Lesson 04 Demo 10

Querying S3 Bucket Operations with Athena

Objective: To execute a query within AWS Athena for performing operations on a designated S3 bucket

Tools required: None

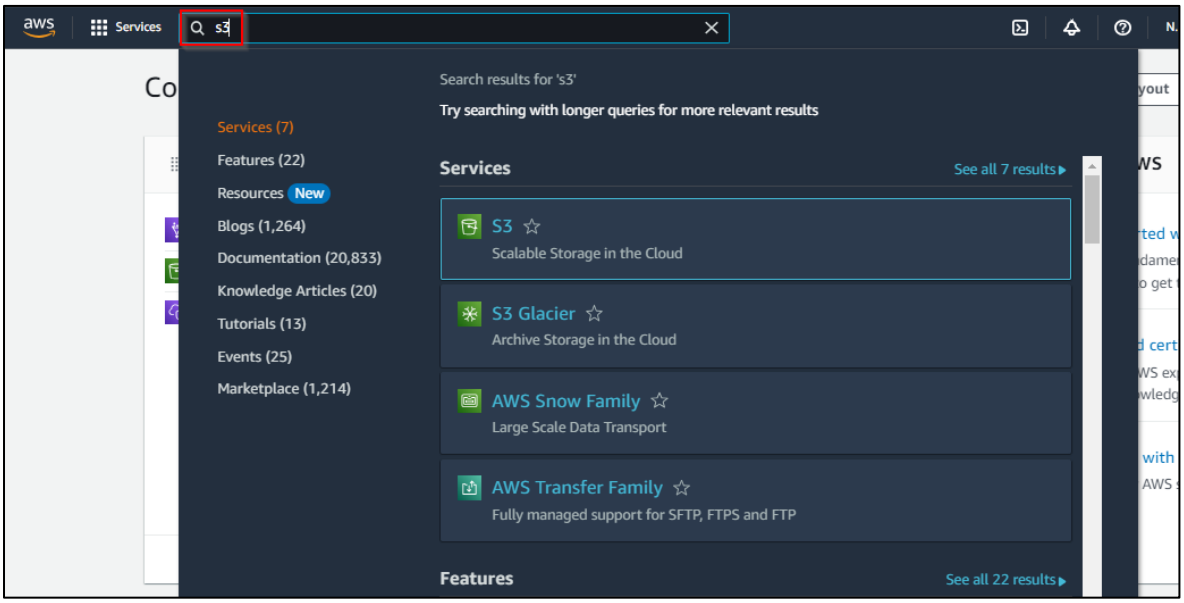
Prerequisites: AWS Lab access with an AWS account created

Steps to be followed:

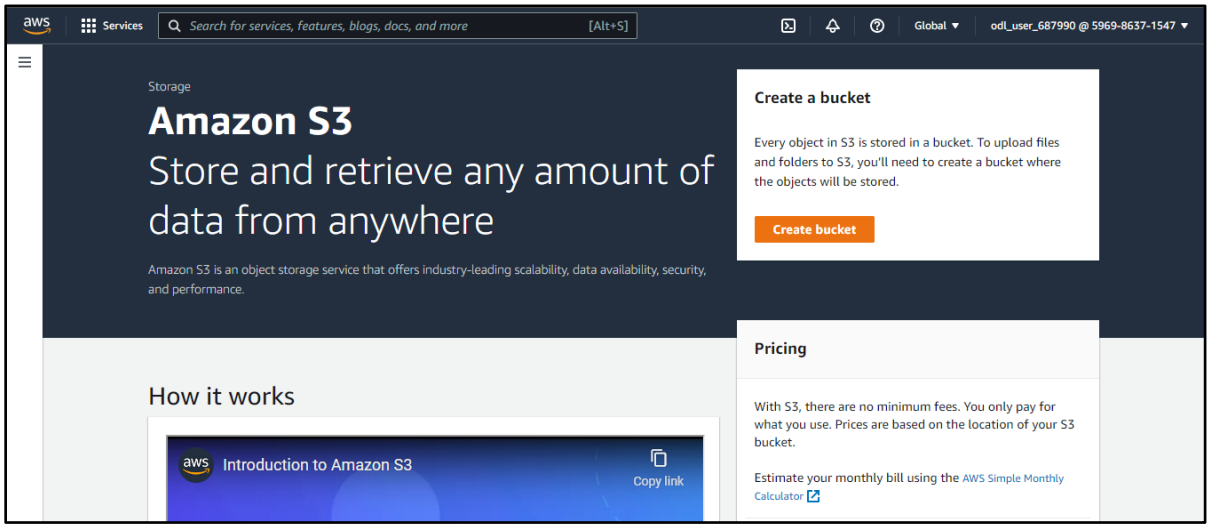
1. Create S3 buckets
2. Execute queries in Athena

Step 1: Create S3 buckets

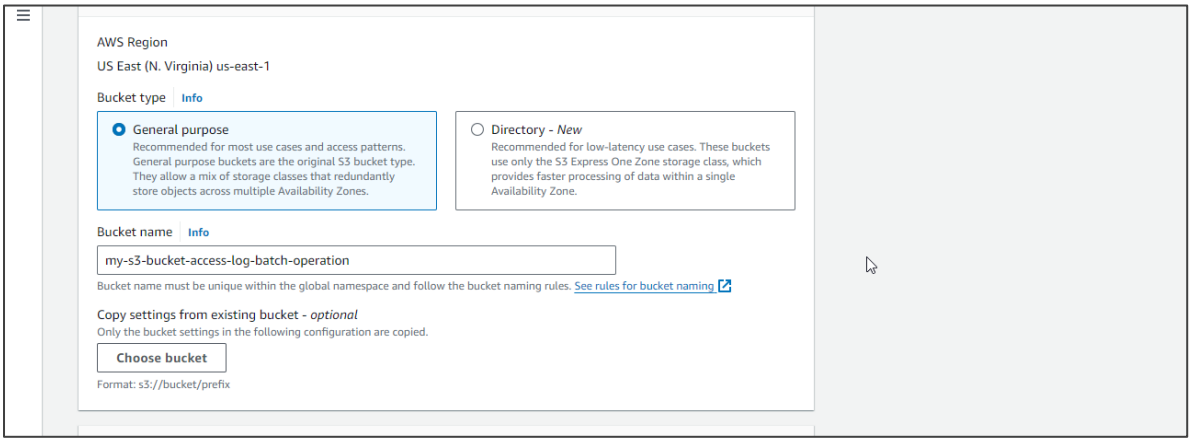
1.1 Open the AWS Management Console and search for **S3 bucket**

