😊 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:
    - Amazon S3
    - Amazon RDS (Relational Databases)
    - Amazon Redshift
    - 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.

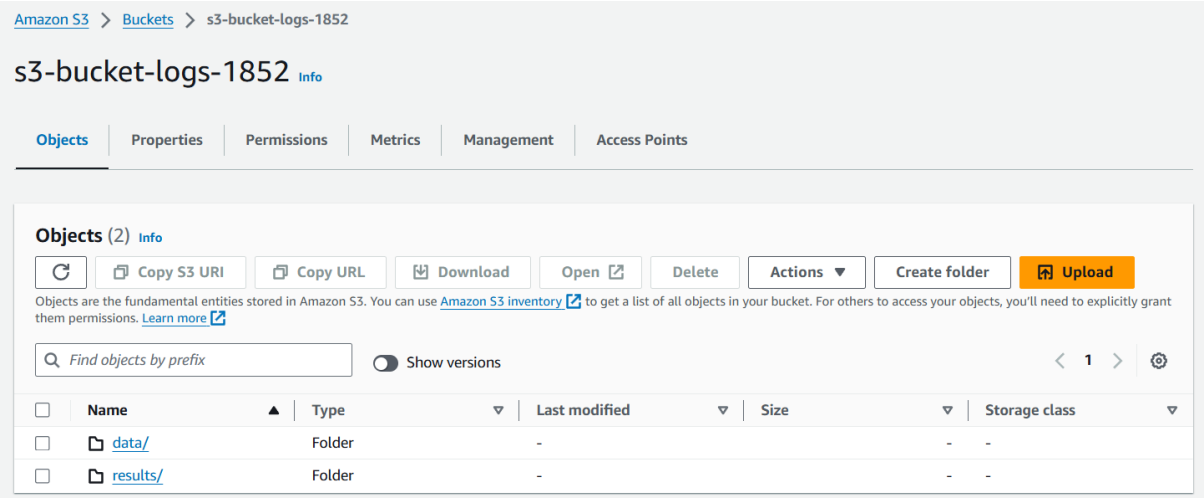
In this lab, the goal is to automatically discover metadata from a CSV file using AWS Glue Crawlers and catalog it for easy querying. The process starts with uploading a CSV file (Titanic dataset) into an Amazon S3 bucket.

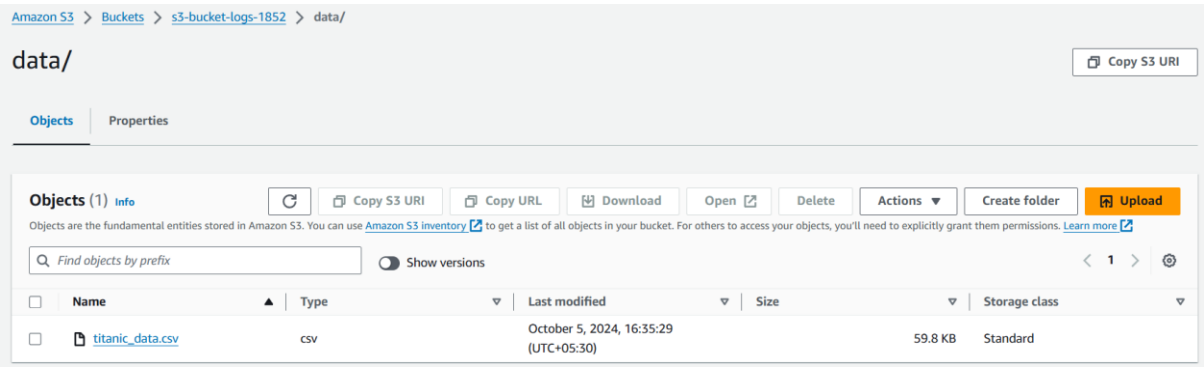
Next, you create a database in AWS Glue and use the crawler to scan the S3 bucket. The crawler inspects the file's contents to automatically infer the structure (schema) of the data. After setting up the crawler, you run it, and it creates a table in the Glue Data Catalog, which contains the metadata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

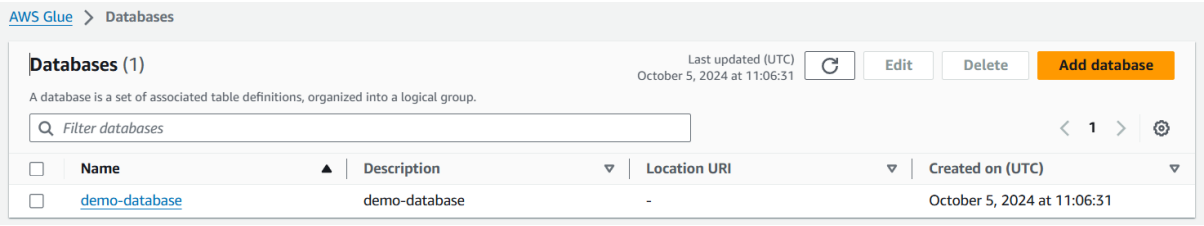
😄 To begin with the Lab:

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.

