**AWS GLUE**

**What is AWS Glue ?**

AWS glue is a serverless cloud service developed by amazon community for the purpose of performing extraction, transformation and loading of data. So, this process is know as ETL.

AWS glue enables user and business organization to extract or pull the data from a source, transform that data by using certain logics and computation and load that into a one place so that whenever user or business house require that data he can pull it off from that source. Usually under this service data are stored in such place where it should be not disrupted/harmed or damaged so this place might be warehouse or inventory. With the help of this glue service user can acquire that data as per their requirement. AWS glue is fully managed serverless cloud service which means user don’t need to manage their underlying resources in order to perform this task everything is taken care by amazon cloud service. So it will provide a flexibility to give faster data access which performs data integration by reducing the time taken for the analysis of your data.

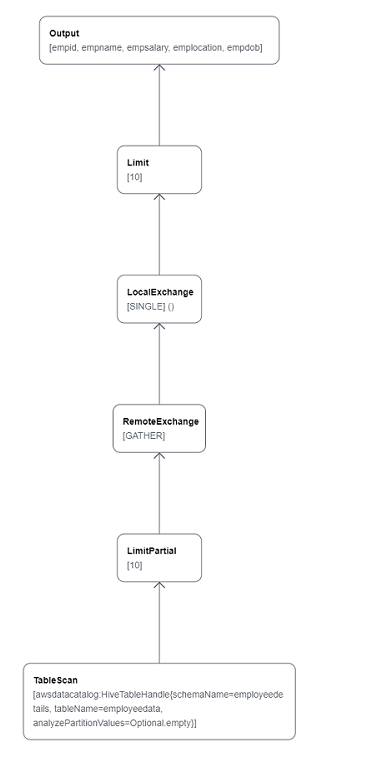
**What it perform in real time scenarios?**

AWS Glue consolidates major data integration capabilities into a single service. These include data discovery, modern ETL, cleansing, transforming, and centralized cataloguing. It’s also serverless, which means there’s no infrastructure to manage. With flexible support for all workloads like ETL, ELT, and streaming in one service, AWS Glue supports users across various workloads and types of users.

It performs provisioning, configuring and scaling of resources that you need to perform the task in glue. You only need to pay the amount for those resources which is required for running the job. It will also automate your data integartion tasks by crawling the data from the data source it identifies the data format and according to that it suggest relevant schemas to store your data.

AWS Glue Studio is a graphical interface that makes it easy to create, run, and monitor data integration jobs in AWS Glue. You can visually compose data transformation workflows and seamlessly run them on the Apache Spark–based serverless ETL engine in AWS Glue. With AWS Glue service, you can rapidly build, test, and run data preparation and analytics applications. It provides a programmatic and visual interface for building and testing extract, transform, and load (ETL) scripts for data preparation.

**Algorithm of AWS Glue**

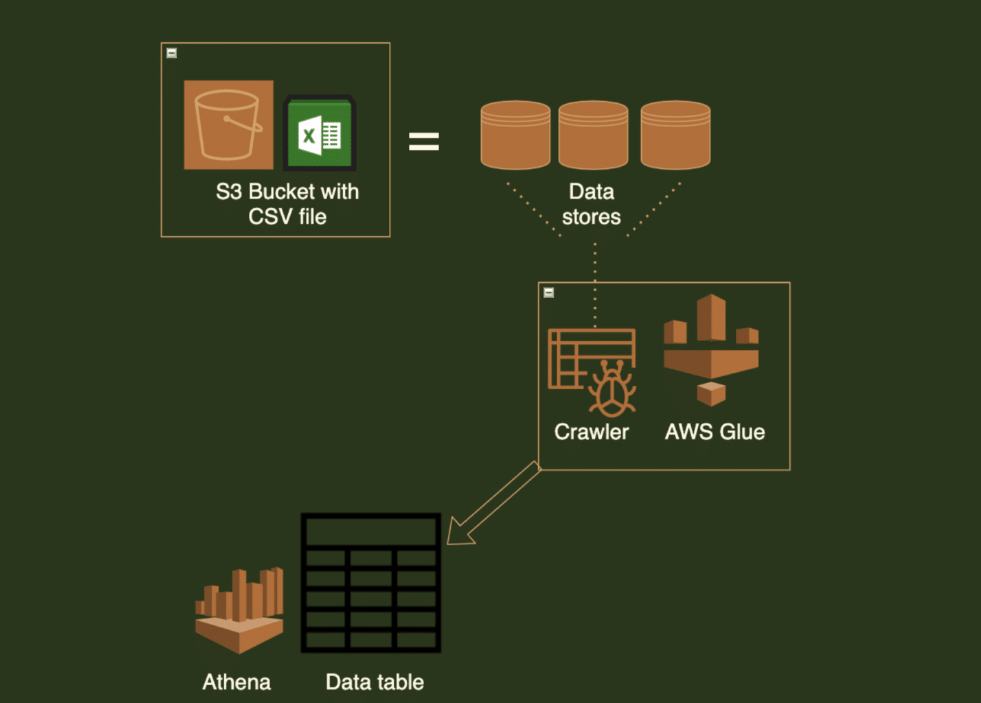


**Features of AWS Glue**

* Discover and organize data.
* Transform, prepare, and clean data for analysis.
* Build and monitor data pipeline.

