Guided Lab: Amazon Athena Data Querying and Table Creation

Description

Welcome to our guided lab, where we'll dive into the world of data analytics using Amazon Athena, AWS S3, and the AWS Glue Data Catalog. This hands-on session is designed to introduce you to the power and flexibility of analyzing structured and semi-structured data stored in S3, utilizing the serverless interactive query service provided by Athena. Whether you're a data analyst, engineer, or just a data enthusiast, this lab will equip you with the knowledge to efficiently query and analyze data at scale.

Overview of Steps

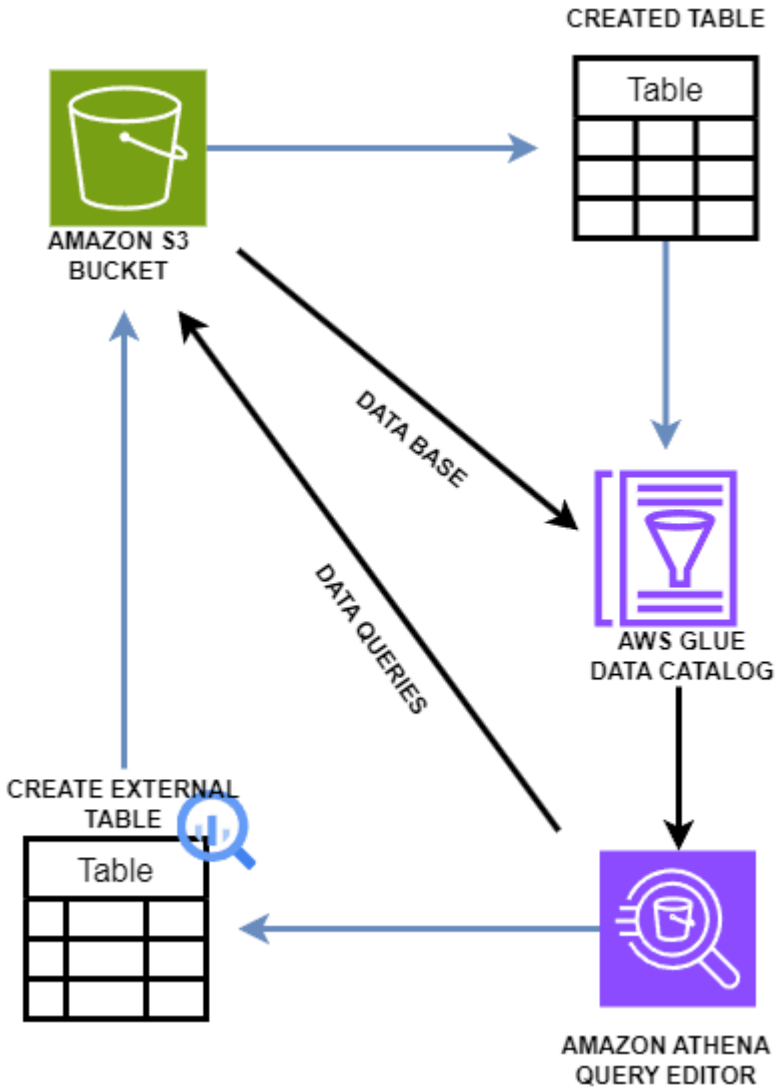
In this lab, you'll go through a series of steps that will take you from setting up your environment to executing queries that will provide insights into your data. Here's a brief outline of what we'll cover:

1. **Environment Setup**: You'll start by creating an S3 bucket and uploading an activity log file, setting the stage for our analysis.

2. **AWS Glue Data Catalog Configuration**: Next, we'll create a database using the AWS Glue Data Catalog. This serves as a central metadata repository that Athena will leverage to understand the structure of your data.

3. **Table Creation in Athena**: With the metadata repository in place, you'll learn how to define and create a table in Athena that maps to the structure of your activity log data stored in S3. This is a critical step that enables Athena to execute SQL queries against your data.