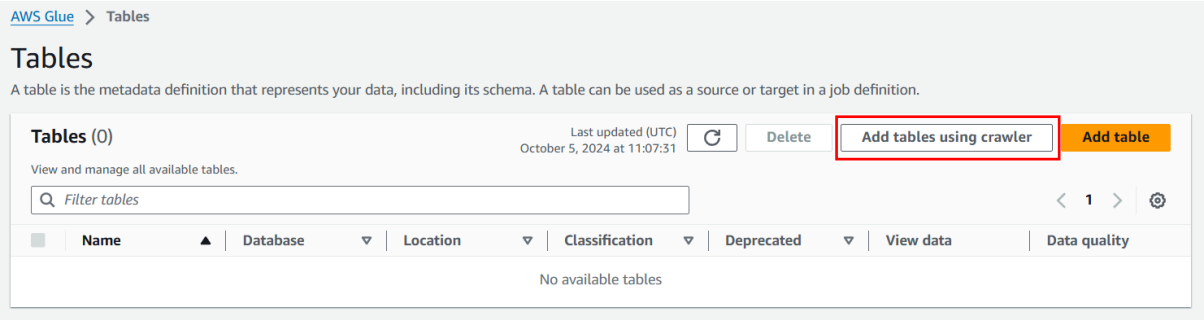




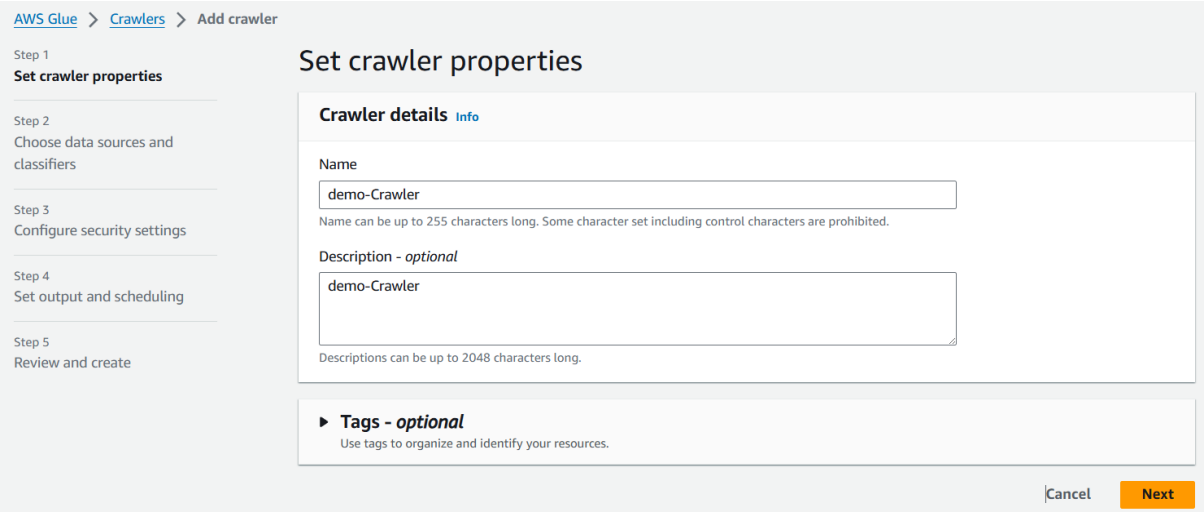
4. Now come to AWS Glue, here you need to go to Databases and create one. For that, you just need to give a name to your Database and create it.