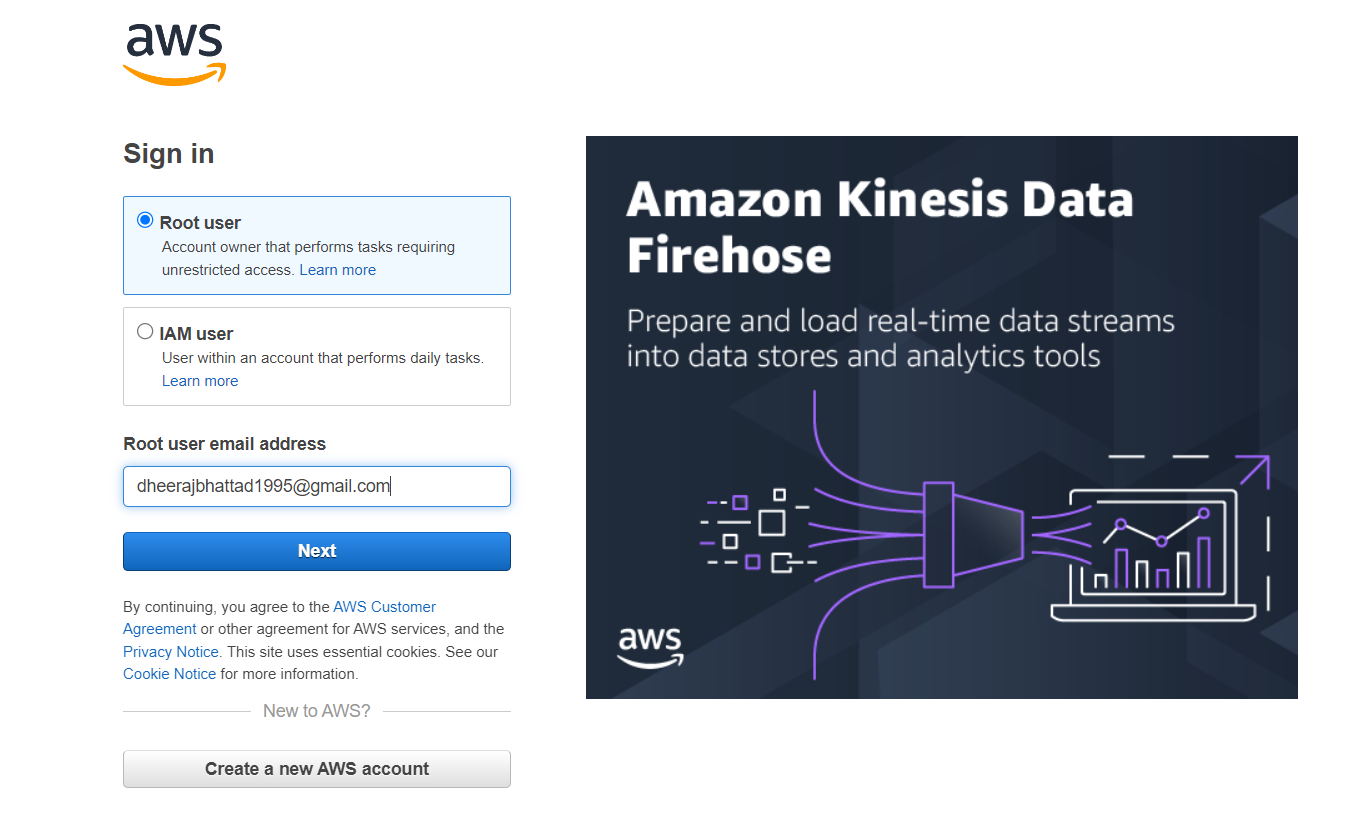
**Workflow of AWS Glue**

**How it works?**

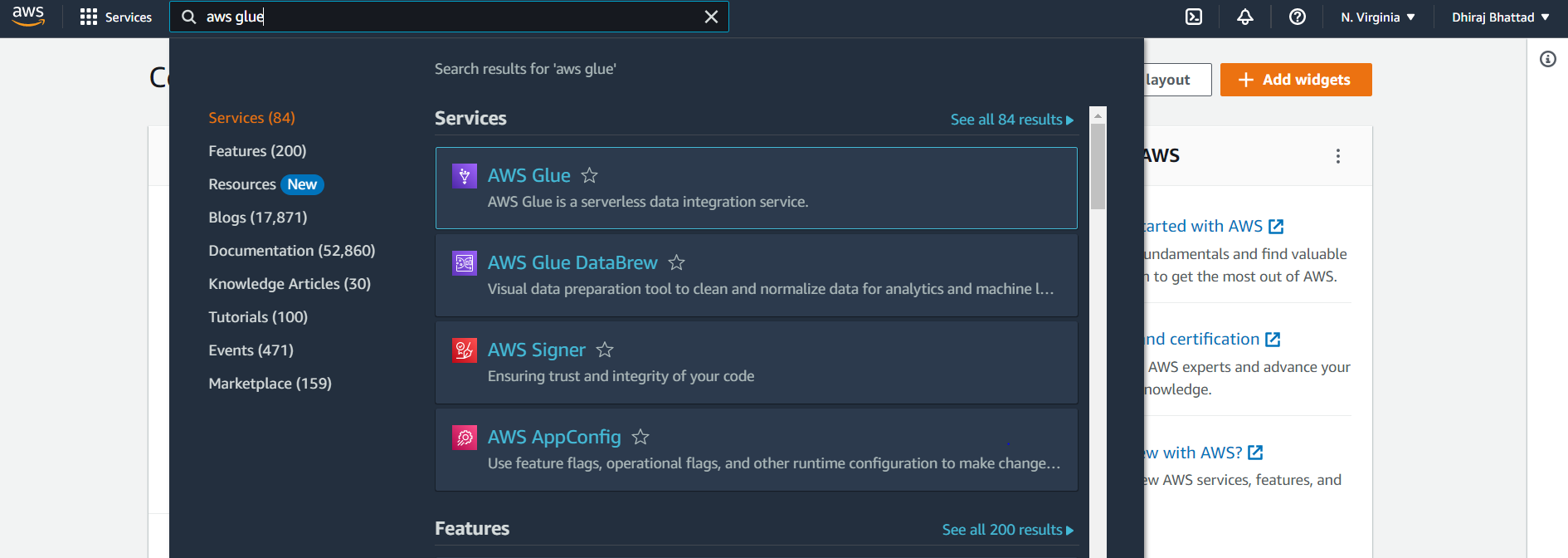