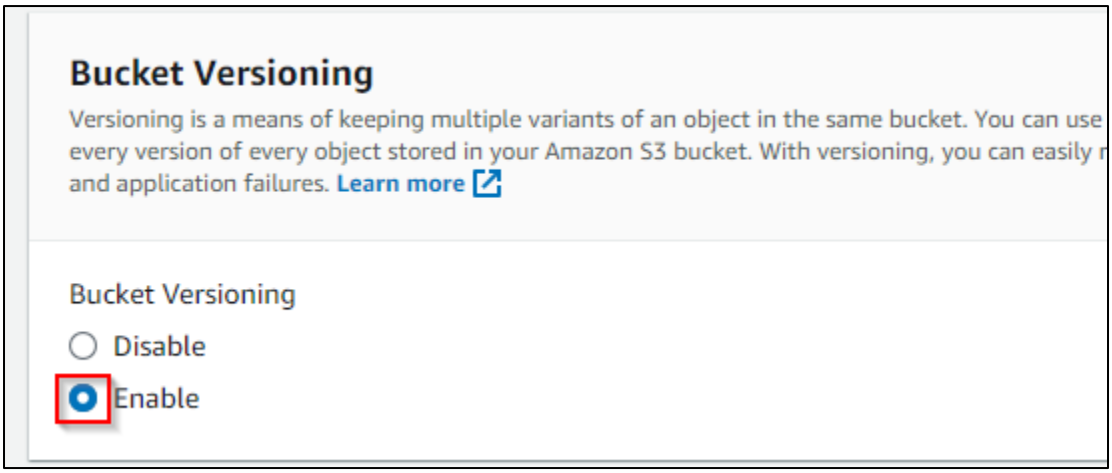


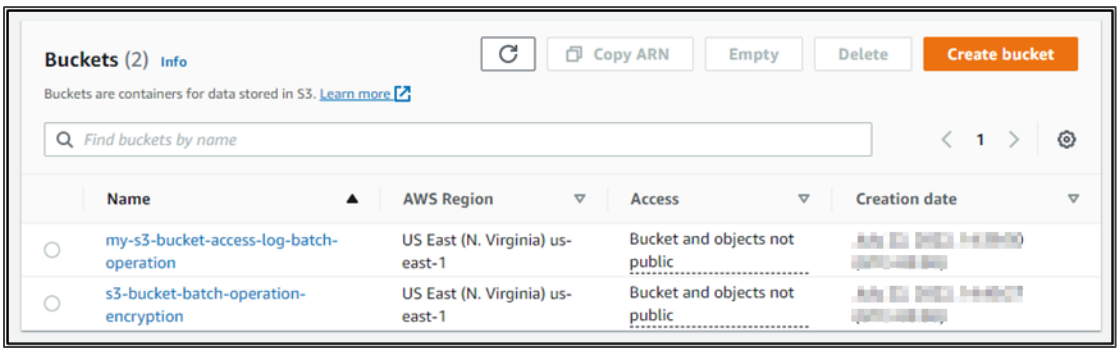
1.2 Click on **Create bucket**



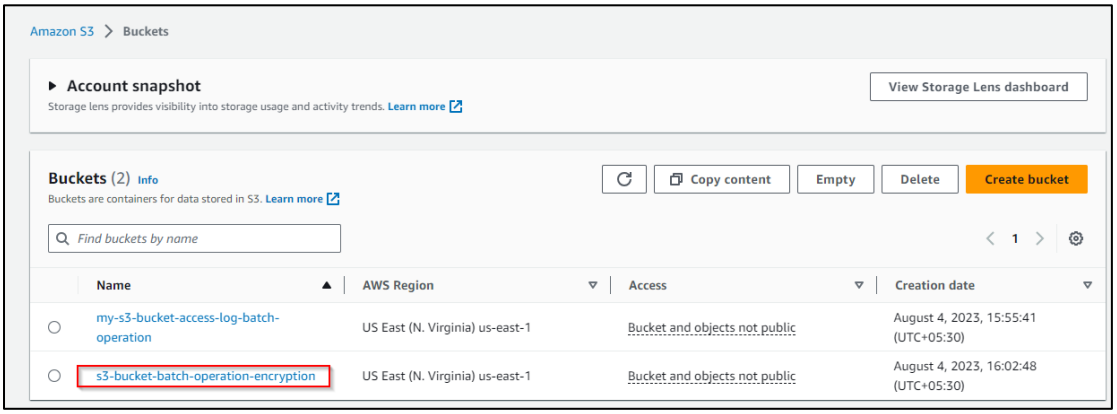
1.3 Create two buckets named **my-s3-bucket-access-log-batch-operation** and **s3-bucket-batch-operation-encryption** to aggregate log files from the **s3-bucket-batch-operation-encryption** bucket. Enable **bucket Versioning** for both.



