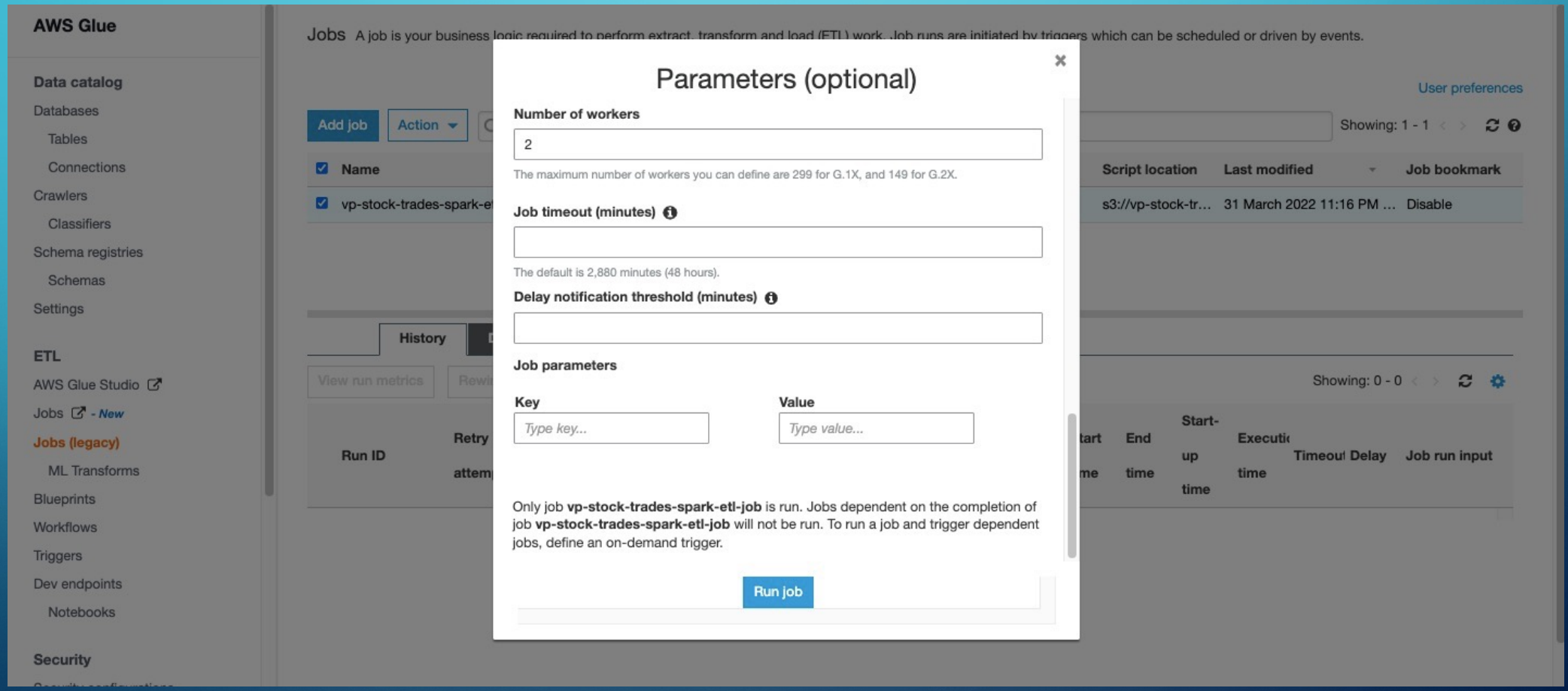
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



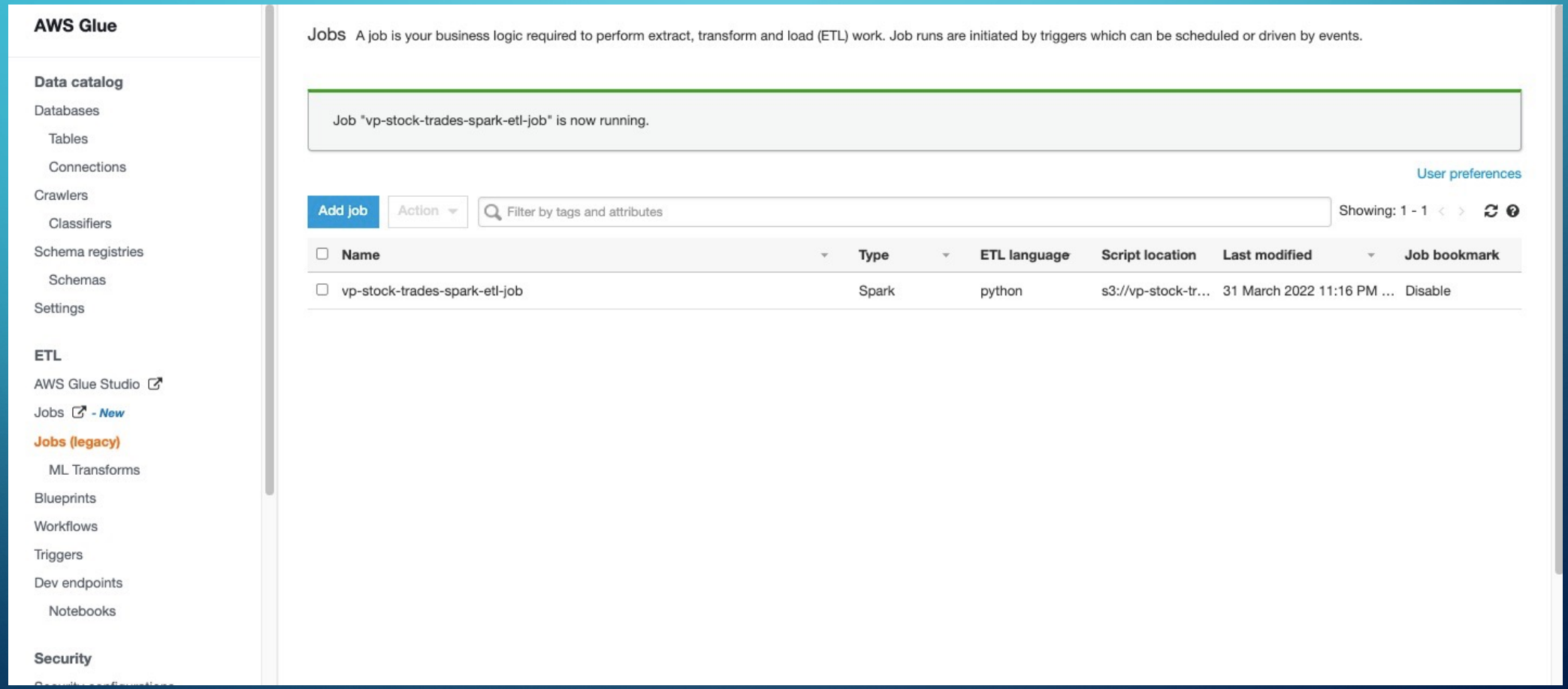
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



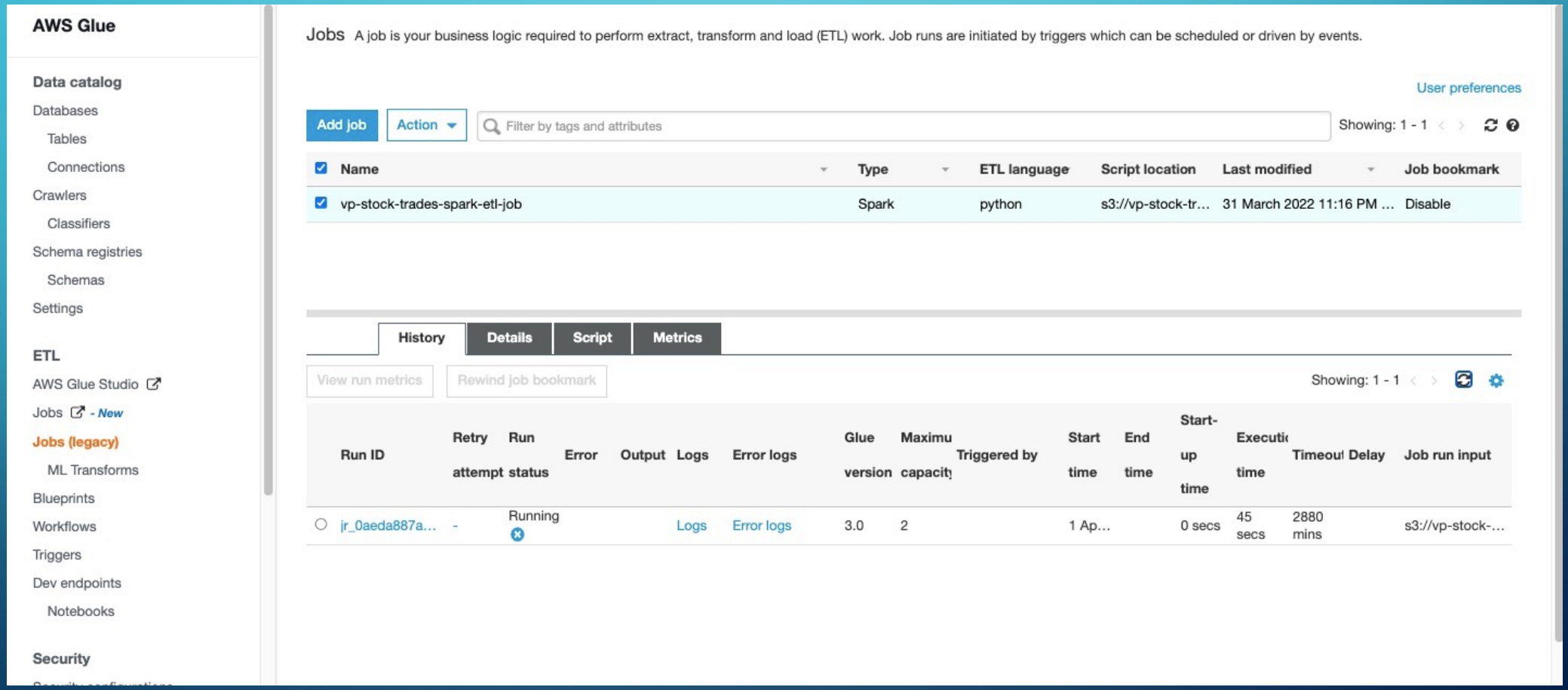
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Transformation Logic will be applied on the source raw data file “trade_transactions.csv” stored in AWS S3 Bucket ‘vp-stock-trades-bucket’



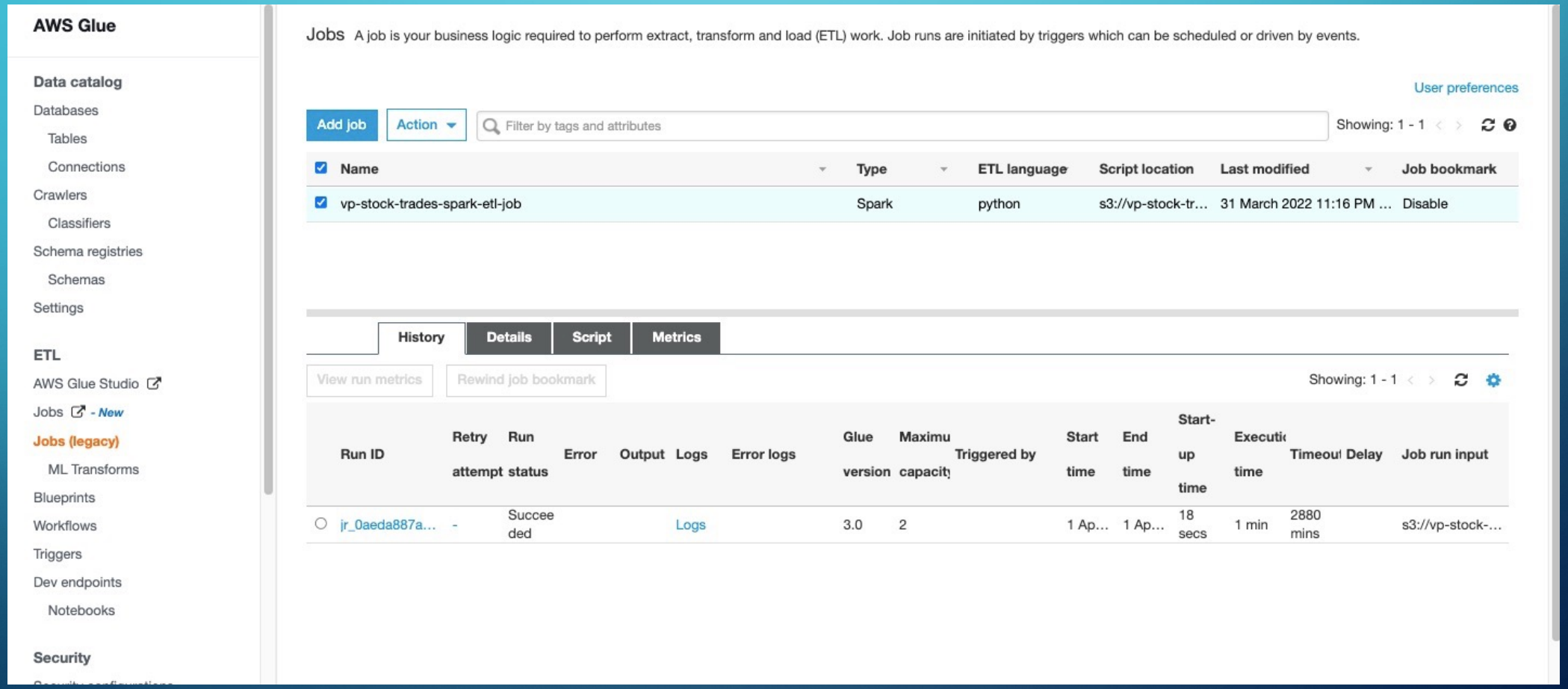
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Job Execution is in Progress



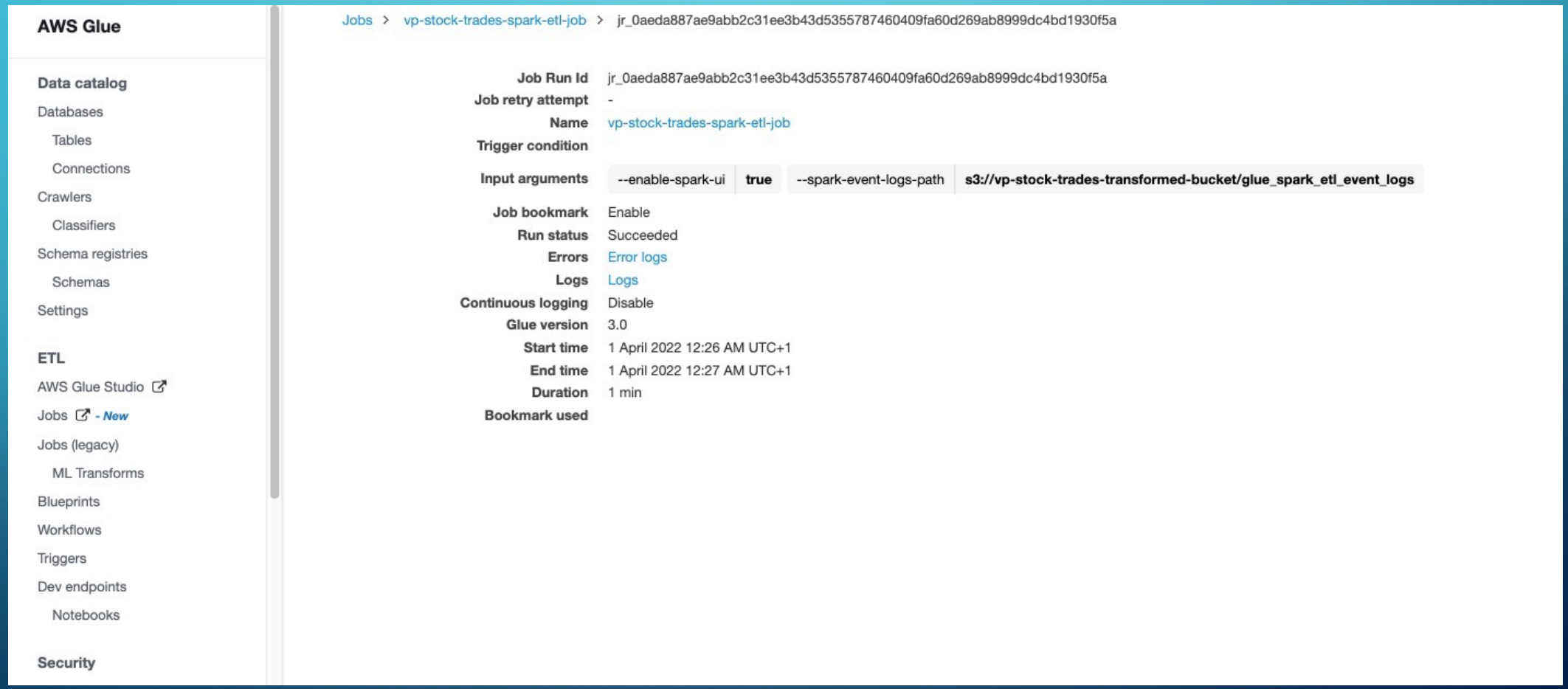
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Job Execution is Completed and Successful



