Guided Lab: Querying Data with Amazon Athena and AWS Glue Crawler Integration

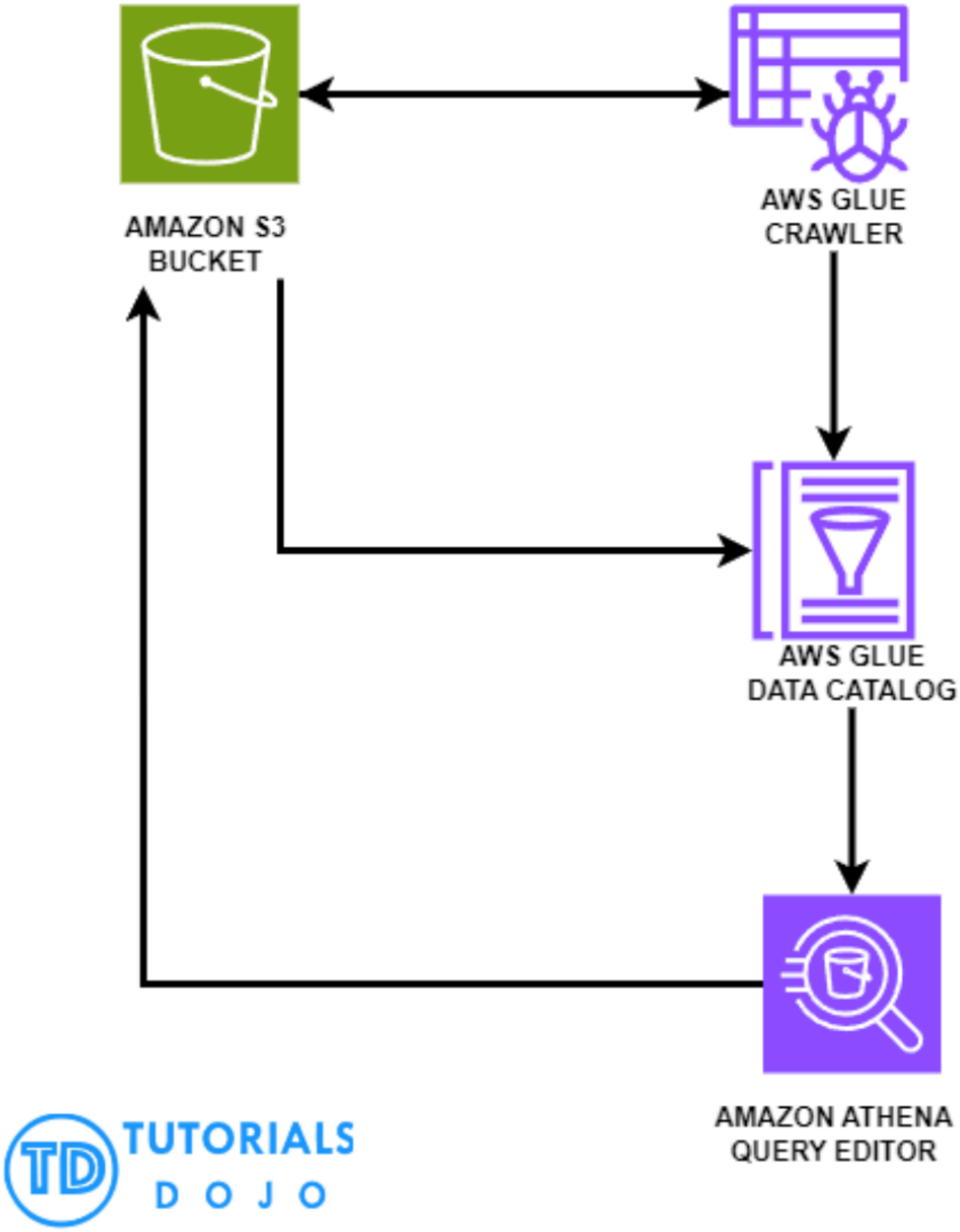
Description

Data analytics has become an indispensable part of business strategy and decision-making. Amazon Web Services (AWS) provides a suite of scalable and flexible services designed for data analytics. Among these services, Amazon S3, Athena, and Glue (for data cataloging and data crawling) stand out for their ability to store massive datasets, query data directly in place, and organize data across various data stores efficiently.

Overview of Steps:

1. **Setting Up Amazon S3 Bucket**: Your data needs a place to reside. Amazon S3 serves as the foundation, providing a secure, scalable, and durable storage solution. Here, you'll store the raw data files that Athena will query.
2. **Creating a Database in AWS Glue Data Catalog**: Think of the database as a container or namespace within which you'll organize your data. It doesn't store data itself but acts as a logical grouping mechanism for your tables, which represent different datasets or aspects of your data.
3. **Adding Tables to the Database**: Tables define the schema or structure of your data (such as columns and data types) and point to the actual data stored in S3. This step is crucial because it tells Athena how to interpret the raw data during queries. You can create tables manually by defining the schema or automatically using crawlers that scan your data in S3 and infer the schema. **In this lab, we will create tables using Glue Crawler.**
4. **Querying Data with Amazon Athena**: With your data in S3, a database to organize your tables, and tables to define your data schema, you're now ready to use Athena to run SQL queries directly

against your data. Athena's serverless nature means you don't manage any infrastructure, focusing solely on analyzing your data.



Prerequisites

This lab assumes you have experience creating an 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.

Objectives

In this lab, you will:

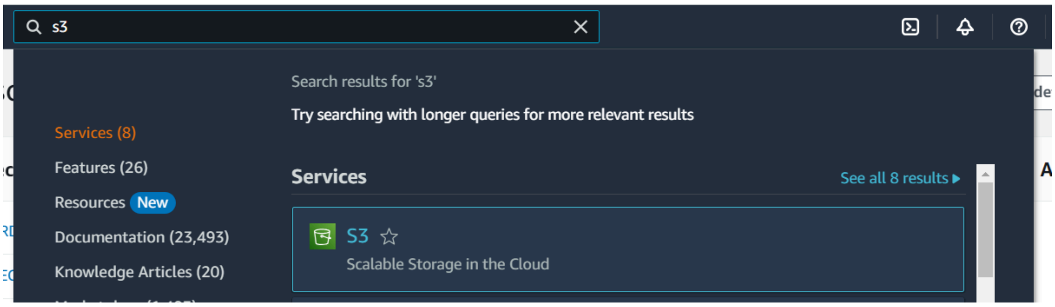
- Learn how to query data directly from S3 using Amazon Athena.
- Use AWS Glue to create a data catalog (database and tables) for organizing data from Amazon S3.

Subscribe to access AWS PlayCloud Labs

Lab Steps

Setting Up Amazon S3 Bucket

1. Log into AWS Management Console and navigate to the S3 service.

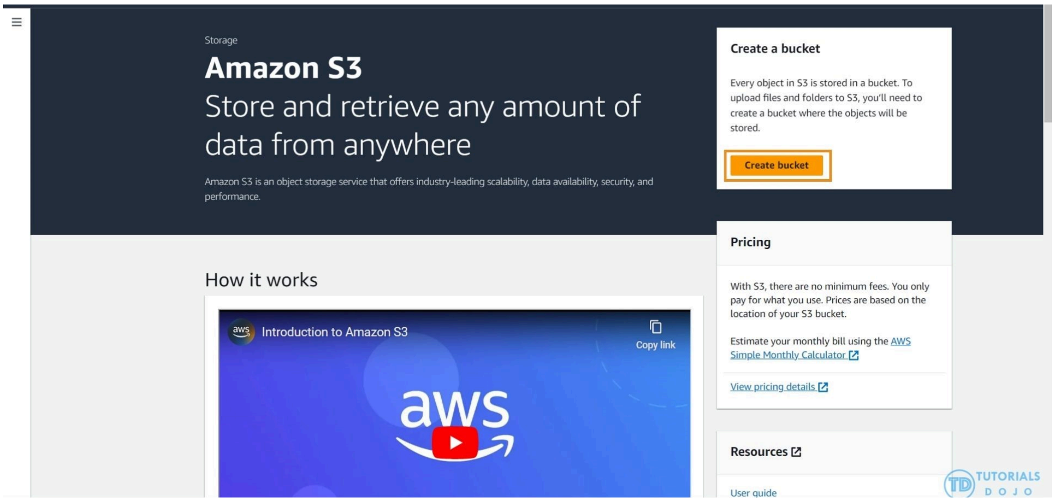


Search results for 's3'

Try searching with longer queries for more relevant results

Scalable Storage in the Cloud

2. Create a new bucket



Amazon S3

Store and retrieve any amount of data from anywhere

Amazon S3 is an object storage service that offers industry-leading scalability, data availability, security, and performance.

Create a bucket

Every object in S3 is stored in a bucket. To upload files and folders to S3, you'll need to create a bucket where the objects will be stored.

Pricing

With S3, there are no minimum fees. You only pay for what you use. Prices are based on the location of your S3 bucket.