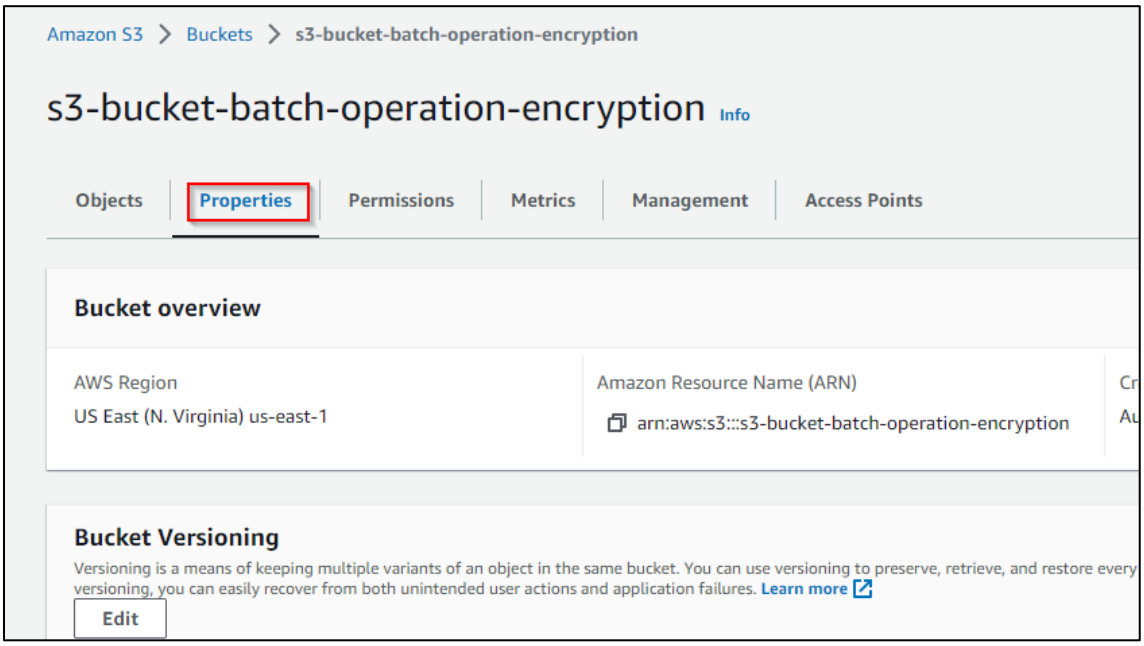




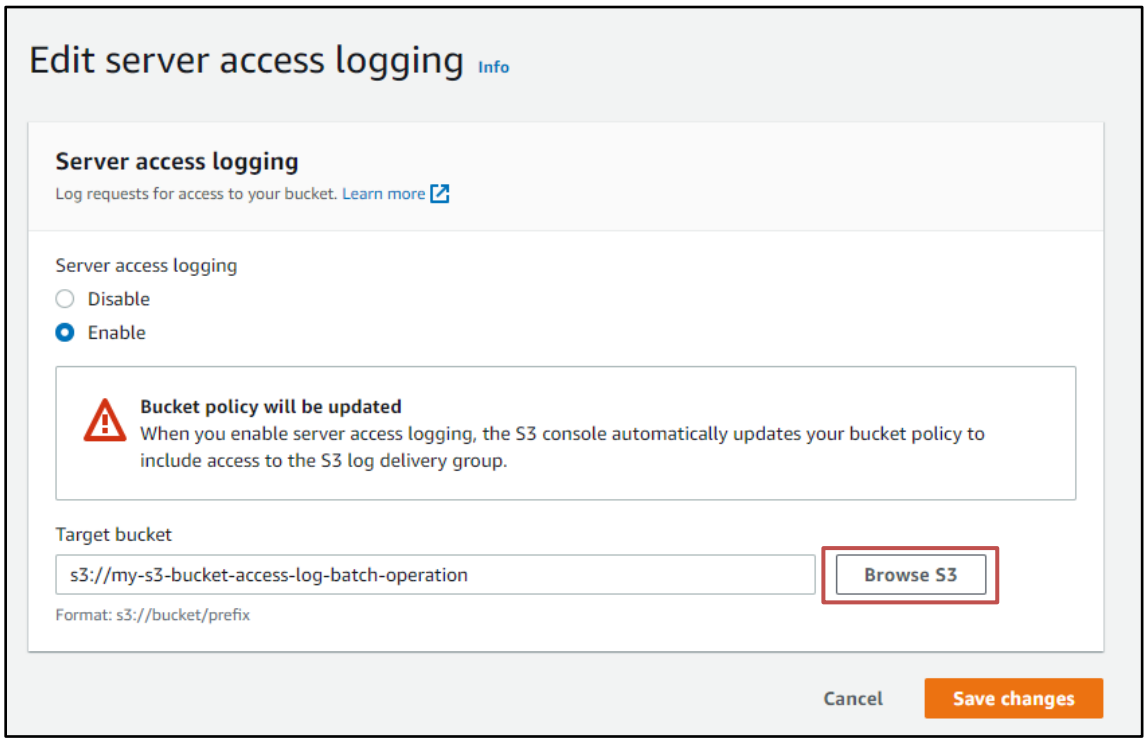
1.4 Access the **s3-bucket-batch-operation-encryption** bucket