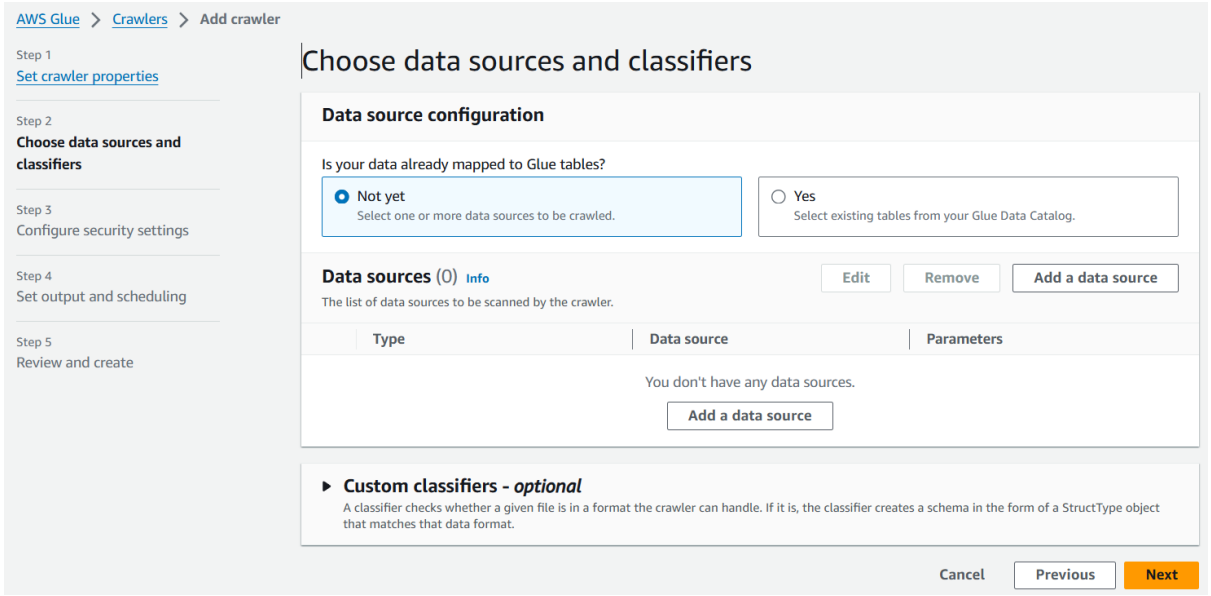
5. Then go to Tables and choose Add tables using crawlers.

