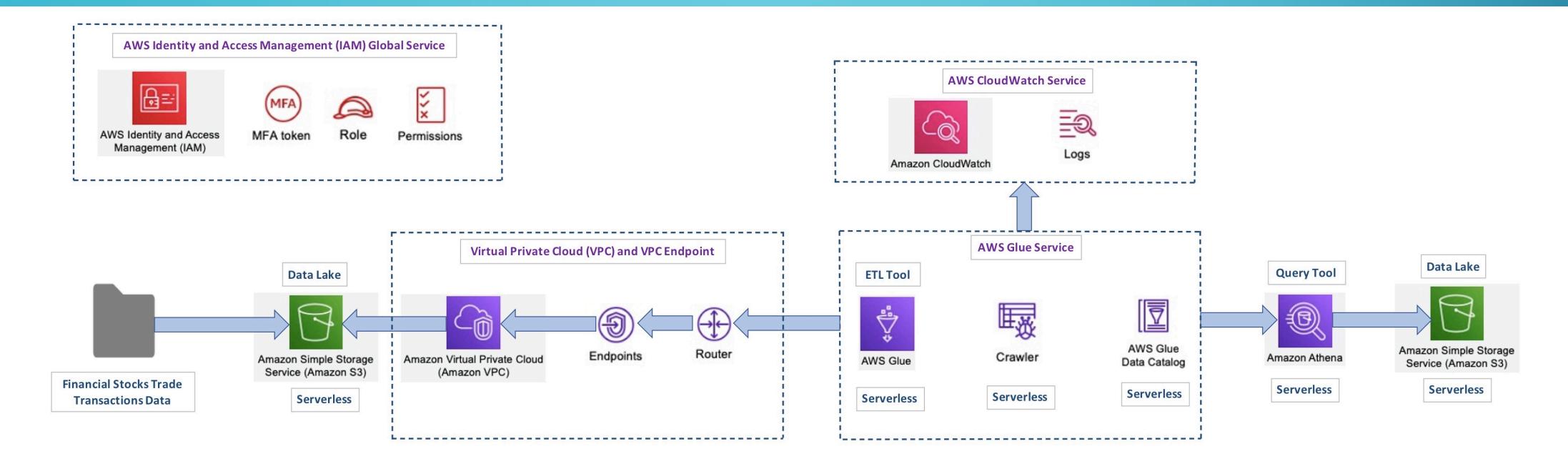


BUILDING A FINANCIAL DATA PIPELINE USING PYTHON AND AWS

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

Summary

User ARN: arn:aws:iam::146871189787:user/VP_S3_User

Path: /

Creation time: 2022-03-28 19:14 UTC+0100

Permissions Groups (1) Tags Security credentials Access Advisor

▼ Permissions policies (1 policy applied)

Add permissions + Add inline policy

Policy name	Policy type
Attached from group	AWS managed policy from group VP_S3_Group
AmazonS3FullAccess	x

Policy summary { } JSON Simulate policy

Filter

Service	Access level	Resource	Request condition
S3	Full access	All resources	None
S3 Object Lambda	Full access	All resources	None

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

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

The screenshot shows the AWS S3 Buckets page. At the top left, it says "Amazon S3 > Buckets". Below that is a section titled "Account snapshot" with a link to "View Storage Lens dashboard". Underneath is a table header for "Buckets (0) [Info](#)". The table has columns for "Name", "AWS Region", "Access", and "Creation date". To the right of the table are buttons for "Create bucket", "Copy ARN", "Empty", and "Delete". Below the table, there's a search bar with the placeholder "Find buckets by name" and a pagination area with a single page indicator "1". A message at the bottom center says "No buckets" and "You don't have any buckets.", with a "Create bucket" button below it.

2. CREATION OF AWS S3 BUCKET (CONTINUED)

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

The screenshot shows the AWS S3 Buckets page. At the top left, there's a navigation bar with three horizontal lines and the text "Amazon S3 > Buckets". On the right side of the top bar, there's a small info icon. Below the top bar, there's a section titled "Account snapshot" with a sub-section "Storage lens provides visibility into storage usage and activity trends. Learn more" and a "View Storage Lens dashboard" button. In the center, there's a table titled "Buckets (1) Info" with a "Find buckets by name" search bar above it. The table has columns: Name, AWS Region, Access, and Creation date. A "Create bucket" button is located at the top right of the table area. The table row shows one bucket: "vp-stock-trades-bucket" in the EU (London) region (eu-west-2), with "Objects can be public" access and a creation date of "March 30, 2022, 16:31:24 (UTC+01:00)".

Name	AWS Region	Access	Creation date
vp-stock-trades-bucket	EU (London) eu-west-2	Objects can be public	March 30, 2022, 16:31:24 (UTC+01:00)

2. CREATION OF AWS S3 BUCKET

The screenshot shows the AWS S3 console interface for the bucket 'vp-stock-trades-bucket'. The top navigation bar includes 'Amazon S3 > Buckets > vp-stock-trades-bucket'. The main heading 'vp-stock-trades-bucket' has an 'Info' link. Below the heading are tabs: 'Objects' (selected), 'Properties', 'Permissions', 'Metrics', 'Management', and 'Access Points'. The 'Objects' section displays 'Objects (0)'. A descriptive text states: '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.' Below this is a 'Learn more' link. A toolbar contains buttons for 'Copy S3 URI', 'Copy URL', 'Download', 'Open', 'Delete', 'Actions', 'Create folder', and 'Upload' (which is highlighted). A search bar says 'Find objects by prefix'. At the bottom, a table header for 'Objects' lists columns: 'Name', 'Type', 'Last modified', 'Size', and 'Storage class'. The message 'No objects' is centered, followed by 'You don't have any objects in this bucket.' and a large 'Upload' button.

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

```
main.py
1 # Stock Trade Transactions Data - ETL Data Pipeline using Python, AWS (CLI, S3, Glue, Athena, QuickSight)
2 # Main Program
3
4 from security import access_key_id, secret_access_key
5 from config import bucket_name, bucket_location, s3_bucket_metadata_file_path, source_data_files_path
6 from s3_bucket_creation import s3_create_bucket
7 from s3_bucket_metadata_retrieval import s3_response_metadata
8 from s3_bucket_data_ingestion import s3_bucket_data_ingest
9
10 # Creating a Bucket in AWS S3
11 s3_response = s3_create_bucket(access_key_id, secret_access_key, bucket_name, bucket_location)
12
13 # Writing the S3 Bucket Metadata Details to a Text File
14 s3_response_metadata(s3_response, s3_bucket_metadata_file_path)
15
16 # Ingesting Source Data Files into AWS S3 Bucket
17 s3_bucket_data_ingest(access_key_id, secret_access_key, bucket_name, source_data_files_path)
18
```

Terminal: Local × +

```
(base) VidhyalshmisAir:source_code vidhyalakshmi.parthasarathy$ clear
(base) VidhyalshmisAir:source_code vidhyalakshmi.parthasarathy$ 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 vidhyalakshmi.parthasarathy$
```

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

The screenshot shows the Amazon S3 console interface for the 'vp-stock-trades-bucket'. The top navigation bar shows 'Amazon S3 > Buckets > vp-stock-trades-bucket'. The main title is 'vp-stock-trades-bucket' with an 'Info' link. Below the title is a navigation bar with tabs: Objects (highlighted in orange), Properties, Permissions, Metrics, Management, and Access Points.

The main content area is titled 'Objects (2)'. A sub-instruction says: '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.' Below this is a 'Learn more' link with a help icon.

Below the sub-instruction is a row of buttons: 'Copy S3 URI' (white), 'Copy URL' (white), 'Download' (white), 'Open' (white), 'Delete' (white), 'Actions' (dropdown), 'Create folder' (white), and 'Upload' (orange). There is also a search bar with the placeholder 'Find objects by prefix'.

The table below lists the objects:

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	financial_stocks_data_csv/	Folder	-	-	-
<input type="checkbox"/>	financial_stocks_data_txt/	Folder	-	-	-

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

The screenshot shows the Amazon S3 console interface. The navigation path is: Amazon S3 > Buckets > vp-stock-trades-bucket > financial_stocks_data_csv/. The current view is the 'Objects' tab. There are two objects listed:

Name	Type	Last modified	Size	Storage class
stock_sectors.csv	csv	March 30, 2022, 16:45:43 (UTC+01:00)	122.7 KB	Standard
trade_transactions.csv	csv	March 30, 2022, 16:45:43 (UTC+01:00)	78.1 KB	Standard

At the top right, there is a 'Copy S3 URI' button. Below the table, there are navigation controls: a search bar with 'Find objects by prefix', a page number '1', and a refresh icon.

3. INGESTING SOURCE DATA FILES IN TO AWS S3 BUCKET

The screenshot shows the Amazon S3 console interface. The navigation path is: Amazon S3 > Buckets > vp-stock-trades-bucket > financial_stocks_data_txt/. The current view is the 'Objects' tab. A single object, 'marketcap.txt', is listed. The object details are as follows:

Name	Type	Last modified	Size	Storage class
marketcap.txt	txt	March 30, 2022, 16:45:43 (UTC+01:00)	8.0 KB	Standard

At the top right of the objects list, there is a 'Copy S3 URI' button. Below the objects list, there is a search bar labeled 'Find objects by prefix'.