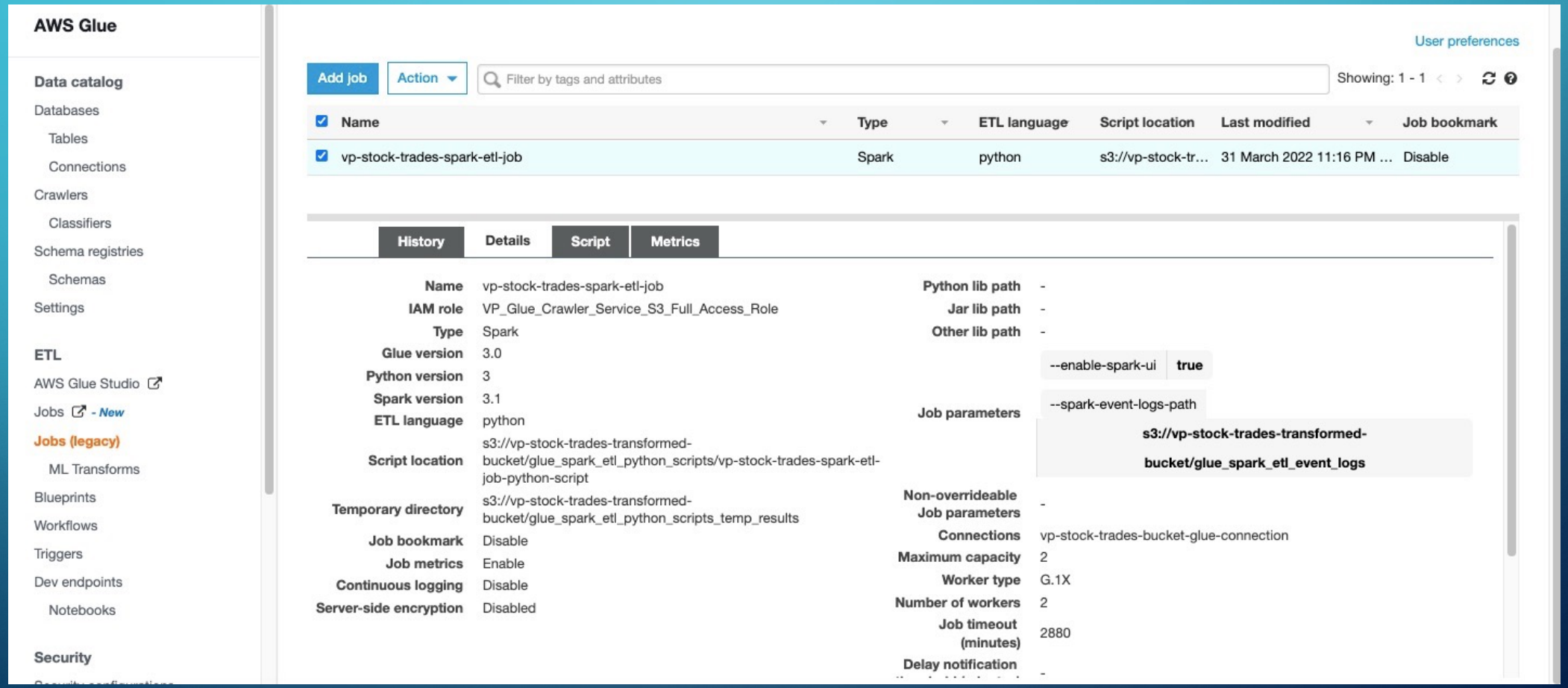
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Job Execution History



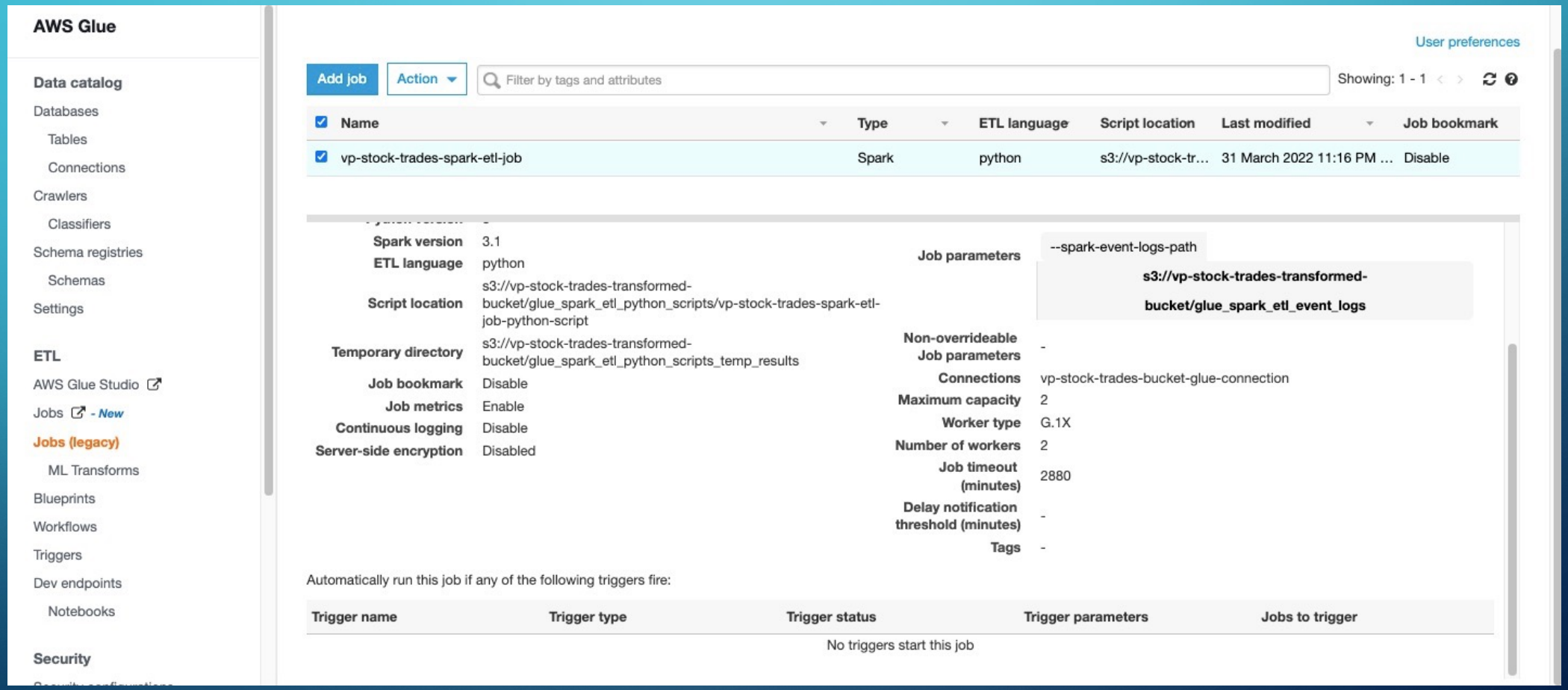
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Job Run Details



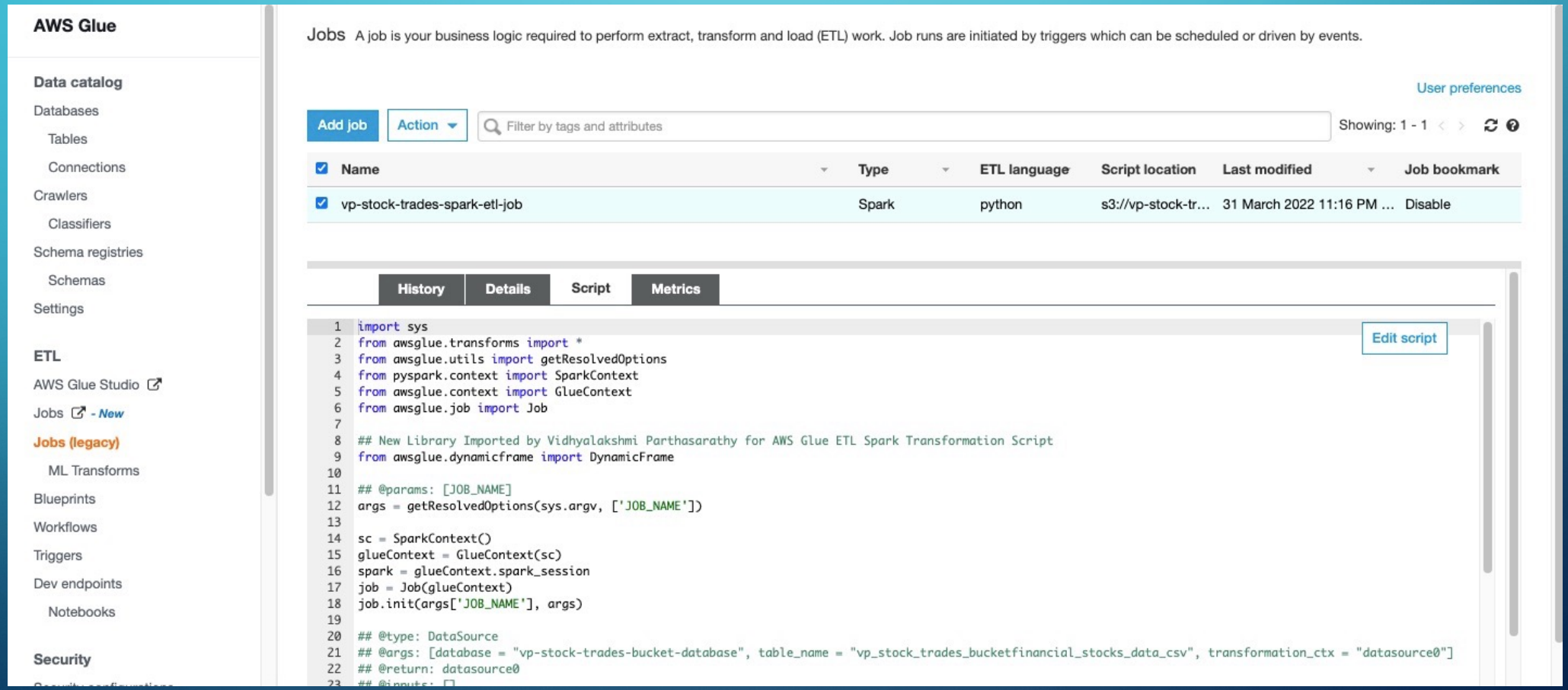
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Job Run Details



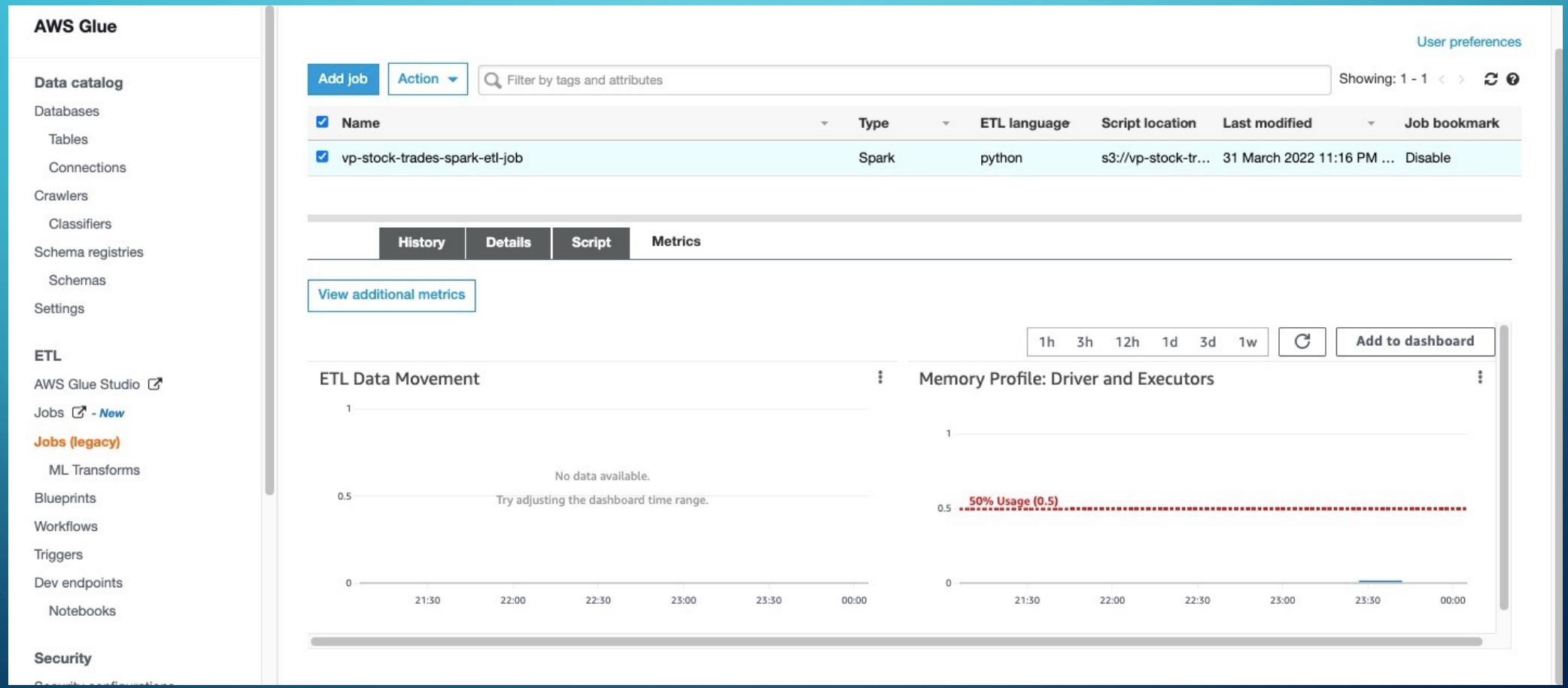
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC *(CONTINUED)*

Copy of the Spark ETL Job Python Script



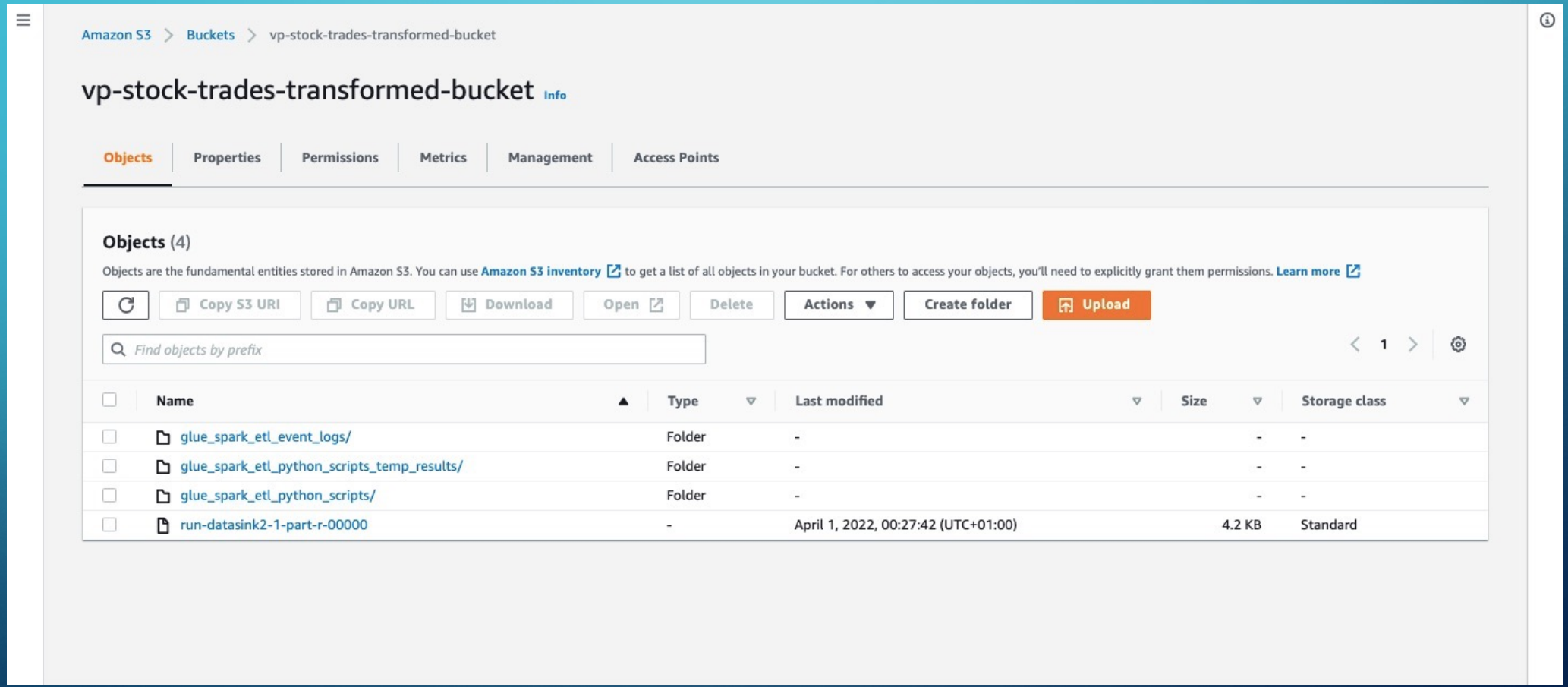
20. EXECUTION OF AWS GLUE ETL SPARK JOB WITH TRANSFORMATION LOGIC

Job Run Metrics



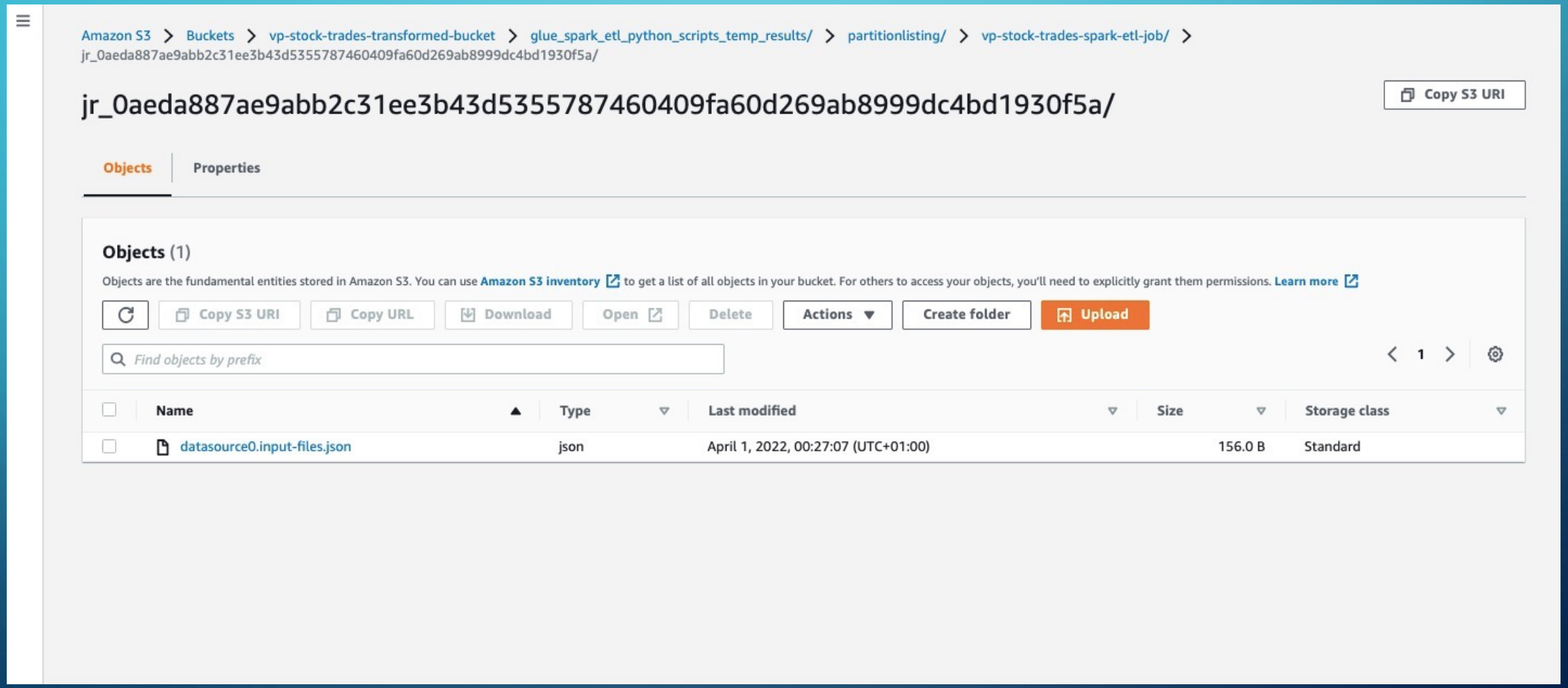
21. FILES SAVED IN THE AWS S3 BUCKET POST COMPLETION OF THE AWS GLUE SPARK ETL JOB *(CONTINUED)*