Estimate your monthly bill using the AWS Simple Monthly Calculator

View pricing details

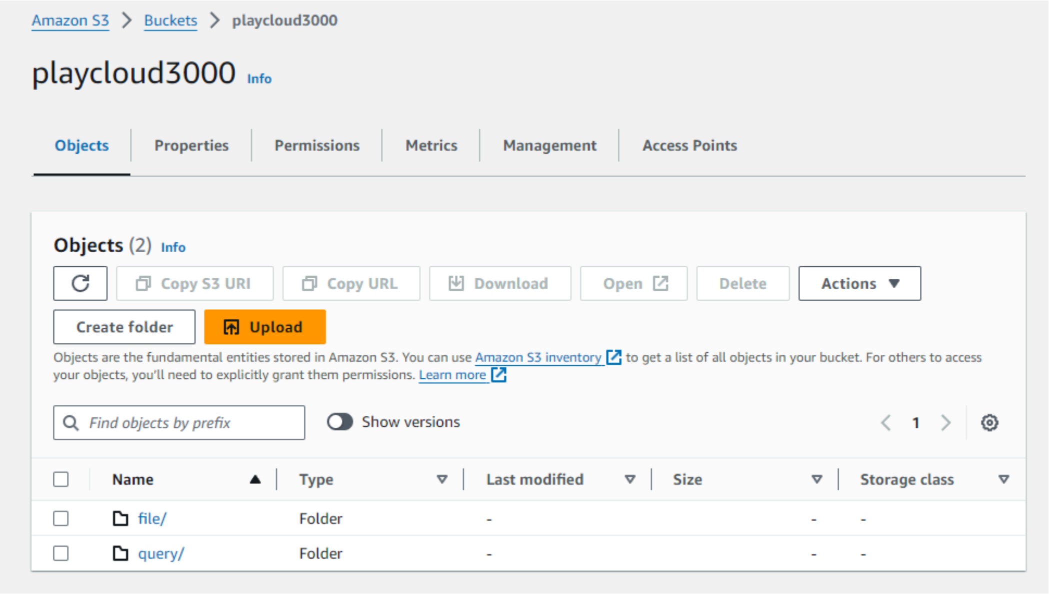
- Provide a unique name
- Leave everything as default settings.
- Click **Create Bucket**

3. Download this file for this lab

https://media.tutorialsdojo.com/public/Philippine_Tourist_Spots.csv

4. Create folders and Upload data files

- Create two folders and name them:
  - file
  - query



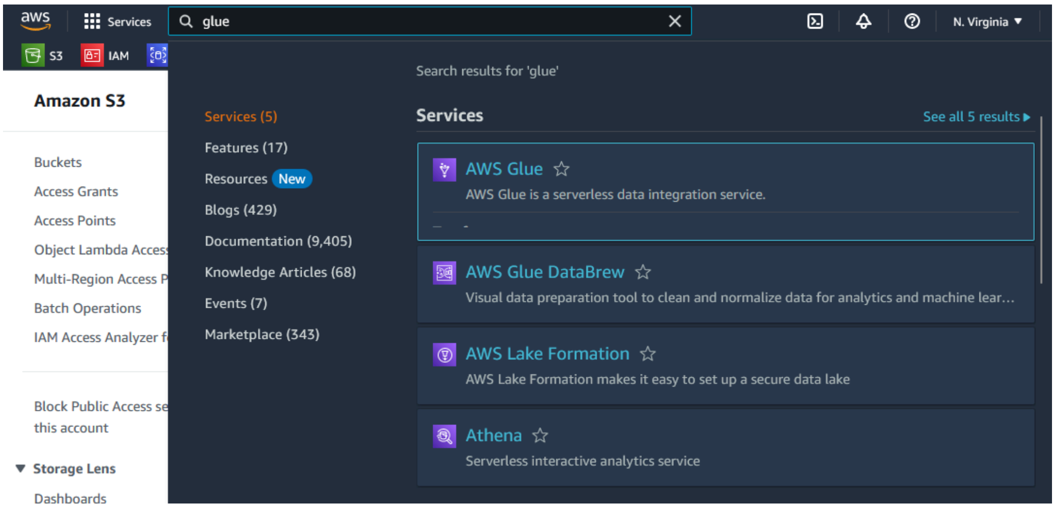
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	file/	Folder	-	-	-
□	query/	Folder	-	-	-

- Upload the file:
  - Open **file/** folder of your bucket.
  - Click on **Upload** and upload the file you downloaded previously.

Setting Up AWS Glue Data Catalog

1. Navigate to the AWS Glue service in the AWS Management Console.



AWS Glue is a serverless data integration service.

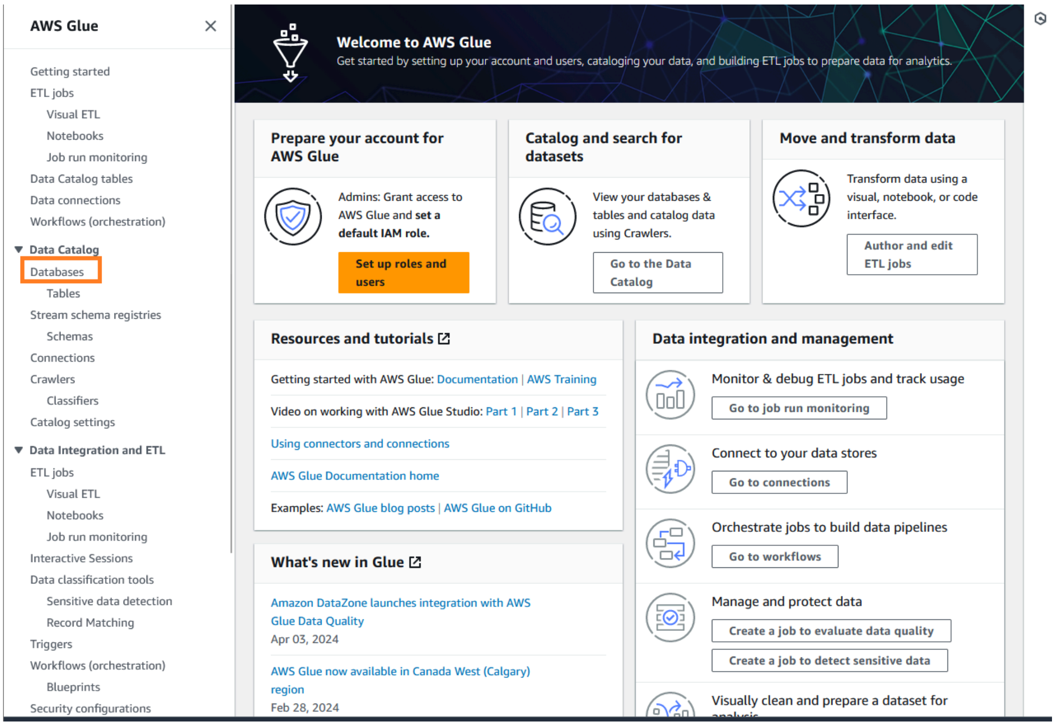
Visual data preparation tool to clean and normalize data for analytics and machine lear…

AWS Lake Formation makes it easy to set up a secure data lake

Serverless interactive analytics service

2. **Creating a database**

- To create a database, you need to
  - Click on **Databases** in the left corner of the window.



Get started by setting up your account and users, cataloging your data, and building ETL jobs to prepare data for analytics.

Admins: Grant access to AWS Glue and set a default IAM role.

View your databases & tables and catalog data using Crawlers.

Transform data using a visual, notebook, or code interface.