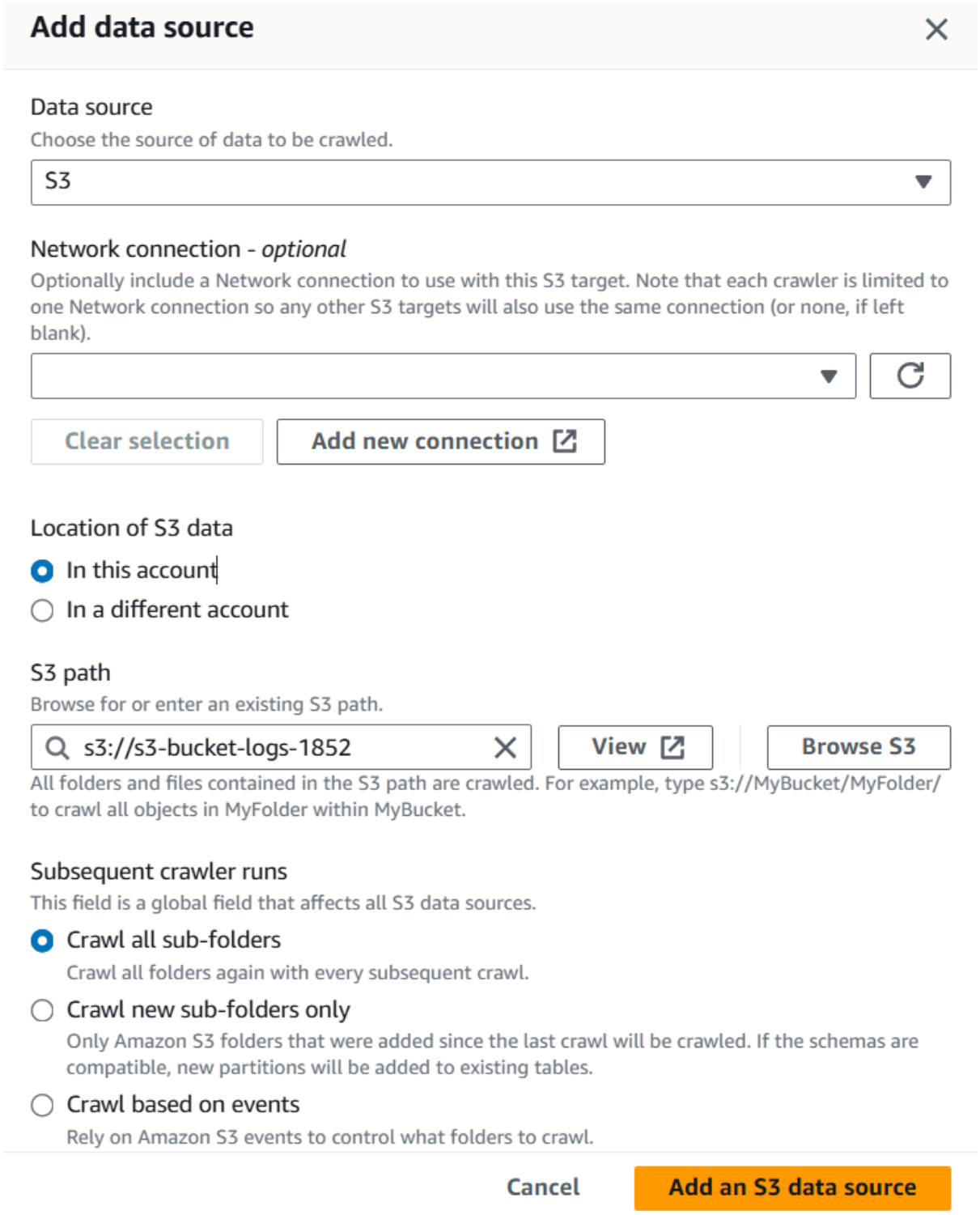


6. On step one, give a name and description to your crawler.

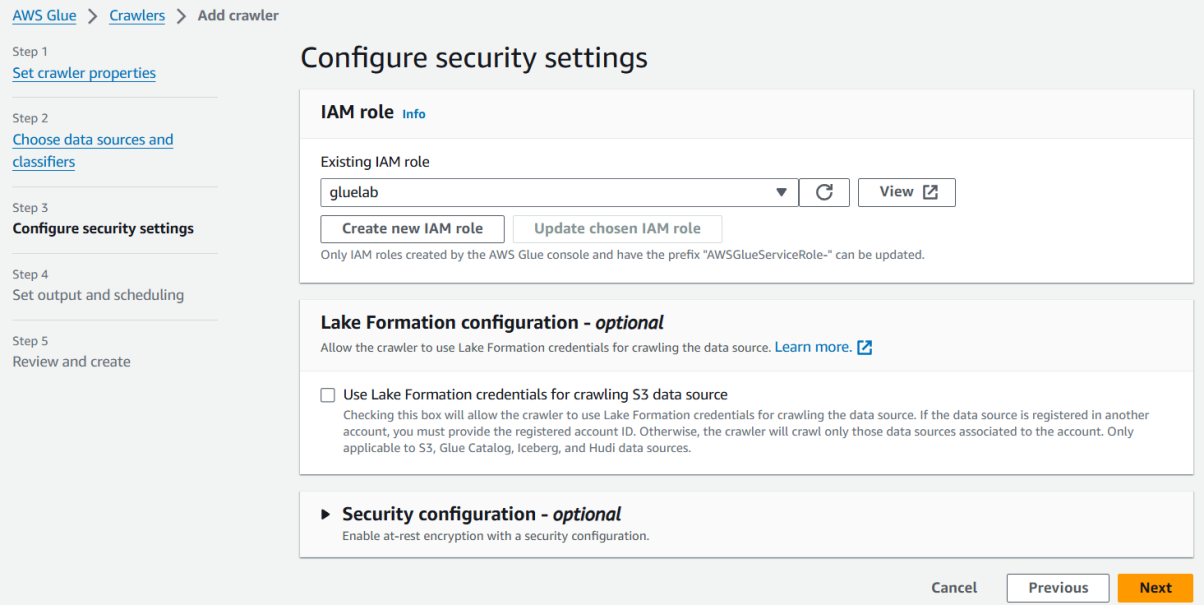


7. Then in step 2 you need to add a data source which is your S3 bucket. Choose the same options as shown in the snapshot.