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

Endpoints (1/1) <small>Info</small>			
<input type="checkbox"/>	Name	VPC endpoint ID	VPC ID
<input checked="" type="checkbox"/>	vp-stock-trades-bucket-vpc-endpoint	vpce-0cef04dedf3d3e365	vpc-0c73fc06befa93a5c vp-stock-trades-bucket-vpc
Endpoint ID <input type="checkbox"/> vpce-0cef04dedf3d3e365	Status Available	Creation time Wednesday, March 30, 2022, 17:05:46 GMT+1	Endpoint type Gateway
VPC ID vpc-0c73fc06befa93a5c (vp-stock-trades-bucket-vpc)	Status message -	Service name <input type="checkbox"/> com.amazonaws.eu-west-2.s3	Private DNS names enabled No

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

The screenshot shows the AWS VPC Endpoint interface details page for the endpoint `vpce-0cef04dedf3d3e365`. The endpoint is associated with the VPC `vpc-0c73fc06befa93a5c` and has an `Endpoint ID` of `vpce-0cef04dedf3d3e365`. It is currently `Available` and was created on `Wednesday, March 30, 2022, 17:05:46 GMT+1`. The service name is `com.amazonaws.eu-west-2.s3`, and it is a `Gateway` type endpoint. Private DNS names are not enabled. The `Route tables` tab is selected, showing one route table named `rtb-04ff62337ed83f7d0` which is associated with 3 subnets.

Details	Actions
Endpoint ID vpce-0cef04dedf3d3e365	Status Available
VPC ID vpc-0c73fc06befa93a5c (vp-stock-trades-bucket-vpc)	Status message -
Creation time Wednesday, March 30, 2022, 17:05:46 GMT+1	Endpoint type Gateway
Service name com.amazonaws.eu-west-2.s3	Private DNS names enabled No

Route tables | Policy | Tags

Route tables (1)			
Name	Route Table ID	Main	Associated Id
-	rtb-04ff62337ed83f7d0	Yes	3 subnets

Filter route tables

Manage route tables

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

VPC > Your VPCs > vpc-0c73fc06befa93a5c

vpc-0c73fc06befa93a5c / vp-stock-trades-bucket-vpc

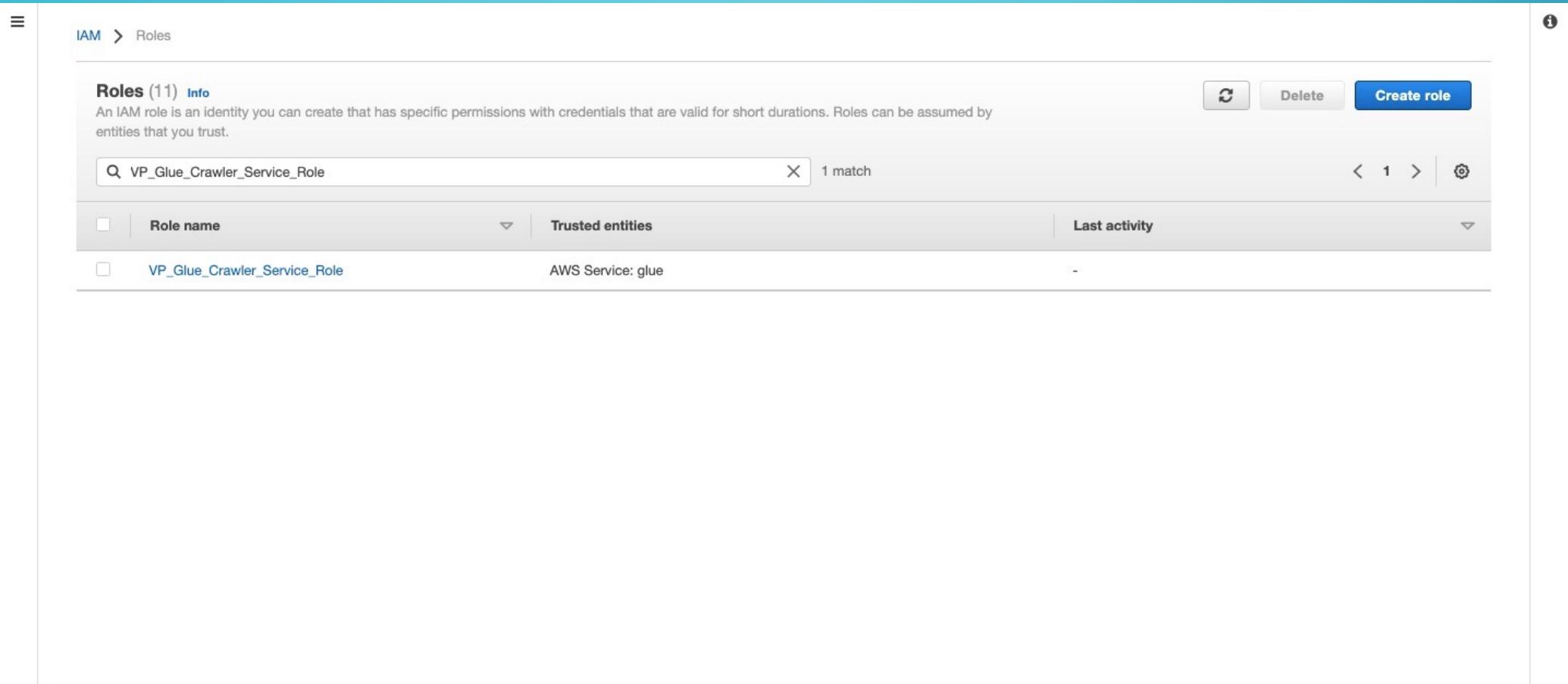
Actions ▾

Details		Info	
VPC ID	<input type="text"/> vpc-0c73fc06befa93a5c	State	Available
Tenancy	Default	DHCP options set	dopt-02273cd8ff19fd048
Default VPC	Yes	IPv4 CIDR	172.31.0.0/16
Route 53 Resolver DNS Firewall rule groups	-	Owner ID	<input type="text"/> 146871189787
DNS hostnames		DNS resolution	
Enabled		Enabled	
Main route table		Main network ACL	
rtb-04ff62337ed83f7d0		acl-0dd545e3e0f094775	
IPv6 pool		IPv6 CIDR (Network border group)	
-		-	

CIDRs | Flow logs | Tags

CIDRs		Info		
Address type	CIDR	Network Border Group	Pool	Status
IPv4	172.31.0.0/16	-	-	Associated

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

The screenshot shows the AWS IAM Roles page with the following details:

Breadcrumbs: IAM > Roles > VP_Glue_Crawler_Service_Role

Role Name: VP_Glue_Crawler_Service_Role

Description: Allows Glue Service with read-only permissions to access the folders and files in the AWS S3 Bucket.

Summary:

Creation date	ARN
March 30, 2022, 17:28 (UTC+01:00)	arn:aws:iam::146871189787:role/VP_Glue_Crawler_Service_Role
Last activity	Maximum session duration
None	1 hour

Permissions: This tab is selected. Other tabs include Trust relationships, Tags, Access Advisor, and Revoke sessions.

Permissions policies (2): You can attach up to 10 managed policies.

Actions: Filter policies by property or policy name and press enter, Refresh, Simulate, Remove, Add permissions ▾, Previous, Next, and Refresh.

Policy name	Type	Description
AWSGlueServiceRole	AWS managed	Policy for AWS Glue service role which allows access to related services including EC2, S3, and Cloudwatch Logs
AmazonS3ReadOnlyAccess	AWS managed	Provides read only access to all buckets via the AWS Management Console.

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

The screenshot shows the AWS Glue Data Catalog interface. On the left, a sidebar menu lists various AWS Glue services: Data catalog, Databases, Tables, Connections, Crawlers, Classifiers, Schema registries, Schemas, Settings, ETL, AWS Glue Studio, Jobs, Jobs (legacy), ML Transforms, Blueprints, Workflows, Triggers, Dev endpoints, Notebooks, and Security. The 'Databases' section is currently selected.

The main content area is titled 'Databases' with the sub-instruction: 'A database is a set of associated table definitions, organized into a logical group.' Below this, there are three buttons: 'Add database' (highlighted in blue), 'View tables', and 'Action ▾'. To the right, a table displays one database entry:

Name	Description
<input type="checkbox"/> vp-stock-trades-bucket-database	This is the AWS Glue Database created to store the Data Catalog Schema Results retrieved by AWS Glue Crawler, by crawling the stock trade transaction raw source data files ingested in AWS S3 Bucket.

At the bottom right of the table, there are links for 'Showing: 1 - 1 < > 🔍 ⓘ'.

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

The screenshot shows the AWS Glue Data Catalog interface. On the left, there is a sidebar with the following navigation options:

- AWS Glue**
- Data catalog**
 - Databases
 - Tables
 - Connections
 - Crawlers
 - Classifiers
 - Schema registries
 - Schemas
 - Settings
- ETL**
 - AWS Glue Studio
 - Jobs
 - Jobs (legacy)
 - ML Transforms
 - Blueprints
 - Workflows
 - Triggers
 - Dev endpoints
 - Notebooks
- Security**
 - Security configurations

The main content area shows the details of a database named "vp-stock-trades-bucket-database". The database was created on "2023-09-01T10:00:00Z". It has the following properties:

- Name:** vp-stock-trades-bucket-database
- Description:** This is the AWS Glue Database created to stored the Data Catalog Schema Results retrieved by AWS Glue Crawler, by crawling the stock trade transaction raw source data files ingested in AWS S3 Bucket.
- Location:** (No location specified)

Below the database details, there is a section titled "Tables in vp-stock-trades-bucket-database" which currently contains no tables.