Getting started with AWS Glue: Documentation | AWS Training

Video on working with AWS Glue Studio: Part 1 | Part 2 | Part 3

Using connectors and connections

AWS Glue Documentation home

Examples: AWS Glue blog posts | AWS Glue on GitHub

Monitor & debug ETL jobs and track usage

Connect to your data stores

Orchestrate jobs to build data pipelines

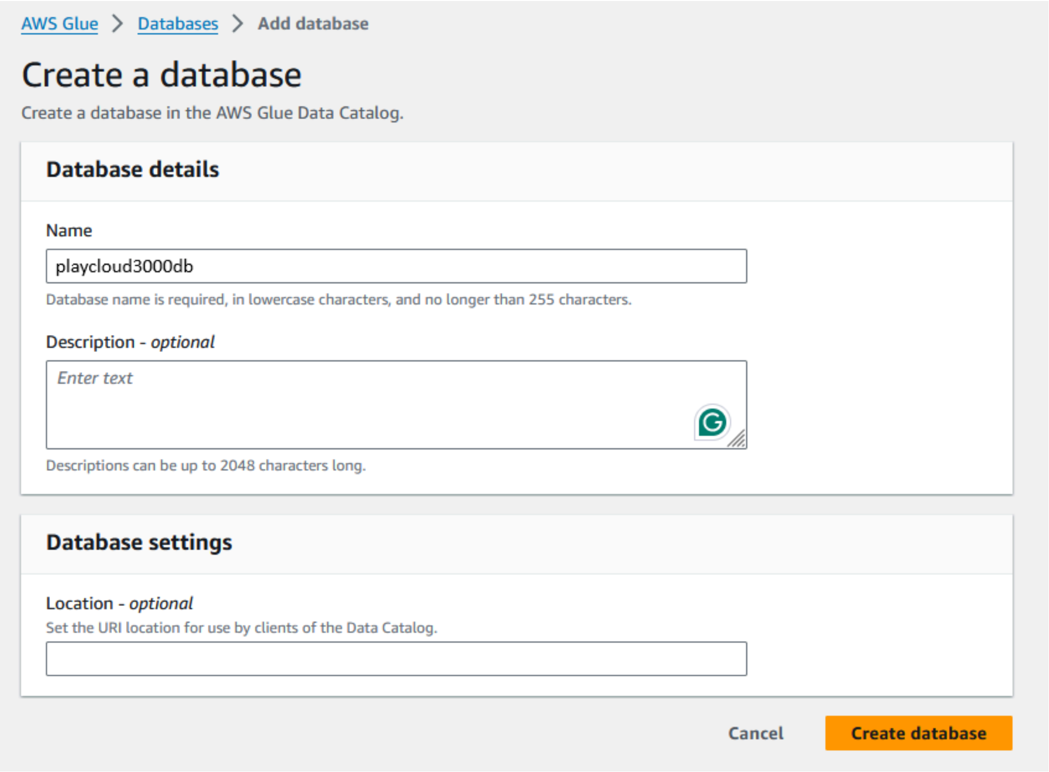
Manage and protect data

Visually clean and prepare a dataset for analysis

Amazon DataZone launches integration with AWS Glue Data Quality
Apr 03, 2024

AWS Glue now available in Canada West (Calgary) region
Feb 28, 2024

- Fill the **Name** with a unique database name and add a **Description** if desired.
- Click on **Create database**



AWS Glue > Databases > Add database

Create a database

Create a database in the AWS Glue Data Catalog.

Database details

Name

playcloud3000db

Database name is required, in lowercase characters, and no longer than 255 characters.

Description - *optional*

Enter text

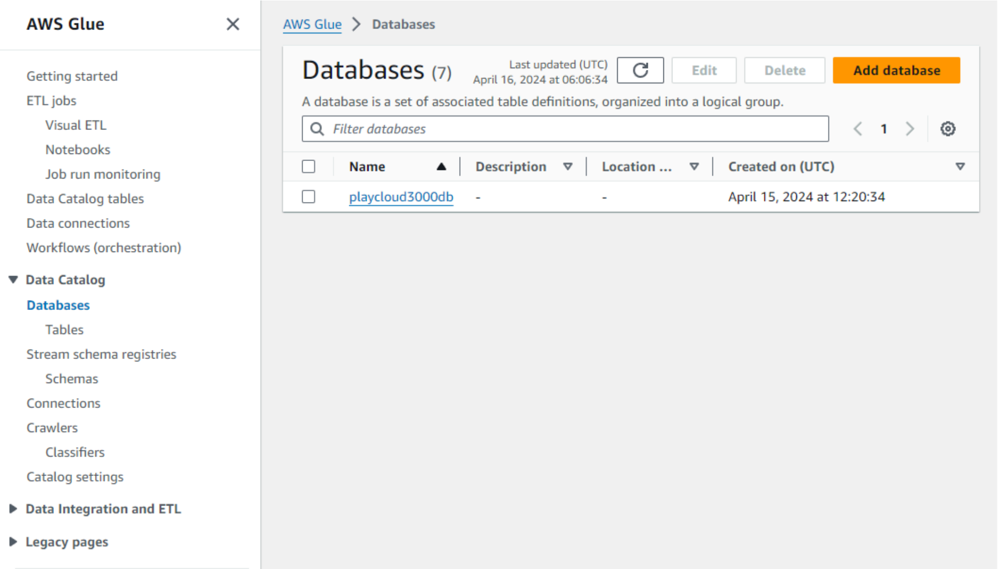
Descriptions can be up to 2048 characters long.

Database settings

Location - *optional*

Set the URI location for use by clients of the Data Catalog.

- You should see your newly created database afterward.



Databases (7)

Last updated (UTC)
April 16, 2024 at 06:06:34

A database is a set of associated table definitions, organized into a logical group.

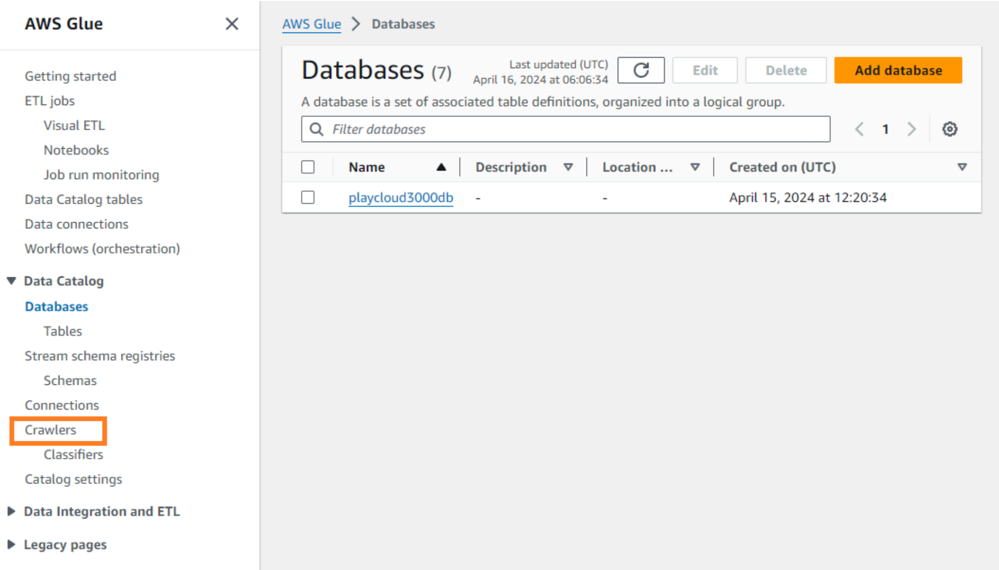
Filter databases

□	Name	Description	Location ...	Created on (UTC)
□	playcloud3000db	-	-	April 15, 2024 at 12:20:34

Setting Up AWS Glue Data Crawler

1. **Adding tables** by Glue Crawler.

- To create a Glue Crawler
  - Click on **Crawlers**



AWS Glue ✕

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

▶ Data Integration and ETL

▶ Legacy pages

AWS Glue > Databases

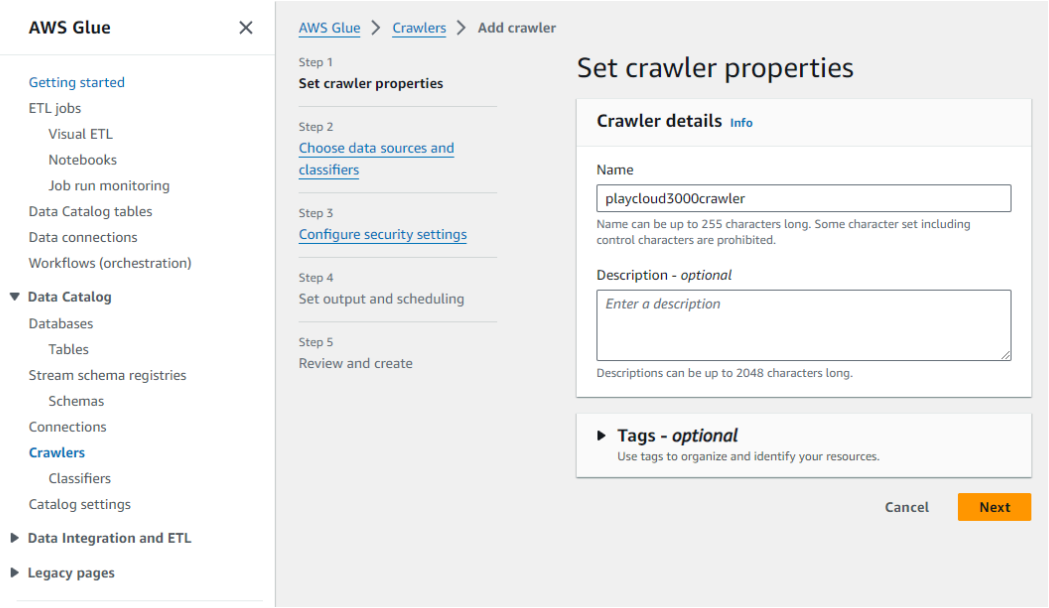
Databases (7) Last updated (UTC) April 16, 2024 at 06:06:34 | Edit | Delete | Add database

A database is a set of associated table definitions, organized into a logical group.

Filter databases

□	Name	Description	Location ...	Created on (UTC)
□	playcloud3000db	-	-	April 15, 2024 at 12:20:34

- Click on **Create crawler**
- Fill the **Name** for your crawler add description if desired.