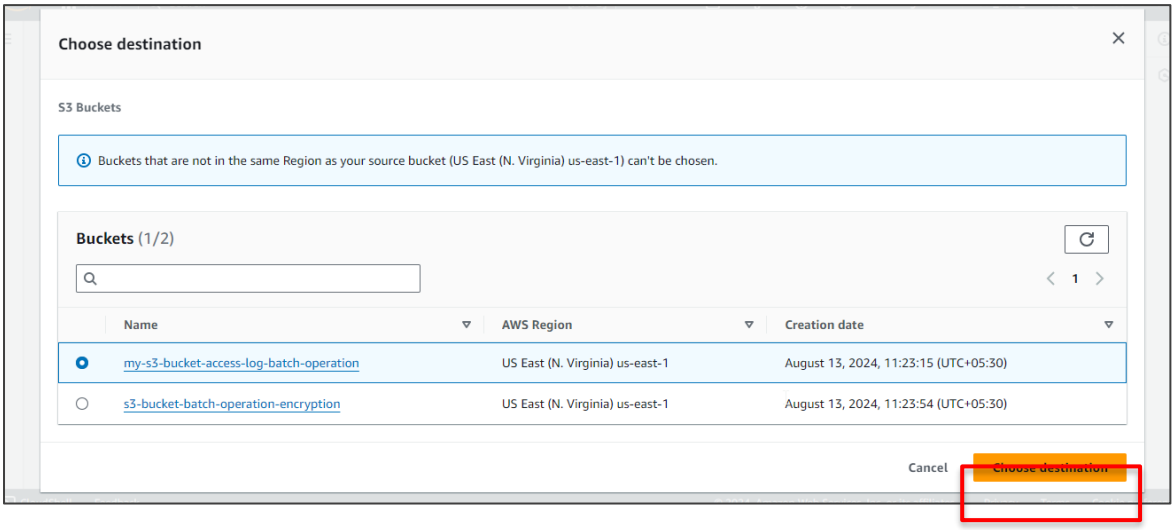
1.5 Navigate to **Properties**



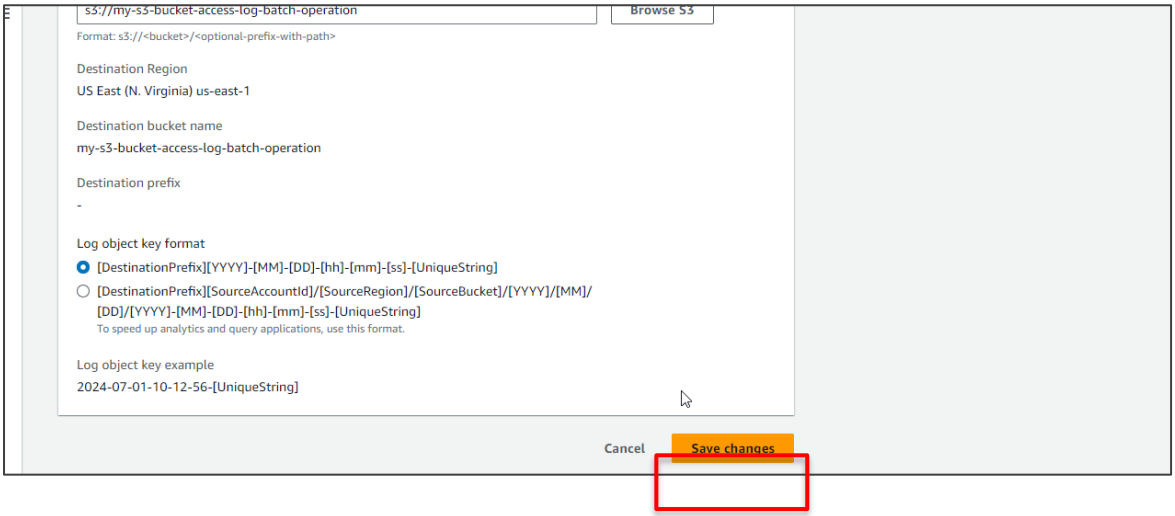
1.6 Scroll down and select **Server access logging**, and then click **Edit** and **Browse S3**



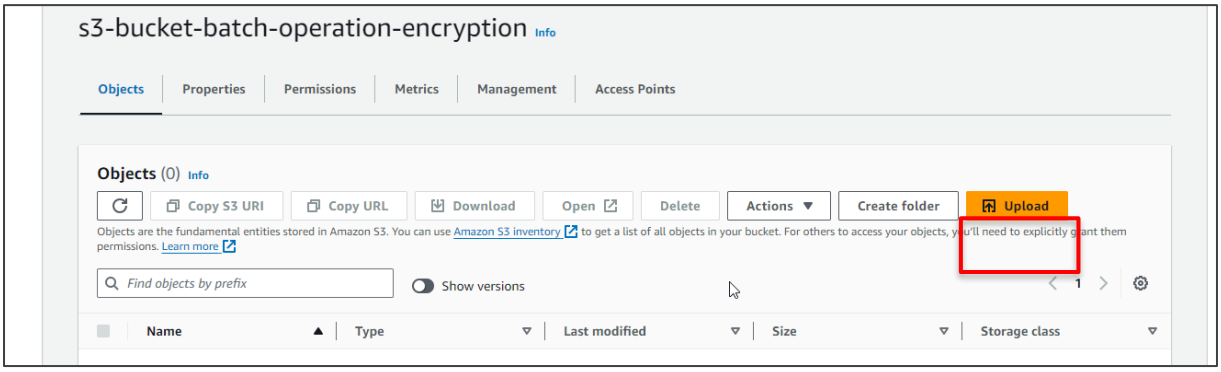
1.7 Choose the log-batch file path and click **Choose destination**