How to perform this task on AWS console

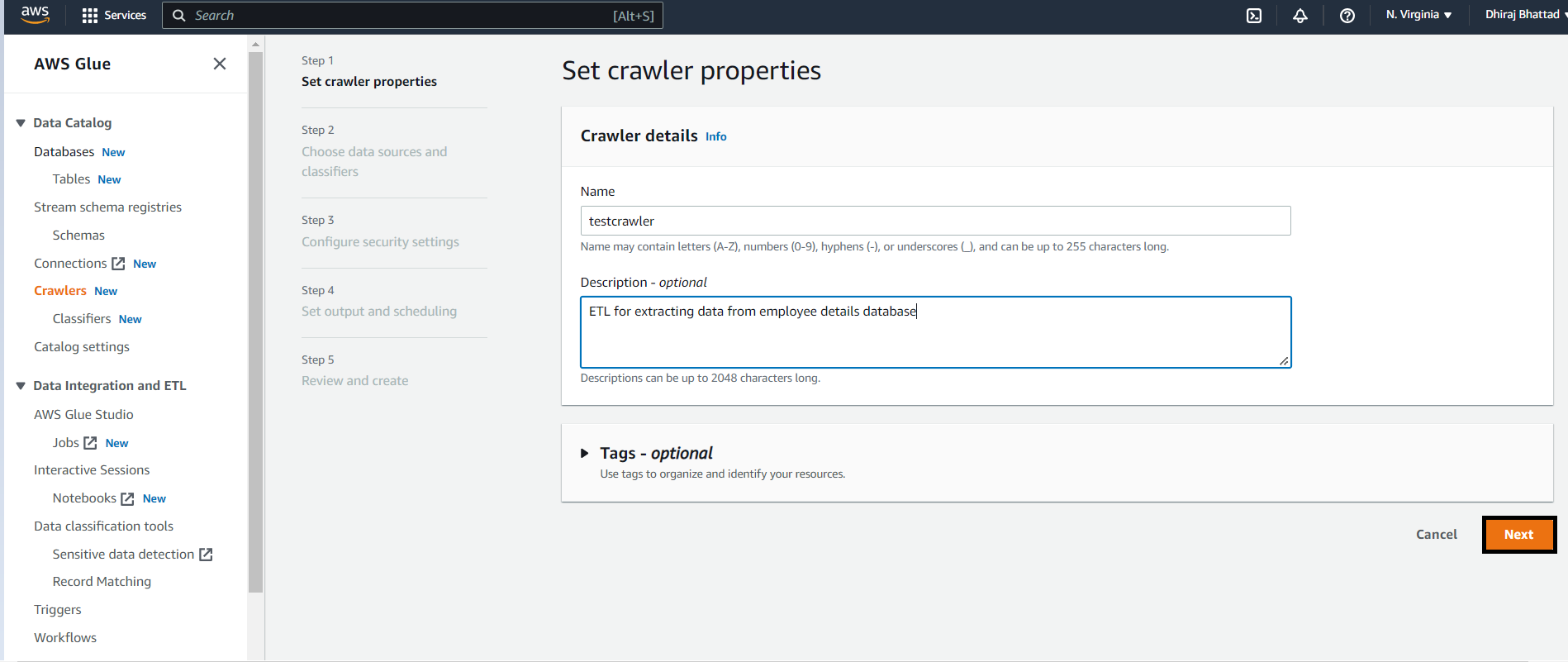
Step 1- Login to your aws account with the credential