4. **Data Querying**: Armed with a structured view of your data, you'll run several SQL queries in Athena. These queries will range from basic data retrieval to more complex aggregation queries, designed to familiarize you with Athena's querying capabilities and help you derive meaningful insights from your dataset.



Prerequisites

This lab assumes you have experience creating Amazon S3 Bucket and are familiar with its basic components.

If you find any gaps in your knowledge, consider taking the following labs:

- Creating an Amazon S3 bucket.
- Querying Data with Amazon Athena and AWS Glue Crawler Integration.

Objectives

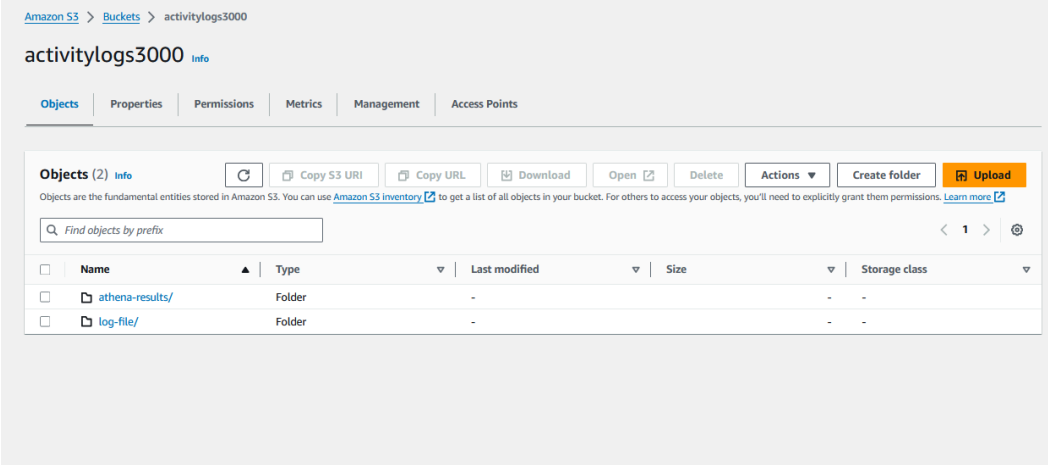
In this lab, you will:

- Learn how to query data directly from S3 using Amazon Athena.
- Learn how to create an external table using Amazon Athena query.
- Use AWS Glue to create a data catalog (database) for organizing data from Amazon S3.

Lab Steps

Prepare Your Environment Prepare Your Environment

1. **Create an S3 Bucket**:
    - Give your bucket a unique name,
    - Leave the rest of the settings as default and click **Create bucket**
2. **Create folders in your S3 Bucket**:
    - Navigate to your newly created bucket.
    - Create two folders
        - **athena-results**
        - **log-file**