List of the folders created earlier



21. FILES SAVED IN THE AWS S3 BUCKET POST COMPLETION OF THE AWS GLUE SPARK ETL JOB *(CONTINUED)*

Temporary Intermediate Results



21. FILES SAVED IN THE AWS S3 BUCKET POST COMPLETION OF THE AWS GLUE SPARK ETL JOB *(CONTINUED)*

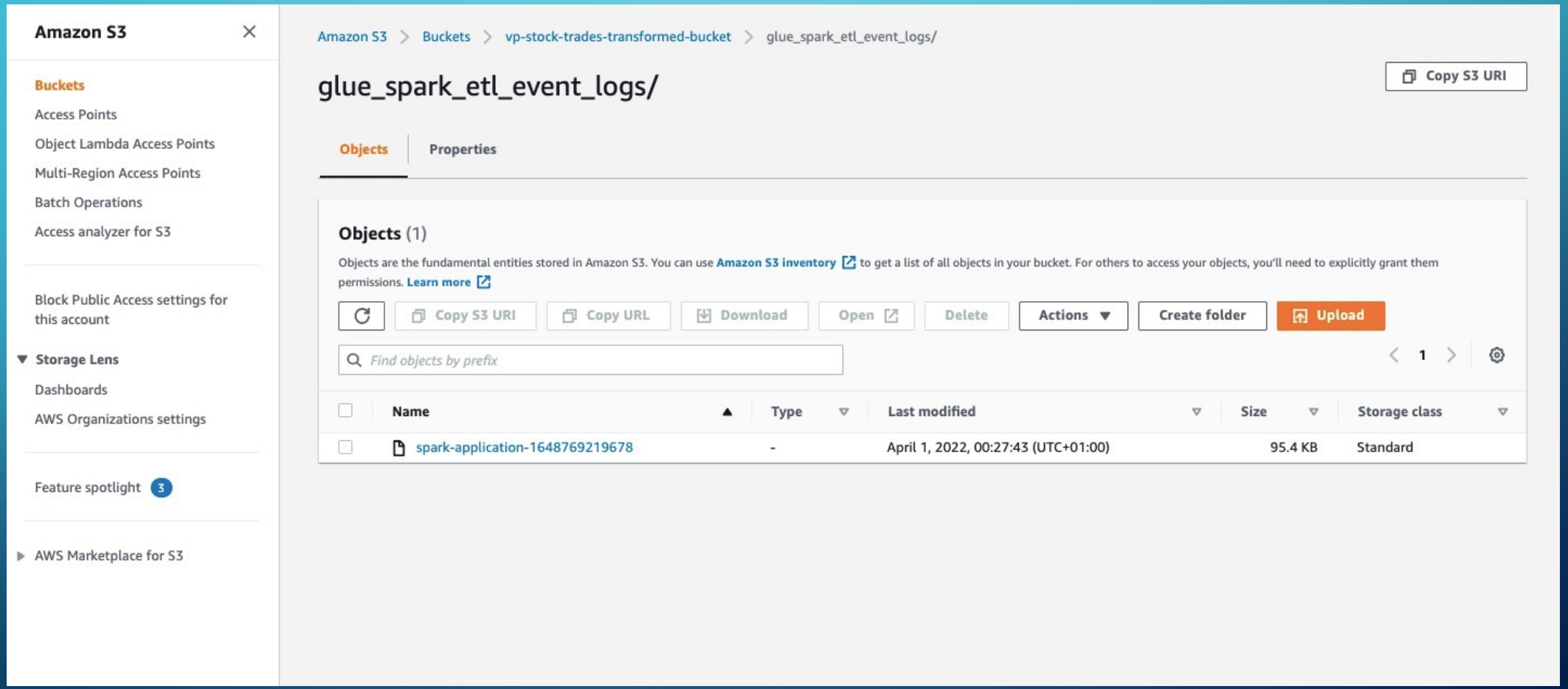
Temporary Intermediate Results

datasource0.input-files.json

[{"path":"s3://vp-stock-trades-bucket/financial_stocks_data_csv/","files":["s3://vp-stock-trades-bucket/financial_stocks_data_csv/trade_transactions.csv"]}]

21. FILES SAVED IN THE AWS S3 BUCKET POST COMPLETION OF THE AWS GLUE SPARK ETL JOB *(CONTINUED)*

Spark Event Logs



21. FILES SAVED IN THE AWS S3 BUCKET POST COMPLETION OF THE AWS GLUE SPARK ETL JOB *(CONTINUED)*

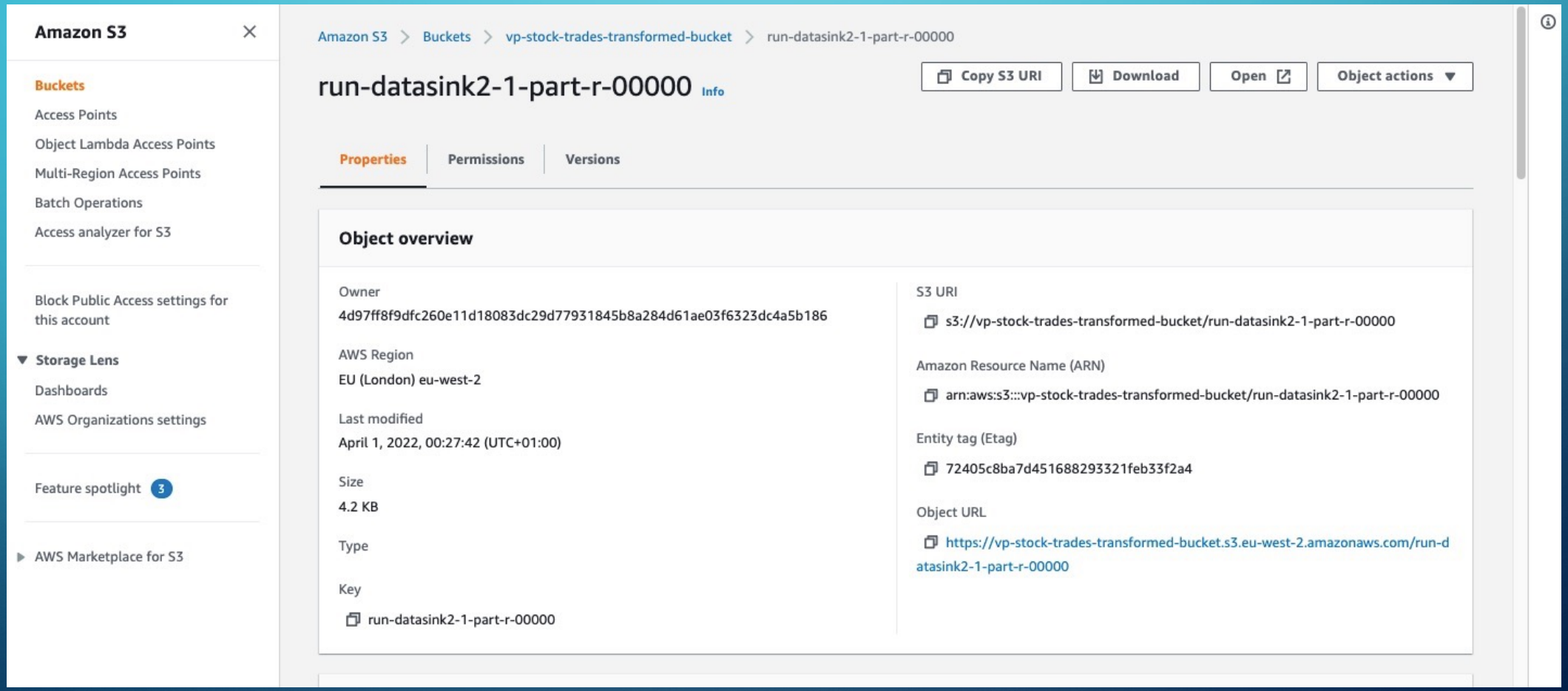
Spark Event Logs

spark-application-1648769219678.txt


```
{"Event":"SparkListenerLogStart","Spark Version":"3.1.1-amzn-0"}
{"Event":"SparkListenerResourceProfileAdded","Resource Profile Id":0,"Executor Resource Requests":{"cores":{"Resource Name":"cores","Amount":4,"Discovery Script":"","Vend
{"Event":"SparkListenerBlockManagerAdded","Block Manager ID":{"Executor ID":"driver","Host":"172.31.25.222","Port":42855},"Maximum Memory":6253707264,"Timestamp":16487692
{"Event":"SparkListenerEnvironmentUpdate","JVM Information":{"Java Home":"/usr/lib/jvm/java-1.8.0-amazon-corretto.x86_64/jre","Java Version":"1.8.0_302 (Amazon.com Inc.)"
{"Event":"SparkListenerApplicationStart","App Name":"nativespark-vp-stock-trades-spark-etl-job-jr_0aeda887ae9abb2c31ee3b43d5355787460409fa60d269ab8999dc4bd1930f5a","App I
{"Event":"SparkListenerJobStart","Job ID":0,"Submission Time":1648769227244,"Stage Infos":[{"Stage ID":0,"Stage Attempt ID":0,"Stage Name":"fromRDD at DynamicFrame.scala:
{"Event":"SparkListenerStageSubmitted","Stage Info":{"Stage ID":0,"Stage Attempt ID":0,"Stage Name":"fromRDD at DynamicFrame.scala:305","Number of Tasks":1,"RDD Info":[{"
{"Event":"SparkListenerExecutorAdded","Timestamp":1648769252106,"Executor ID":"1","Executor Info":{"Host":"172.31.31.118","Total Cores":4,"Log Urls":{},"Attributes":{},"R
{"Event":"SparkListenerBlockManagerAdded","Block Manager ID":{"Executor ID":"1","Host":"172.31.31.118","Port":38497},"Maximum Memory":6253707264,"Timestamp":1648769252217
{"Event":"SparkListenerTaskStart","Stage ID":0,"Stage Attempt ID":0,"Task Info":{"Task ID":0,"Index":0,"Attempt":0,"Launch Time":1648769252282,"Executor ID":"1","Host":"1
{"Event":"SparkListenerTaskEnd","Stage ID":0,"Stage Attempt ID":0,"Task Type":"ResultTask","Task End Reason":{"Reason":"Success"},"Task Info":{"Task ID":0,"Index":0,"Atte
{"Event":"SparkListenerStageCompleted","Stage Info":{"Stage ID":0,"Stage Attempt ID":0,"Stage Name":"fromRDD at DynamicFrame.scala:305","Number of Tasks":1,"RDD Info":[{"
{"Event":"SparkListenerJobEnd","Job ID":0,"Completion Time":1648769255144,"Job Result":{"Result":"JobSucceeded"}}
{"Event":"SparkListenerJobStart","Job ID":1,"Submission Time":1648769259376,"Stage Infos":[{"Stage ID":1,"Stage Attempt ID":0,"Stage Name":"runJob at HadoopWriters.scala:
{"Event":"SparkListenerStageSubmitted","Stage Info":{"Stage ID":1,"Stage Attempt ID":0,"Stage Name":"runJob at HadoopWriters.scala:129","Number of Tasks":1,"RDD Info":[{"
{"Event":"SparkListenerTaskStart","Stage ID":1,"Stage Attempt ID":0,"Task Info":{"Task ID":1,"Index":0,"Attempt":0,"Launch Time":1648769259406,"Executor ID":"1","Host":"1
{"Event":"SparkListenerTaskEnd","Stage ID":1,"Stage Attempt ID":0,"Task Type":"ResultTask","Task End Reason":{"Reason":"Success"},"Task Info":{"Task ID":1,"Index":0,"Atte
{"Event":"SparkListenerStageCompleted","Stage Info":{"Stage ID":1,"Stage Attempt ID":0,"Stage Name":"runJob at HadoopWriters.scala:129","Number of Tasks":1,"RDD Info":[{"
{"Event":"SparkListenerJobEnd","Job ID":1,"Completion Time":1648769262069,"Job Result":{"Result":"JobSucceeded"}}
{"Event":"SparkListenerApplicationEnd","Timestamp":1648769262161}
```

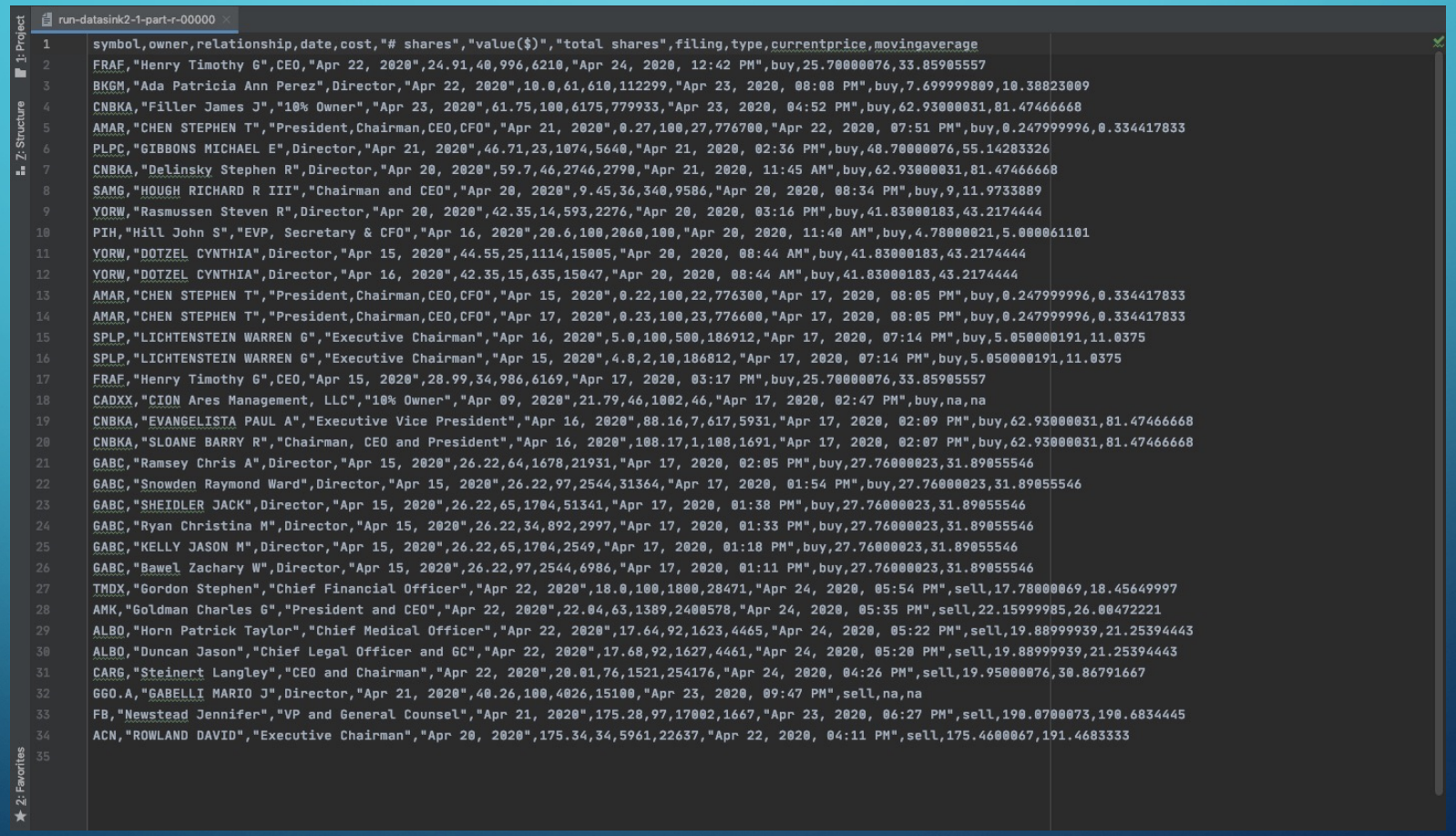
21. FILES SAVED IN THE AWS S3 BUCKET POST COMPLETION OF THE AWS GLUE SPARK ETL JOB *(CONTINUED)*