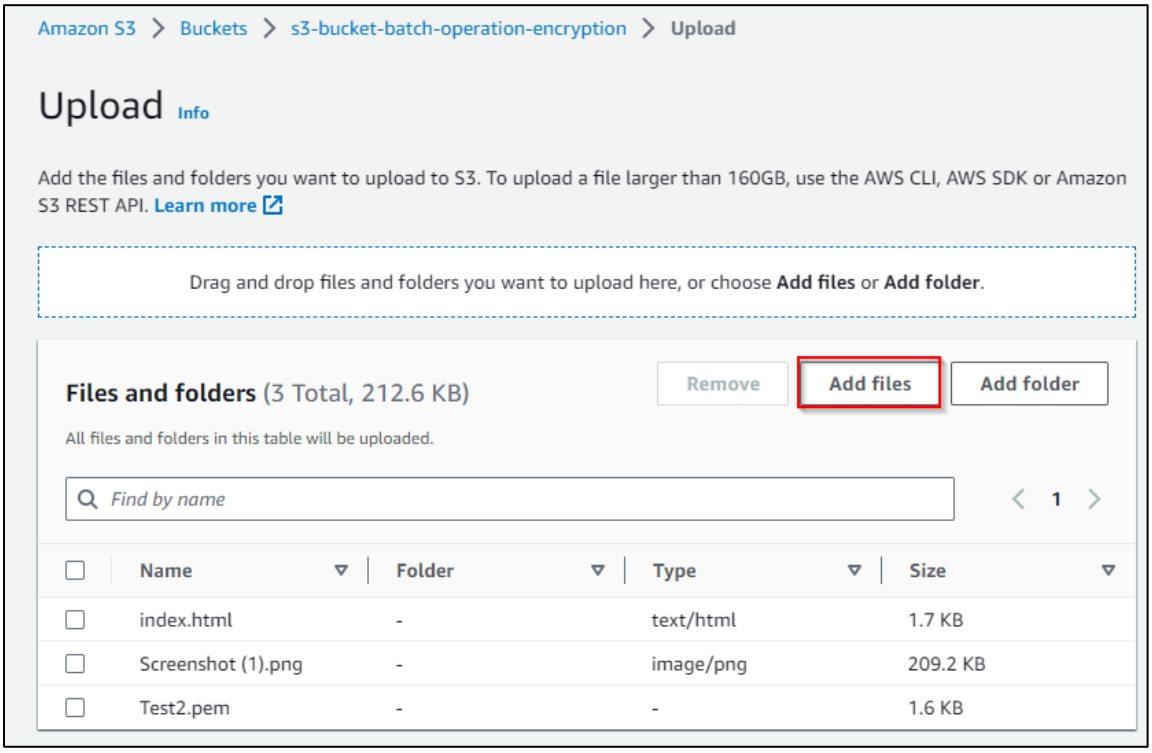


1.8 Click **Save changes**



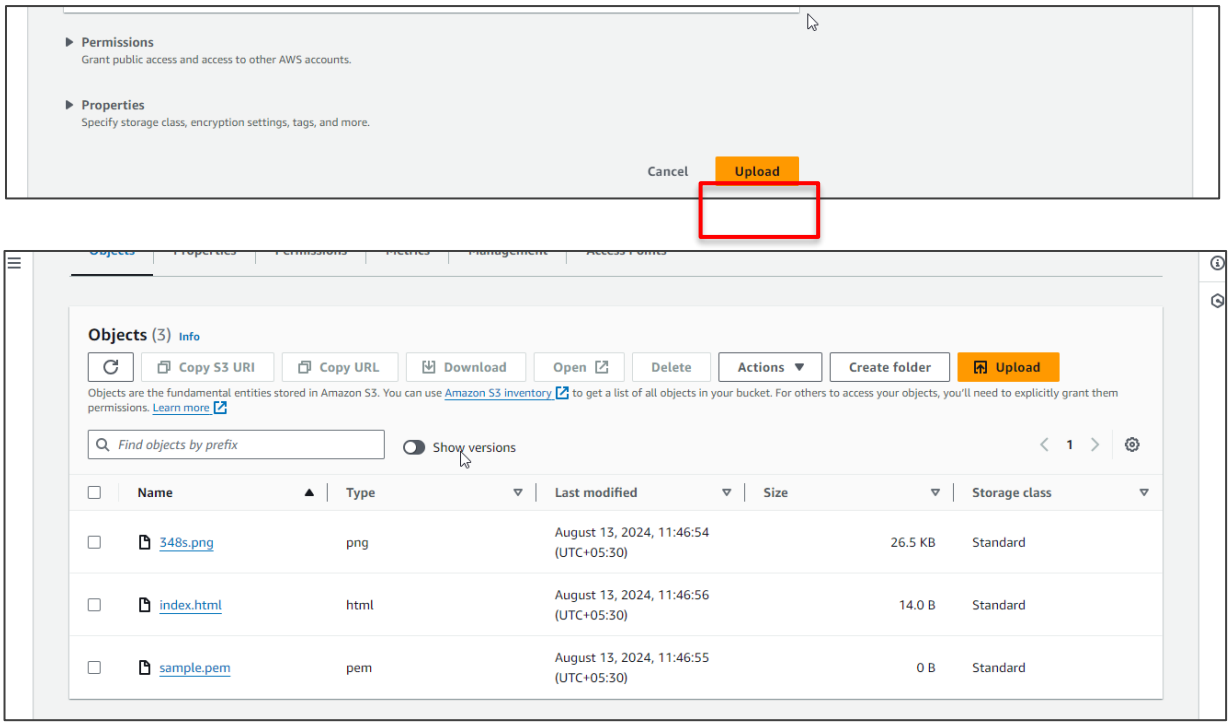
1.9 Add HTML, PNG, and PEM files to the bucket using the **Upload** option, and click **Add files**





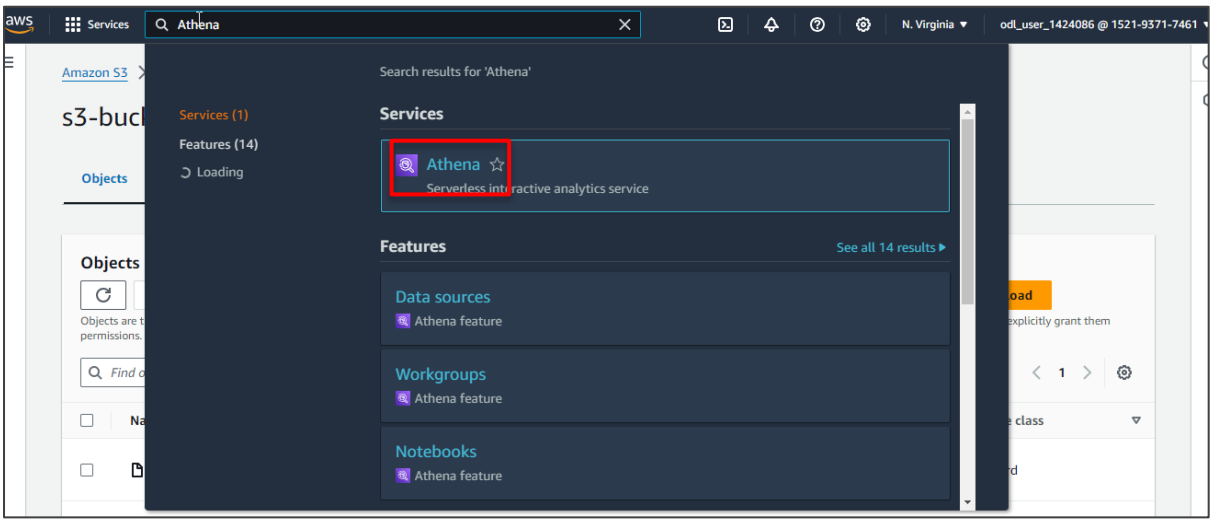
Note: These are sample files that can be created or sourced for demonstration purposes. For example, you can create a simple HTML file, use any image for the PNG file, and generate a PEM file using OpenSSL or any other tool.