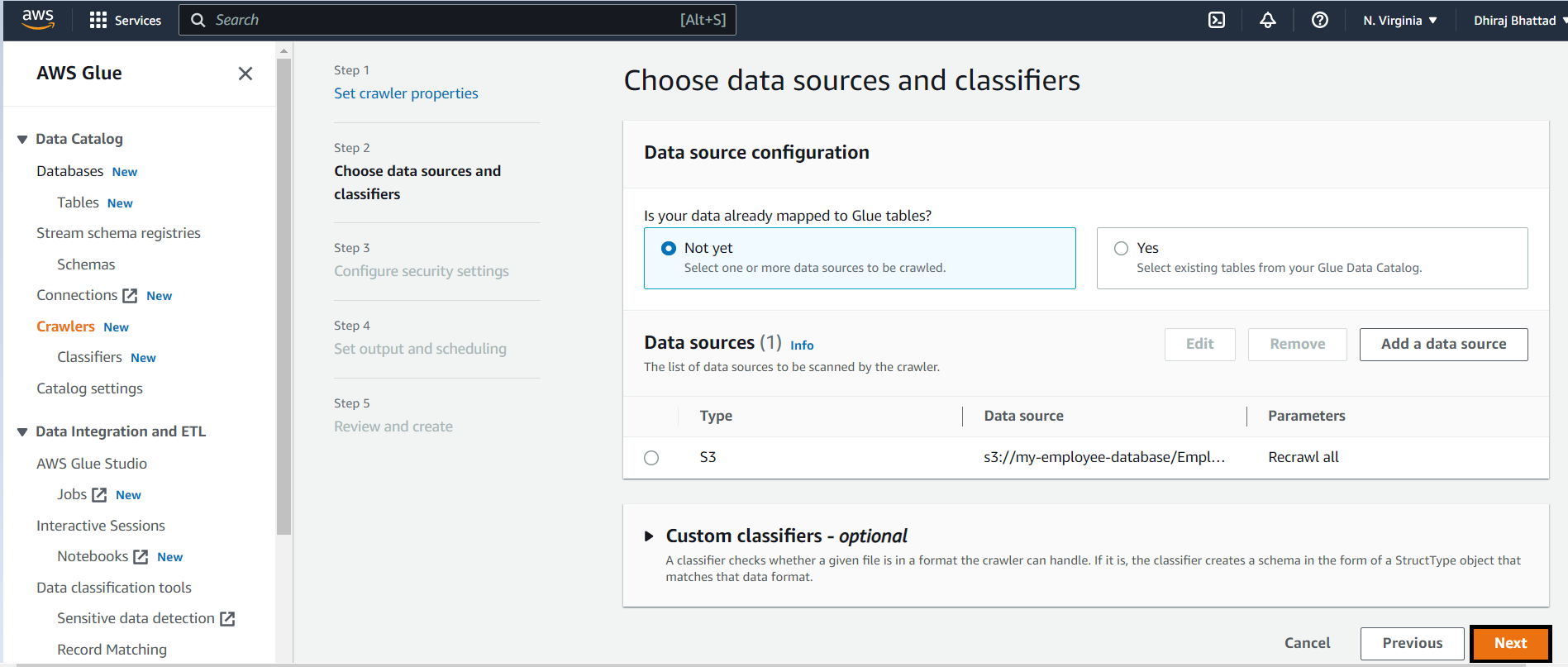
Step 2:- Search it with AWS Glue in service section.