Transformed Data File



21. FILES SAVED IN THE AWS S3 BUCKET POST COMPLETION OF THE AWS GLUE SPARK ETL JOB

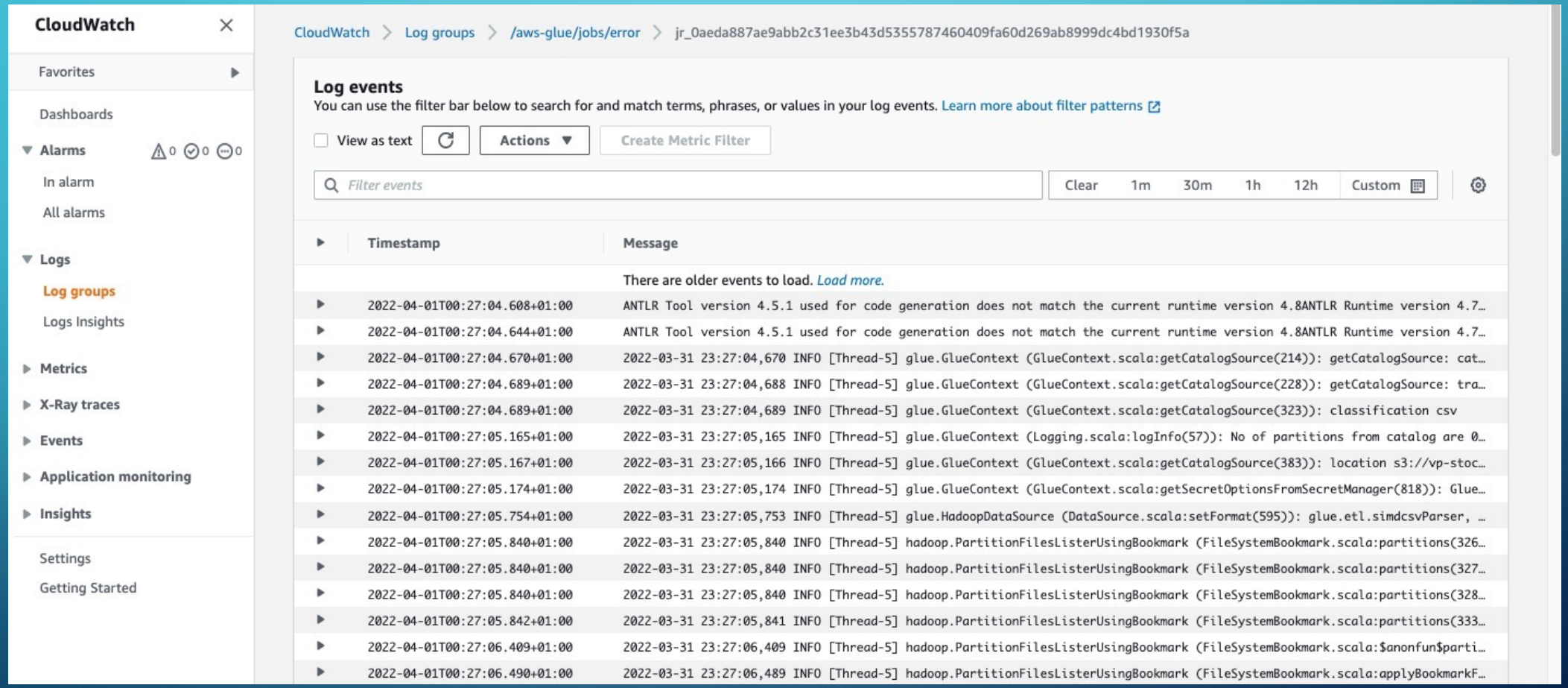
- Transformed Data File
- Transformation Logic to filter the trade transactions with the total number of shares (‘# Shares’) as less than or equal to 100. Expected Results achieved as 33 Trade Transactions (Rows) in the output transformed data file.



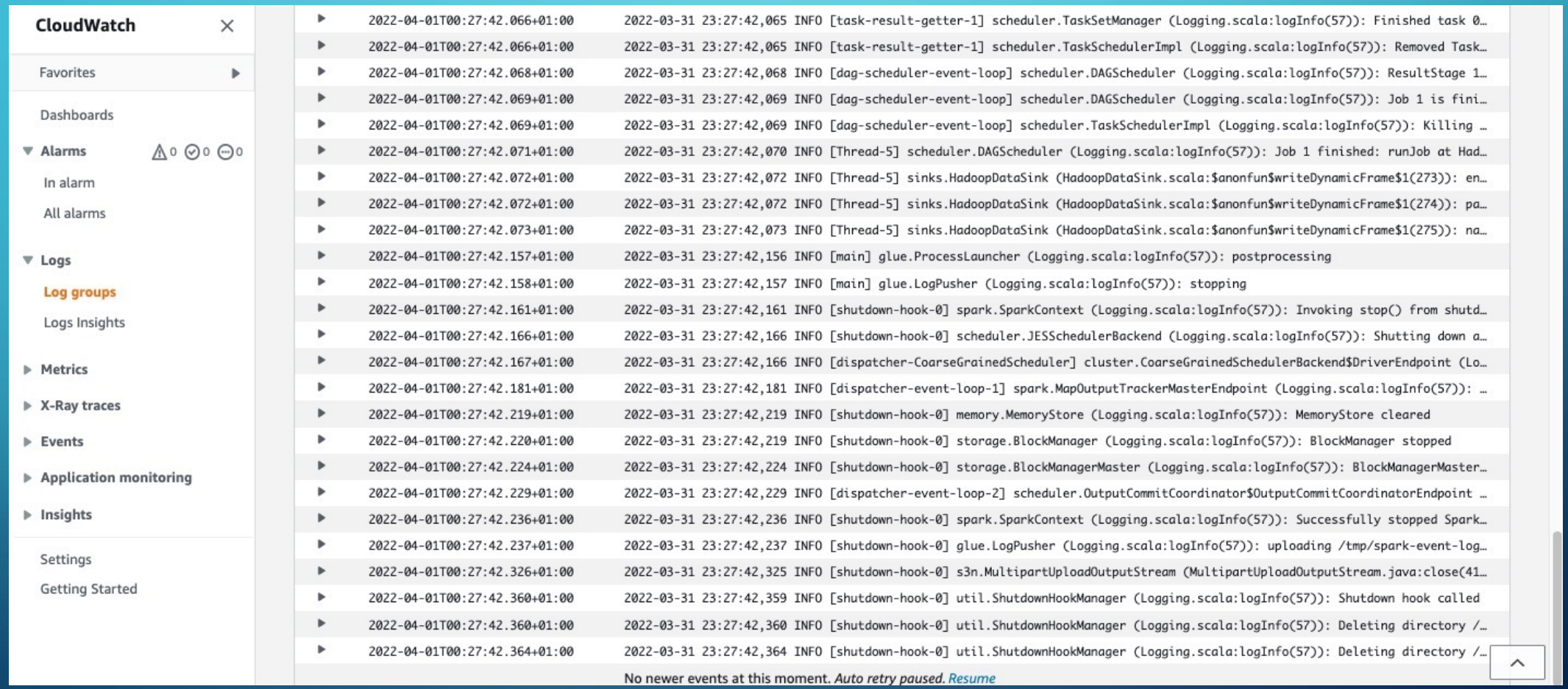
run-datasink2-1-part-r-00000

```
symbol,owner,relationship,date,cost,"# shares","value($)","total shares",filing,type,currentprice,movingaverage
FRAF,"Henry Timothy G",CEO,"Apr 22, 2020",24.91,40,996,6210,"Apr 24, 2020, 12:42 PM",buy,25.70000076,33.85905557
BKGM,"Ada Patricia Ann Perez",Director,"Apr 22, 2020",10.0,61,610,112299,"Apr 23, 2020, 08:08 PM",buy,7.699999809,10.38823009
CNBKA,"Filler James J","10% Owner","Apr 23, 2020",61.75,100,6175,779933,"Apr 23, 2020, 04:52 PM",buy,62.93000031,81.47466668
AMAR,"CHEN STEPHEN T","President,Chairman,CEO,CFO","Apr 21, 2020",0.27,100,27,776700,"Apr 22, 2020, 07:51 PM",buy,0.247999996,0.334417833
PLPC,"GIBBONS MICHAEL E",Director,"Apr 21, 2020",46.71,23,1074,5640,"Apr 21, 2020, 02:36 PM",buy,48.70000076,55.14283326
CNBKA,"Delinsky Stephen R",Director,"Apr 20, 2020",59.7,46,2746,2790,"Apr 21, 2020, 11:45 AM",buy,62.93000031,81.47466668
SAMG,"HOUGH RICHARD R III","Chairman and CEO","Apr 20, 2020",9.45,36,340,9586,"Apr 20, 2020, 08:34 PM",buy,9,11.9733889
YORW,"Rasmussen Steven R",Director,"Apr 20, 2020",42.35,14,593,2276,"Apr 20, 2020, 03:16 PM",buy,41.83000183,43.2174444
PIH,"Hill John S","EVP, Secretary & CFO","Apr 16, 2020",20.6,100,2060,100,"Apr 20, 2020, 11:40 AM",buy,4.78000021,5.000061101
YORW,"DOTZEL CYNTHIA",Director,"Apr 15, 2020",44.55,25,1114,15005,"Apr 20, 2020, 08:44 AM",buy,41.83000183,43.2174444
YORW,"DOTZEL CYNTHIA",Director,"Apr 16, 2020",42.35,15,635,15047,"Apr 20, 2020, 08:44 AM",buy,41.83000183,43.2174444
AMAR,"CHEN STEPHEN T","President,Chairman,CEO,CFO","Apr 15, 2020",0.22,100,22,776300,"Apr 17, 2020, 08:05 PM",buy,0.247999996,0.334417833
AMAR,"CHEN STEPHEN T","President,Chairman,CEO,CFO","Apr 17, 2020",0.23,100,23,776600,"Apr 17, 2020, 08:05 PM",buy,0.247999996,0.334417833
SPLP,"LICHTENSTEIN WARREN G","Executive Chairman","Apr 16, 2020",5.0,100,500,186912,"Apr 17, 2020, 07:14 PM",buy,5.050000191,11.0375
SPLP,"LICHTENSTEIN WARREN G","Executive Chairman","Apr 15, 2020",4.8,2,10,186812,"Apr 17, 2020, 07:14 PM",buy,5.050000191,11.0375
FRAF,"Henry Timothy G",CEO,"Apr 15, 2020",28.99,34,986,6169,"Apr 17, 2020, 03:17 PM",buy,25.70000076,33.85905557
CADXX,"CION Ares Management, LLC","10% Owner","Apr 09, 2020",21.79,46,1002,46,"Apr 17, 2020, 02:47 PM",buy,na,na
CNBKA,"EVANGELISTA PAUL A","Executive Vice President","Apr 16, 2020",88.16,7,617,5931,"Apr 17, 2020, 02:09 PM",buy,62.93000031,81.47466668
CNBKA,"SLOANE BARRY R","Chairman, CEO and President","Apr 16, 2020",108.17,1,108,1691,"Apr 17, 2020, 02:07 PM",buy,62.93000031,81.47466668
GABC,"Ramsey Chris A",Director,"Apr 15, 2020",26.22,64,1678,21931,"Apr 17, 2020, 02:05 PM",buy,27.76000023,31.89055546
GABC,"Snowden Raymond Ward",Director,"Apr 15, 2020",26.22,97,2544,31364,"Apr 17, 2020, 01:54 PM",buy,27.76000023,31.89055546
GABC,"SHEIDLER JACK",Director,"Apr 15, 2020",26.22,65,1704,51341,"Apr 17, 2020, 01:38 PM",buy,27.76000023,31.89055546
GABC,"Ryan Christina M",Director,"Apr 15, 2020",26.22,34,892,2997,"Apr 17, 2020, 01:33 PM",buy,27.76000023,31.89055546
GABC,"KELLY JASON M",Director,"Apr 15, 2020",26.22,65,1704,2549,"Apr 17, 2020, 01:18 PM",buy,27.76000023,31.89055546
GABC,"Bawel Zachary W",Director,"Apr 15, 2020",26.22,97,2544,6986,"Apr 17, 2020, 01:11 PM",buy,27.76000023,31.89055546
TMDX,"Gordon Stephen","Chief Financial Officer","Apr 22, 2020",18.0,100,1800,28471,"Apr 24, 2020, 05:54 PM",sell,17.78000069,18.45649997
AMK,"Goldman Charles G","President and CEO","Apr 22, 2020",22.04,63,1389,2400578,"Apr 24, 2020, 05:35 PM",sell,22.15999985,26.00472221
ALBO,"Horn Patrick Taylor","Chief Medical Officer","Apr 22, 2020",17.64,92,1623,4465,"Apr 24, 2020, 05:22 PM",sell,19.88999939,21.25394443
ALBO,"Duncan Jason","Chief Legal Officer and GC","Apr 22, 2020",17.68,92,1627,4461,"Apr 24, 2020, 05:20 PM",sell,19.88999939,21.25394443
CARG,"Steinert Langley","CEO and Chairman","Apr 22, 2020",20.01,76,1521,254176,"Apr 24, 2020, 04:26 PM",sell,19.95000076,30.86791667
GGO.A,"GABELLI MARIO J",Director,"Apr 21, 2020",40.26,100,4026,15100,"Apr 23, 2020, 09:47 PM",sell,na,na
FB,"Newstead Jennifer","VP and General Counsel","Apr 21, 2020",175.28,97,17002,1667,"Apr 23, 2020, 06:27 PM",sell,190.0700073,190.6834445
ACN,"ROWLAND DAVID","Executive Chairman","Apr 20, 2020",175.34,34,5961,22637,"Apr 22, 2020, 04:11 PM",sell,175.4600067,191.4683333
```

22. AWS CLOUDWATCH LOGS OF THE AWS GLUE SPARK ETL SPARK TRANSFORMATION JOB *(CONTINUED)*



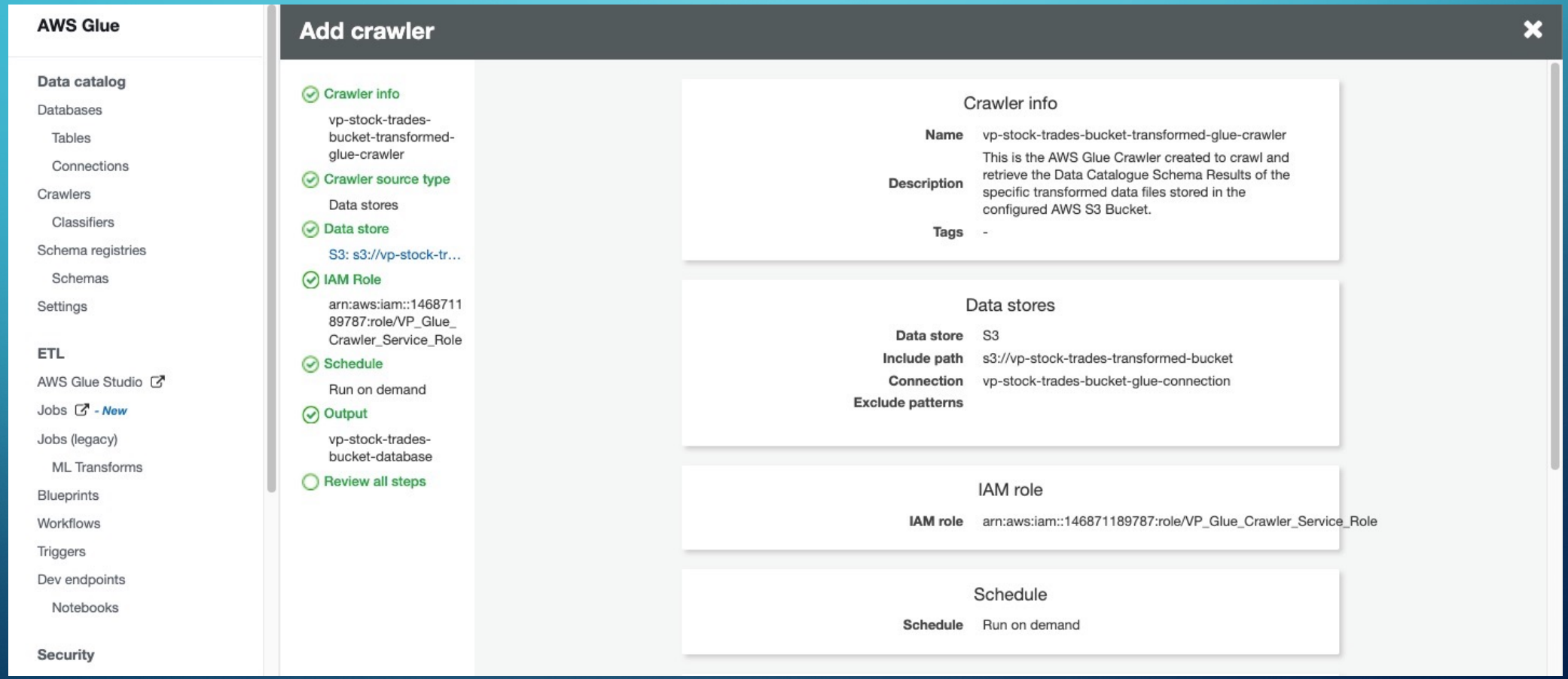
22. AWS CLOUDWATCH LOGS OF THE AWS GLUE SPARK ETL SPARK TRANSFORMATION JOB *(CONTINUED)*