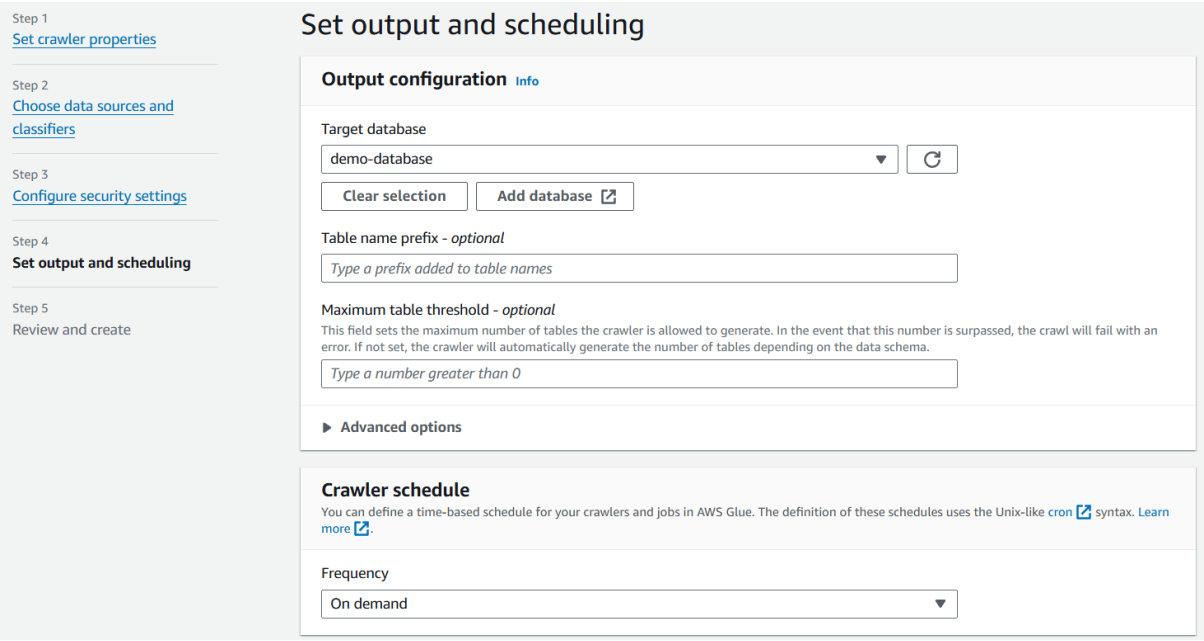

AWS Glue > Crawlers > Add crawler
Step 1
Set crawler properties
Step 2
Choose data sources and classifiers
Step 3
Configure security settings
Step 4
Set output and scheduling
Step 5
Review and create
Choose data sources and classifiers
Data source configuration
Is your data already mapped to Glue tables?
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.
Edit
Remove
Add a data source
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



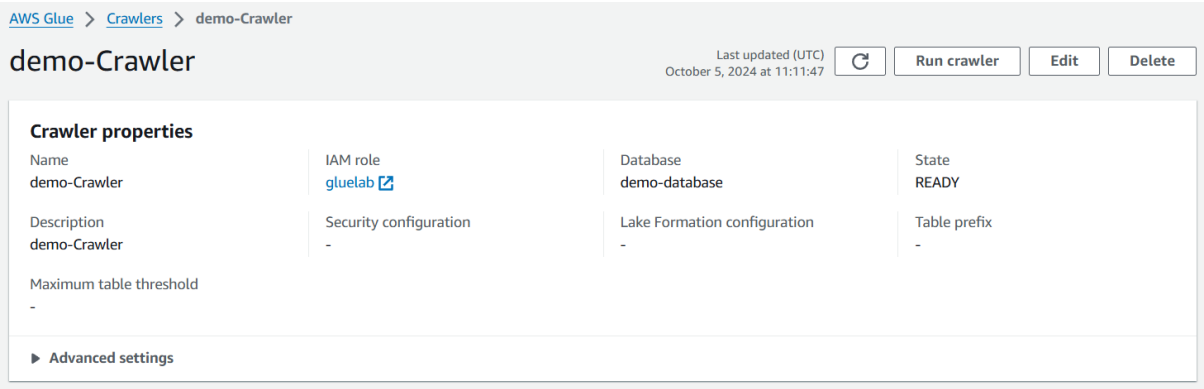
8. Now on the step 3 you need to choose an IAM role if you have, or you can click on Create new IAM role and create a new one.