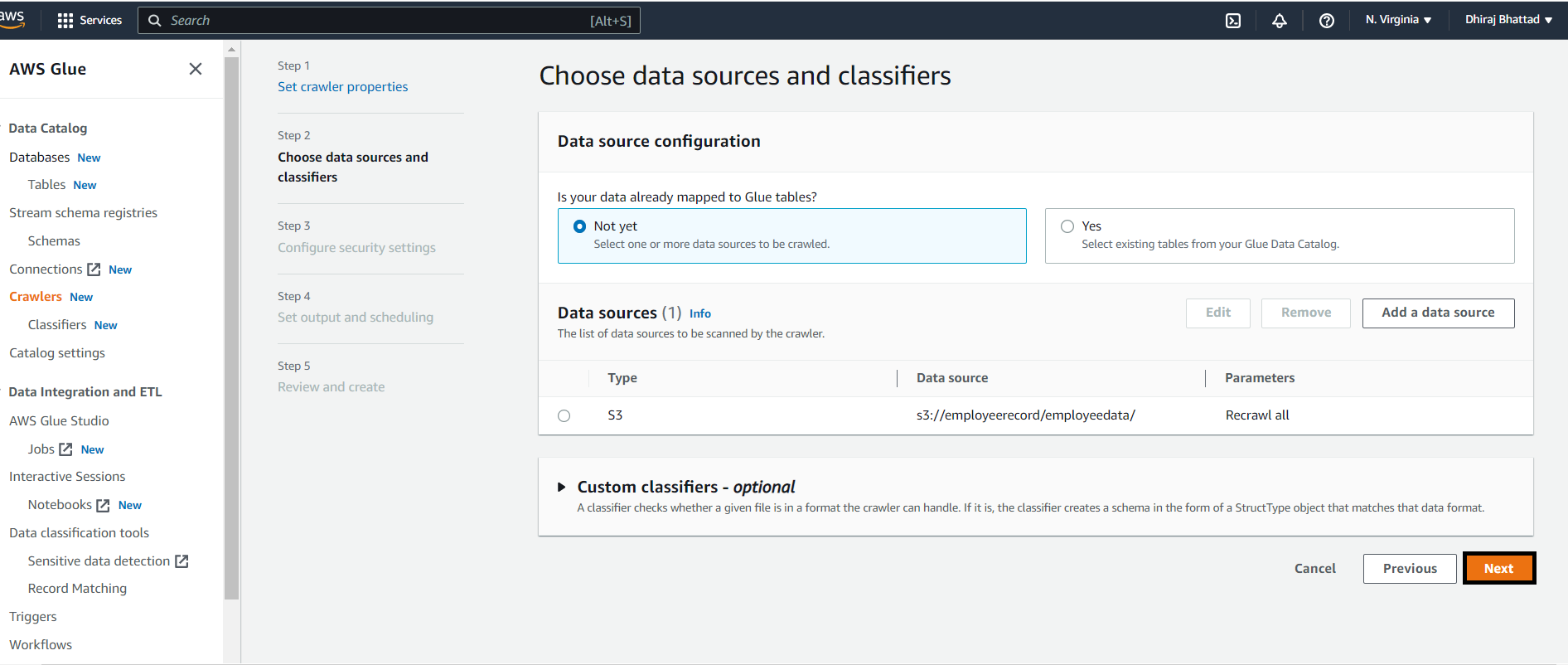
Step 3:- Once it is logged in you to need to first create a crawler for the AWS glue in the sidebar and click on create crawler from the option below. Now you need to give name to your crawler and click on next option. Note- (It’s a updated dashboard of AWS glue crawler).



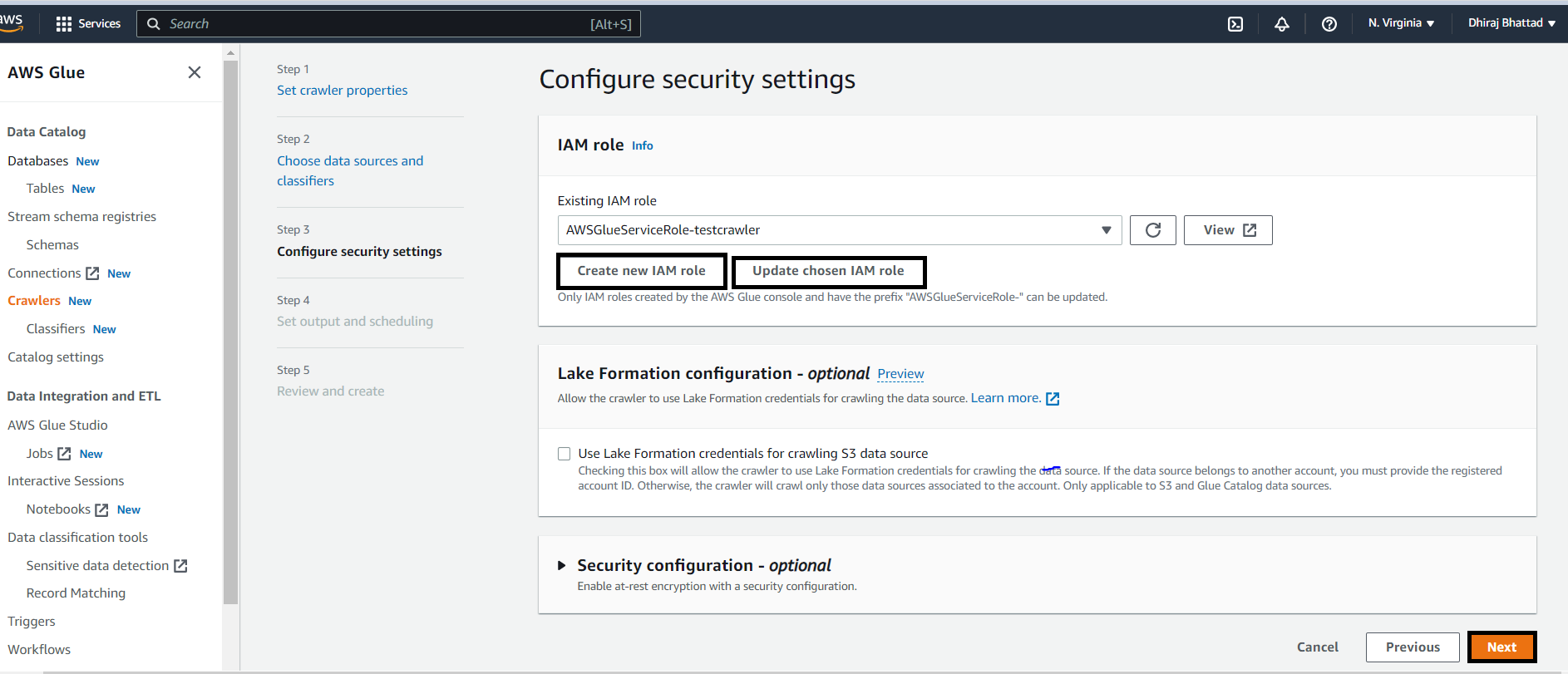
Step 4:- Now it will directed to data source option for that you choose default option for crawling the data of all folders.