CATALOGUE AND QUERY THE TRANSFORMED DATA

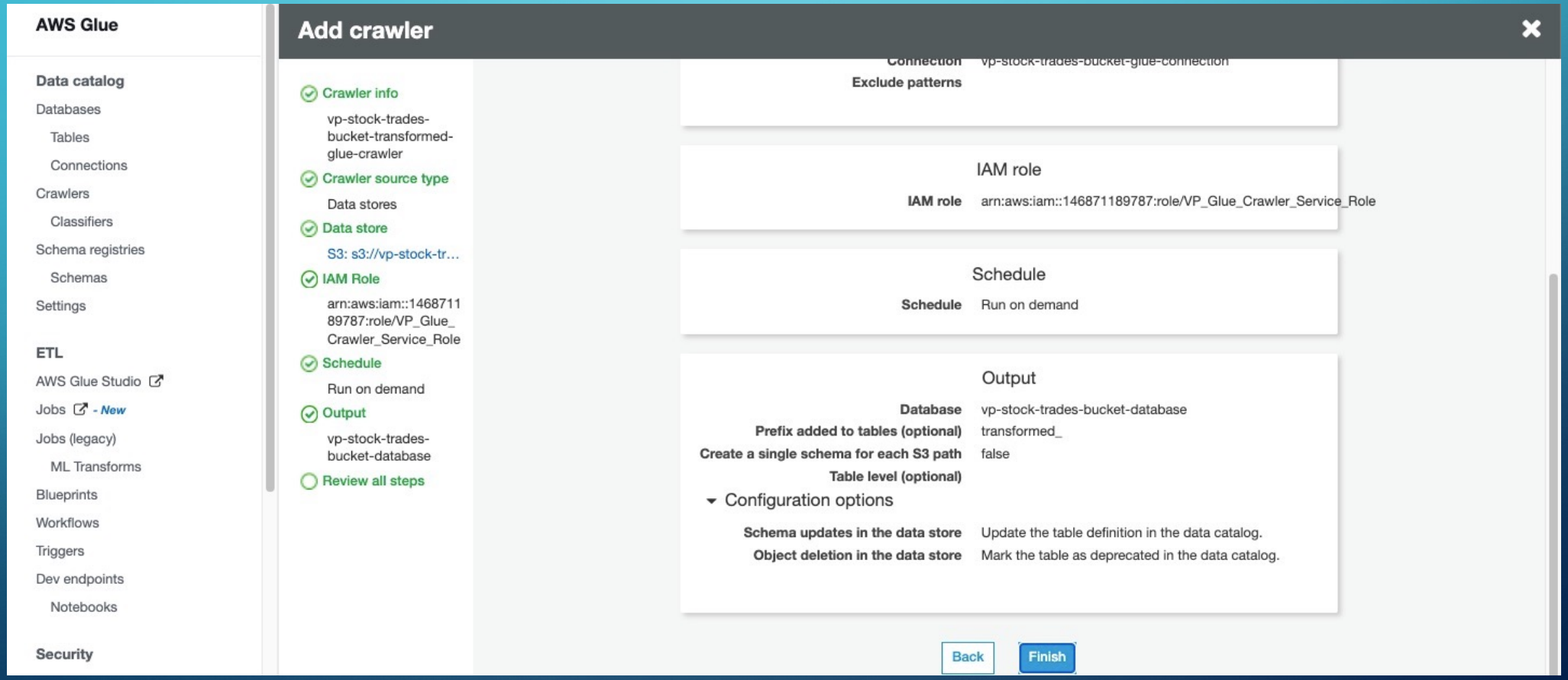
23. CREATING AWS GLUE CRAWLER TO CRAWL THE TRANSFORMED DATA STORED IN THE AWS S3 BUCKET *(CONTINUED)*

The transformed data file ‘run-datasink2-1-part-r-00000’ stored in the AWS S3 Bucket ’vp-stocks-trade-transformed-bucket’ will be crawled to extract the data catalog information



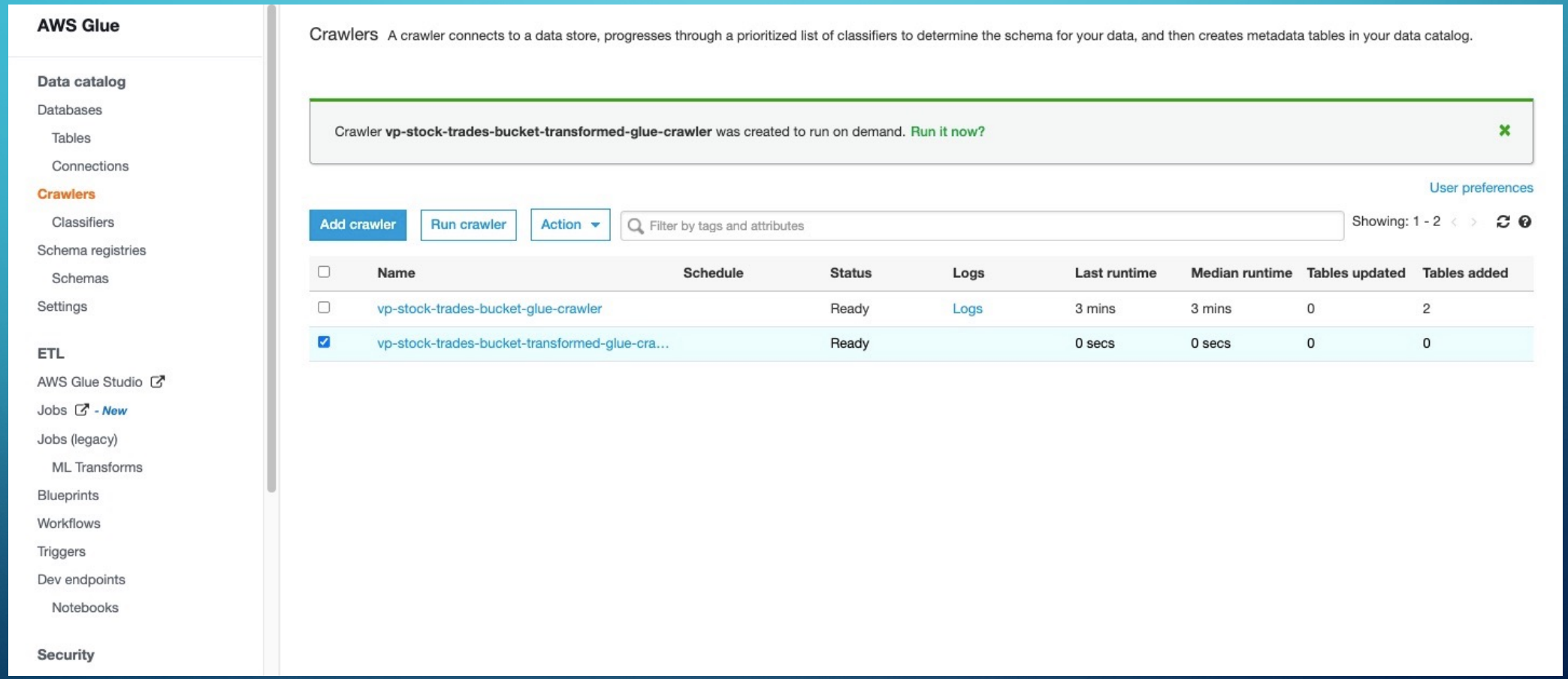
23. CREATING AWS GLUE CRAWLER TO CRAWL THE TRANSFORMED DATA STORED IN THE AWS S3 BUCKET *(CONTINUED)*

The transformed data file ‘run-datasink2-1-part-r-00000’ stored in the AWS S3 Bucket ’vp-stocks-trade-transformed-bucket’ will be crawled to extract the data catalog information



23. CREATING AWS GLUE CRAWLER TO CRAWL THE TRANSFORMED DATA STORED IN THE AWS S3 BUCKET *(CONTINUED)*

The transformed data file 'run-datasink2-1-part-r-00000' stored in the AWS S3 Bucket 'vp-stocks-trade-transformed-bucket' will be crawled to extract the data catalog information

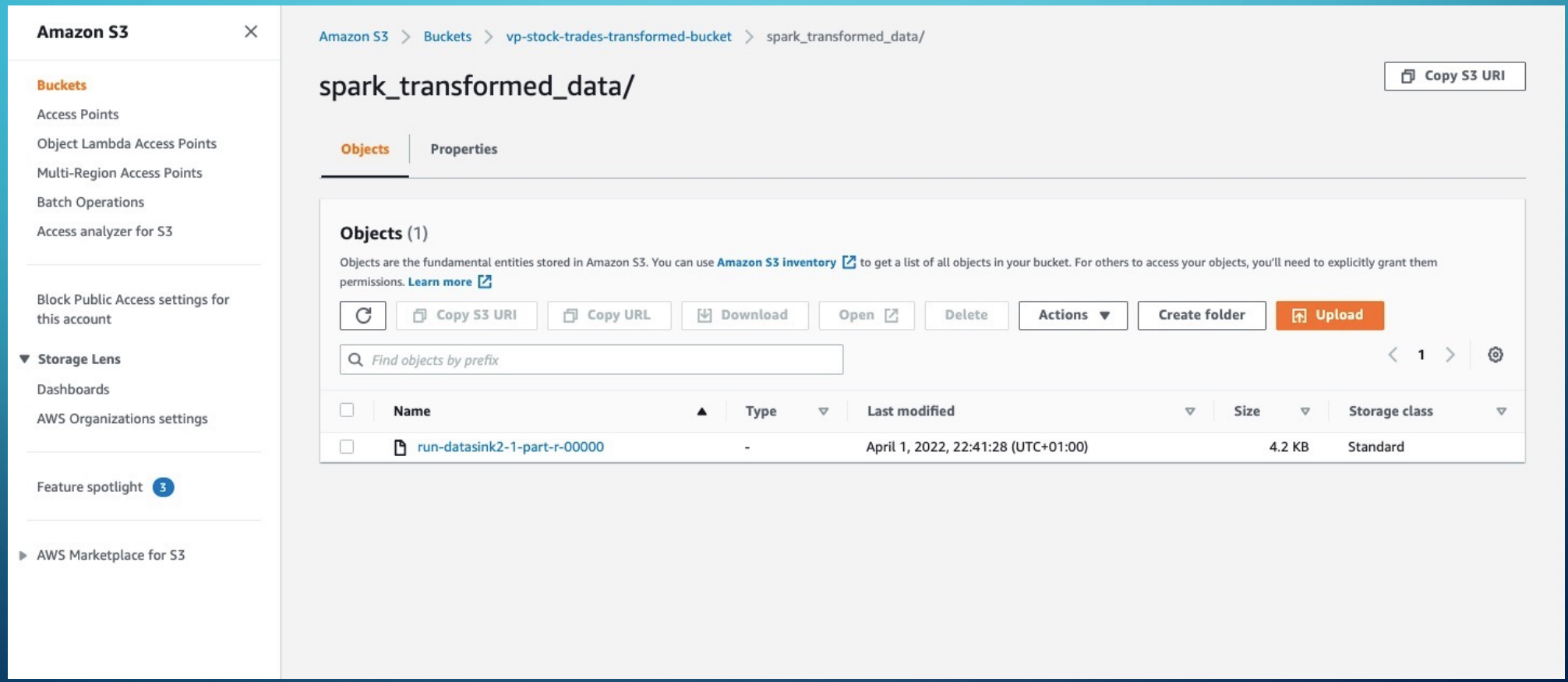


23. CREATING AWS GLUE CRAWLER TO CRAWL THE TRANSFORMED DATA STORED IN THE AWS S3 BUCKET

The transformed data file ‘run-datasink2-1-part-r-00000’ stored in the AWS S3 Bucket ’vp-stocks-trade-transformed-bucket’ will be crawled to extract the data catalog information

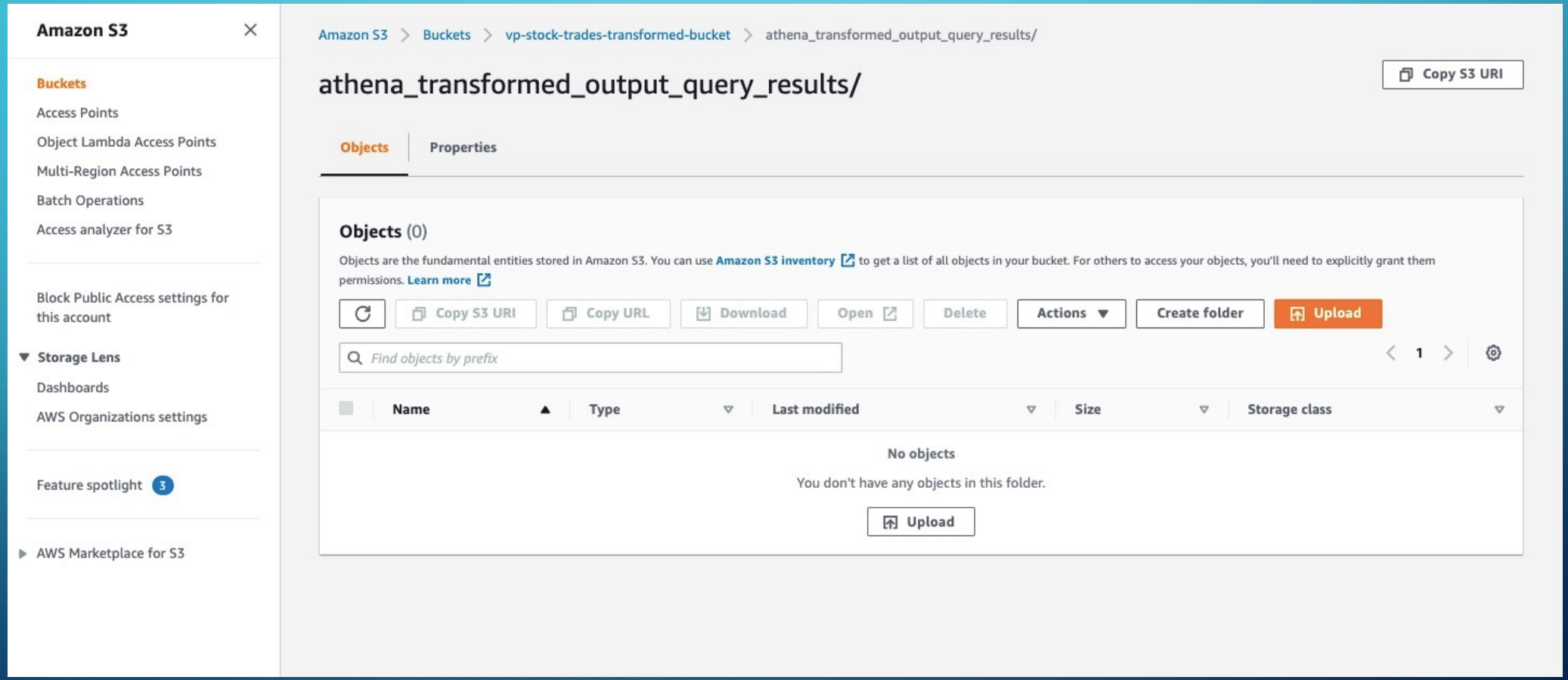
24. COPIED THE TRANSFORMED DATA FILE INTO A FOLDER WITHIN THE AWS S3 TRANSFORMED BUCKET BEFORE EXECUTING THE AWS GLUE CRAWLER

This is followed as a best practice when we have multiple sub-folders within the given AWS S3 Bucket to be crawled and avoid having querying issues with AWS Athena for those files that are placed directly within the AWS S3 Bucket without any sub-folder structure



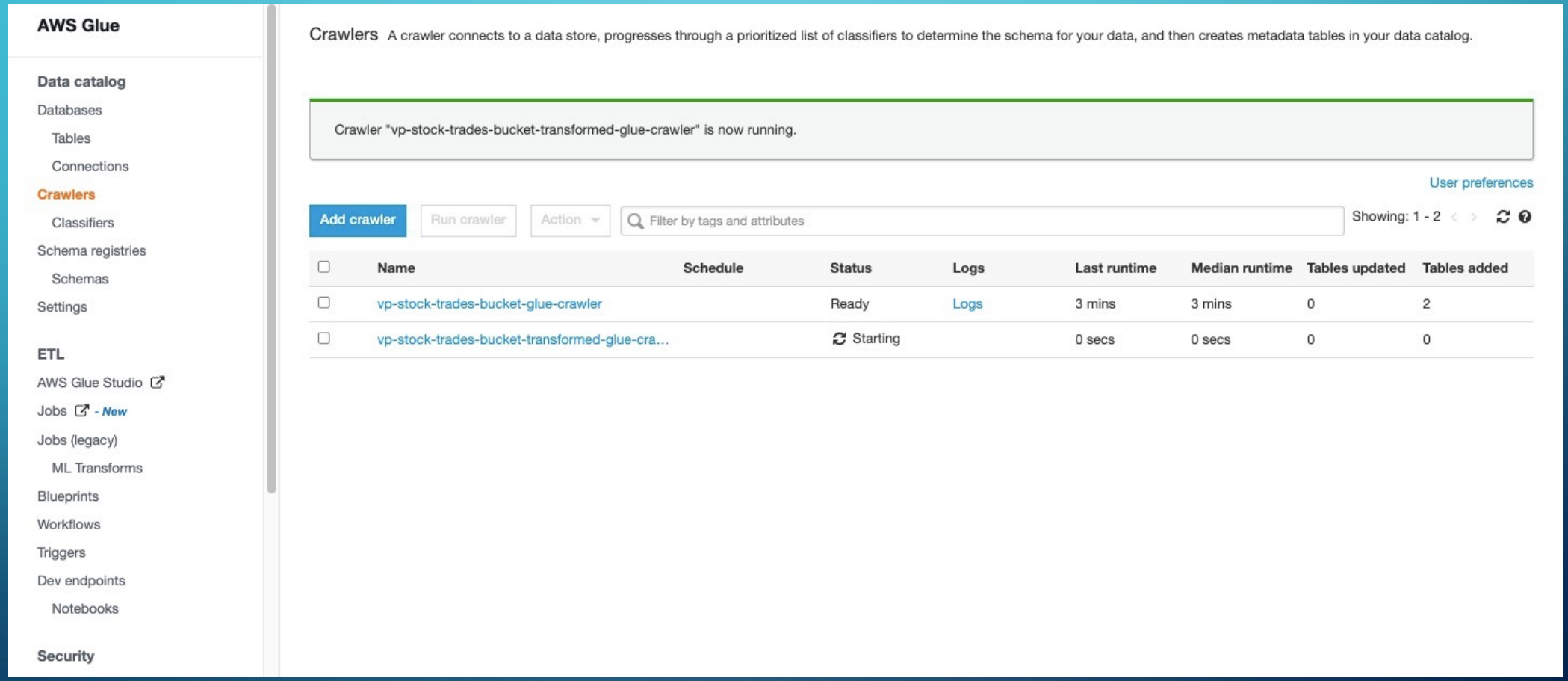
25. CREATING A SUB-FOLDER IN THE AWS S3 TRANSFORMED BUCKET TO STORES THE AWS ATHENA QUERY RESULTS