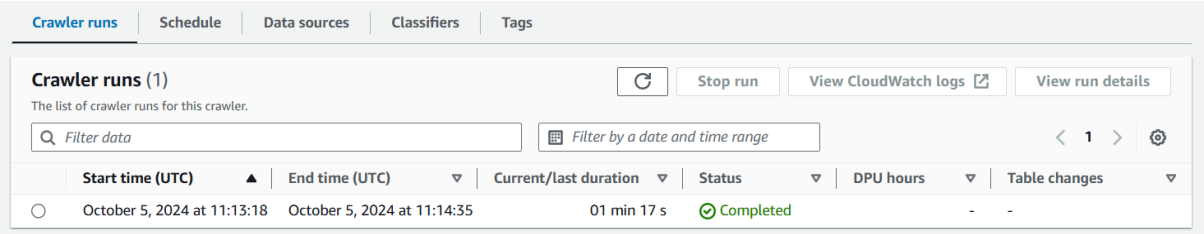
9. In step 4, choose your target database and for the crawler schedule choose on demand. Then just move ahead and create your crawler.



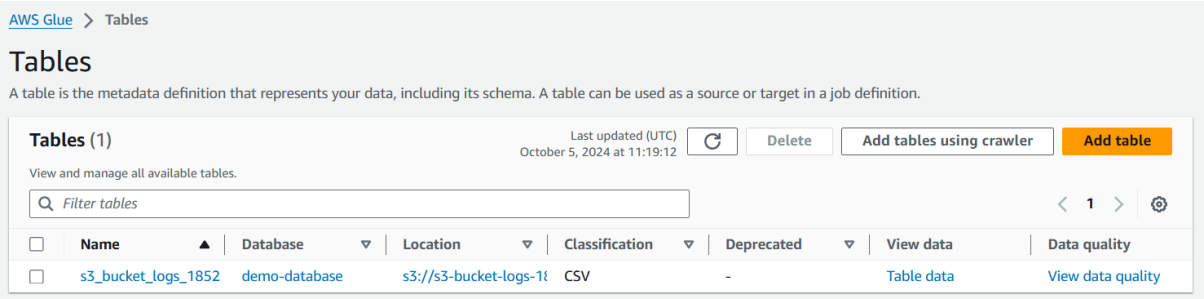
10. Below you can see that your crawler is ready, so click on run crawler and wait for it to get completed.



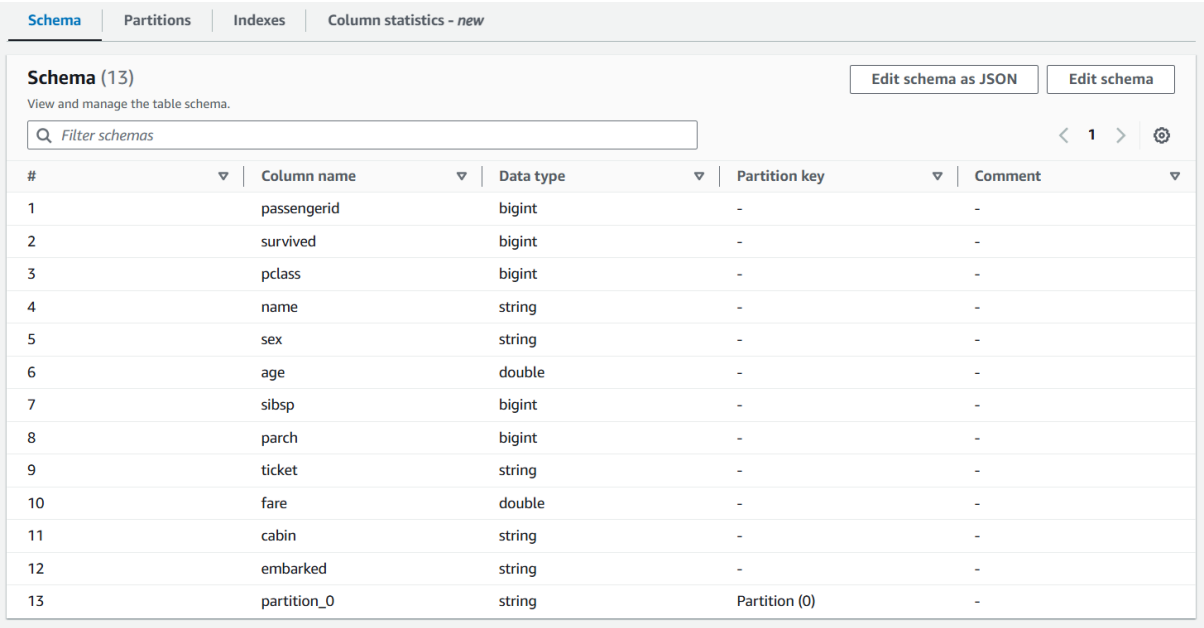
11. You can see that our crawler run has been completed. Now if you to tables.



12. You will see a table has been created.



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

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