3. **Download this sample activity log file:**

https://media.tutorialsdojo.com/public/ActivityLog

4. **Upload the file in the /log-file of the s3 bucket folder you created**

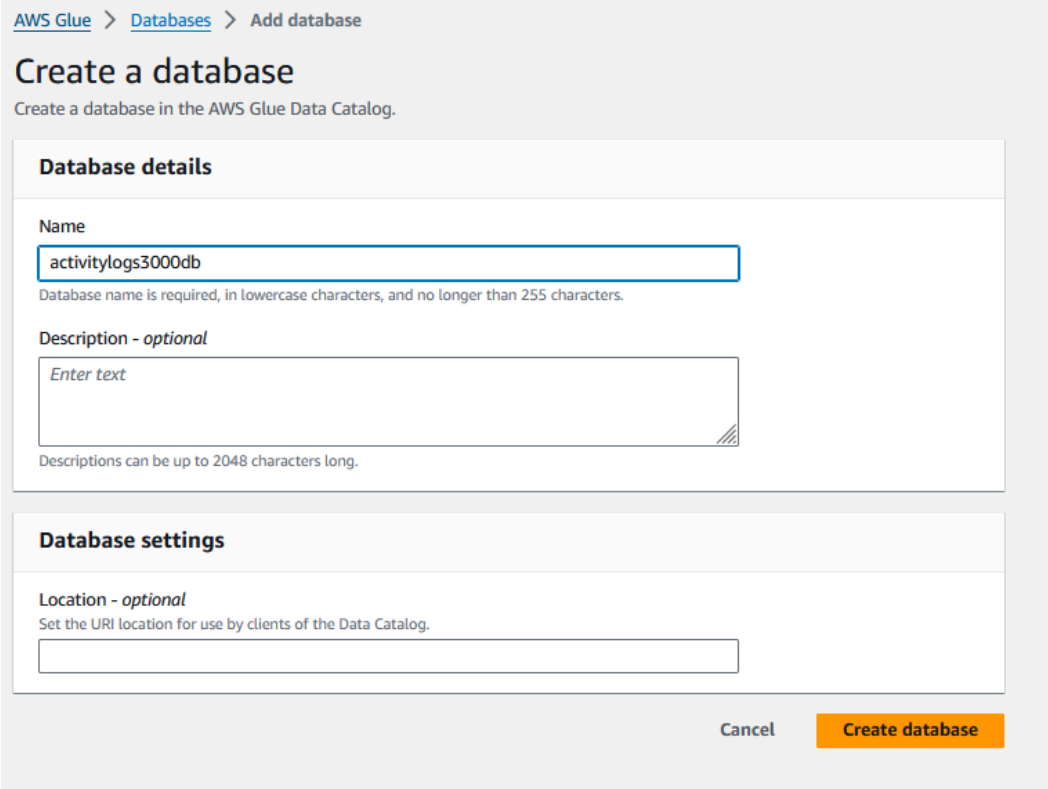
Set Up AWS Glue Data Catalog

1. **Navigate to AWS Glue**: In the AWS Management Console, find and select AWS Glue.

2. **Create a Database**:

    - In the Glue Console, select **"Databases"** from the left-hand menu, then click **"Add database"**.

    - Name your database uniquely, like **activitylogs3000db**, and provide a description if desired.



- Click Create

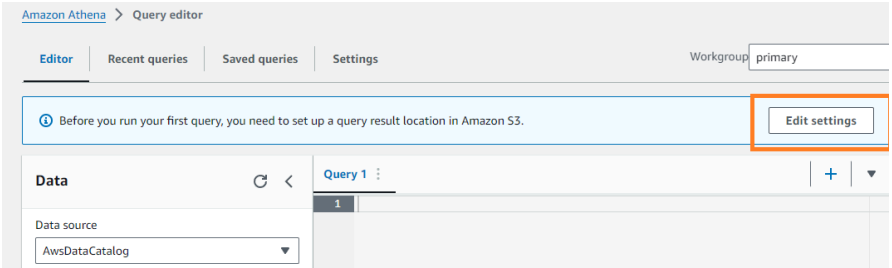
Create a Table in Athena Using the Glue Data Catalog

1. **Open Athena**: Navigate to the Athena service in the AWS Console.

2. **Set Up Query Result Location in S3**:

   - In Athena, you'll see a query editor. Before executing any queries, you need to specify an S3 bucket location where query results will be saved.