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

The screenshot shows the AWS Glue Connections page. On the left, there's a sidebar with navigation links for Data catalog, ETL, and AWS Glue Studio. The main area is titled "Connections" with a sub-instruction: "A connection contains the properties needed to connect to your data." It includes buttons for "Add connection", "Test connection", and "Action". A table lists the existing connections, showing one entry:

<input type="checkbox"/>	Name	Type	Date created	Last updated	Updated by
<input type="checkbox"/>	vp-stock-trades-bucket-glue-connection	Network	30 March 2022 6:01 PM UTC+1	30 March 2022 6:01 PM UTC+1	user/vidhyala...

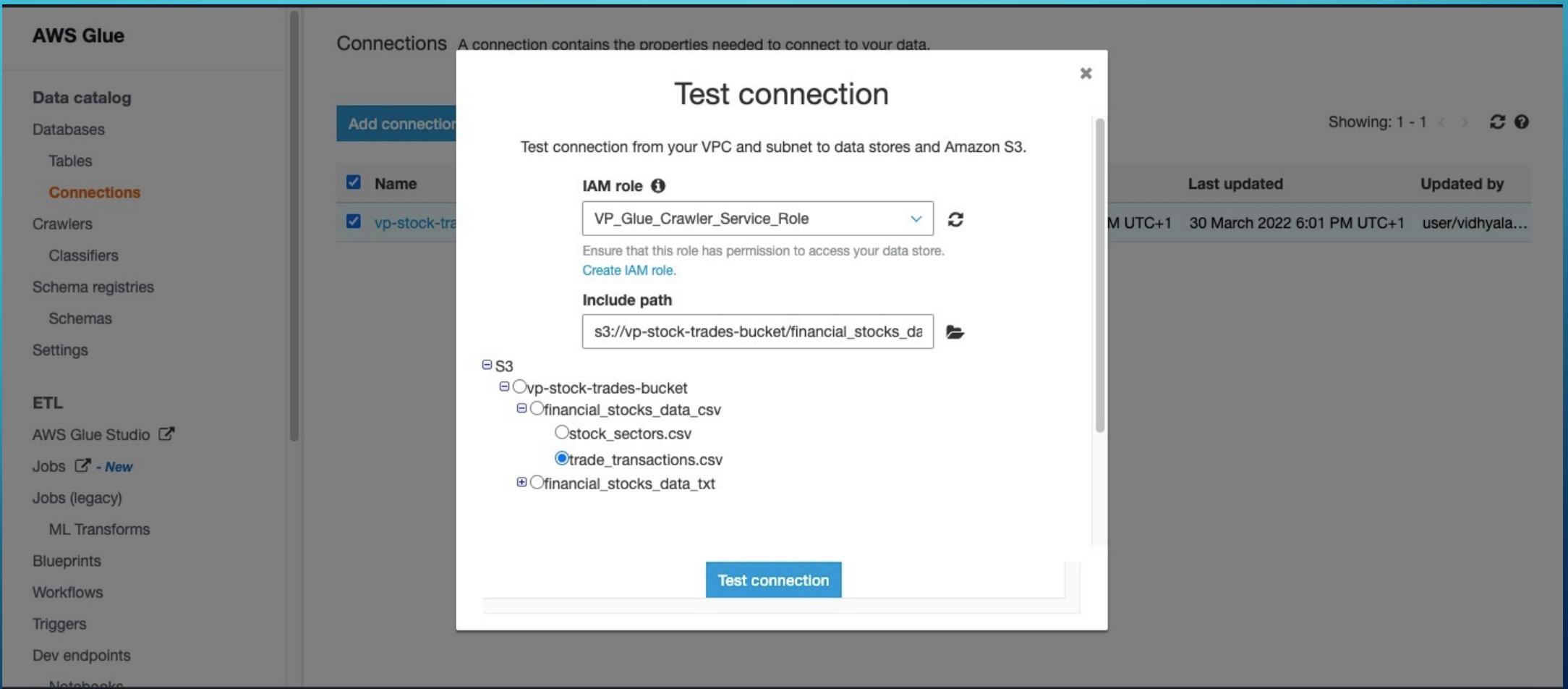
At the top right, there's a status message "Showing: 1 - 1" and some navigation icons.

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

The screenshot shows the AWS Glue Connections page. On the left, there's a sidebar with 'AWS Glue' and 'ETL' sections. Under 'AWS Glue', there are links for Data catalog, Databases, Tables, Connections, Crawlers, Classifiers, Schema registries, Schemas, and Settings. Under 'ETL', there are links for AWS Glue Studio, Jobs (New), Jobs (legacy), ML Transforms, Blueprints, Workflows, Triggers, Dev endpoints, and Notebooks. The main content area shows a connection named 'vp-stock-trades-bucket-glue-connection'. The 'Edit' button is highlighted with a red box. The connection details are as follows:

Setting	Value
Type	Network
VPC Id	vpc-0c73fc06befa93a5c
Subnet	subnet-018f5d5d635679cd6
Security groups	sg-06ce986cc8a1ca2ab
Require SSL connection	false
Description	Establish a network connection from AWS Glue to AWS S3 Bucket.
Created	30 March 2022 6:01 PM UTC+1
Last modified	30 March 2022 6:01 PM UTC+1

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)

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Schema registries

Schemas

Settings

ETL

AWS Glue Studio

Jobs - New

Jobs (legacy)

ML Transforms

Blueprints

Workflows

Triggers

Dev endpoints

Notebooks

Connections A connection contains the properties needed to connect to your data.

Testing vp-stock-trades-bucket-glue-connection access to your data store is in progress. This can take a few moments.

Add connection Test connection Action ▾ Showing: 1 - 1

<input checked="" type="checkbox"/> Name	Type	Date created	Last updated	Updated by
<input checked="" type="checkbox"/> vp-stock-trades-bucket-glue-connection	Network	30 March 2022 6:01 PM UTC+1	30 March 2022 6:01 PM UTC+1	user/vidhyala...

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

The screenshot shows the AWS Glue Connections page. On the left sidebar, under the 'Connections' section, the 'Connections' link is highlighted in orange. The main content area displays a success message: 'vp-stock-trades-bucket-glue-connection connected successfully to your instance.' Below this message is a table listing connections. The table has columns: Name, Type, Date created, Last updated, and Updated by. One row is visible, showing the connection 'vp-stock-trades-bucket-glue-connection' which is a 'Network' type connection created on 30 March 2022 at 6:01 PM UTC+1, last updated on the same date at 6:01 PM UTC+1, and updated by 'user/vidhyala...'. The 'Test connection' button is visible above the table.

Name	Type	Date created	Last updated	Updated by
vp-stock-trades-bucket-glue-connection	Network	30 March 2022 6:01 PM UTC+1	30 March 2022 6:01 PM UTC+1	user/vidhyala...

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

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a sidebar with navigation links for AWS Glue, Data catalog, Databases, Tables, Connections, Crawlers, Classifiers, Schema registries, Schemas, Settings, ETL (AWS Glue Studio, Jobs, Jobs (legacy), ML Transforms, Blueprints, Workflows, Triggers, Dev endpoints, Notebooks), and a link to the AWS Glue Studio documentation.

The main content area is titled "Crawlers". It contains a brief description: "A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog." Below this is a search bar with placeholder text "Filter by tags and attributes".

There are three buttons at the top of the crawler list: "Add crawler" (highlighted in blue), "Run crawler", and "Action ▾". To the right of the search bar are "User preferences" and "Showing: 1 - 1 < > ⌂ ⓘ".

The crawler list table has the following columns: Name, Schedule, Status, Logs, Last runtime, Median runtime, Tables updated, and Tables added. One row is visible in the table:

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
vp-stock-trades-bucket-glue-crawler		Ready		0 secs	0 secs	0	0

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

The screenshot shows the AWS Glue console interface. On the left, there's a navigation sidebar with sections like Data catalog, ETL, and Security. The main area is titled 'Crawlers > vp-stock-trades-bucket-glue-crawler'. It displays the configuration for this specific crawler.

Crawler Details:

- Name:** vp-stock-trades-bucket-glue-crawler
- Description:** This is an AWS Glue Crawler created to crawl and retrieve the Data Catalogue Schema Results of the specific source raw data files in the configured AWS S3 Bucket.
- Create a single schema for each S3 path:** false
- Table level:**
- Security configuration:**
- Tags:** -
- State:** Ready
- Schedule:**
- Last updated:** Wed Mar 30 23:56:02 GMT+100 2022
- Date created:** Wed Mar 30 19:29:49 GMT+100 2022
- Database:** vp-stock-trades-bucket-database
- Table prefix:** vp_stock_trades_bucket
- Service role:** VP_Glue_Crawler_Service_Role
- Selected classifiers:**
- Data store:** S3
- Include path:** s3://vp-stock-trades-bucket
- Connection:** vp-stock-trades-bucket-glue-connection
- Exclude patterns:**

Configuration options:

- Schema updates in the data store:** Update the table definition in the data catalog.
- Object deletion in the data store:** Mark the table as deprecated in the data catalog.