AWS Glue ✕

Getting started
ETL jobs
Visual ETL
Notebooks
Job run monitoring
Data Catalog tables
Data connections
Workflows (orchestration)

▼ Data Catalog
Databases
Tables
Stream schema registries
Schemas
Connections
Crawlers
Classifiers
Catalog settings

▶ Data Integration and ETL

▶ Legacy pages

AWS Glue > Crawlers > Add crawler

Step 1
Set crawler properties

Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and create

Set crawler properties

Crawler details Info

Name

playcloud3000crawler

Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - *optional*

Enter a description

Descriptions can be up to 2048 characters long.

▶ Tags - *optional*

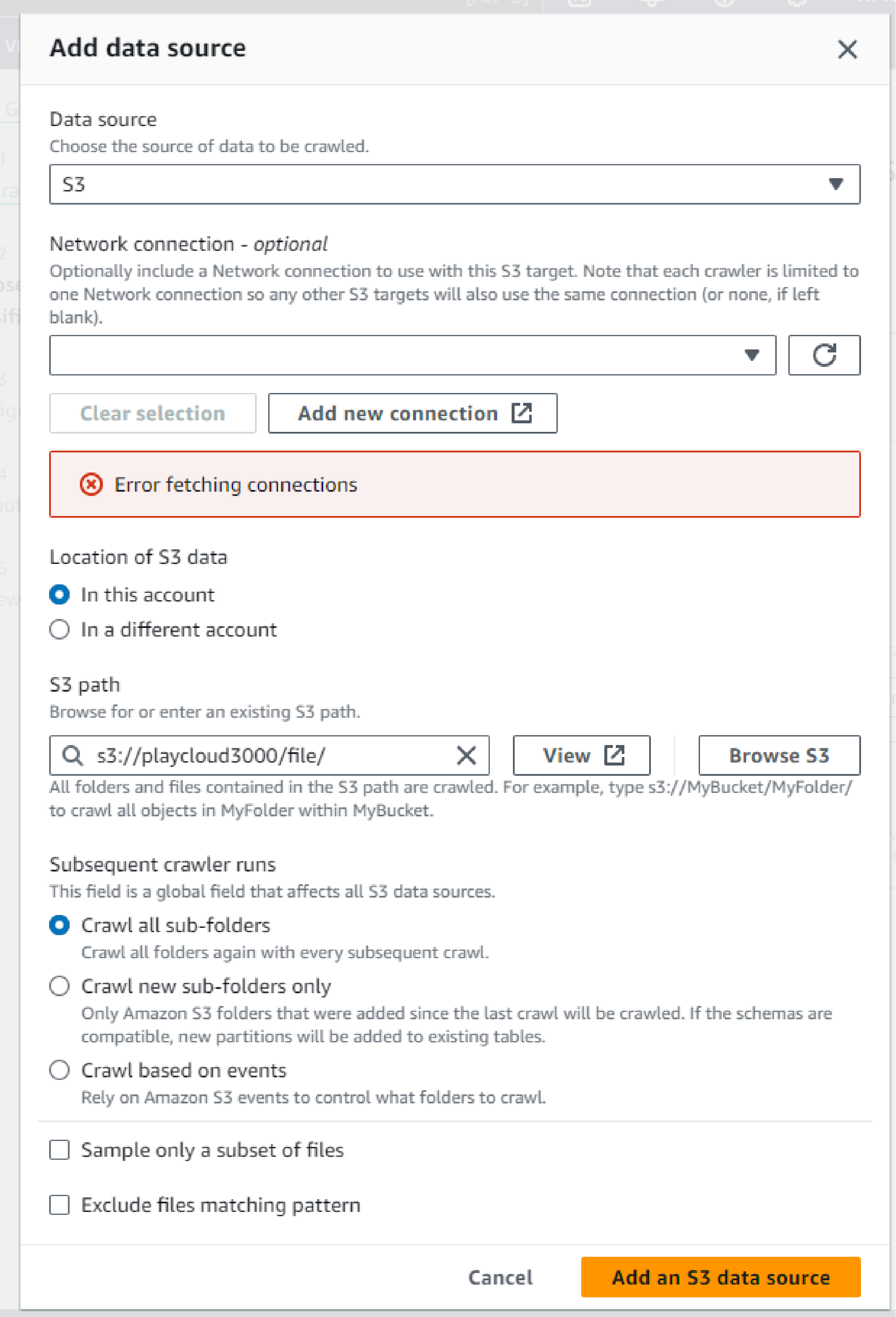
Use tags to organize and identify your resources.

Cancel | Next

- Click on **Next**
- Click on **Add a data source** under the Data source.
  - Add these details:
    - Data source: **S3**
    - S3 path:

```
s3://<name-of-your-s3-bucket>/file/
```

Remember to change the placeholder ***<name-of-your-s3-bucket>*** *with the name of your S3 bucket*



Add data source

Data source

Choose the source of data to be crawled.

S3

Network connection - *optional*

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Error fetching connections

Location of S3 data

● In this account

○ In a different account

S3 path

Browse for or enter an existing S3 path.

s3://playcloud3000/file/

All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Subsequent crawler runs

This field is a global field that affects all S3 data sources.

● Crawl all sub-folders

Crawl all folders again with every subsequent crawl.

○ Crawl new sub-folders only

Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.

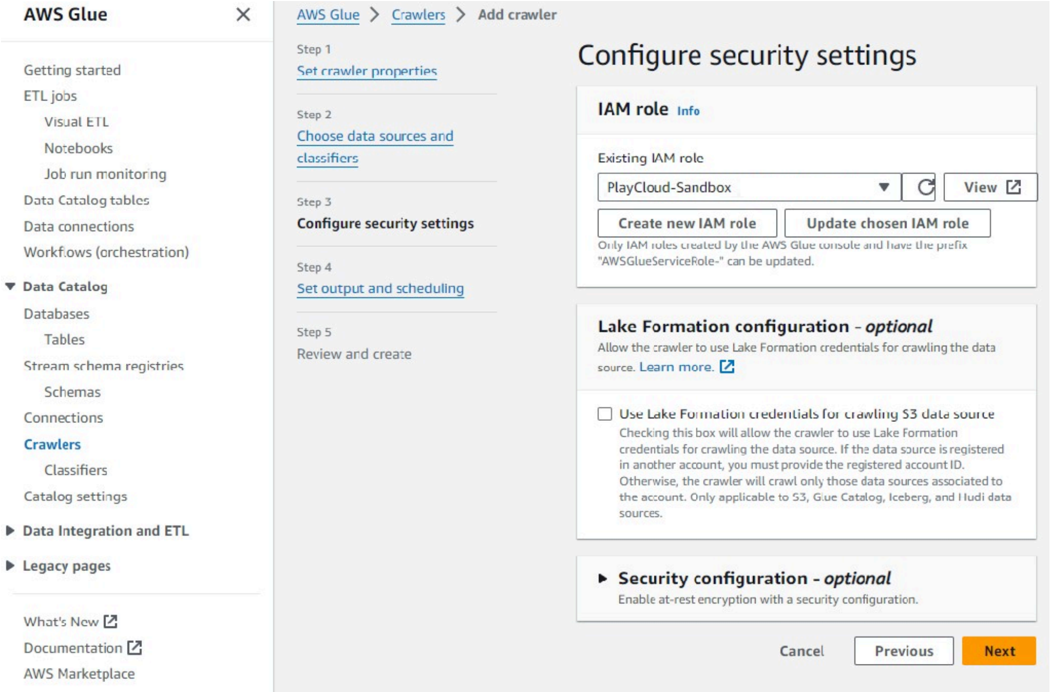
○ Crawl based on events

Rely on Amazon S3 events to control what folders to crawl.

□ Sample only a subset of files

□ Exclude files matching pattern

- Click in **Add an S3 data source**
- Click on **Next**
- Under IAM role, Select **PlayCLoud-Sandbox**



Configure security settings

IAM role Info

Existing IAM role: PlayCloud-Sandbox

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Lake Formation configuration - *optional*

Allow the crawler to use Lake Formation credentials for crawling the data source. Learn more.