1.10 Click on **Upload**

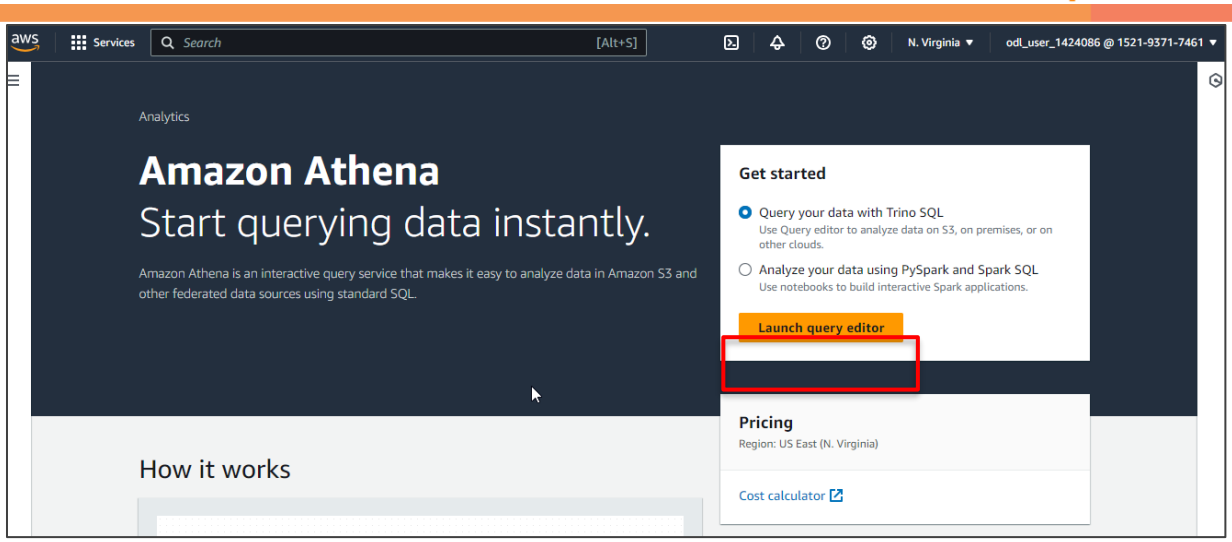


Step 2: Execute queries in Athena

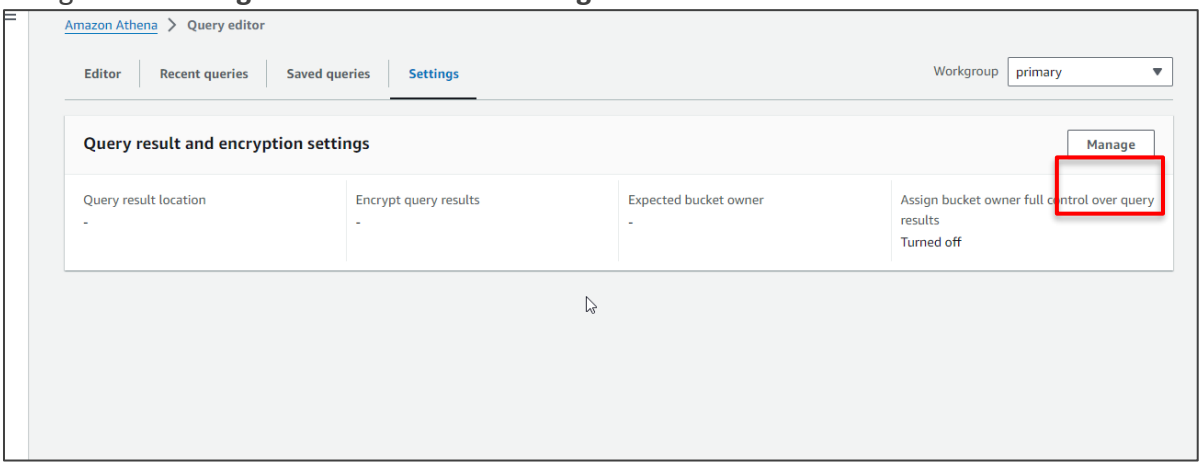
2.1 Navigate to the AWS Management Console and search for **Athena**



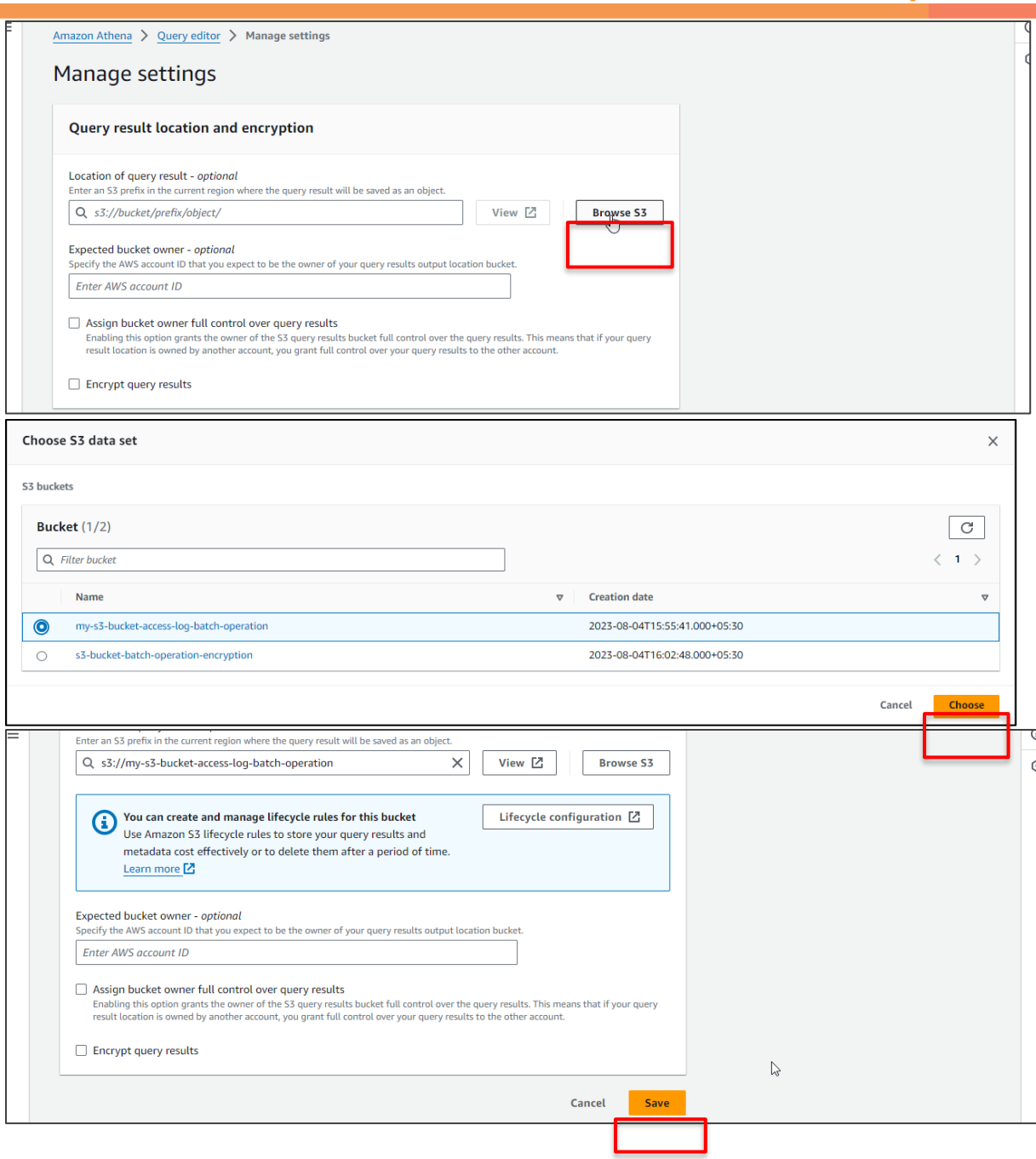
2.2 Click on **Launch query editor** to launch the query editor in Athena