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

The screenshot shows the AWS Glue console interface. On the left, there is a navigation sidebar with the following menu items:

- Data catalog
- Databases
- Tables
- Connections
- Crawlers** (highlighted in orange)
- Classifiers
- Schema registries
- Schemas
- Settings

Below the main content area, there is another sidebar for ETL:

- AWS Glue Studio
- Jobs - New
- Jobs (legacy)
- ML Transforms
- Blueprints
- Workflows
- Triggers
- Dev endpoints
- Notebooks

The main content area is titled "Crawlers" and contains the following information:

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawler "vp-stock-trades-bucket-glue-crawler" is now running.

Below this message, there is a search bar with the placeholder "Filter by tags and attributes" and a "User preferences" link. The table below shows the status of the crawler:

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	vp-stock-trades-bucket-glue-crawler		⌚ Starting		0 secs	0 secs	0	0

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

AWS Glue

Data catalog

- Databases
- Tables
- Connections

Crawlers

- Classifiers
- Schema registries
- Schemas
- Settings

ETL

- AWS Glue Studio
- Jobs - New
- Jobs (legacy)
- ML Transforms
- Blueprints
- Workflows
- Triggers
- Dev endpoints
- Notebooks

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawler "vp-stock-trades-bucket-glue-crawler" completed and made the following changes: 1 tables created, 0 tables updated. See the tables created in database [vp-stock-trades-bucket-database](#).

[User preferences](#)

Add crawler Run crawler Action ▾ Filter by tags and attributes Showing: 1 - 1

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	vp-stock-trades-bucket-glue-crawler		Ready	Logs	2 mins	2 mins	0	1

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

The screenshot shows the AWS CloudWatch Log Events interface. The left sidebar contains navigation links for CloudWatch, Favorites, Dashboards, Alarms, Logs (Log groups, Logs Insights), Metrics, X-Ray traces, Events, Application monitoring, Insights, Settings, and Getting Started. The main content area displays the log group path: CloudWatch > Log groups > /aws-glue/crawlers > vp-stock-trades-bucket-glue-crawler. The title is "Log events". A search bar contains the query "[401653b4-4416-40e2-a222-a29dd54e74b9]". Below the search bar are filter options: Clear, 1m, 30m, 1h, 12h, Custom, and a refresh icon. The log table has columns: ▶, Timestamp, and Message. The table lists the following log entries:

▶	Timestamp	Message
▶	2022-03-30T19:34:19.411+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Running Start Crawl for Crawler vp-stock-trades-buck...
▶	2022-03-30T19:35:01.821+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] INFO : S3 ConnectionName is vp-stock-trades-bucket-glue-connecti...
▶	2022-03-30T19:36:17.365+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Classification complete, writing results to database...
▶	2022-03-30T19:36:17.365+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] INFO : Crawler configured with SchemaChangePolicy {"UpdateBehavi...
▶	2022-03-30T19:36:29.056+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] INFO : Created table vp_stock_trades_buckettrade_transactions_cs...
▶	2022-03-30T19:36:29.068+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] INFO : !!398!!: Optional[{"namespace":"vp-stock-trades-bucket-da...
▶	2022-03-30T19:36:31.776+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Finished writing to Catalog
▶	2022-03-30T19:37:39.094+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Crawler has finished running and is in state READY

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

The screenshot shows the AWS CloudWatch Log Events interface. The left sidebar navigation includes:

- Favorites
- Dashboards
- Alarms (0 alarms, 0 events, 0 metrics)
- Logs
 - Log groups** (highlighted in orange)
 - Logs Insights
- Metrics
- X-Ray traces
- Events
- Application monitoring
- Insights
- Settings
- Getting Started

The main content area displays log events for the log group `/aws-glue/crawlers` under the crawler `vp-stock-trades-bucket-glue-crawler`. The log events table has columns for **Timestamp** and **Message**. The first few log entries are:

Timestamp	Message
2022-03-30T19:34:19.411+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Running Start Crawl for Crawler vp-stock-trades-bu... [401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Running Start Crawl for Crawler vp-stock-trades-bucket-glue-crawler
2022-03-30T19:35:01.821+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] INFO : S3 ConnectionName is vp-stock-trades-bucket-glue-connec... [401653b4-4416-40e2-a222-a29dd54e74b9] INFO : S3 ConnectionName is vp-stock-trades-bucket-glue-connection
2022-03-30T19:36:17.365+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Classification complete, writing results to database... [401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Classification complete, writing results to database vp-stock-trades-bucket-database
2022-03-30T19:36:17.365+01:00	[401653b4-4416-40e2-a222-a29dd54e74b9] INFO : Crawler configured with SchemaChangePolicy {"UpdateBeha... [401653b4-4416-40e2-a222-a29dd54e74b9] INFO : Crawler configured with SchemaChangePolicy { "UpdateBehavior": "UPDATE_IN_DATABASE", "DeleteBehavior": "DEPRECATE_IN_DATABASE" }

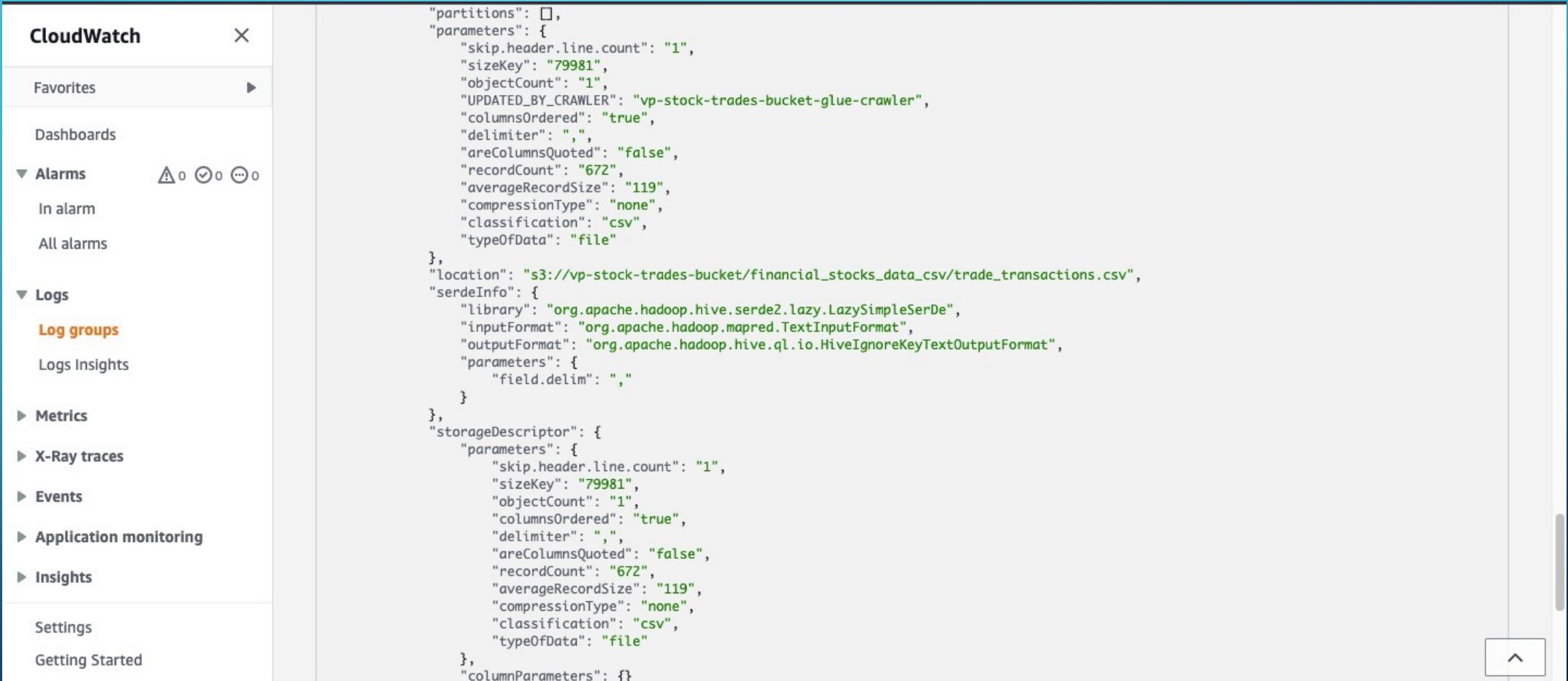
Filtering options include a search bar with placeholder `"401653b4-4416-40e2-a222-a29dd54e74b9"`, time range buttons (Clear, 1m, 30m, 1h, 12h, Custom), and a refresh button.