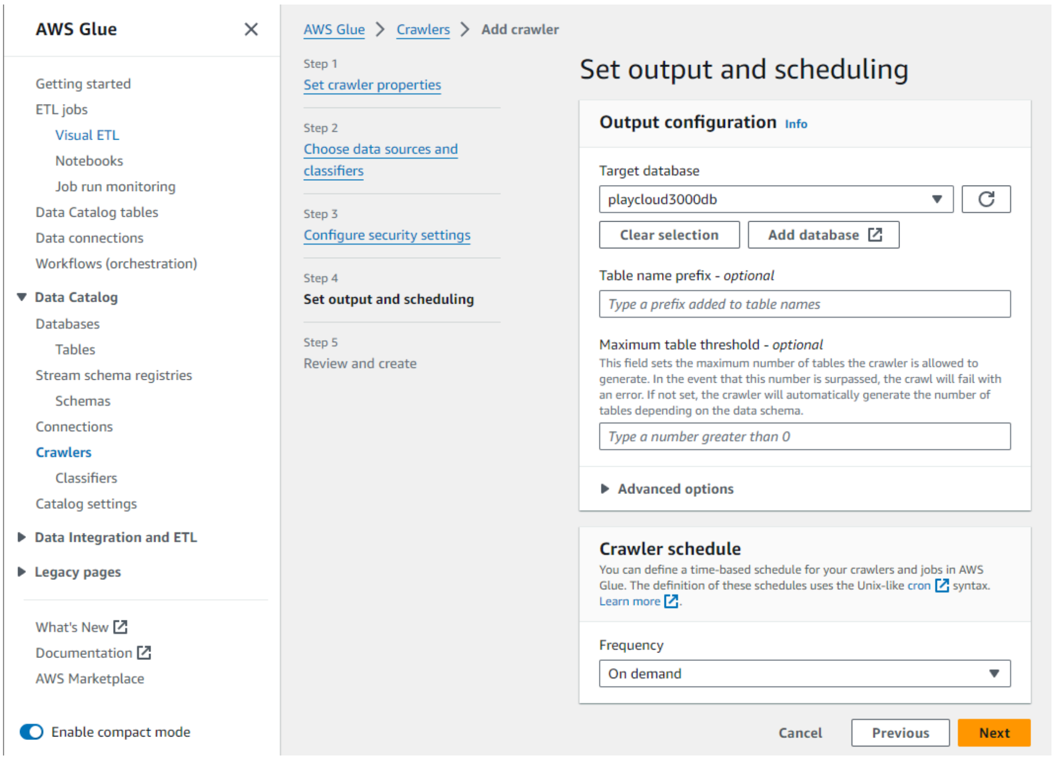
☐ Use Lake Formation credentials for crawling S3 data source

Checking this box will allow the crawler to use Lake Formation credentials for crawling the data source. If the data source is registered in another account, you must provide the registered account ID. Otherwise, the crawler will crawl only those data sources associated to the account. Only applicable to S3, Glue Catalog, Iceberg, and Hudi data sources.

Security configuration - *optional*

Enable at-rest encryption with a security configuration.

- Click on **Next**
- Under Target Database, Select the database you created.



Set output and scheduling

Output configuration Info

Target database: playcloud3000db

Table name prefix - *optional*: *Type a prefix added to table names*

Maximum table threshold - *optional*

This field sets the maximum number of tables the crawler is allowed to generate. In the event that this number is surpassed, the crawl will fail with an error. If not set, the crawler will automatically generate the number of tables depending on the data schema.

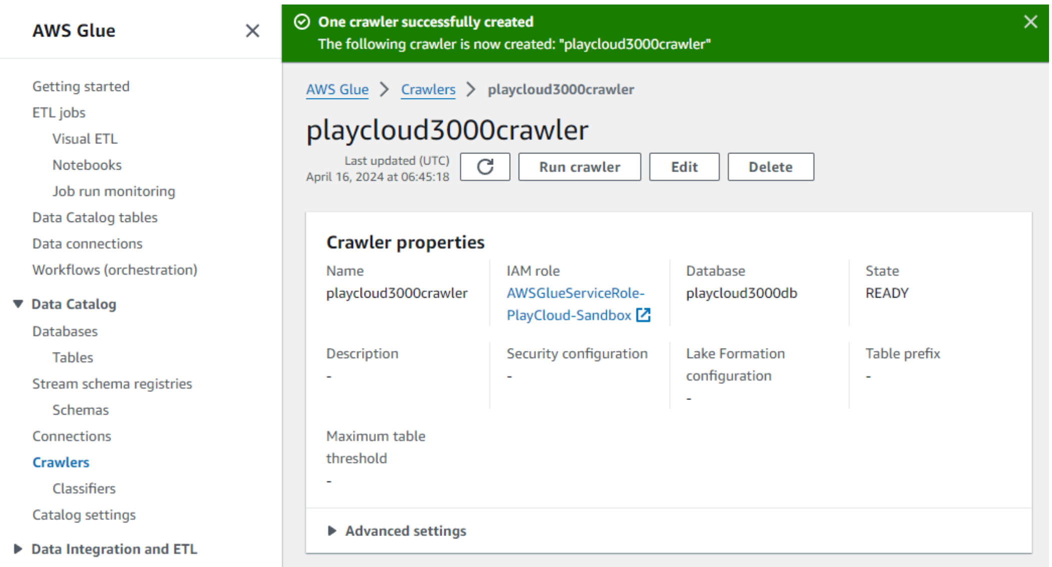
Type a number greater than 0

Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. Learn more.

Frequency: On demand

- Click on **Next**
- Review all the details under the **Review and create**
- Click on **Create crawler**
- You should be seeing a successful window and redirected to a window similar to the image below



One crawler successfully created
The following crawler is now created: "playcloud3000crawler"

playcloud3000crawler

Last updated (UTC)
April 16, 2024 at 06:45:18

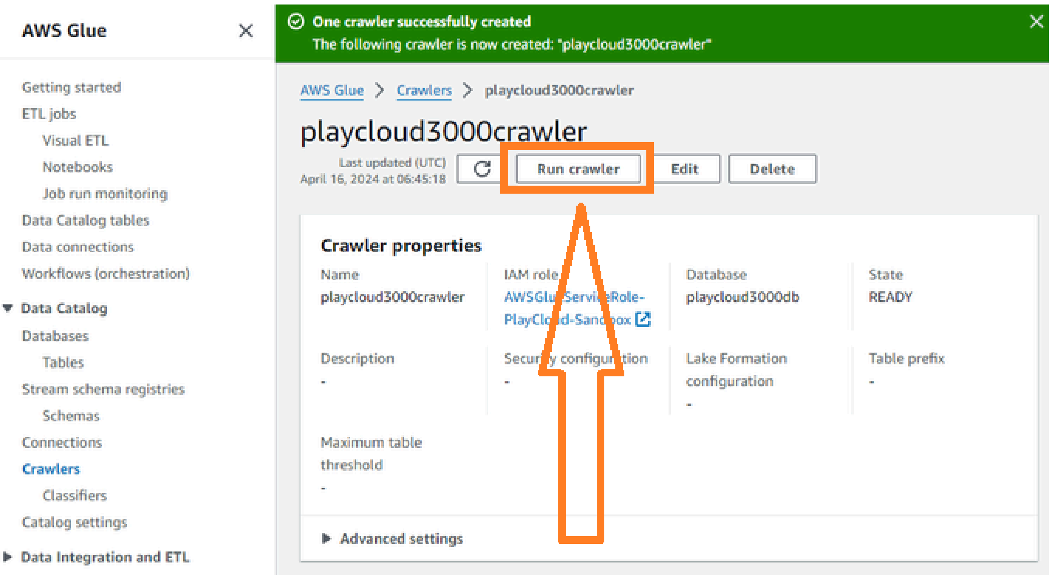
Crawler properties

Name	IAM role	Database	State
playcloud3000crawler	AWSGlueServiceRole-PlayCloud-Sandbox	playcloud3000db	READY
Description	**Security configuration**	**Lake Formation configuration**	**Table prefix**
-	-	-	-
Maximum table threshold			
-			

2. Run the Crawler

- After creating the crawler, Click on **Run Crawler**

  NOTE: It will take a few minutes for the crawler to crawl the AWS S3 Bucket



One crawler successfully created
The following crawler is now created: "playcloud3000crawler"

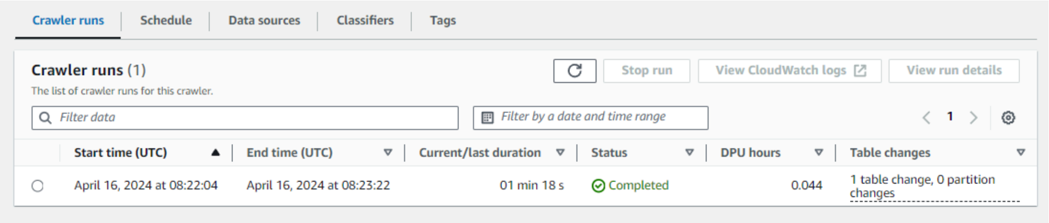
playcloud3000crawler

Last updated (UTC)
April 16, 2024 at 06:45:18

Crawler properties

Name	IAM role	Database	State
playcloud3000crawler	AWSGlueServiceRole-PlayCloud-Sandbox	playcloud3000db	READY
Description	**Security configuration**	**Lake Formation configuration**	**Table prefix**
-	-	-	-
Maximum table threshold			
-			

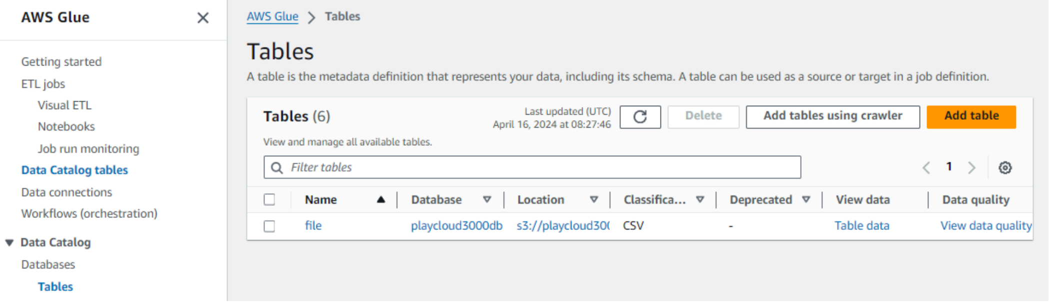
- When the crawler finishes crawling, you should be able to see similar image below



Crawler runs (1)
The list of crawler runs for this crawler.