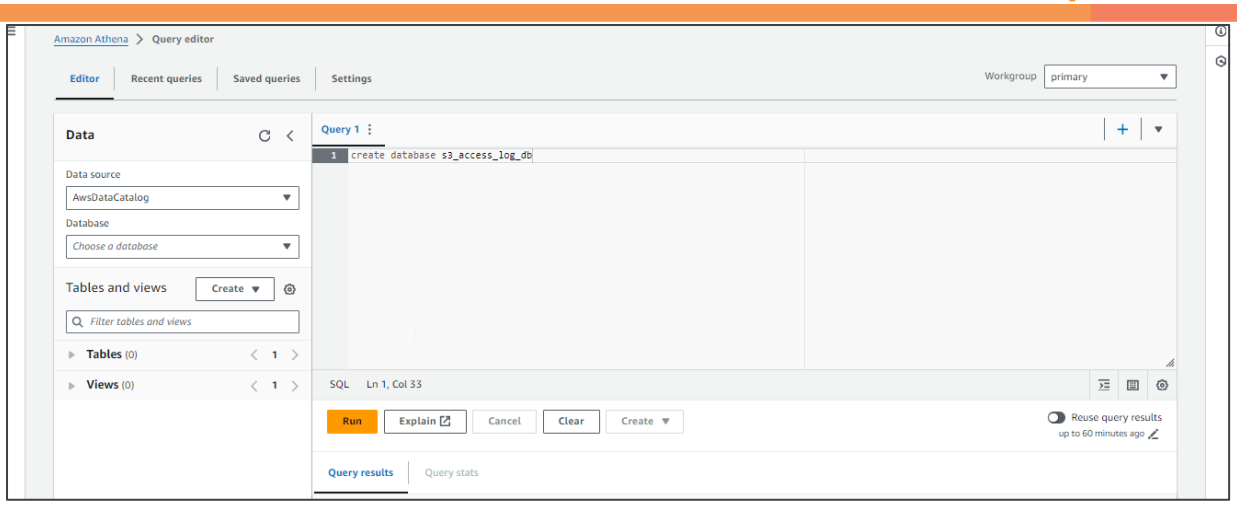
2.3 Navigate to **Settings** and then click on **Manage**



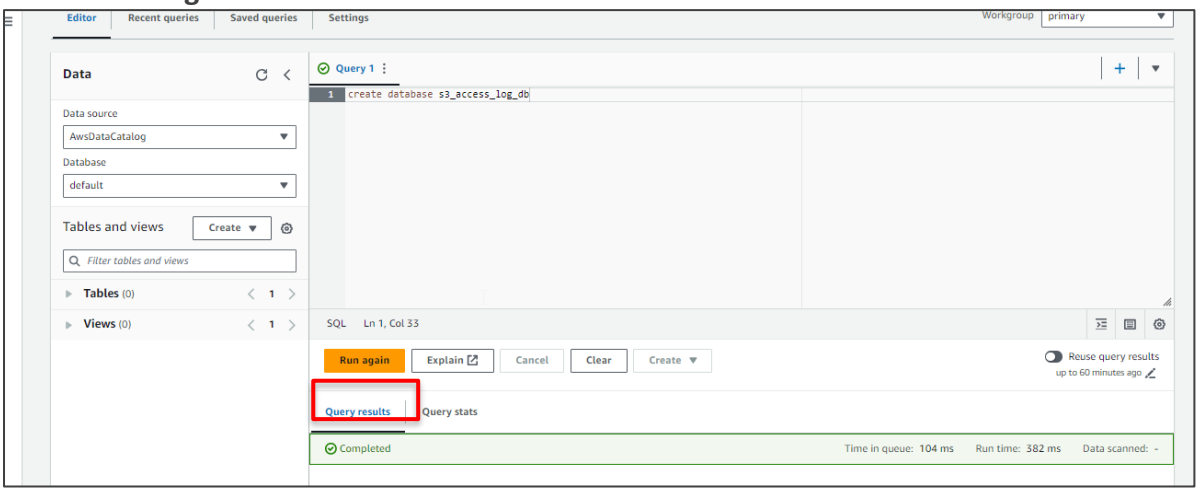
2.4 Click on **Browse S3**, select one of the path locations, and then click **Save**



2.5 In the editor, execute the query **create database s3_access_log_db;**



2.6 Click on **Run again**

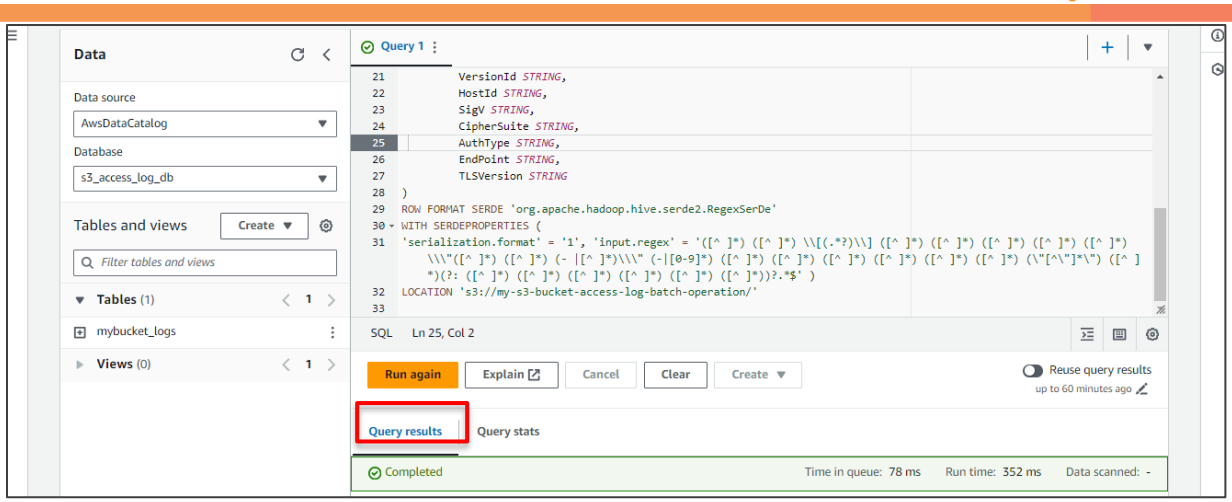


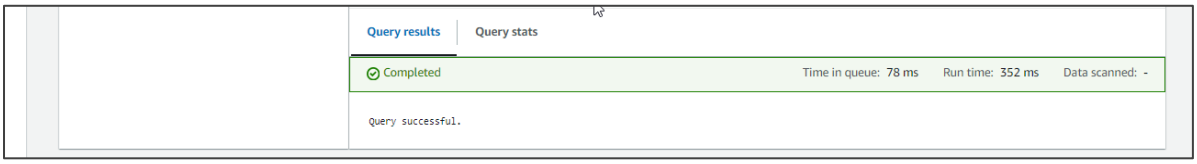
2.7 Execute the given query to create an external table:

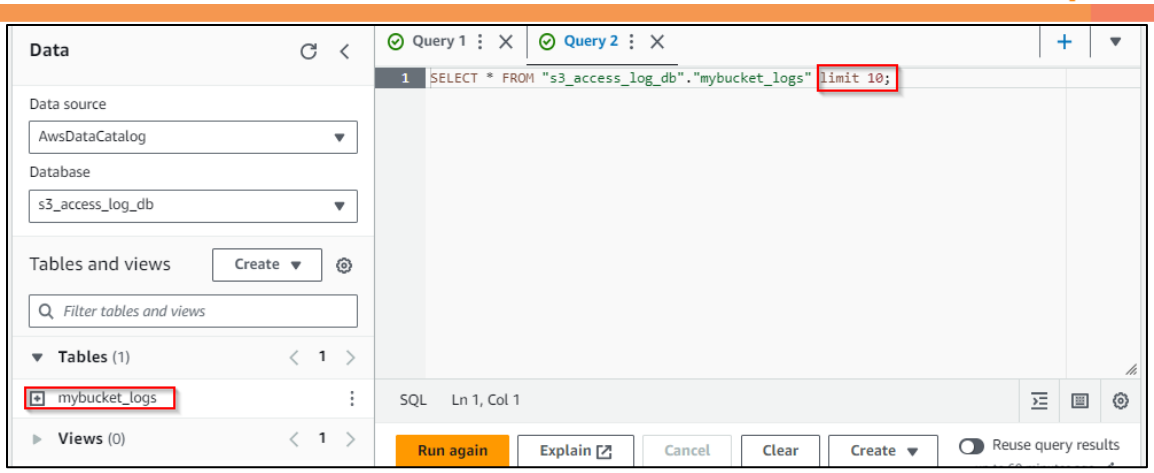
CREATE EXTERNAL TABLE IF NOT EXISTS s3_access_log_db.mybucket_logs(

```
  BucketOwner STRING,
  Bucket STRING,
  RequestDateTime STRING,
  RemoteIP STRING,
  Requester STRING,
  RequestID STRING,
  Operation STRING,
  Key STRING,
  RequestURI_operation STRING,
  RequestURI_key STRING,
  RequestURI_httpProtoversion STRING,
  HTTPstatus STRING,
  ErrorCode STRING,
  BytesSent BIGINT,
  ObjectSize BIGINT,
  TotalTime STRING,
  TurnAroundTime STRING,
  Referrer STRING,
  UserAgent STRING,
  VersionId STRING,
  HostId STRING,
  SigV STRING,
  CipherSuite STRING,
  AuthType STRING,
  EndPoint STRING,
  TLSVersion STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
  'serialization.format' = '1', 'input.regex' = '([^ ]*) ([^ ]*) \\[(.*?)\\] ([^ ]*) ([^ ]*) ([^ ]*)
([^ ]*) ([^ ]*) \\"([^ ]*) ([^ ]*) (- |[^ ]*)\\" (-|[0-9]*) ([^ ]*) ([^ ]*) ([^ ]*) ([^ ]*) ([^ ]*)
([^ ]*) (\"[^\"]*\") ([^ ]*)(?: ([^ ]*) ([^ ]*) ([^ ]*) ([^ ]*) ([^ ]*) ([^ ]*))?.*$' )
LOCATION 's3://my-s3-bucket-access-log-batch-operation/'
```

2.8 Execute the code by selecting **s3_access_log_db**, and then click **Run again**


Data
Data source
AwsDataCatalog
Database
s3_access_log_db
Tables and views
Create
Filter tables and views
Tables (1)
mybucket_logs
Views (0)
Query 1
VersionId STRING,
HostId STRING,
SigV STRING,
CipherSuite STRING,
AuthType STRING,
EndPoint STRING,
TLSVersion STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
'serialization.format' = '1', 'input.regex' = '([^]*) ([^]*) \\[(.*?)\\] ([^]*) ([^]*) ([^]*) ([^]*) ([^]*) ([^]*) ([^]*) (\"[^\"]*\") ([^]*)(?: ([^]*) ([^]*) ([^]*) ([^]*) ([^]*) ([^]*))?.*$')
LOCATION 's3://my-s3-bucket-access-log-batch-operation/'
SQL
Ln 25, Col 2
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 78 ms
Run time: 352 ms
Data scanned: -

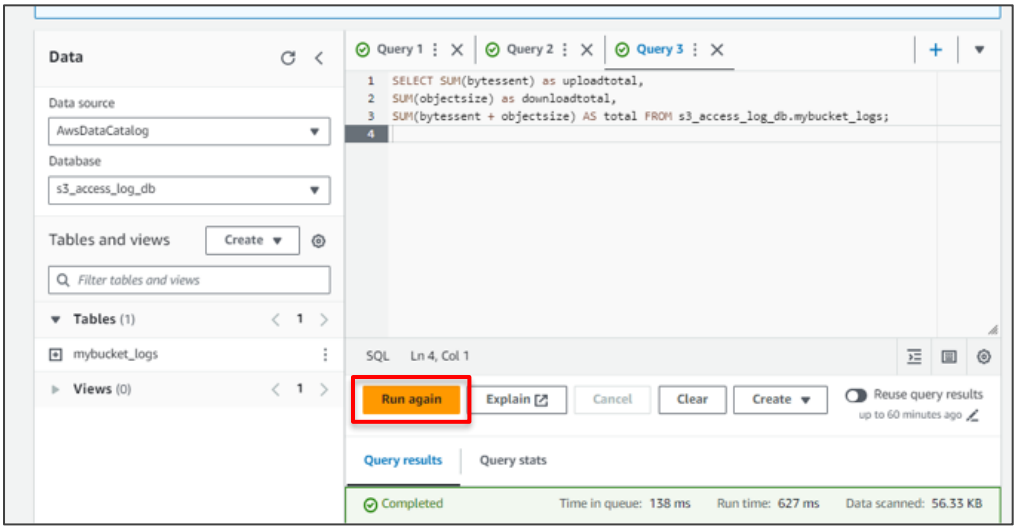

Query results
Query stats
Completed
Time in queue: 78 ms
Run time: 352 ms
Data scanned: -
Query successful.



Note: A predefined query with a limit of 10 rows will be available.

2.9 Execute the given query to find the exact amount of data uploaded to and downloaded from the monitored bucket:

SELECT SUM(bytessent) as uploadtotal,
SUM(objectsize) as downloadtotal,
SUM(bytessent + objectsize) AS total FROM s3_access_log_db.mybucket_logs;



By following these steps, you can now execute queries in AWS Athena, allowing you to analyze operations on a designated S3 bucket and gain insights into data usage patterns.