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

The screenshot shows the AWS CloudWatch interface with the 'Logs' section selected. On the left, a sidebar lists various monitoring options: Favorites, Dashboards, Alarms, Logs (selected), Log groups (highlighted in orange), Metrics, X-Ray traces, Events, Application monitoring, Insights, Settings, and Getting Started. The main pane displays log entries for a crawler named 'vp-stock-trades-buckettrade_transactions'. The first entry shows the crawler creating a table in a database. The second entry shows the schema definition for the table, which includes fields for 'Symbol', 'Owner', and 'Relationship', each with specific data types and properties.

```
[{"time": "2022-03-30T19:36:29.056+01:00", "log": "[401653b4-4416-40e2-a222-a29dd54e74b9] INFO : Created table vp_stock_trades_buckettrade_transactions_csv in database vp-stock-trades-bucket-database"}, {"time": "2022-03-30T19:36:29.068+01:00", "log": "[401653b4-4416-40e2-a222-a29dd54e74b9] INFO : !!398!!: Optional[{\n    \"namespace\": \"vp-stock-trades-bucket-database\",\n    \"tblName\": \"vp_stock_trades_buckettrade_transactions_csv\",\n    \"schema\": {\n        \"dataType\": \"struct\",\n        \"fields\": [\n            {\n                \"name\": \"Symbol\",\n                \"container\": {\n                    \"dataType\": \"string\",\n                    \"properties\": {}\n                },\n                \"properties\": {}\n            },\n            {\n                \"name\": \"Owner\",\n                \"container\": {\n                    \"dataType\": \"string\",\n                    \"properties\": {}\n                },\n                \"properties\": {}\n            },\n            {\n                \"name\": \"Relationship\",\n                \"container\": {\n                    \"dataType\": \"string\",\n                    \"properties\": {}\n                },\n                \"properties\": {}\n            }\n        ]\n    }\n}]}"}]
```

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



The screenshot shows the AWS CloudWatch console interface. The left sidebar contains navigation links: Favorites, Dashboards, Alarms (with 0 items), Logs (selected), Log groups, Logs Insights, Metrics, X-Ray traces, Events, Application monitoring, Insights, Settings, and Getting Started. The main content area displays a JSON log entry for a crawler execution. The log includes details such as partitions, parameters (including skip.header.line.count: "1", sizeKey: "79981", objectCount: "1", UPDATED_BY_CRAWLER: "vp-stock-trades-bucket-glue-crawler", columnsOrdered: "true", delimiter: ",", areColumnsQuoted: "false", recordCount: "672", averageRecordSize: "119", compressionType: "none", classification: "csv", typeOfData: "file"), location (s3://vp-stock-trades-bucket/financial_stocks_data_csv/trade_transactions.csv), serdeInfo (library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe, inputFormat: org.apache.hadoop.mapred.TextInputFormat, outputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat), parameters (field.delim: ","), and storageDescriptor (parameters: skip.header.line.count: "1", sizeKey: "79981", objectCount: "1", columnsOrdered: "true", delimiter: ",", areColumnsQuoted: "false", recordCount: "672", averageRecordSize: "119", compressionType: "none", classification: "csv", typeOfData: "file"). A scroll bar is visible on the right side of the content area.

```
    "partitions": [],
    "parameters": {
        "skip.header.line.count": "1",
        "sizeKey": "79981",
        "objectCount": "1",
        "UPDATED_BY_CRAWLER": "vp-stock-trades-bucket-glue-crawler",
        "columnsOrdered": "true",
        "delimiter": ",",
        "areColumnsQuoted": "false",
        "recordCount": "672",
        "averageRecordSize": "119",
        "compressionType": "none",
        "classification": "csv",
        "typeOfData": "file"
    },
    "location": "s3://vp-stock-trades-bucket/financial_stocks_data_csv/trade_transactions.csv",
    "serdeInfo": {
        "library": "org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe",
        "inputFormat": "org.apache.hadoop.mapred.TextInputFormat",
        "outputFormat": "org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat",
        "parameters": {
            "field.delim": ","
        }
    },
    "storageDescriptor": {
        "parameters": {
            "skip.header.line.count": "1",
            "sizeKey": "79981",
            "objectCount": "1",
            "columnsOrdered": "true",
            "delimiter": ",",
            "areColumnsQuoted": "false",
            "recordCount": "672",
            "averageRecordSize": "119",
            "compressionType": "none",
            "classification": "csv",
            "typeOfData": "file"
        }
    },
    "columnParameters": {}
}
```

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

The screenshot shows the AWS CloudWatch console interface. On the left, a sidebar menu lists various CloudWatch services: Favorites, Dashboards, Alarms, Logs (selected), Metrics, X-Ray traces, Events, Application monitoring, Insights, Settings, and Getting Started. The Logs section is expanded, showing Log groups and Log Insights.

The main pane displays the execution logs for a crawler. The log entries are:

```
        "sizeKey": "79981",
        "objectCount": "1",
        "columnsOrdered": "true",
        "delimiter": ",",
        "areColumnsQuoted": "false",
        "recordCount": "672",
        "averageRecordSize": "119",
        "compressionType": "none",
        "classification": "csv",
        "typeOfData": "file"
    },
    "columnParameters": {}
},
"hiveCompatible": false,
"description": null,
"columnComments": {},
"partitionComments": {},
"additionalLocations": null,
"registeredWithLakeFormation": false,
"classification": {
    "present": true
},
"deprecated": false
}
] CatalogDataForLocation(location=s3://vp-stock-trades-bucket/financial_stocks_data_csv/trade_transactions.csv,
simpleName=vp_stock_trades_buckettrade_transactions_csv,
compositeName=vp_stock_trades_buckettrade_transactions_csv_3f23fa671f8a592aab69a81b91a7038e, existingTable=Optional.empty,
tableForSimpleNamePresent=false)

```

2022-03-30T19:36:31.776+01:00 [401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Finished writing to Catalog

2022-03-30T19:37:39.094+01:00 [401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Finished writing to Catalog

2022-03-30T19:37:39.094+01:00 [401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Crawler has finished running and is in state READY

2022-03-30T19:37:39.094+01:00 [401653b4-4416-40e2-a222-a29dd54e74b9] BENCHMARK : Crawler has finished running and is in state READY

Copy Copy ^

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

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a sidebar with sections for Data catalog, ETL, and Dev endpoints. Under Data catalog, the 'Tables' section is selected, indicated by an orange highlight. The main content area is titled 'Tables' with a sub-instruction: 'A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.' Below this, there are buttons for 'Add tables' and 'Action', a search bar with placeholder 'Filter by attributes or search by keyword', and a 'Save view' button. The table header includes columns for Name, Database, Location, Classification, Last updated, and Deprecated. A message at the bottom states, 'You don't have any tables defined in your data catalog.', accompanied by a grid icon. A blue button labeled 'Add tables using a crawler' is visible.

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

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a sidebar with navigation links for Data catalog, ETL, and Security. Under Data catalog, the 'Tables' link is highlighted in orange. The main area is titled 'Tables' with a sub-instruction: 'A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.' Below this is a search bar with a placeholder 'Name : vp_stock_trades_bucketfinancial_stocks...', a filter bar, and a 'Save view' button. The table itself has columns: Name, Database, Location, Classification, Last updated, and Deprecated. There is one row visible:

Name	Database	Location	Classification	Last updated	Deprecated
vp_stock_trades_bucketfinancial_stocks_data_csv	vp-stock-trades-bucket-database	s3://vp-stock-trades-bucket/fin...	csv	31 March 2022 12:00 PM UTC+1	

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

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a sidebar with navigation links for Data catalog, ETL, and other services like AWS Glue Studio, Jobs, Workflows, etc. The main area displays a table named "vp_stock_trades_bucketfinancial_stocks_data_csv". The table details are as follows:

Name	vp_stock_trades_bucketfinancial_stocks_data_csv
Description	
Database	vp-stock-trades-bucket-database
Classification	csv
Location	s3://vp-stock-trades-bucket/financial_stocks_data_csv/
Connection	
Deprecated	No
Last updated	Thu Mar 31 00:00:10 GMT+100 2022
Input format	org.apache.hadoop.mapred.TextInputFormat
Output format	org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Serde parameters	field.delim , skip.header.line.count 1 sizeKey 79981 objectCount 1
Table properties	UPDATED_BY_CRAWLER vp-stock-trades-bucket-glue-crawler CrawlerSchemaSerializerVersion 1.0 recordCount 672 averageRecordSize 119 CrawlerSchemaDeserializerVersion 1.0 compressionType none columnsOrdered true areColumnsQuoted false delimiter , typeOfData file

On the right, there's a "Versions" section showing one version entry:

Version	Created	Created by
0	31 March 2022 1...	role/VP_Glue_Cr... 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

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a sidebar with navigation links for Data catalog, Databases, Tables, Connections, Crawlers, Classifiers, Schema registries, Schemas, Settings, ETL (AWS Glue Studio, Jobs, Jobs (legacy), ML Transforms, Blueprints, Workflows, Triggers, Dev endpoints, Notebooks), and Security. The main area shows a table named "vp_stock_trades_bucketfinancial_stocks_data_csv". The table properties are listed as follows:

Name	vp_stock_trades_bucketfinancial_stocks_data_csv
Description	
Database	vp-stock-trades-bucket-database
Classification	csv
Location	s3://vp-stock-trades-bucket/financial_stocks_data_csv/
Connection	
Deprecated	No
Last updated	Thu Mar 31 00:00:10 GMT+100 2022
Input format	org.apache.hadoop.mapred.TextInputFormat
Output format	org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Serde serialization lib	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
Serde parameters	field.delim : , skip.header.line.count : 1 sizeKey : 79981 objectCount : 1
Table properties	UPDATED_BY_CRAWLER : vp-stock-trades-bucket-glue-crawler CrawlerSchemaSerializerVersion : 1.0 recordCount : 672 averageRecordSize : 119 CrawlerSchemaDeserializerVersion : 1.0 compressionType : none columnsOrdered : true areColumnsQuoted : false delimiter : , typeOfData : file

At the bottom right, it says "Showing: 1 - 12 of 12 < >".