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
○	April 16, 2024 at 08:22:04	April 16, 2024 at 08:23:22	01 min 18 s	Completed	0.044	1 table change, 0 partition changes

- In the **Data Catalog Tables**, you can see that a new table should be added.



Tables

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.

Tables (6)

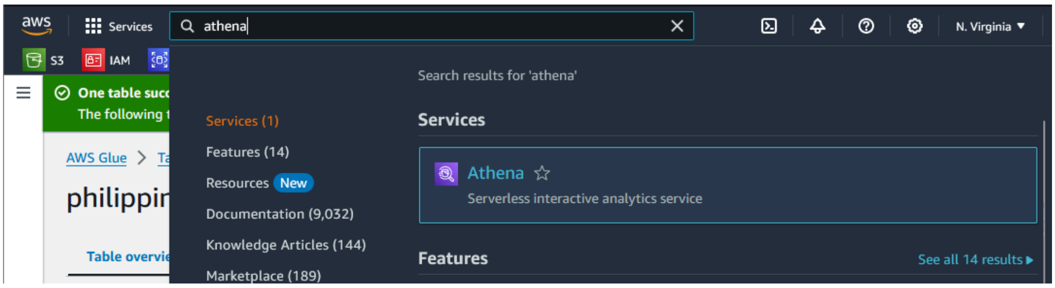
Last updated (UTC) April 16, 2024 at 08:27:46

View and manage all available tables.

Name	Database	Location	Classifica...	Deprecated	View data	Data quality
file	playcloud3000db	s3://playcloud30(	CSV	-	Table data	View data quality

Querying Data with Amazon Athena

1. Navigate to Amazon Athena in the AWS Management Console.



Search results for 'athena'

Services

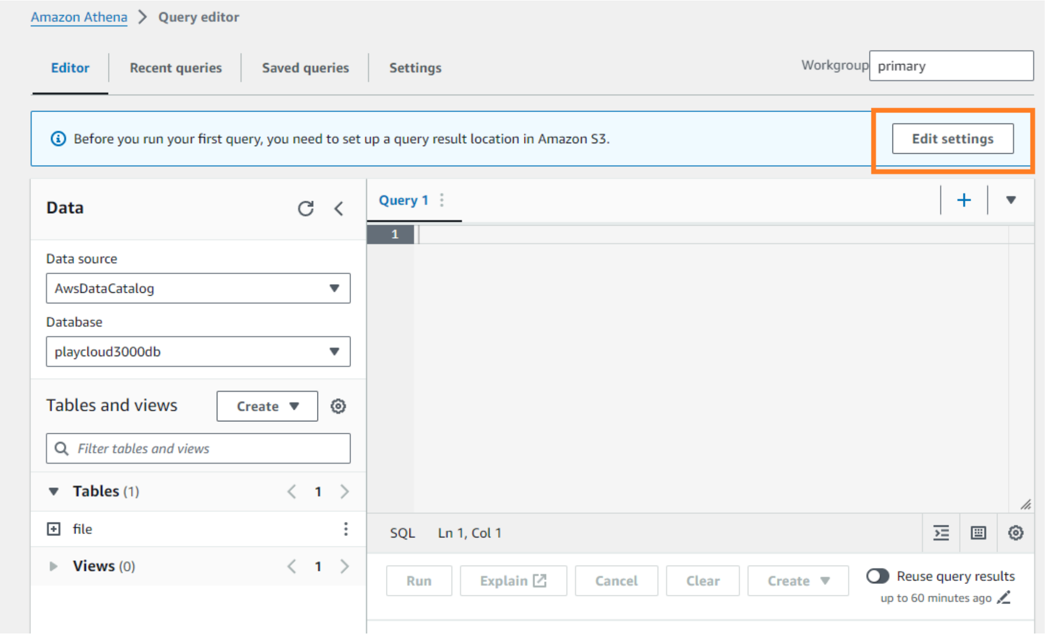
Athena

Serverless interactive analytics service

Features

2. **Set up a query location** in Athena settings to specify an S3 bucket for storing query results.

- Click on **Edit settings**



Workgroup: primary

Before you run your first query, you need to set up a query result location in Amazon S3.

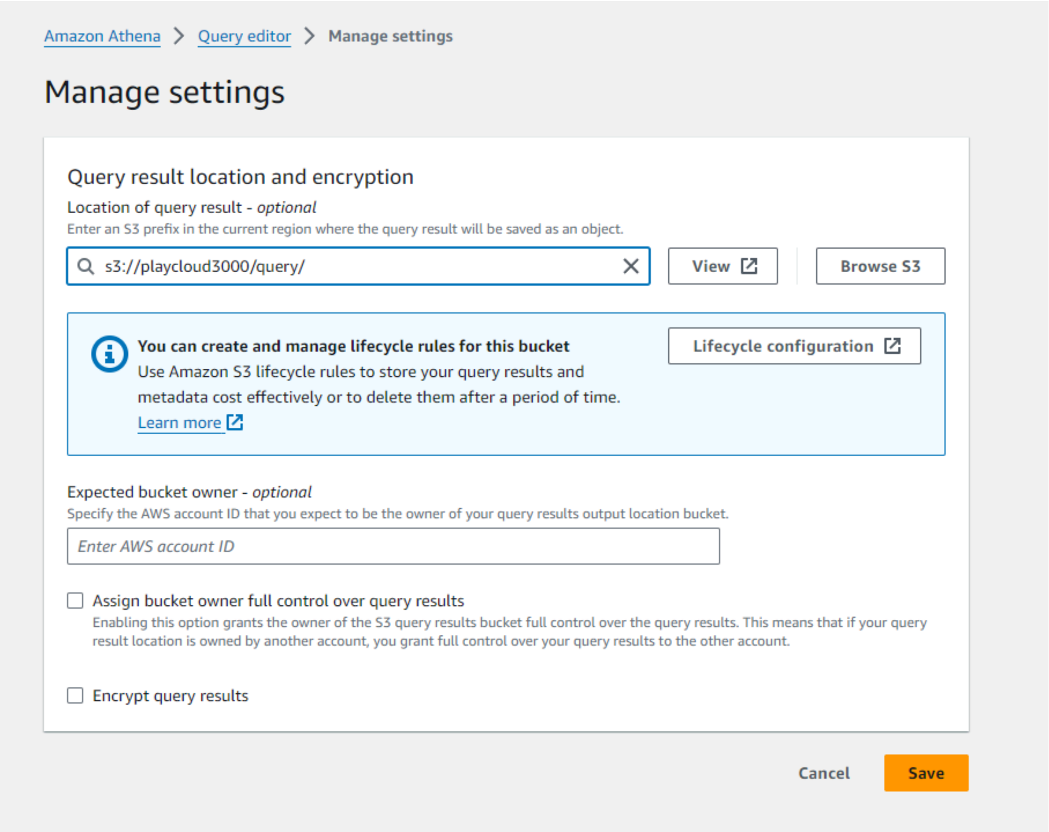
Data source: AwsDataCatalog

Database: playcloud3000db

- In the Query result location and encryption
    - Add

```
s3://<name-of-your-s3-bucket>/query/
```

Remember to change the placeholder <name-of-your-s3-bucket> with the name of your S3 bucket



Amazon Athena > Query editor > Manage settings

Manage settings

Query result location and encryption

Location of query result - *optional*

Enter an S3 prefix in the current region where the query result will be saved as an object.

s3://playcloud3000/query/

You can create and manage lifecycle rules for this bucket
Use Amazon S3 lifecycle rules to store your query results and metadata cost effectively or to delete them after a period of time.
Learn more

Expected bucket owner - *optional*

Specify the AWS account ID that you expect to be the owner of your query results output location bucket.

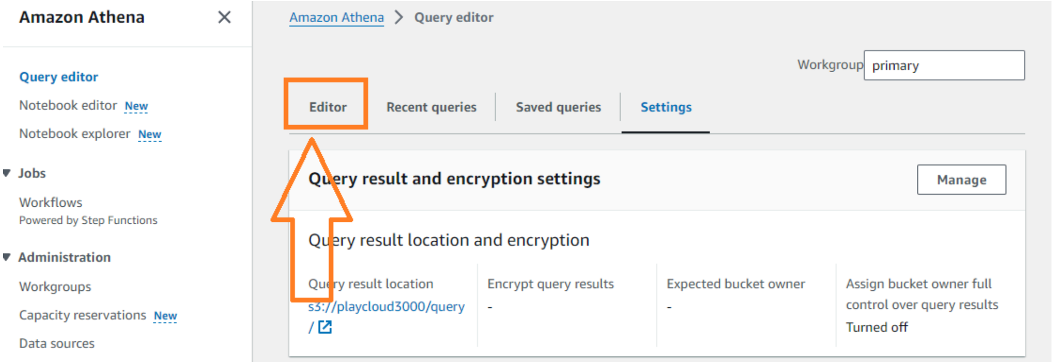
Enter AWS account ID

□ Assign bucket owner full control over query results

Enabling this option grants the owner of the S3 query results bucket full control over the query results. This means that if your query result location is owned by another account, you grant full control over your query results to the other account.