Sub-folder named ‘athena_transformed_output_query_results’



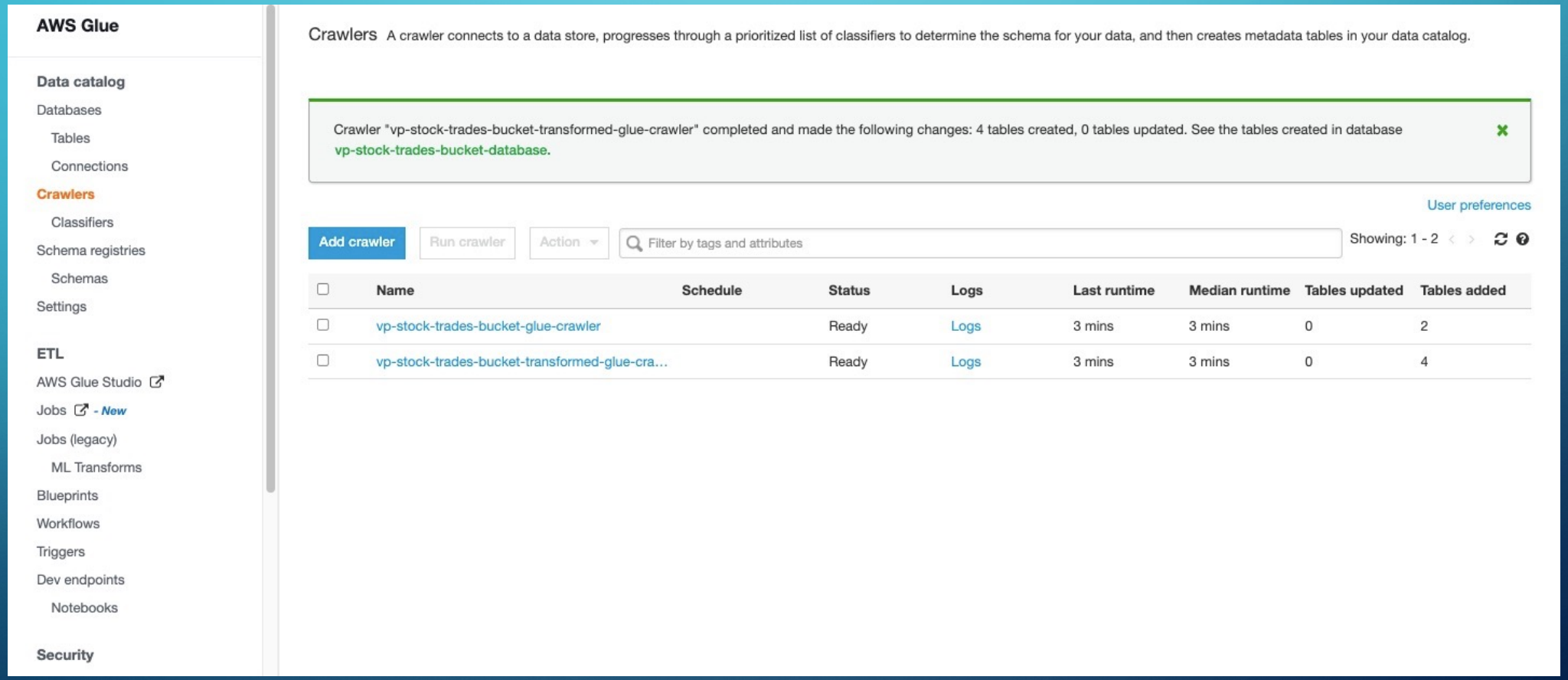
26. EXECUTING AWS GLUE CRAWLER TO CRAWL THE TRANSFORMED DATA STORED IN THE AWS S3 BUCKET *(CONTINUED)*

AWS Glue Crawler execution started and running



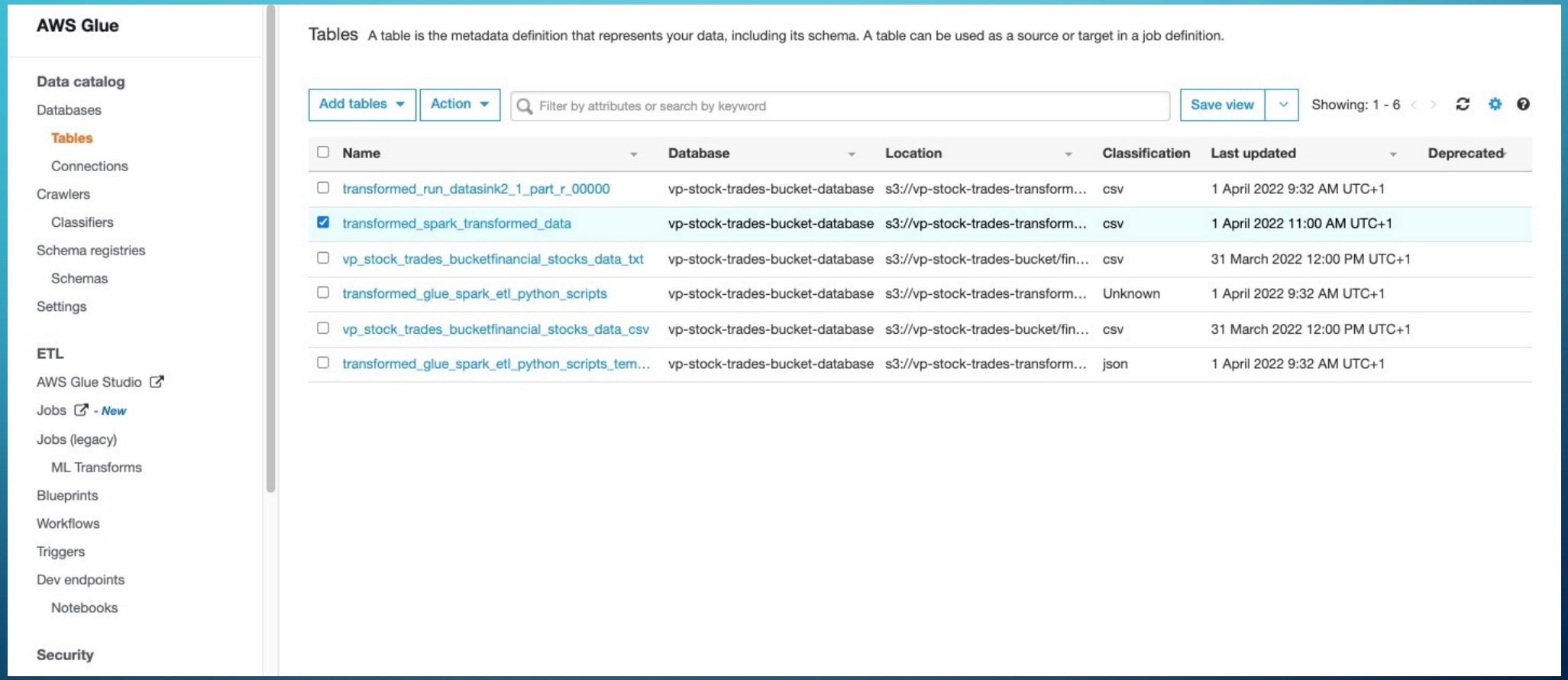
26. EXECUTING AWS GLUE CRAWLER TO CRAWL THE TRANSFORMED DATA STORED IN THE AWS S3 BUCKET

AWS Glue Crawler execution is completed, and the transformed Data Catalogue Tables are created and added to the AWS Glue database specified



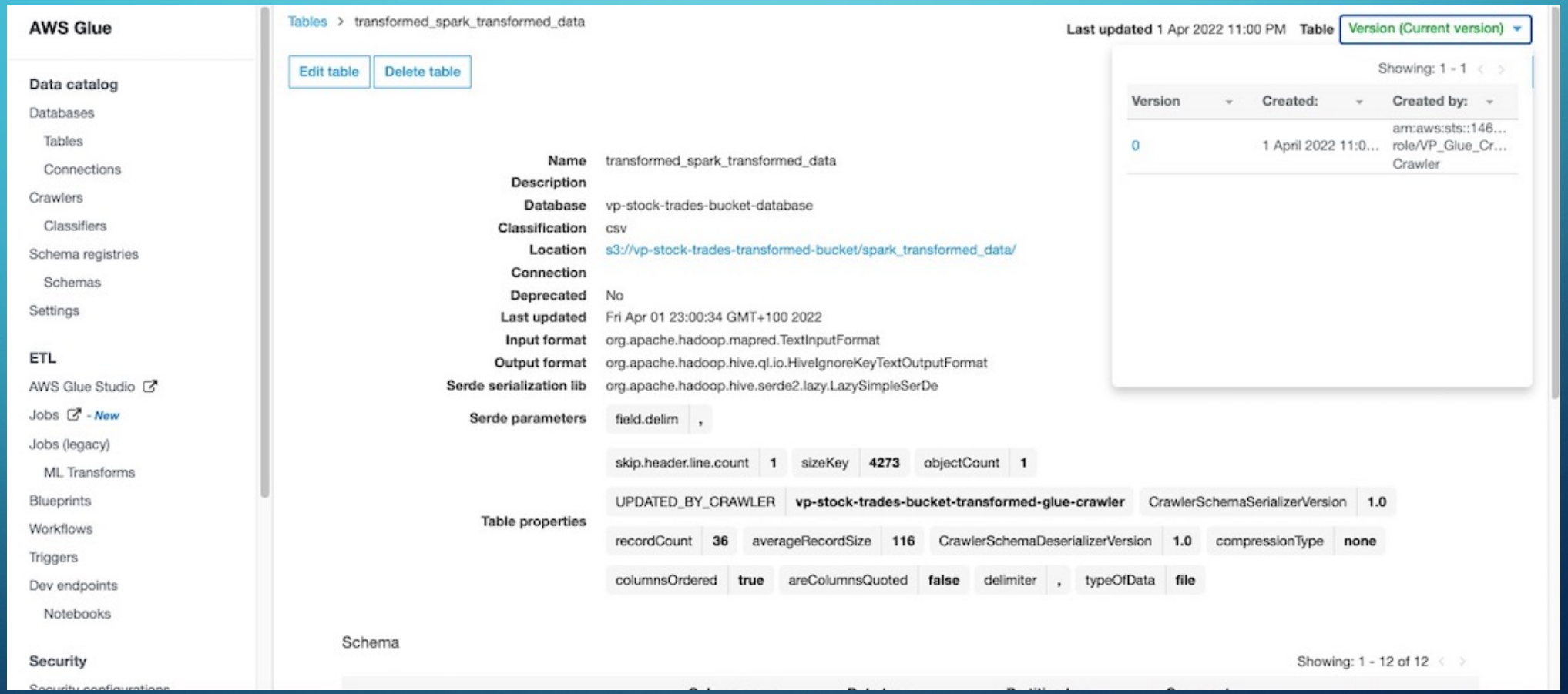
27. VIEWING THE DATA CATALOGUE TABLE CREATED FOR THE TRANSFORMED DATA *(CONTINUED)*

Data Catalogue Table named ‘transformed_spark_transformed_data’



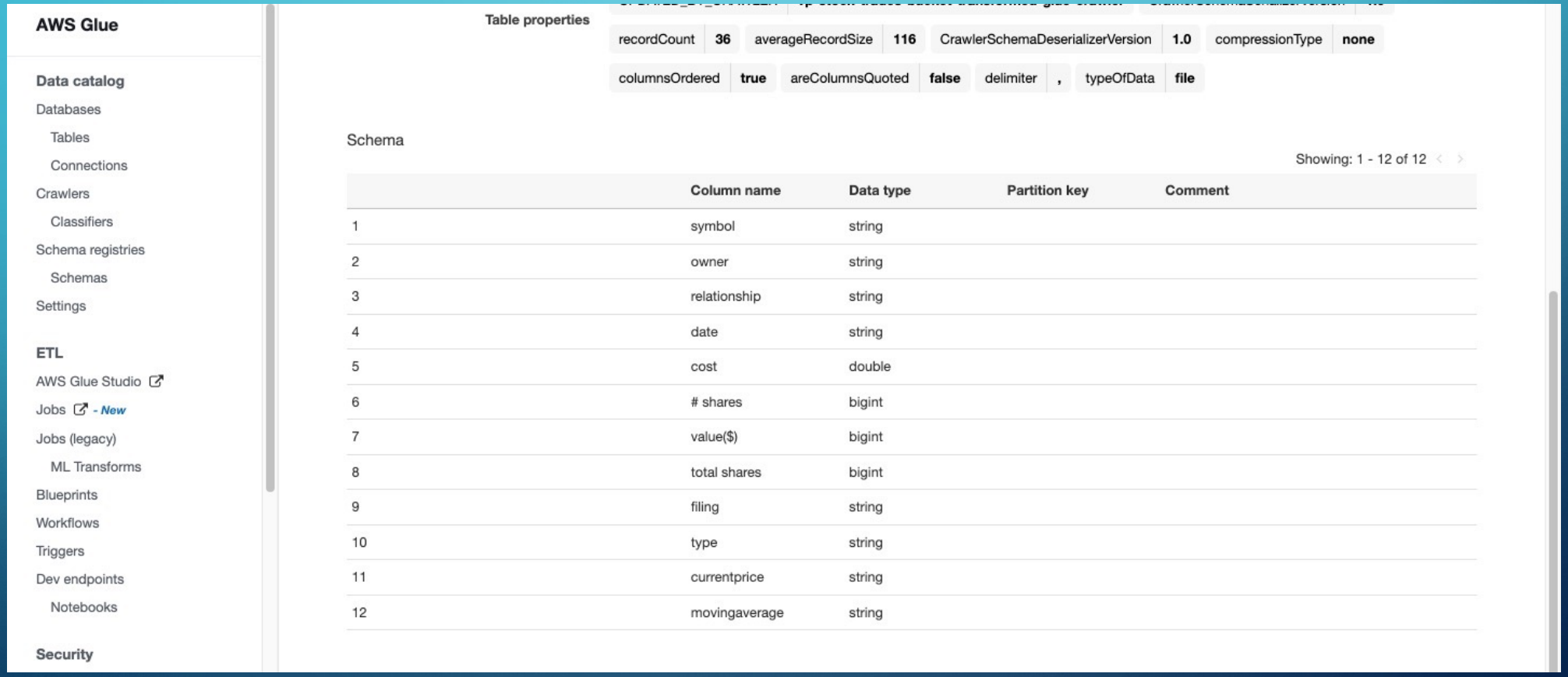
27. VIEWING THE DATA CATALOGUE TABLE CREATED FOR THE TRANSFORMED DATA *(CONTINUED)*

Data Catalogue Table named ‘transformed_spark_transformed_data’