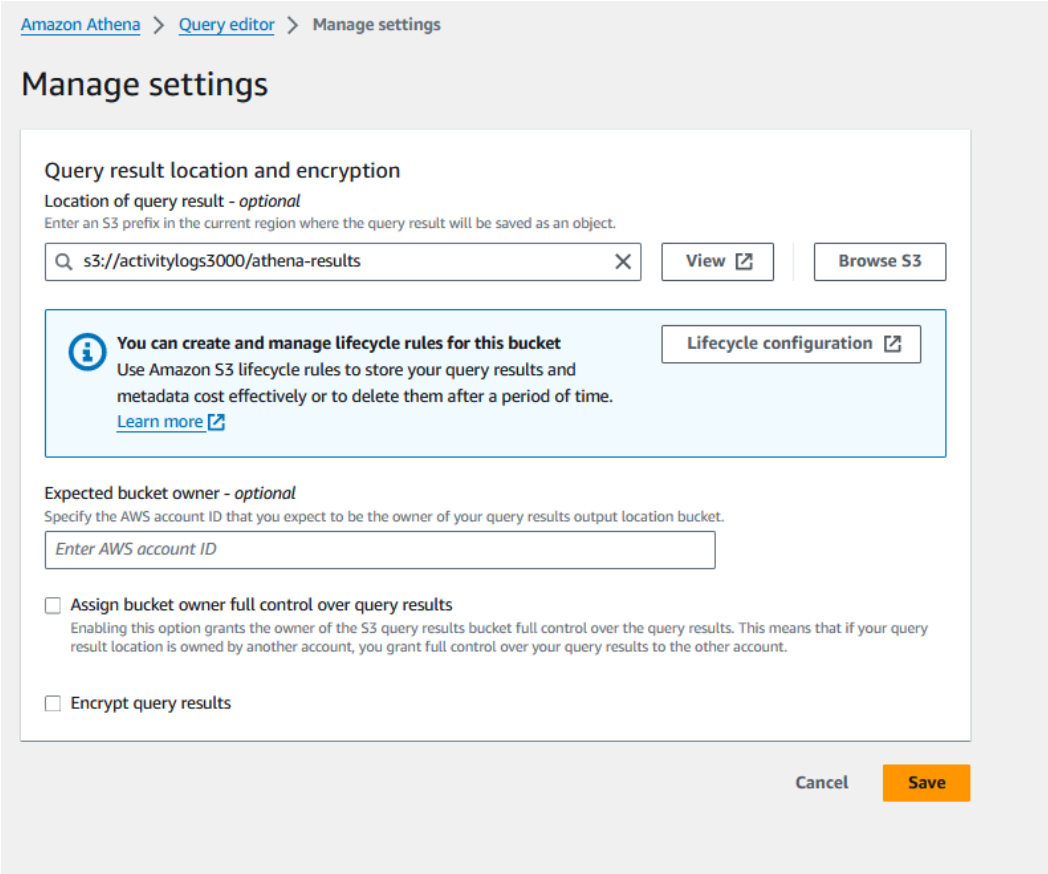
     - Click on the **Edit setting**



- Click in **Manage**

- Browse and select on the S3 Bucket folder athena-results or enter in the **Location of query result**

```
s3://your-bucket-name/athena-results/
```



- Click on **Save**

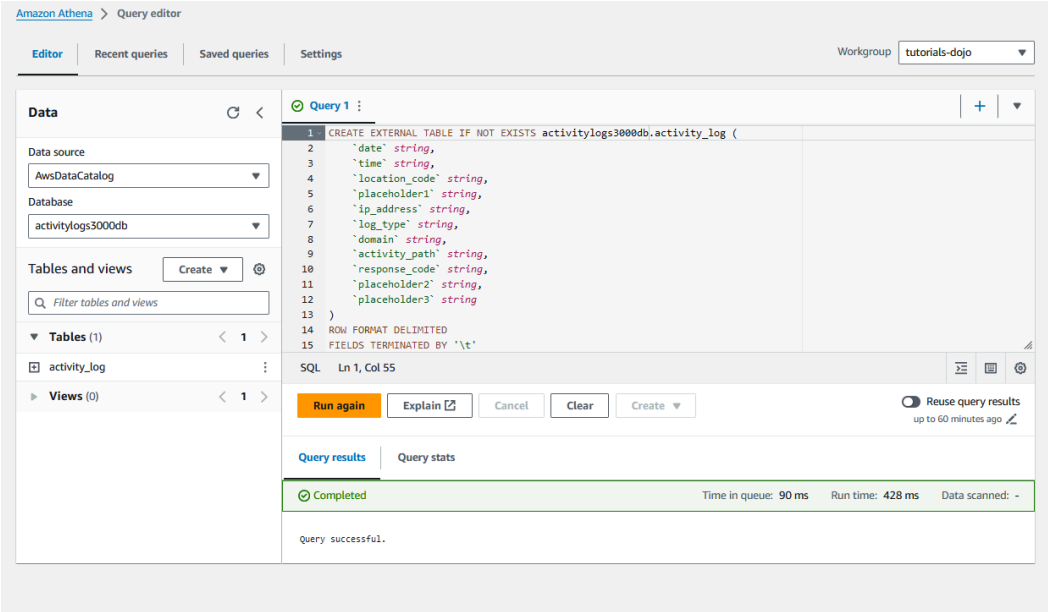
3. **Create Table**:

    - Navigate back to the **Editor**'s tab, you'll create a table that references your activity log in S3. Use a DDL statement like the following, adjusting paths, table names, and column definitions as needed:

```
CREATE EXTERNAL TABLE IF NOT EXISTS
activitylogs3000db.activity_log (
     `date` date,
     `time` string,
     `location_code` string,
     `placeholder1` string,
     `ip_address` string,
     `log_type` string,
     `domain` string,
     `activity_path` string,
     `response_code` bigint,
     `placeholder2` string,
     `placeholder3` string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
STORED AS TEXTFILE
LOCATION 's3://your-bucket-name/log-file/';
```

REMEMBER TO CHANGE THE PLACEHOLDER `your-bucket-name` *IN THE* `s3://your-bucket-name/log-file/`

- Click **Run**



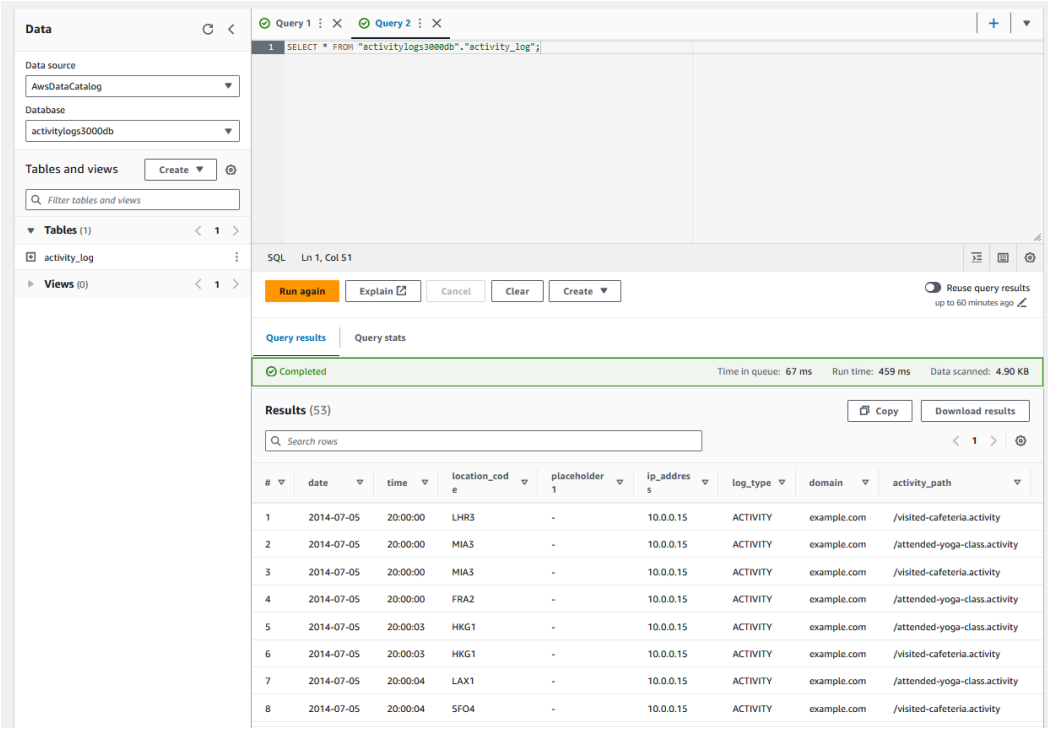
Query Your Data

Now, you can **clear** the SQL Editor by clicking **Clear** bottom of the SQL editor or create a new SQL editor tab by click the **+ (plus) sign** in the upper right of the SQL editor to add and run simple queries to analyze your activity

log. Here are a few to get you started:

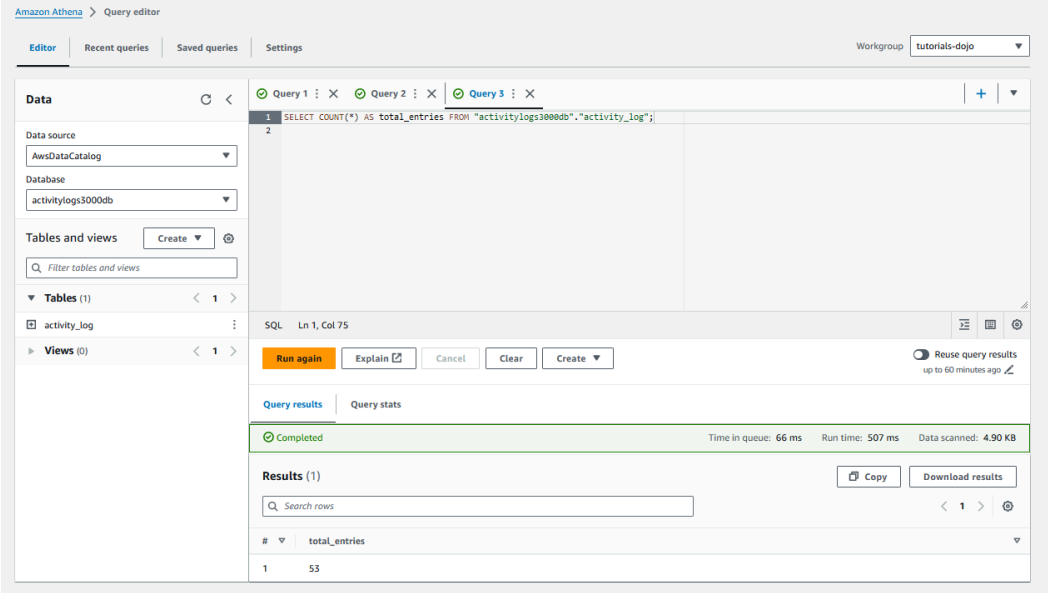
1. **To Select Everything from the Table**
   This query retrieves all the data from your activity log table. It's a good starting point to see the entire dataset you're working with.

```
SELECT * FROM "activitylogs3000db"."activity_log";
```



2. **To Count Total Entries**
   To understand the scale of your data, you can count the total number of entries in your activity log.

```
SELECT COUNT(*) AS total_entries FROM
"activitylogs3000db"."activity_log";
```



3. **Activity Frequency**

To see which activities are most common, you can count the number of occurrences for each `activity_path`.

```
SELECT activity_path, COUNT(*) AS frequency
FROM "activitylogs3000db"."activity_log"
GROUP BY activity_path
ORDER BY frequency DESC;
```

4. **Activities Over Time**

If you're interested in how activity changes over time, you could count the number of log entries per day.

```
SELECT date, COUNT(*) AS daily_activity_count
FROM "activitylogs3000db"."activity_log"
GROUP BY date
ORDER BY date;
```

Congratulations! You have learned how to prepare your environment, create a database and table for your data, and run queries to extract valuable insights. This knowledge will serve as a solid foundation for your future data analytics projects, empowering you to leverage AWS's scalable and serverless data analytics services efficiently. Whether for personal projects or professional tasks, the skills acquired today will enable you to tackle data analysis challenges with confidence and expertise.

One last thing! It is a good practice to clean up the resources created during this lab. Not only will it make you a better professional, but you will also become a more organized person. Happy learning!