□ Encrypt query results

- click on **Save**
- Navigate back to the Amazon Athena **Editor** tab



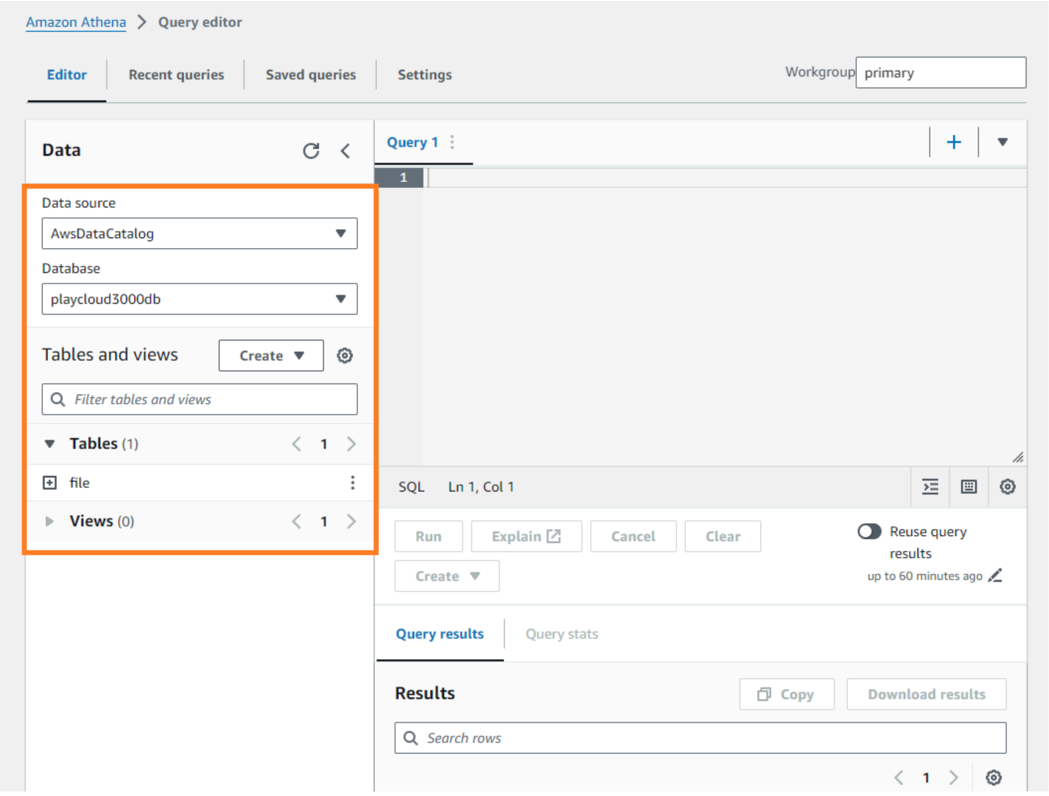
Query result and encryption settings

Query result location and encryption

Query result location	Encrypt query results	Expected bucket owner	Assign bucket owner full control over query results
s3://playcloud3000/query/	-	-	Turned off

3. **Select the database** created in the AWS Glue Data Catalog.

- Follow the configuration below
  - Data source: **AwsDataCatalog**
  - Database: **(select the name of the database you created)**
  - The Tables should be automatically with the tables you created a while ago.



Workgroup: primary

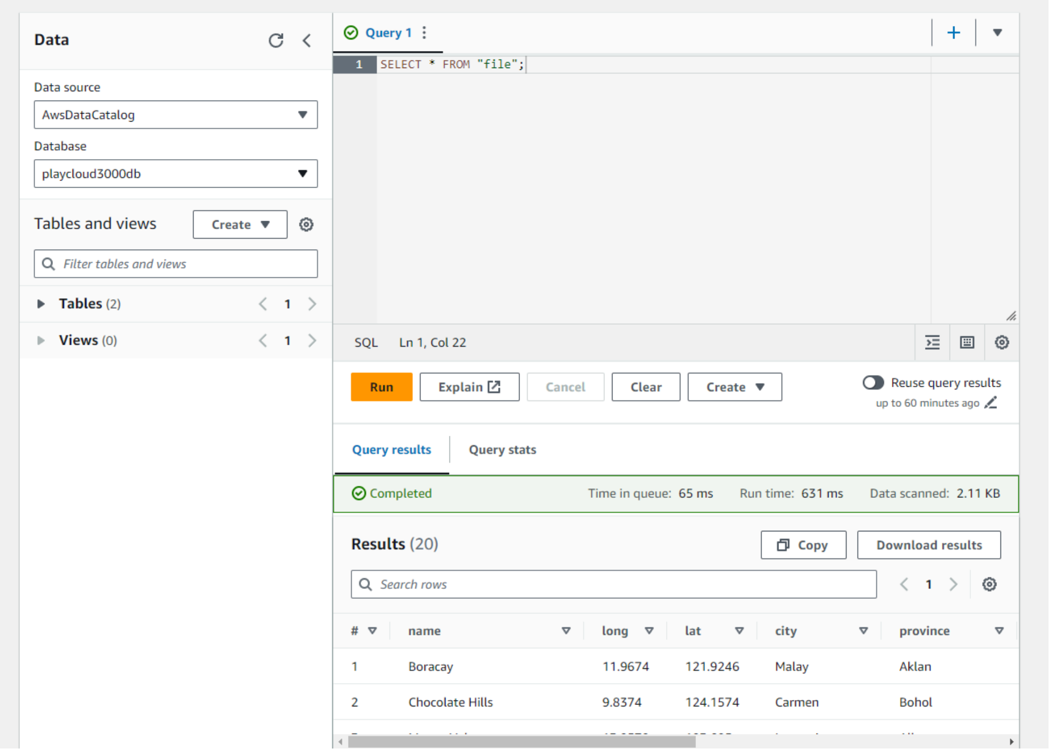
Data source: AwsDataCatalog

Database: playcloud3000db

4. **Write and run SQL queries** to analyze your data. You can use the standard SQL syntax.

- **To view all records:**
  - Copy & Paste. Then, **Run** the following SQL query and check the results afterward.

```
SELECT * FROM "file";
```



Data source: AwsDataCatalog

Database: playcloud3000db

```
SELECT * FROM "file";
```

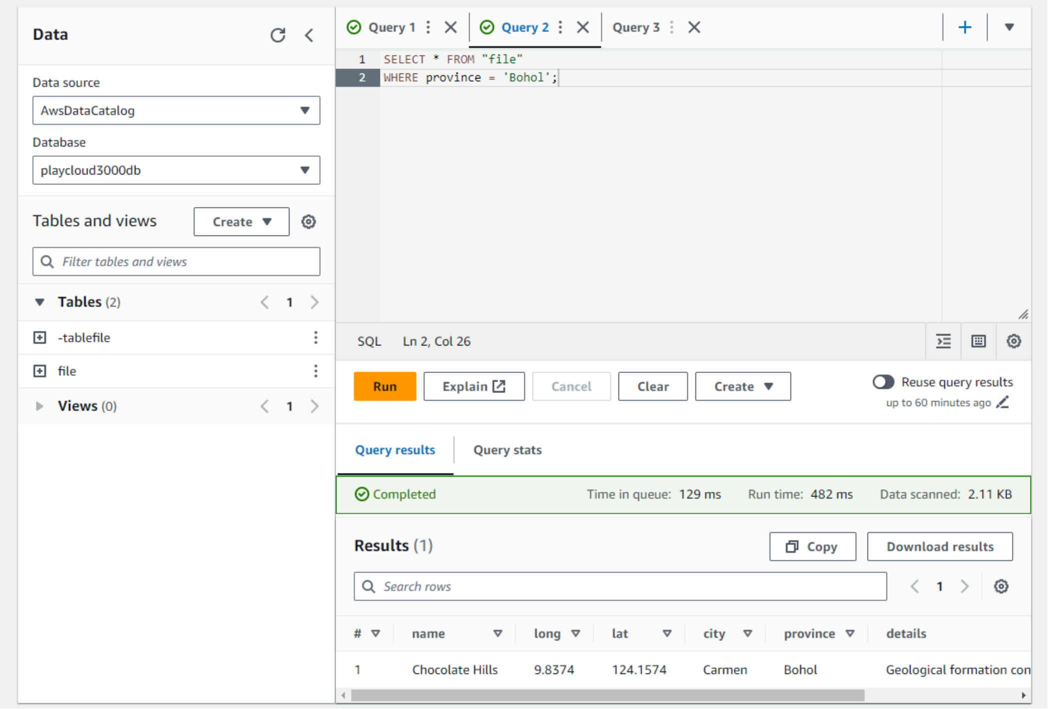
Completed   Time in queue: 65 ms   Run time: 631 ms   Data scanned: 2.11 KB