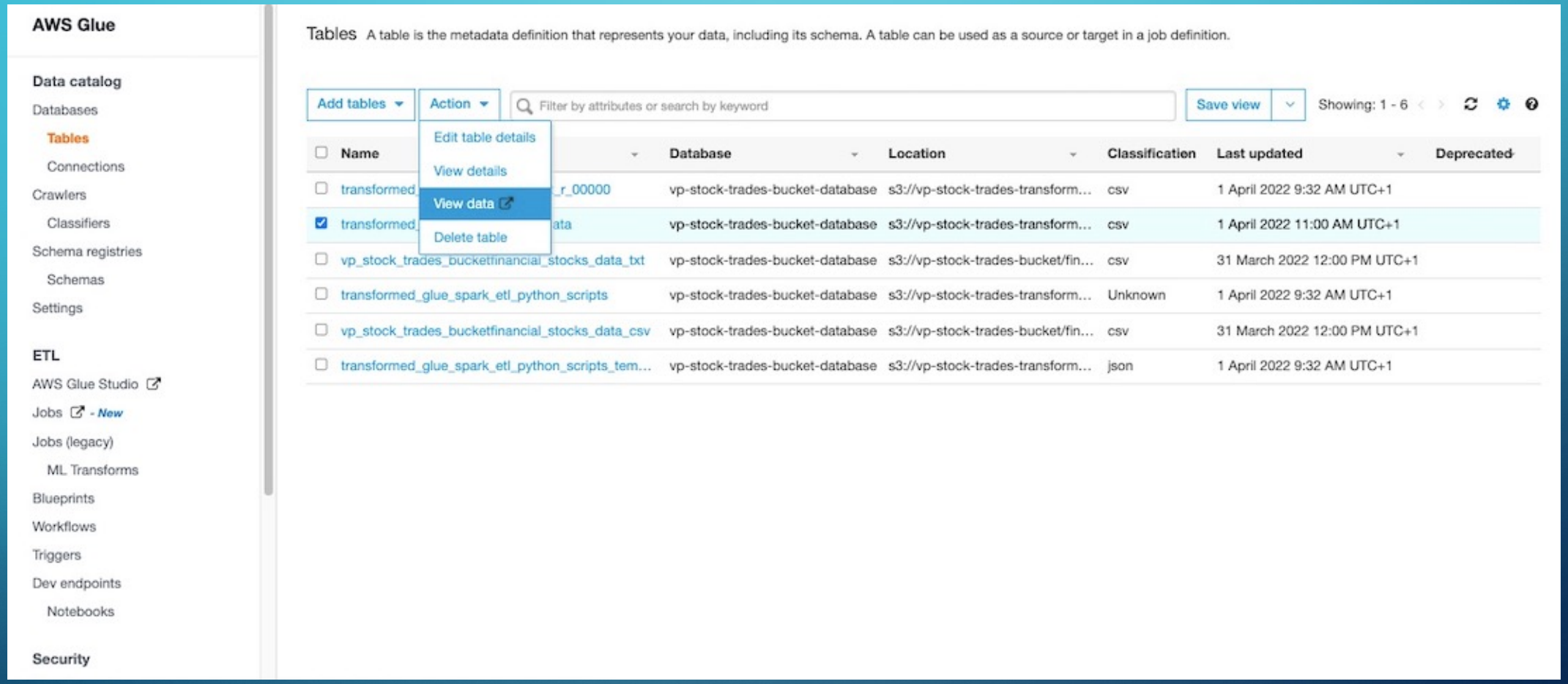


27. VIEWING THE DATA CATALOGUE TABLE CREATED FOR THE TRANSFORMED DATA

As we have not performed any modifications/transformations to the data schema for the transformed data in the Spark Transformation Python Job; hence the data schema remains same as the source data schema

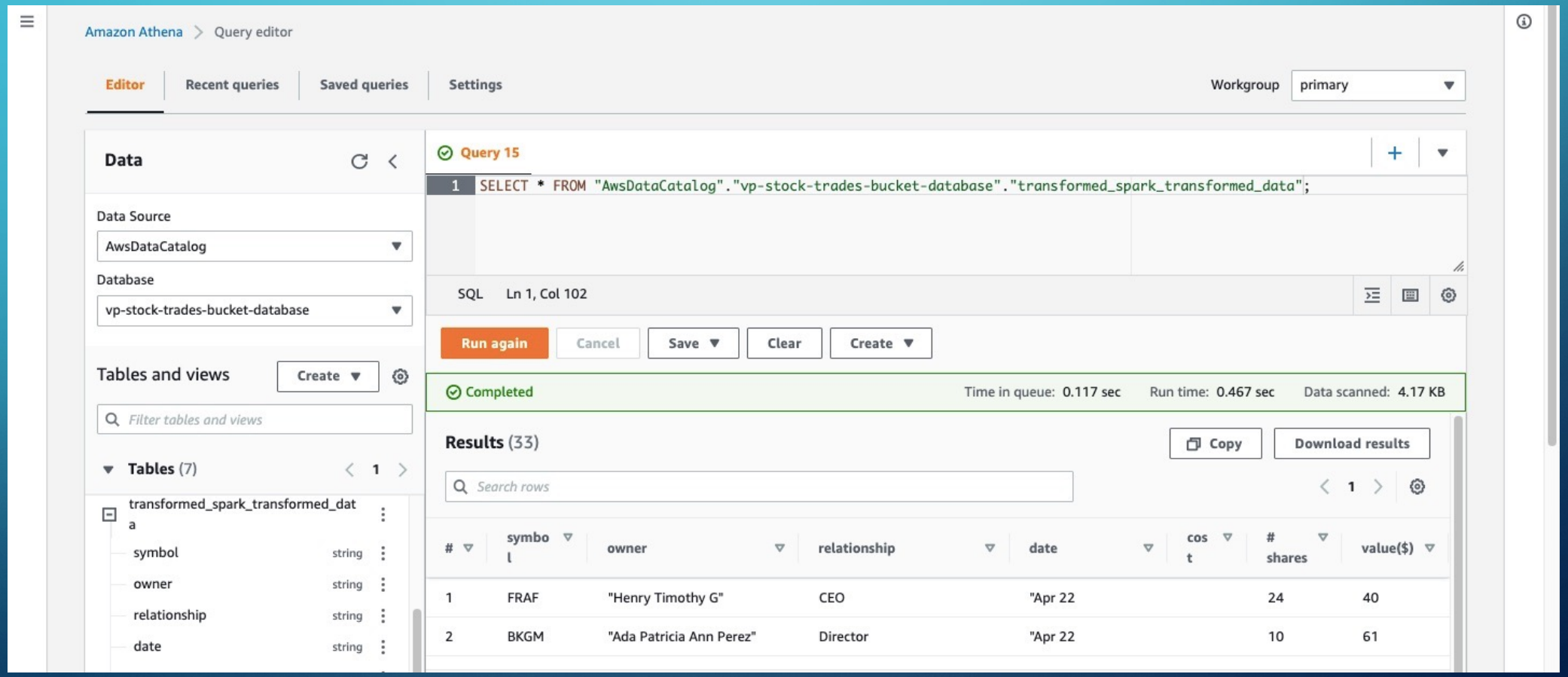


28. QUERYING THE TRANSFORMED DATA IN AWS ATHENA *(CONTINUED)*



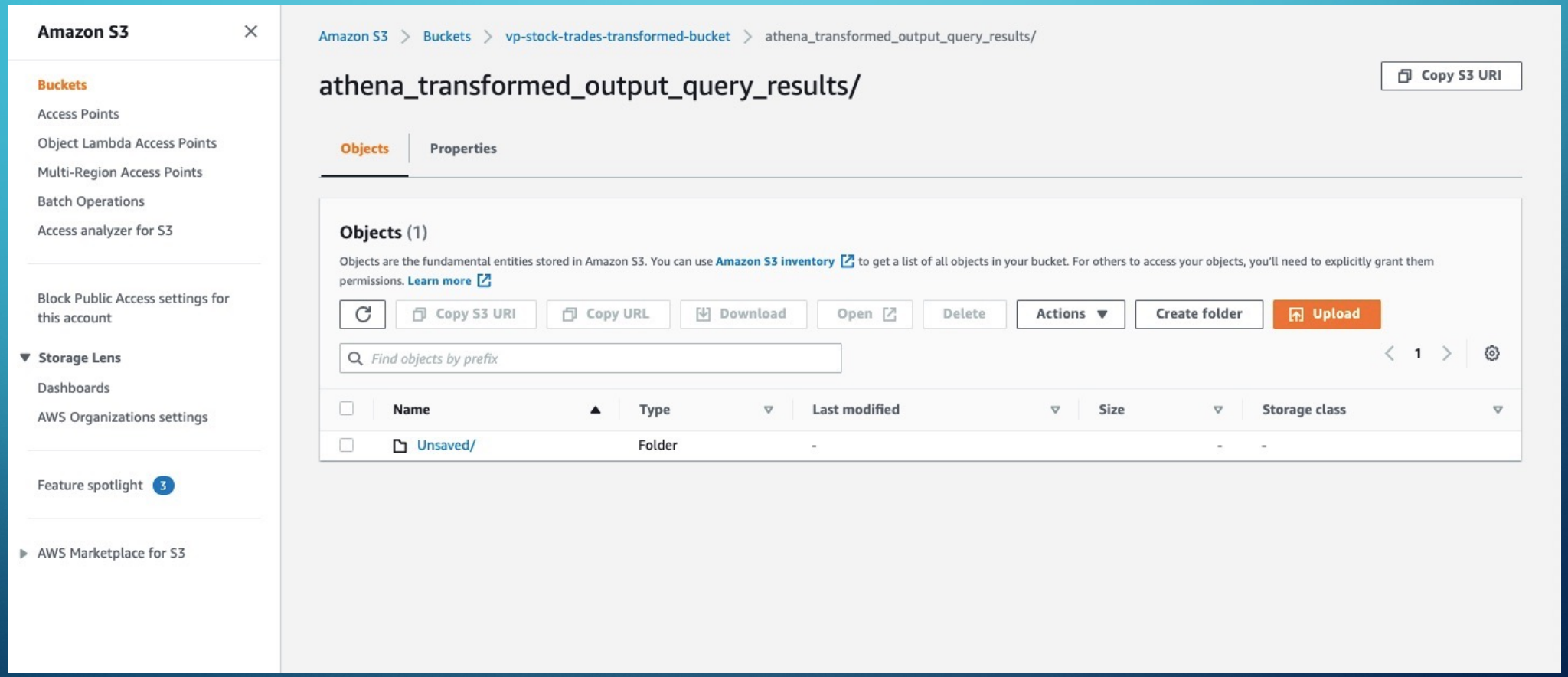
28. QUERYING THE TRANSFORMED DATA IN AWS ATHENA

After executing the SQL query to display all the trade transactions from the transformed data file. The query displays a total of 33 trade transactions as held in the transformed data file as per the spark transformation logic implemented



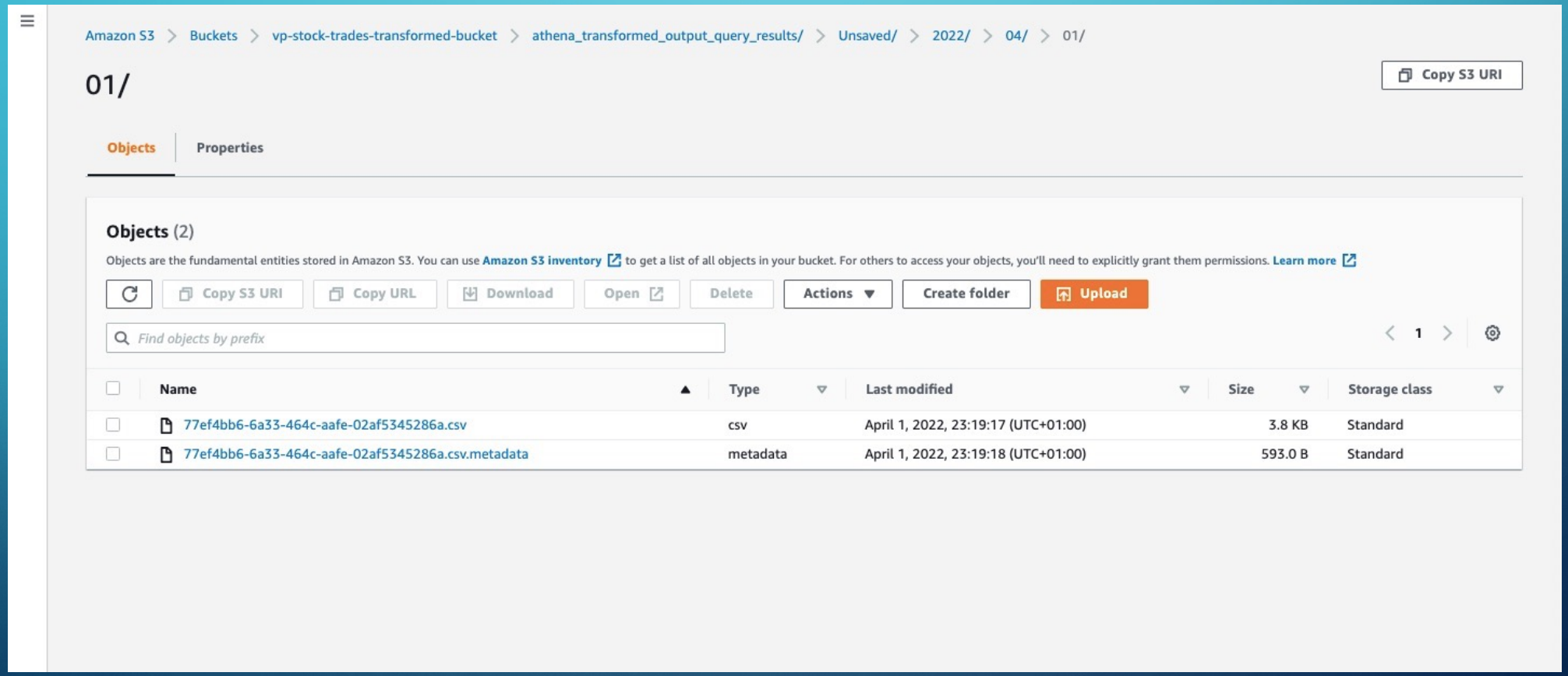
29. VIEWING THE AWS ATHENA TRANSFORMED DATA QUERY RESULTS STORED IN AWS S3 TRANSFORMED DATA BUCKET *(CONTINUED)*

AWS Athena SQL query results are saved within a new folder 'Unsaved' created automatically within the 'athena_transformed_output_query_results' sub-folder



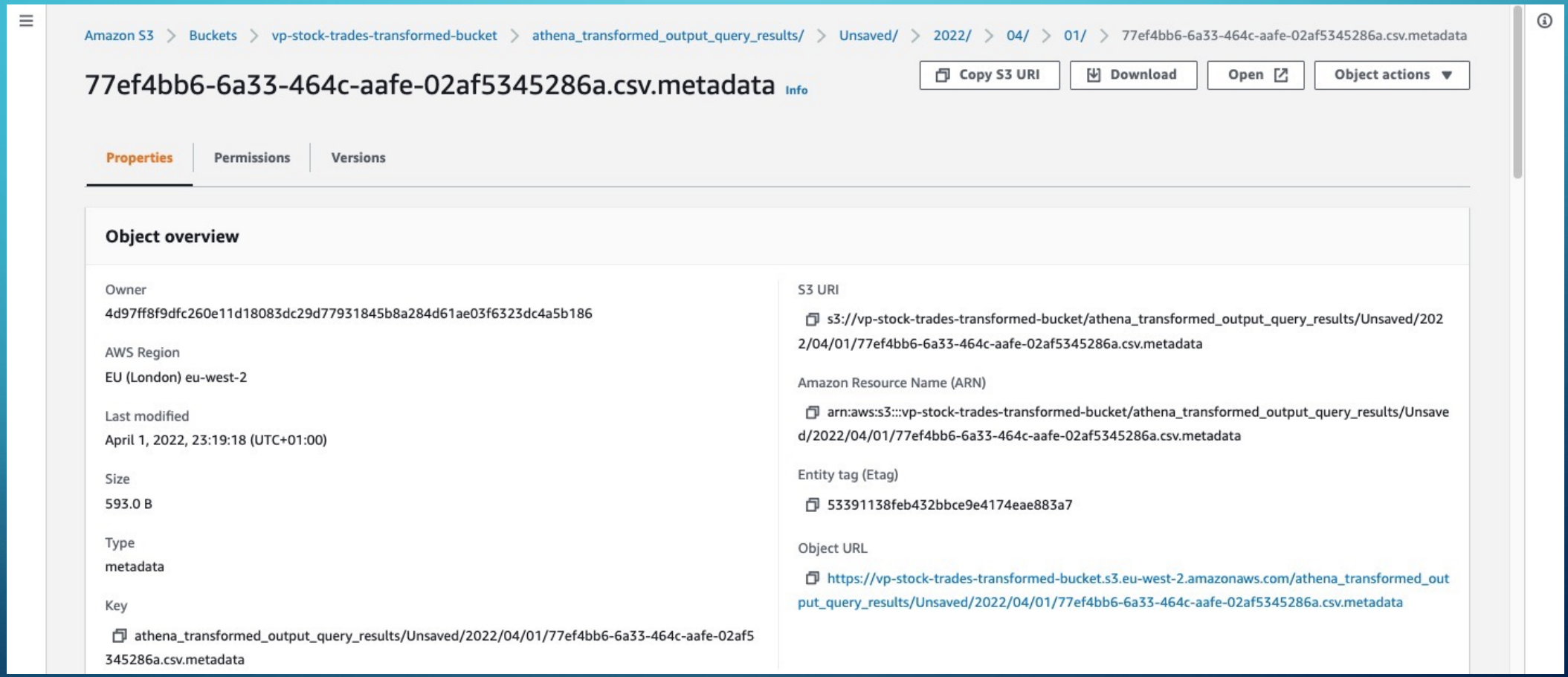
29. VIEWING THE AWS ATHENA TRANSFORMED DATA QUERY RESULTS STORED IN AWS S3 TRANSFORMED DATA BUCKET *(CONTINUED)*

AWS Athena SQL query results are saved within a new folder 'Unsaved' created automatically within the 'athena_transformed_output_query_results' sub-folder