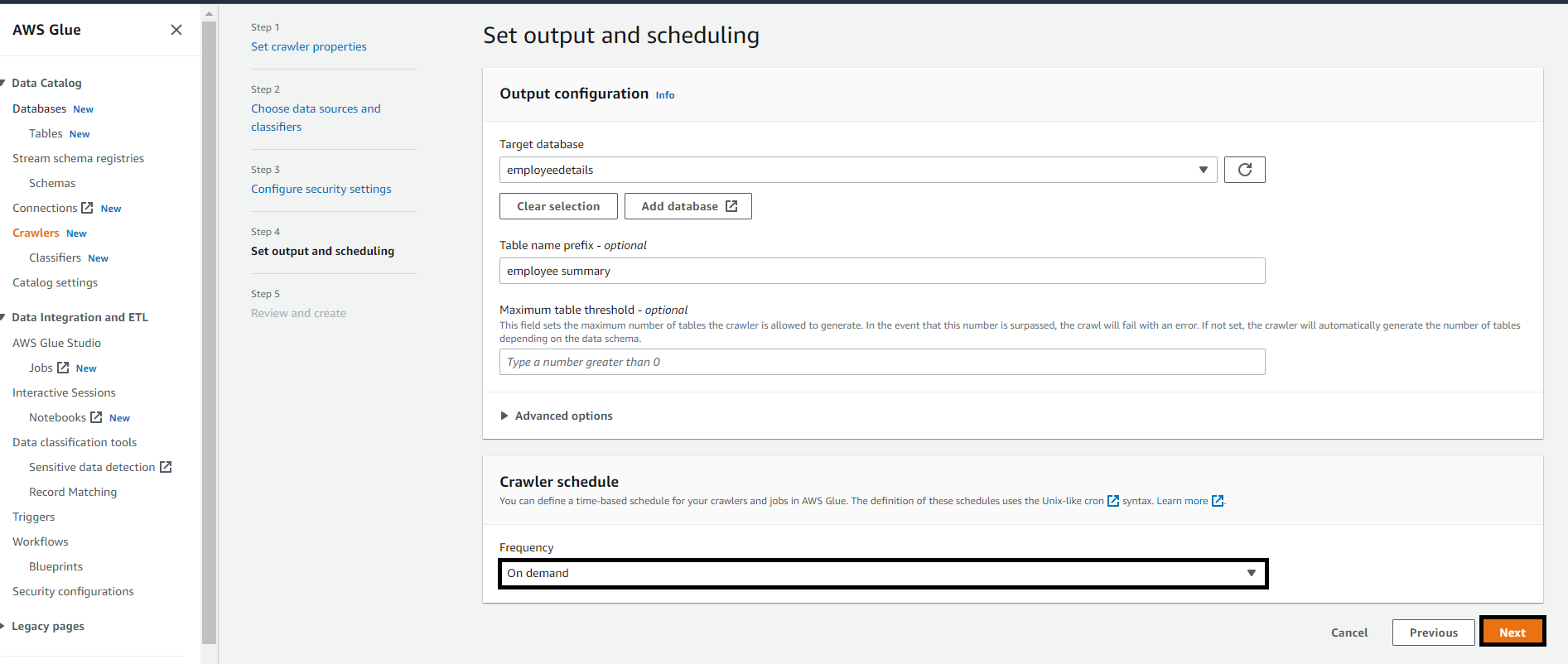
Step 5:- Choose existing data source for your table or you can create new one with which you need to provide storage directory path of s3 in standardized format.



