



AWS Glue Crawlers

In AWS, a **crawler** is a tool provided by AWS Glue that automatically discovers and catalogs metadata about your data, making it easier to query and analyze. A crawler connects to your data source, such as files in an Amazon S3 bucket, and infers the structure (schema) of the data by analyzing its content. The crawler then stores this metadata in the AWS Glue Data Catalog, which acts as a centralized metadata repository.

Key Features:

1. **Schema Discovery:** The crawler automatically infers the schema (columns, data types, etc.) of the data by inspecting the contents of the files. This is useful for handling semi-structured data like JSON, Parquet, or CSV files.
2. **Multiple Data Sources:** Crawlers support various data sources, such as:
 - o Amazon S3
 - o Amazon RDS (Relational Databases)
 - o Amazon Redshift
 - o JDBC-compatible data stores
3. **Automated Metadata Cataloging:** Once the crawler runs and analyzes the data, it updates the AWS Glue Data Catalog, making the metadata available for querying using services like **Amazon Athena**, **Amazon Redshift Spectrum**, and **AWS Glue ETL jobs**.
4. **Incremental Updates:** Crawlers can be scheduled to run periodically, which means they can incrementally add new partitions or detect changes in the schema of the data, keeping the catalog up to date.

Use Cases:

- **Data Lakes:** For organizing and preparing data in S3 for analysis using tools like Amazon Athena or Redshift Spectrum.
- **ETL Processes:** Before transforming data using AWS Glue jobs, the crawler ensures that the metadata is available and up to date.
- **Big Data Analytics:** Helps prepare raw data for analysis by automatically understanding its structure.

In short, an AWS Glue crawler simplifies the process of organizing and preparing data for analysis by automatically discovering and cataloging the necessary metadata.

In this lab, the goal is to automatically discover metadata from a CSV file using AWS Glue Crawlers and catalog it for easy querying. The process starts with uploading a CSV file (Titanic dataset) into an Amazon S3 bucket.

Next, you create a database in AWS Glue and use the crawler to scan the S3 bucket. The crawler inspects the file's contents to automatically infer the structure (schema) of the data. After setting up the crawler, you run it, and it creates a table in the Glue Data Catalog, which contains the metadata (e.g., columns, data types). You can then view the schema in the table that was generated.

The end goal is to demonstrate how AWS Glue Crawlers simplify metadata discovery, making it easier to work with and analyze data in S3 by automatically cataloging it in a structured format.

💡 To begin with the Lab:

1. In this lab we are going to see how we can perform the automatic discovery of the metadata with the help of AWS Glue Crawlers.
2. Here we also have a CSV file named Titanic which we are going to upload in S3 Bucket.
3. In your S3 bucket you need to create 2 folders with the same name and in the data folder you need to upload the CSV file.

Amazon S3 > Buckets > s3-bucket-logs-1852

s3-bucket-logs-1852 [Info](#)

Objects Properties Permissions Metrics Management Access Points

Objects (2) [Info](#)

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

Amazon S3 > Buckets > s3-bucket-logs-1852 > data/

data/ [Copy S3 URI](#)

Objects Properties

Objects (1) [Info](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	titanic_data.csv	csv	October 5, 2024, 16:35:29 (UTC+05:30)	59.8 KB	Standard

4. Now come to AWS Glue, here you need to go to Databases and create one. For that, you just need to give a name to your Database and create it.

AWS Glue > Databases

Databases (1)					
A database is a set of associated table definitions, organized into a logical group.					
Last updated (UTC) Edit Delete Add database					
Filter databases < 1 > ⚙					
□ Name	▲ Description	▼ Location URI	▼ Created on (UTC)	▼	▼
demo-database	demo-database	-	October 5, 2024 at 11:06:31		

5. Then go to Tables and choose Add tables using crawlers.

AWS Glue > Tables

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 (0)					
Last updated (UTC) Edit Delete Add tables using crawler Add table					
View and manage all available tables. Filter tables					
□ Name	▲ Database	▼ Location	▼ Classification	▼ Deprecated	▼ View data Data quality
No available tables					

6. On step one, give a name and description to your crawler.

AWS Glue > Crawlers > Add crawler

Set crawler properties

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

Crawler details Info

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

Description - *optional* Descriptions can be up to 2048 characters long.

▶ Tags - *optional* Use tags to organize and identify your resources.

Cancel Next

7. Then in step 2 you need to add a data source which is your S3 bucket. Choose the same options as shown in the snapshot.

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

Choose data sources and classifiers

Data source configuration

Is your data already mapped to Glue tables?

Not yet
Select one or more data sources to be crawled.

Yes
Select existing tables from your Glue Data Catalog.

Data sources (0) Info
The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
You don't have any data sources.		

[Add a data source](#)

► **Custom classifiers - optional**
A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

[Cancel](#) [Previous](#) **Next**

Add data source

X

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).

▼ C

[Clear selection](#)

[Add new connection](#)

Location of S3 data

- In this account
- In a different account

S3 path

Browse for or enter an existing S3 path.

s3://s3-bucket-logs-1852



[View](#)

[Browse S3](#)

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.

[Cancel](#)

[Add an S3 data source](#)

8. Now on the step 3 you need to choose an IAM role if you have, or you can click on Create new IAM role and create a new one.

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

Configure security settings

IAM role [Info](#)

Existing IAM role
gluelab [View](#)

[Create new IAM role](#) [Update chosen IAM role](#)

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.

[Cancel](#) [Previous](#) [Next](#)

9. In step 4, choose your target database and for the crawler schedule choose on demand. Then just move ahead and create your 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 output and scheduling

Output configuration [Info](#)

Target database
demo-database [View](#)

[Clear selection](#) [Add database](#)

Table name prefix - *optional*

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.

Advanced options

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

10. Below you can see that your crawler is ready, so click on run crawler and wait for it to get completed.

AWS Glue > Crawlers > demo-Crawler

demo-Crawler

Last updated (UTC)
October 5, 2024 at 11:11:47

Run crawler Edit Delete

Crawler properties			
Name demo-Crawler	IAM role gluelab	Database demo-database	State READY
Description demo-Crawler	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			
► Advanced settings			

11. You can see that our crawler run has been completed. Now if you to tables.

Crawler runs | Schedule | Data sources | Classifiers | Tags

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
October 5, 2024 at 11:13:18	October 5, 2024 at 11:14:35	01 min 17 s	Completed	-	-

Filter data Filter by a date and time range < 1 > ⚙

12. You will see a table has been created.

AWS Glue > Tables

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 (1)

Last updated (UTC)
October 5, 2024 at 11:19:12

View and manage all available tables.

Name	Database	Location	Classification	Deprecated	View data	Data quality
s3_bucket_logs_1852	demo-database	s3://s3-bucket-logs-1#	CSV	-	Table data	View data quality

Filter tables < 1 > ⚙ Add tables using crawler Add table

13. If you go inside of it, you will be able to see the schema.

Schema | Partitions | Indexes | Column statistics - new

Schema (13)

View and manage the table schema.

#	Column name	Data type	Partition key	Comment
1	passengerid	bigint	-	-
2	survived	bigint	-	-
3	pclass	bigint	-	-
4	name	string	-	-
5	sex	string	-	-
6	age	double	-	-
7	sibsp	bigint	-	-
8	parch	bigint	-	-
9	ticket	string	-	-
10	fare	double	-	-
11	cabin	string	-	-
12	embarked	string	-	-
13	partition_0	string	Partition (0)	-

Edit schema as JSON | Edit schema | < 1 > | ⚙️

- Once you are done just delete your database, table, and crawler. Also, empty and delete your S3 bucket.