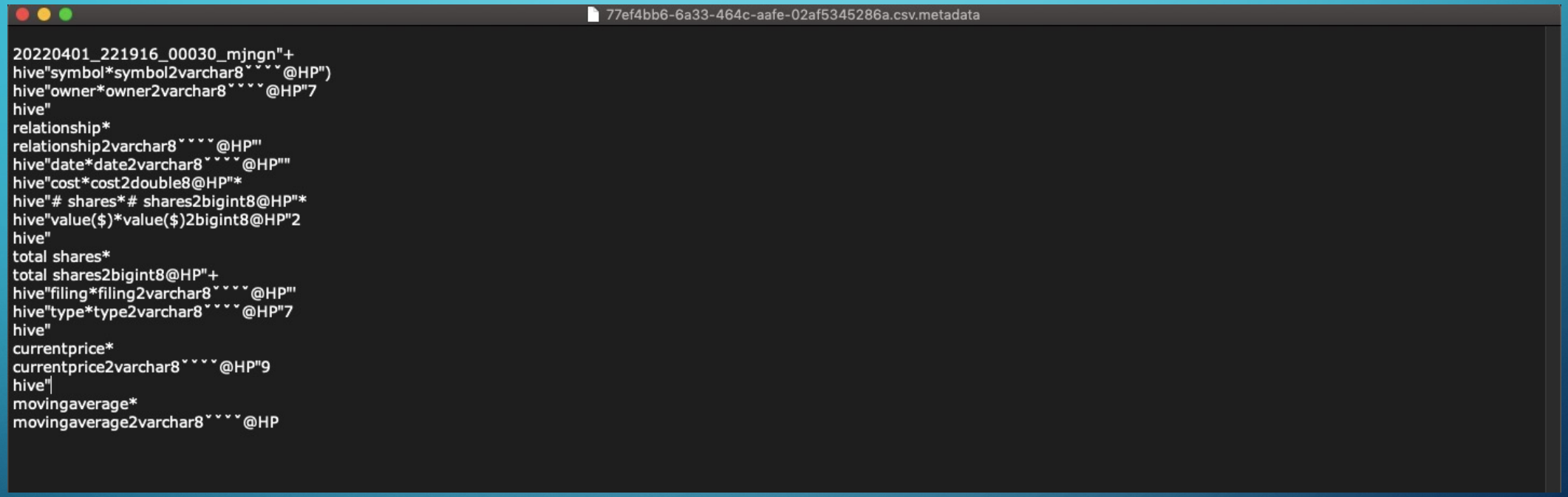
29. VIEWING THE AWS ATHENA TRANSFORMED DATA QUERY RESULTS STORED IN AWS S3 TRANSFORMED DATA BUCKET *(CONTINUED)*

Metadata file



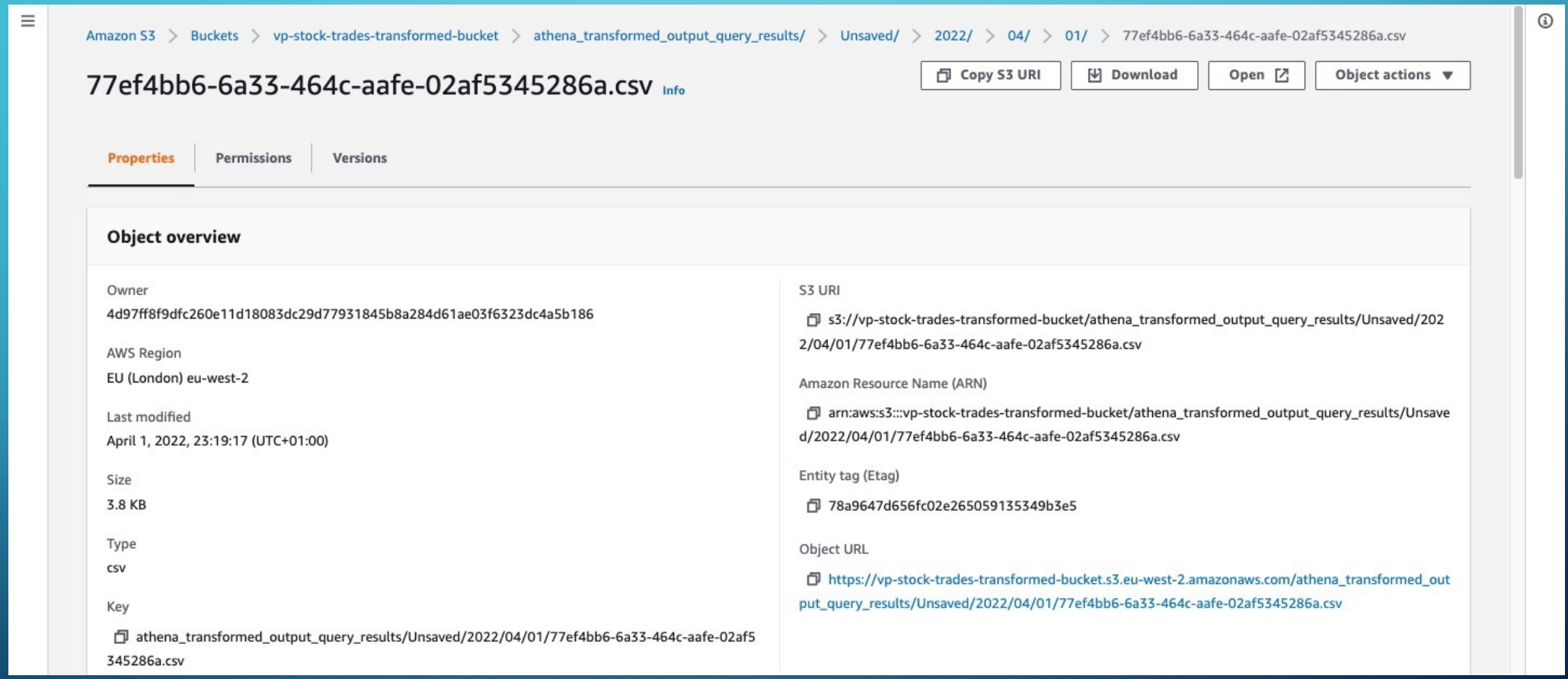
29. VIEWING THE AWS ATHENA TRANSFORMED DATA QUERY RESULTS STORED IN AWS S3 TRANSFORMED DATA BUCKET *(CONTINUED)*

Metadata file



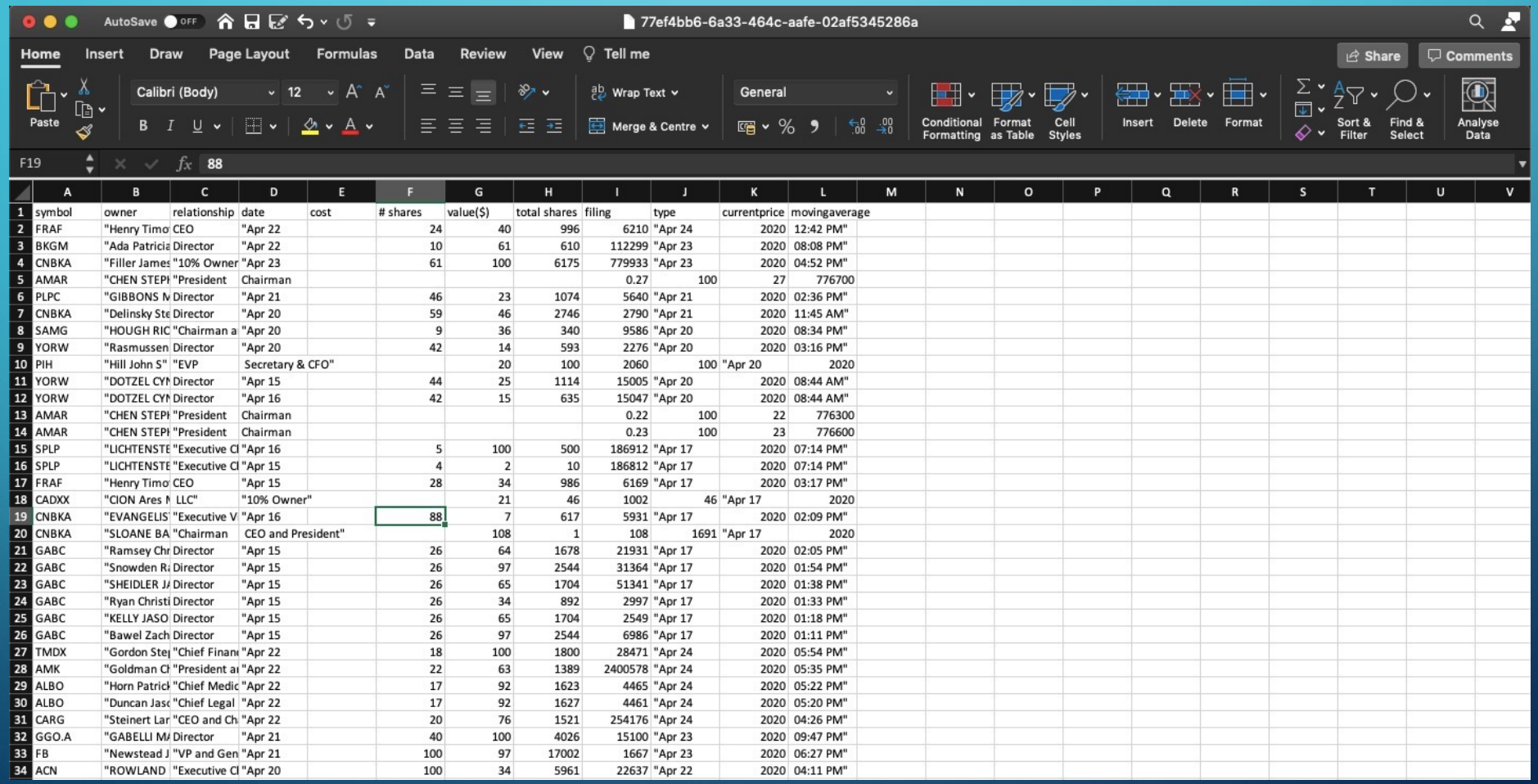
29. VIEWING THE AWS ATHENA TRANSFORMED DATA QUERY RESULTS STORED IN AWS S3 TRANSFORMED DATA BUCKET *(CONTINUED)*

Transformed Data SQL query output results csv file



29. VIEWING THE AWS ATHENA TRANSFORMED DATA QUERY RESULTS STORED IN AWS S3 TRANSFORMED DATA BUCKET

Transformed Data SQL query output results csv file



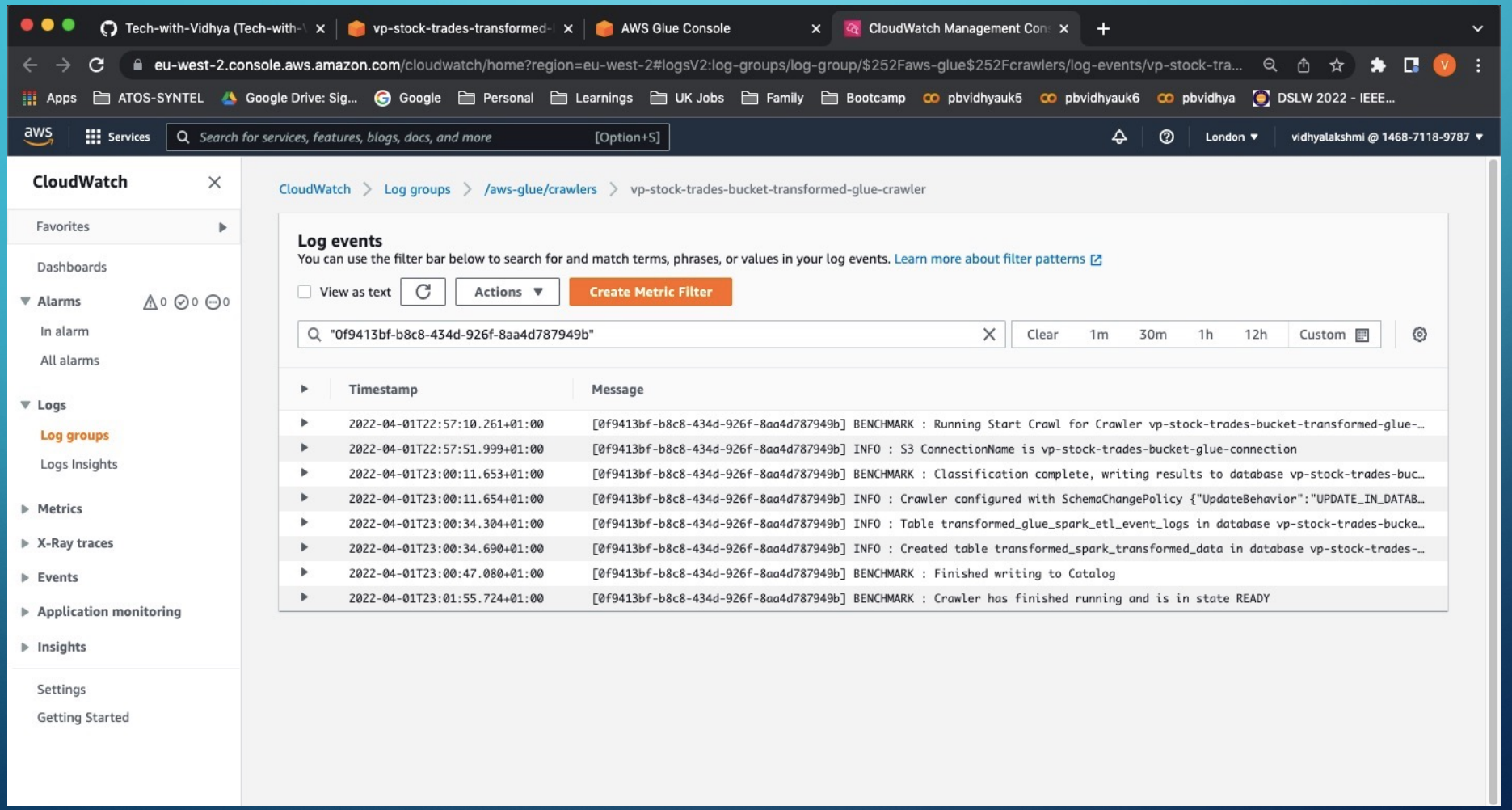
77ef4bb6-6a33-464c-aafe-02af5345286a

F19 | 88

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	symbol	owner	relationship	date	cost	# shares	value($)	total shares	filing	type	currentprice	movingaverage										
2	FRAF	"Henry Timo	CEO	"Apr 22		24	40	996	6210	"Apr 24	2020	12:42 PM"										
3	BKGM	"Ada Patricia	Director	"Apr 22		10	61	610	112299	"Apr 23	2020	08:08 PM"										
4	CNBKA	"Filler James	"10% Owner	"Apr 23		61	100	6175	779933	"Apr 23	2020	04:52 PM"										
5	AMAR	"CHEN STEPI	"President	Chairman					0.27	100	27	776700										
6	PLPC	"GIBBONS N	Director	"Apr 21		46	23	1074	5640	"Apr 21	2020	02:36 PM"										
7	CNBKA	"Delinsky Ste	Director	"Apr 20		59	46	2746	2790	"Apr 21	2020	11:45 AM"										
8	SAMG	"HOUGH RIC	"Chairman a	"Apr 20		9	36	340	9586	"Apr 20	2020	08:34 PM"										
9	YORW	"Rasmussen	Director	"Apr 20		42	14	593	2276	"Apr 20	2020	03:16 PM"										
10	PIH	"Hill John S"	"EVP	Secretary & CFO"			20	100	2060	100	"Apr 20	2020										
11	YORW	"DOTZEL CYI	Director	"Apr 15		44	25	1114	15005	"Apr 20	2020	08:44 AM"										
12	YORW	"DOTZEL CYI	Director	"Apr 16		42	15	635	15047	"Apr 20	2020	08:44 AM"										
13	AMAR	"CHEN STEPI	"President	Chairman					0.22	100	22	776300										
14	AMAR	"CHEN STEPI	"President	Chairman					0.23	100	23	776600										
15	SPLP	"LICHTENSTE	"Executive Cl	"Apr 16		5	100	500	186912	"Apr 17	2020	07:14 PM"										
16	SPLP	"LICHTENSTE	"Executive Cl	"Apr 15		4	2	10	186812	"Apr 17	2020	07:14 PM"										
17	FRAF	"Henry Timo	CEO	"Apr 15		28	34	986	6169	"Apr 17	2020	03:17 PM"										
18	CADXX	"CION Ares M	LLC"	"10% Owner"			21	46	1002	46	"Apr 17	2020										
19	CNBKA	"EVANGELIS	"Executive V	"Apr 16		88	7	617	5931	"Apr 17	2020	02:09 PM"										
20	CNBKA	"SLOANE BA	"Chairman	CEO and President"			108	1	108	1691	"Apr 17	2020										
21	GABC	"Ramsey Chr	Director	"Apr 15		26	64	1678	21931	"Apr 17	2020	02:05 PM"										
22	GABC	"Snowden Ri	Director	"Apr 15		26	97	2544	31364	"Apr 17	2020	01:54 PM"										
23	GABC	"SHEIDLER JA	Director	"Apr 15		26	65	1704	51341	"Apr 17	2020	01:38 PM"										
24	GABC	"Ryan Christi	Director	"Apr 15		26	34	892	2997	"Apr 17	2020	01:33 PM"										
25	GABC	"KELLY JASO	Director	"Apr 15		26	65	1704	2549	"Apr 17	2020	01:18 PM"										
26	GABC	"Bawel Zach	Director	"Apr 15		26	97	2544	6986	"Apr 17	2020	01:11 PM"										
27	TMDX	"Gordon Ste	"Chief Finan	"Apr 22		18	100	1800	28471	"Apr 24	2020	05:54 PM"										
28	AMK	"Goldman Cl	"President a	"Apr 22		22	63	1389	2400578	"Apr 24	2020	05:35 PM"										
29	ALBO	"Horn Patrick	"Chief Medic	"Apr 22		17	92	1623	4465	"Apr 24	2020	05:22 PM"										
30	ALBO	"Duncan Jasc	"Chief Legal	"Apr 22		17	92	1627	4461	"Apr 24	2020	05:20 PM"										
31	CARG	"Steinert Lar	"CEO and Ch	"Apr 22		20	76	1521	254176	"Apr 24	2020	04:26 PM"										
32	GGO.A	"GABELLI MA	Director	"Apr 21		40	100	4026	15100	"Apr 23	2020	09:47 PM"										
33	FB	"Newstead J	"VP and Gen	"Apr 21		100	97	17002	1667	"Apr 23	2020	06:27 PM"										
34	ACN	"ROWLAND	"Executive Cl	"Apr 20		100	34	5961	22637	"Apr 22	2020	04:11 PM"										

30. VIEWING THE AWS CLOUDWATCH LOGS FOR THE TRANSFORMATION AWS GLUE CRAWLER EXECUTED

Transformation AWS Glue Crawler named ‘vp-stock-trades-bucket-transformed-glue-crawler’



END OF THE PROJECT