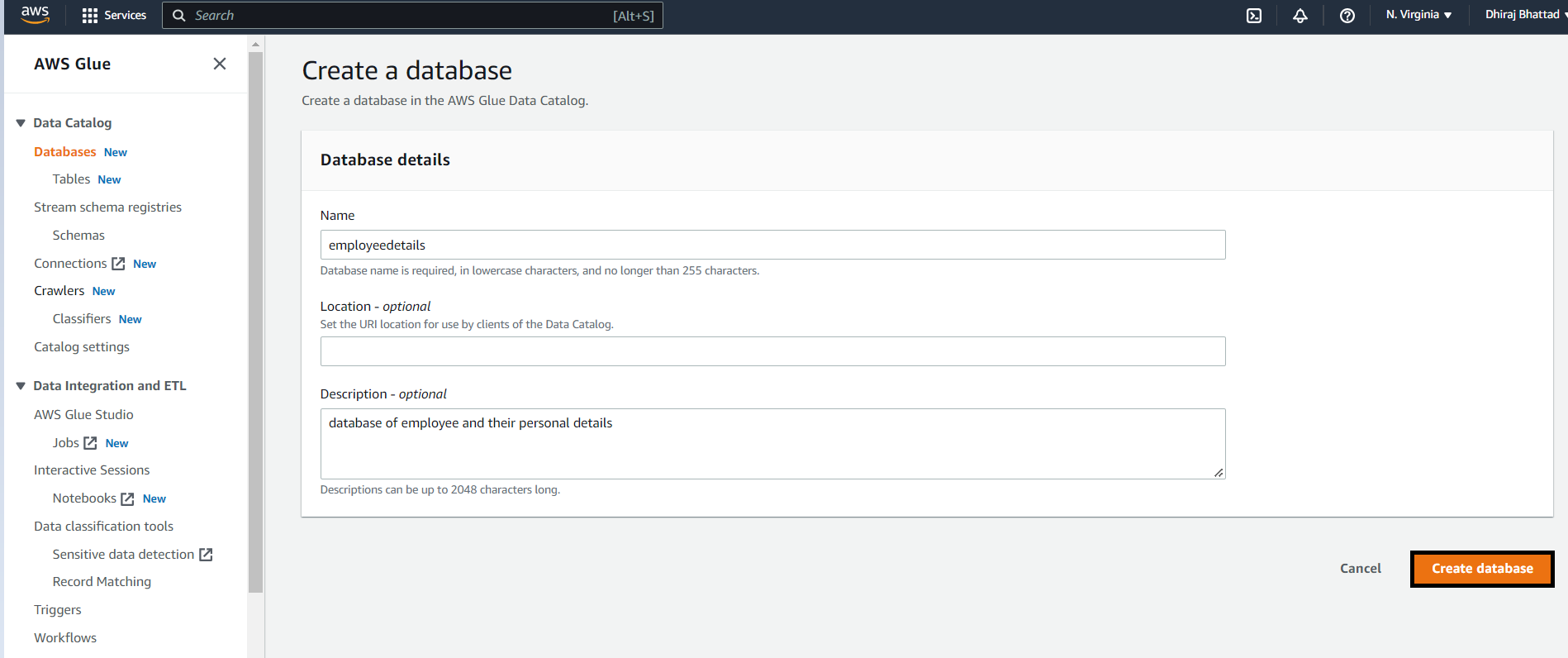
Step 6:-Then you need to choose new IAM role and you can choose new IAM rule or either you can choose existing **IAM role** also.



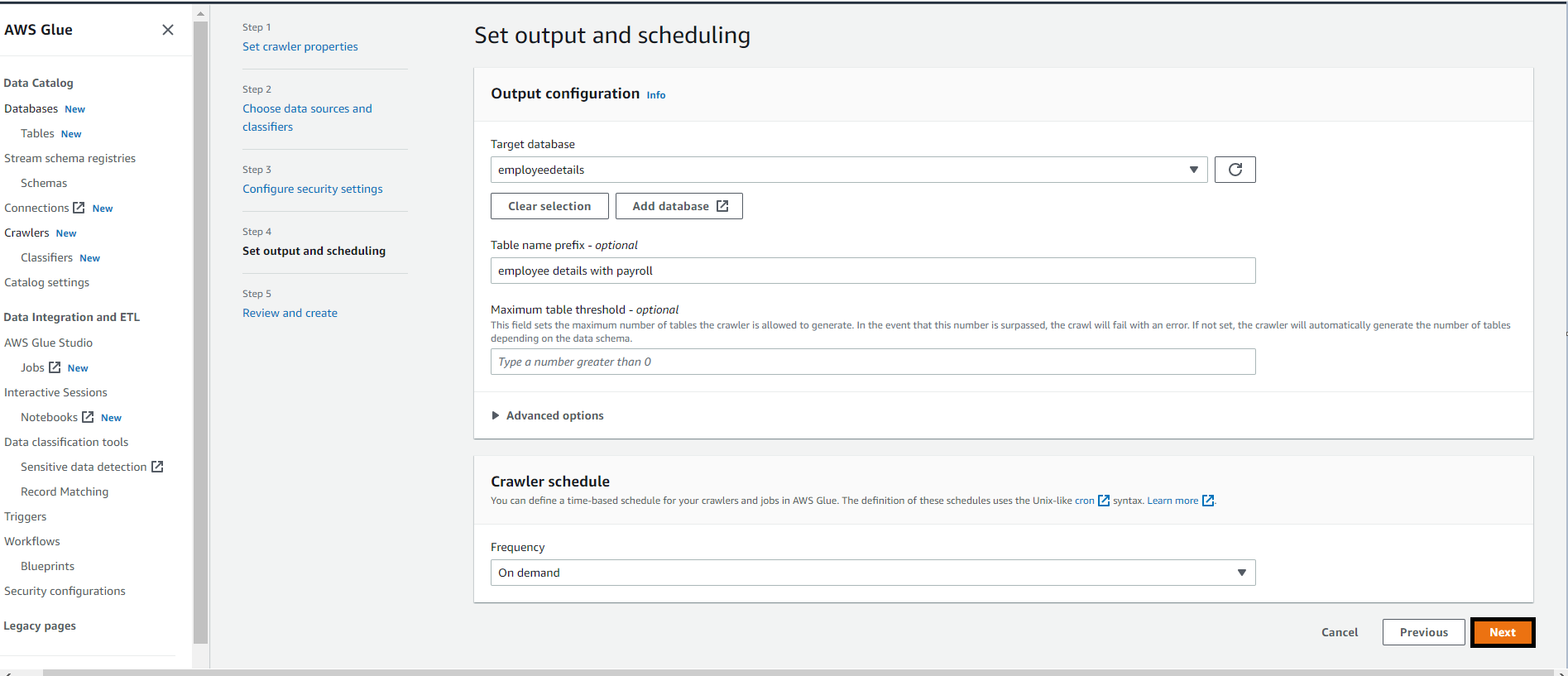
Step 7:- Now it will ask where you want to store your crawler output. For that you need to create new database/existing database to your crawler for it so click on **next** option now.