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

AWS Glue Database Table after executing the AWS Glue Crawler

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a sidebar with navigation links for AWS Glue, Data catalog, ETL, and Security. The main area displays a table named 'vp_stock_trades' with the following properties:

Table properties	Value
UPDATED_BY_CRAWLER	vp_stock_trades bucket glue crawler
CrawlerSchemaDeserializerVersion	1.0
compressionType	none
columnsOrdered	true
averageRecordSize	119
areColumnsQuoted	false
delimiter	,
typeOfData	file

The table has 12 columns, each with a unique ID from 1 to 12. The schema details are as follows:

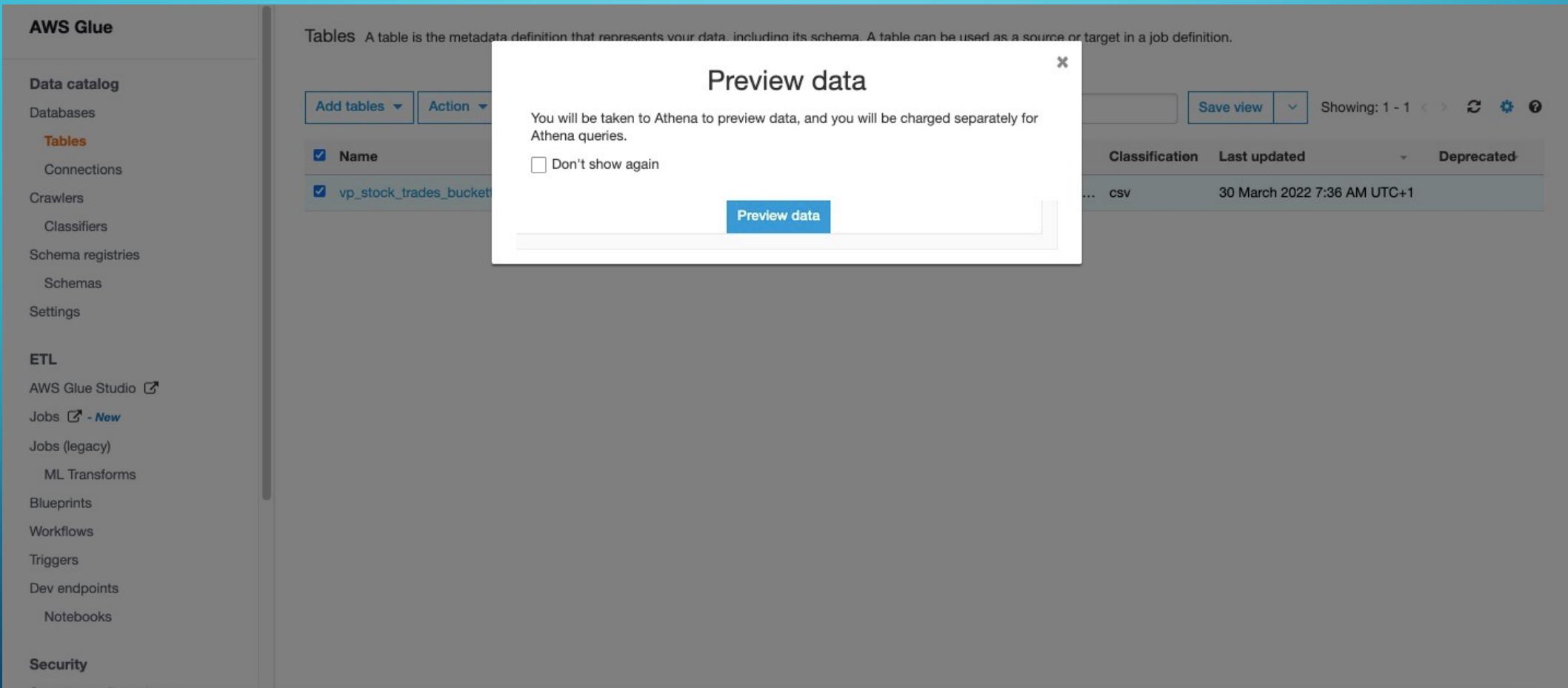
	Column name	Data type	Partition key	Comment
1	symbol	string		
2	owner	string		
3	relationship	string		
4	date	string		
5	cost	double		
6	# shares	bigint		
7	value(\$)	string		
8	total shares	bigint		
9	filing	string		
10	type	string		
11	currentprice	string		
12	movingaverage	string		

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

The screenshot shows the AWS Glue Data Catalog interface. On the left, there's a sidebar with navigation links for Data catalog, ETL, and Security. Under Data catalog, the 'Tables' link is highlighted. The main area displays a table of tables. One row is selected, and a context menu is open over it, showing options like 'Edit table details', 'View details', 'View data' (which is highlighted with a blue background), and 'Delete table'. The table itself has columns for Name, Database, Location, Classification, Last updated, and Deprecated. The selected row shows 'actions_csv' as the name, 'vp-stock-trades-bucket-database' as the database, 's3://vp-stock-trades-bucket/fin...' as the location, 'csv' as the classification, and '30 March 2022 7:36 AM UTC+1' as the last updated time.

Name	Database	Location	Classification	Last updated	Deprecated
actions_csv	vp-stock-trades-bucket-database	s3://vp-stock-trades-bucket/fin...	csv	30 March 2022 7:36 AM UTC+1	

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)

The screenshot shows the Amazon Athena Query editor interface. The top navigation bar includes 'Amazon Athena > Query editor', tabs for 'Editor' (selected), 'Recent queries', 'Saved queries', and 'Settings', and a 'Workgroup' dropdown set to 'primary'. The main area is divided into two panes: 'Data' on the left and 'Query' on the right.

Data Pane: Contains fields for 'Data Source' (set to 'AwsDataCatalog') and 'Database' (set to 'vp-stock-trades-bucket-database'). Below these are sections for 'Tables and views' (with a 'Create' button) and a table listing. The table section shows 'Tables (2)' with one entry: 'vp_stock_trades_bucketfinancial_stocks_data_csv'. This table has columns: symbol (string), owner (string), relationship (string), date (string), cost (double), # shares (int), value(\$), and total shares (int).

Query Pane: Shows 'Query 7' (disabled) and 'Query 8' (selected). The SQL query in 'Query 8' is:

```
1 SELECT * FROM "AwsDataCatalog"."vp-stock-trades-bucket-database"."vp_stock_trades_bucketfinancial_stocks_data_csv" limit 10;
```

The status bar indicates the query is 'Completed' with a run time of 0.424 sec and data scanned of 78.11 KB. The results pane displays 10 rows of data:

#	symbol	owner	relationship	date	cost	# shares	value(\$)	total shares
1	EVR	Walsh Robert B	Principal Financial Officer	"Apr 23"	50	2000	100280	
2	DSNY	Graber Mark A	10% Owner	"Apr 20"	0	13000	7800	

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

The screenshot shows the Amazon S3 console interface. On the left, the navigation pane includes 'Buckets', 'Access Points', 'Object Lambda Access Points', 'Multi-Region Access Points', 'Batch Operations', 'Access analyzer for S3', 'Block Public Access settings for this account', 'Storage Lens' (with 'Dashboards' and 'AWS Organizations settings'), 'Feature spotlight' (with a '3' notification), and 'AWS Marketplace for S3'. The main content area shows the 'vp-stock-trades-bucket' bucket details. The 'Objects' tab is selected, displaying three objects: 'athena_output_query_results/' (selected), 'financial_stocks_data_csv/', and 'financial_stocks_data_txt/'. Below the table are buttons for 'Upload', 'Find objects by prefix', and navigation controls.