#	name	long	lat	city	province
1	Boracay	11.9674	121.9246	Malay	Aklan
2	Chocolate Hills	9.8374	124.1574	Carmen	Bohol

- You can click on **Clear** to clear the current cell contents and click on the plus button (**+**) to add a new cell besides the current cell

- **Filtering Records with a WHERE Clause**
  - Copy & Paste. Then, **Run** the following SQL query and check the results afterward.

```
SELECT * FROM "file"
WHERE province = 'Bohol';
```



```
SELECT * FROM "file"
WHERE province = 'Bohol';
```

Data source: AwsDataCatalog

Database: playcloud3000db

Tables (2)
- -tablefile
- file

Views (0)

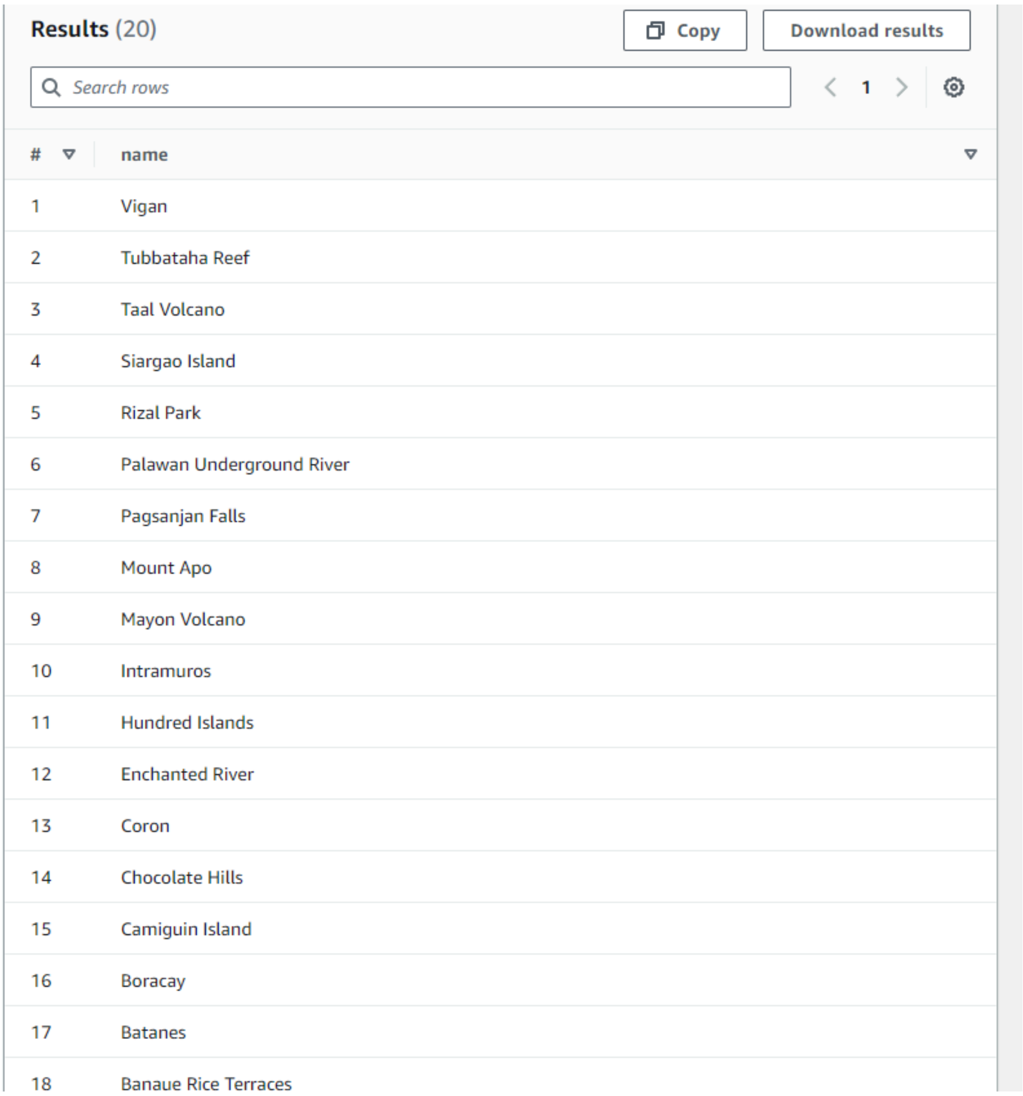
Completed Time in queue: 129 ms Run time: 482 ms Data scanned: 2.11 KB

Results (1)

#	name	long	lat	city	province	details
1	Chocolate Hills	9.8374	124.1574	Carmen	Bohol	Geological formation con

- **Sorting Results**
  - Copy & Paste. Then, **Run** the following SQL query and check the results afterward.

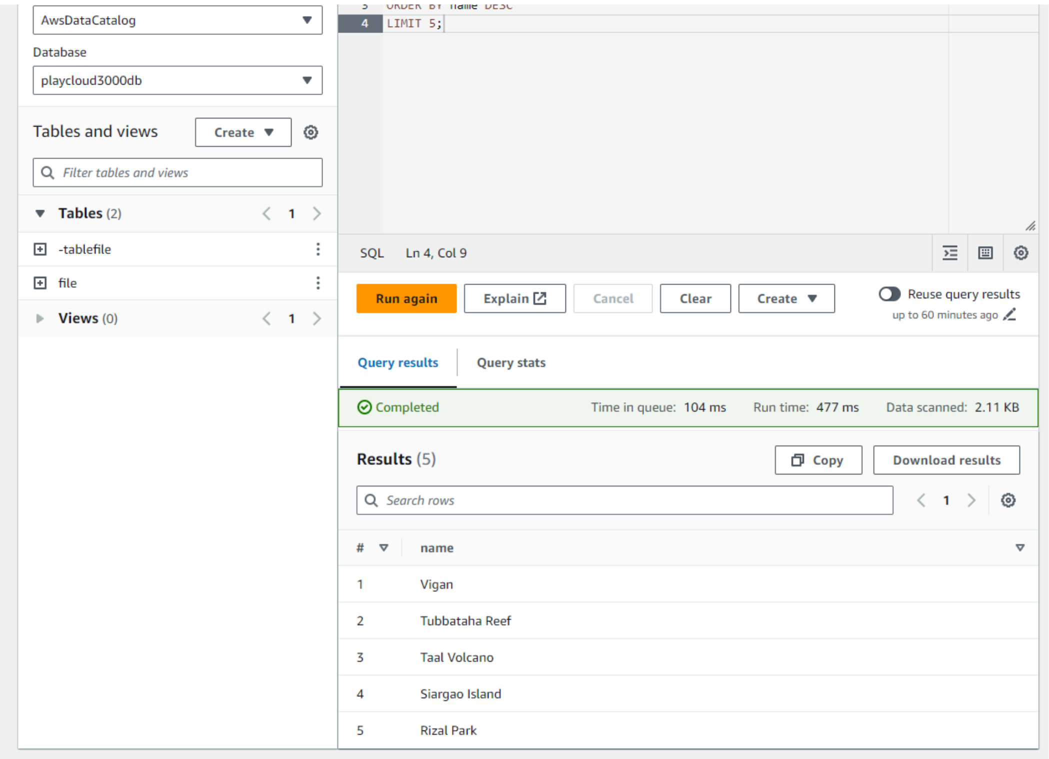
```
SELECT name
FROM "file"
ORDER BY name DESC;
```



#	name
1	Vigan
2	Tubbataha Reef
3	Taal Volcano
4	Siargao Island
5	Rizal Park
6	Palawan Underground River
7	Pagsanjan Falls
8	Mount Apo
9	Mayon Volcano
10	Intramuros
11	Hundred Islands
12	Enchanted River
13	Coron
14	Chocolate Hills
15	Camiguin Island
16	Boracay
17	Batanes
18	Banaue Rice Terraces

- **Limiting Results**
  - Copy & Paste. Then, **Run** the following SQL query and check the results afterward.

```
SELECT name
FROM "file"
ORDER BY name
DESC LIMIT 5;
```



AwsDataCatalog

Database: playcloud3000db

Tables (2)
- -tablefile
- file

Views (0)

```
ORDER BY name DESC
LIMIT 5;
```

SQL Ln 4, Col 9

Completed Time in queue: 104 ms Run time: 477 ms Data scanned: 2.11 KB

Results (5)

#	name
1	Vigan
2	Tubbataha Reef
3	Taal Volcano
4	Siargao Island
5	Rizal Park

That’s it! Congratulations!

You just learned how to use AWS Glue to create a data catalog (database and tables) for organizing data from Amazon S3 and query data directly from S3 using Amazon Athena. This lab serves as a foundational step into the world of cloud-based data analytics, empowering you to explore more complex data analytics scenarios.

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!