Name	Type	Last modified	Size	Storage class
athena_output_query_results/	Folder	-	-	-
financial_stocks_data_csv/	Folder	-	-	-
financial_stocks_data_txt/	Folder	-	-	-

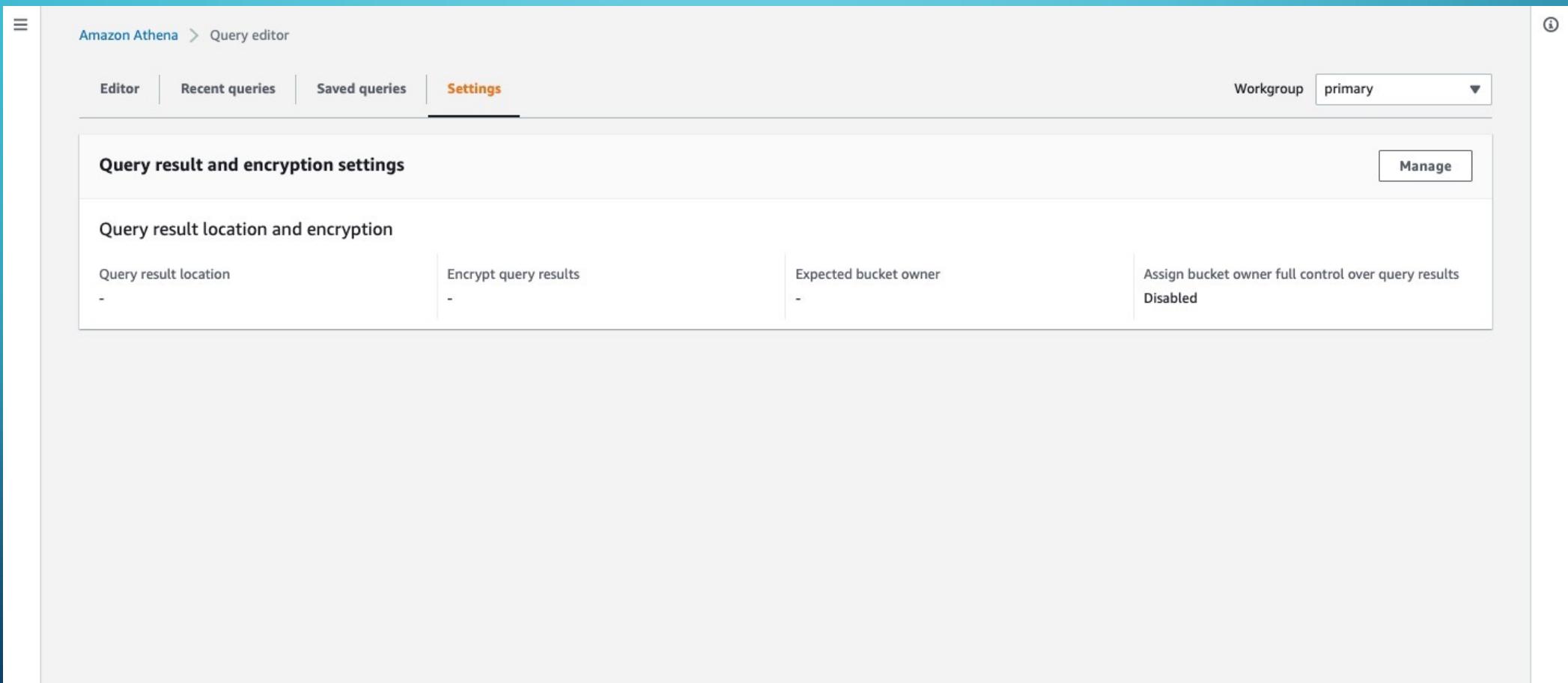
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA (CONTINUED)

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

The screenshot shows the Amazon S3 console interface. On the left, the navigation pane includes 'Buckets', 'Access Points', 'Object Lambda Access Points', 'Multi-Region Access Points', 'Batch Operations', 'Access analyzer for S3', 'Block Public Access settings for this account', 'Storage Lens' (with 'Dashboards' and 'AWS Organizations settings' sub-options), 'Feature spotlight' (with a '3' notification), and 'AWS Marketplace for S3'. The main content area shows the path 'Amazon S3 > Buckets > vp-stock-trades-bucket > athena_output_query_results/'. The folder name 'athena_output_query_results/' is displayed prominently. A 'Copy S3 URI' button is located in the top right. Below it, tabs for 'Objects' and 'Properties' are shown, with 'Objects' selected. The 'Objects (0)' section contains instructions about objects and permissions, along with buttons for 'Upload', 'Copy S3 URI', 'Copy URL', 'Download', 'Open', 'Delete', 'Actions', and 'Create folder'. A search bar labeled 'Find objects by prefix' is also present. At the bottom, a table header for 'Name', 'Type', 'Last modified', 'Size', and 'Storage class' is shown, followed by the message 'No objects' and 'You don't have any objects in this folder.' A final 'Upload' button is at the bottom right.

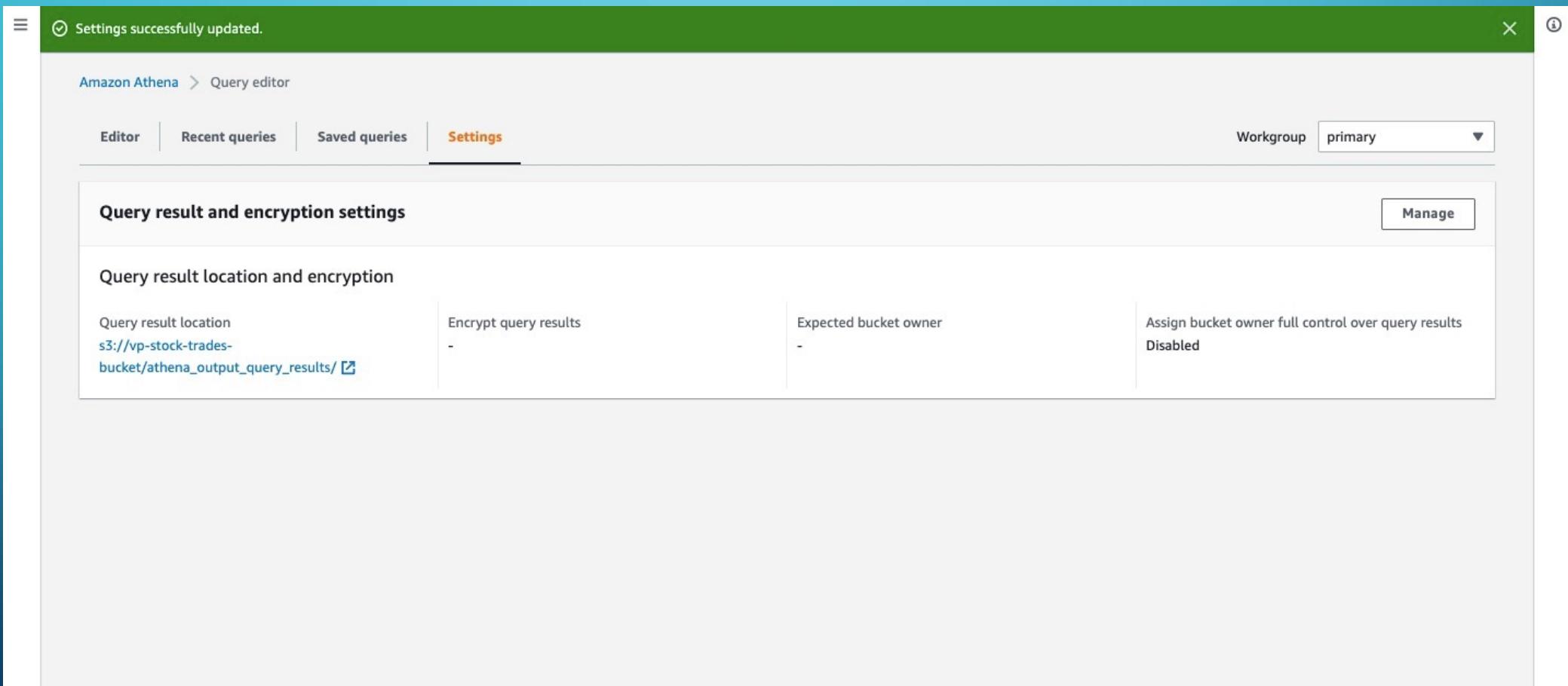
13. QUERYING THE RAW SOURCE DATA STORED IN THE AWS S3 BUCKET THROUGH AWS GLUE INTERFACE USING AWS ATHENA (CONTINUED)

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

The screenshot shows the Amazon Athena Query editor interface. The top navigation bar includes tabs for 'Editor' (which is selected), 'Recent queries', 'Saved queries', and 'Settings'. A dropdown for 'Workgroup' is set to 'primary'. On the left, the 'Data' sidebar shows the 'Data Source' as 'AwsDataCatalog' and the 'Database' as 'vp-stock-trades-bucket-database'. Under 'Tables and views', there are two tables listed: 'vp_stock_trades_bucketfinancial_stoc' and 'ks_data_csv'. The 'vp_stock_trades_bucketfinancial_stoc' table has columns: symbol (string), owner (string), relationship (string), date (string), and cost (double). The main workspace displays 'Query 7' with the following SQL query:

```
1 SELECT * FROM "AwsDataCatalog"."vp-stock-trades-bucket-database"."vp_stock_trades_bucketfinancial_stocks_data_csv" WHERE "# shares" < 10
```

The query is currently at 'Ln 1, Col 1'. Below the query are buttons for 'Run', 'Cancel', 'Save', 'Clear', and 'Create'. The results section shows 'Results (0)' with a note: 'No results' and 'Run a query to view results'. There are also 'Copy' and 'Download results' buttons.

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

The screenshot shows the Amazon Athena Query Editor interface. The left sidebar displays the Data Source (AwsDataCatalog) and Database (vp-stock-trades-bucket-database). The Tables and views section shows two tables: vp_stock_trades_bucketfinancial_stoc ks_data_csv. The main area displays a completed query named "Query 7". The SQL code is:

```
1 SELECT * FROM "AwsDataCatalog"."vp-stock-trades-bucket-database"."vp_stock_trades_bucketfinancial_stocks_data_csv" WHERE "# shares" < 10
```

The results section shows the output of the query, which includes columns: #, symbol, owner, relationship, date, cost, # shares, and value(\$). There are two rows of data:

#	symbol	owner	relationship	date	cost	# shares	value(\$)
1	DSNY	Graber Mark A	10% Owner	"Apr 20		0	13000
2	RVP	SHAW THOMAS J	President and CEO	"Apr 24		3	800

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

Amazon S3 > Buckets > vp-stock-trades-bucket > athena_output_query_results/

athena_output_query_results/

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 more](#)

Actions ▾

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	Unsaved/	Folder	-	-	-

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

The screenshot shows the AWS S3 console interface. The navigation path is: Amazon S3 > Buckets > vp-stock-trades-bucket > athena_output_query_results/ > Unsaved/ > 2022/ > 03/ > 31/. The current view is for the folder '31/'. There are two objects listed:

Name	Type	Last modified	Size	Storage class
bfe0ca20-2dfc-4d44-8c50-469783b52534.csv	csv	March 31, 2022, 00:41:53 (UTC+01:00)	25.4 KB	Standard
bfe0ca20-2dfc-4d44-8c50-469783b52534.csv.metadata	metadata	March 31, 2022, 00:41:53 (UTC+01:00)	598.0 B	Standard

Below the table, there is a note: "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](#)".

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

The screenshot shows the AWS S3 Object Overview page for the file `bfe0ca20-2dfc-4d44-8c50-469783b52534.csv.metadata`. The page displays various metadata details such as Owner, AWS Region, Last modified, Size, Type, and Key. It also provides links for S3 URI, Amazon Resource Name (ARN), Entity tag (Etag), and Object URL.

Object overview

Key	Value
Owner	4d97ff8f9dfc260e11d18083dc29d77931845b8a284d61ae03f6323dc4a5b186
AWS Region	EU (London) eu-west-2
Last modified	March 31, 2022, 00:41:53 (UTC+01:00)
Size	598.0 B
Type	metadata
Key	athena_output_query_results/Unsaved/2022/03/31/bfe0ca20-2dfc-4d44-8c50-469783b52534.csv.metadata

S3 URI

s3://vp-stock-trades-bucket/athena_output_query_results/Unsaved/2022/03/31/bfe0ca20-2dfc-4d44-8c50-469783b52534.csv.metadata

Amazon Resource Name (ARN)

arn:aws:s3:::vp-stock-trades-bucket/athena_output_query_results/Unsaved/2022/03/31/bfe0ca20-2dfc-4d44-8c50-469783b52534.csv.metadata

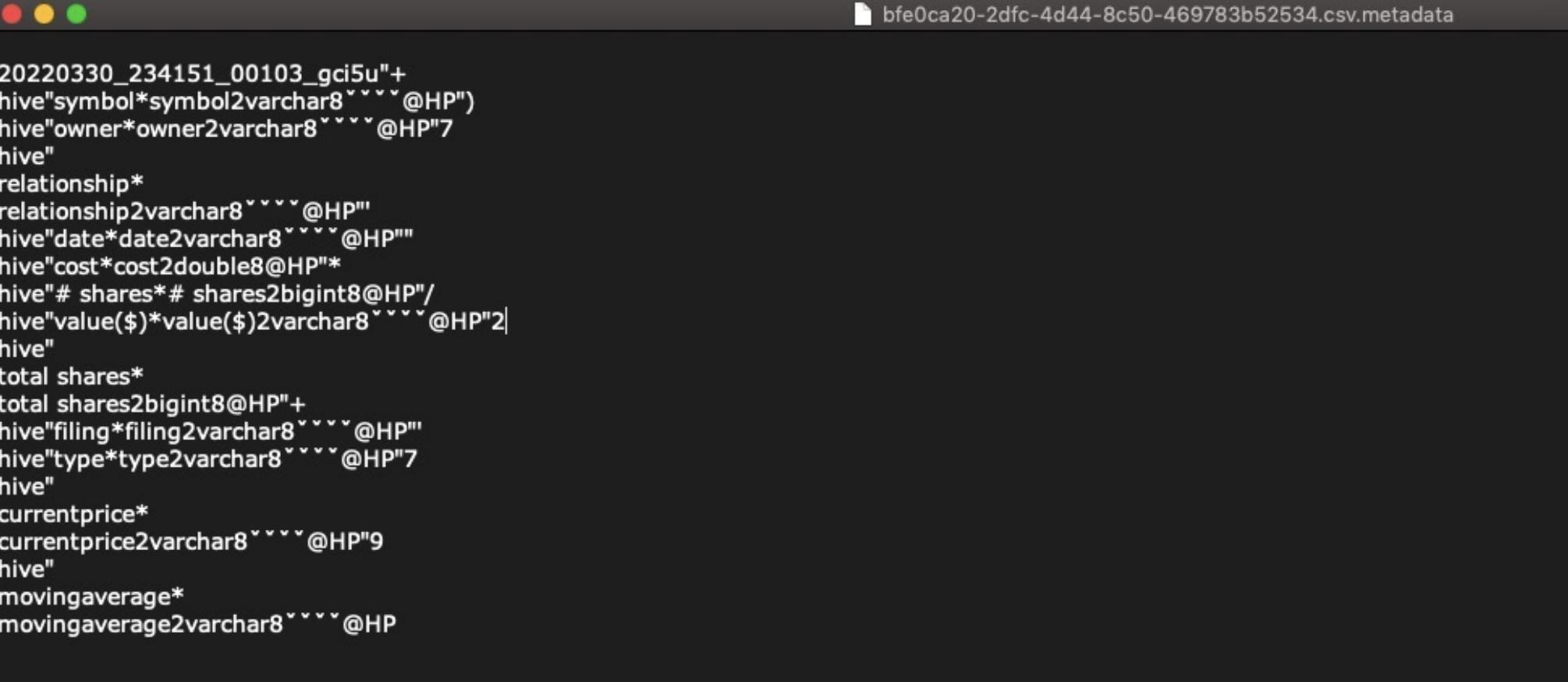
Entity tag (Etag)

[fb8f77e04d3aea7a6d0cf10230b703c9](#)

Object URL

https://vp-stock-trades-bucket.s3.eu-west-2.amazonaws.com/athena_output_query_results/Unsaved/2022/03/31/bfe0ca20-2dfc-4d44-8c50-469783b52534.csv.metadata

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



The screenshot shows a terminal window with a dark background and light-colored text. The title bar of the window reads "bfe0ca20-2dfc-4d44-8c50-469783b52534.csv.metadata". The window contains the following text:

```
20220330_234151_00103_gci5u"+
hive"symbol*symbol2varchar8***@HP")
hive"owner*owner2varchar8***@HP"7
hive"
relationship*
relationship2varchar8***@HP"
hive"date*date2varchar8***@HP"""
hive"cost*cost2double8@HP"*
hive"# shares*# shares2bigint8@HP"/
hive"value($)*value($)2varchar8***@HP"2
hive"
total shares*
total shares2bigint8@HP"+
hive"filing*filing2varchar8***@HP"
hive"type*type2varchar8***@HP"7
hive"
currentprice*
currentprice2varchar8***@HP"9
hive"
movingaverage*
movingaverage2varchar8***@HP
```

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

The screenshot shows the AWS S3 Object Properties page for a file named `bfe0ca20-2dfc-4d44-8c50-469783b52534.csv`. The file was generated by an Athena query and is stored in the `athena_output_query_results/Unsaved/2022/03/31/` directory of the `vp-stock-trades-bucket`.

Object overview

Attribute	Value
Owner	4d97ff8f9dfc260e11d18083dc29d77931845b8a284d61ae03f6323dc4a5b186
AWS Region	EU (London) eu-west-2
Last modified	March 31, 2022, 00:41:53 (UTC+01:00)
Size	25.4 KB
Type	csv
Key	athena_output_query_results/Unsaved/2022/03/31/bfe0ca20-2dfc-4d44-8c50-469783b52534.csv

Properties | Permissions | Versions

S3 URI
s3://vp-stock-trades-bucket/athena_output_query_results/Unsaved/2022/03/31/bfe0ca20-2dfc-4d44-8c50-469783b52534.csv

Amazon Resource Name (ARN)
[arn:aws:s3:::vp-stock-trades-bucket/athena_output_query_results/Unsaved/2022/03/31/bfe0ca20-2dfc-4d44-8c50-469783b52534.csv](#)

Entity tag (Etag)
[3f68334897281d32a7232ce975d55a43](#)

Object URL
https://vp-stock-trades-bucket.s3.eu-west-2.amazonaws.com/athena_output_query_results/Unsaved/2022/03/31/bfe0ca20-2dfc-4d44-8c50-469783b52534.csv

Copy S3 URI | Download | Open | Object actions ▾

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

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 Warren	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 & TREA	"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	Hermann 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 GIF1	RICHARD C D	Director	"Apr 22		2	500	1496	4625	"Apr 24	2020	12:15 PM"				
19 PAAC	Chou Shih-Ct	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 & Chair	"Apr 22		0	2805	1543	1384758	"Apr 23	2020	06:34 PM"				
23 XELB	D LOREN RO	CEO & Chair	"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 MMPL	MARTIN RUI	President an	"Apr 21		1	7684	8927	442587	"Apr 23	2020	03:59 PM"				
30 MMPL	Shoup Scot A	Senior VP Of	"Apr 21		1	3111	3614	15037	"Apr 23	2020	03:59 PM"				
31 MMPL	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"				



END OF THE PROJECT