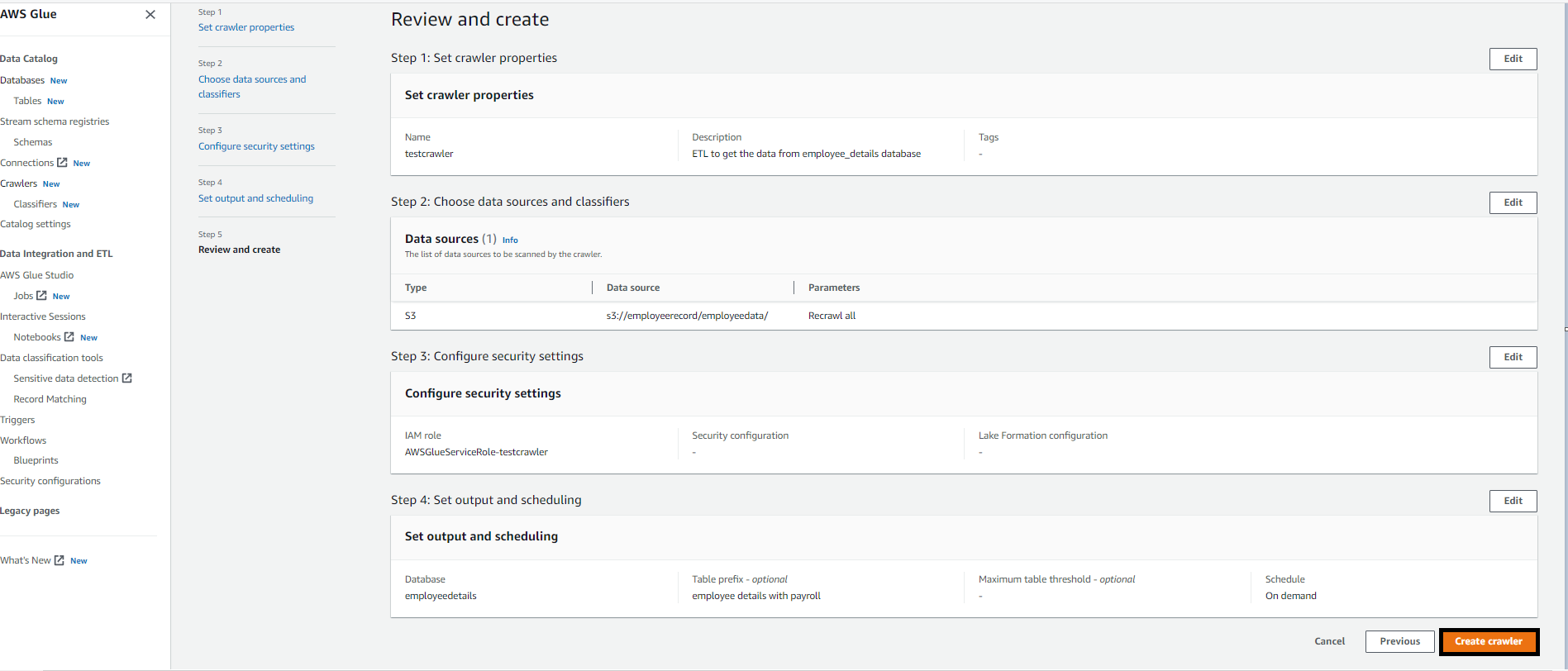
Step 8:- Add the database name as per your preference and keep all other field left as it is or you can enter the details and click on create database option.



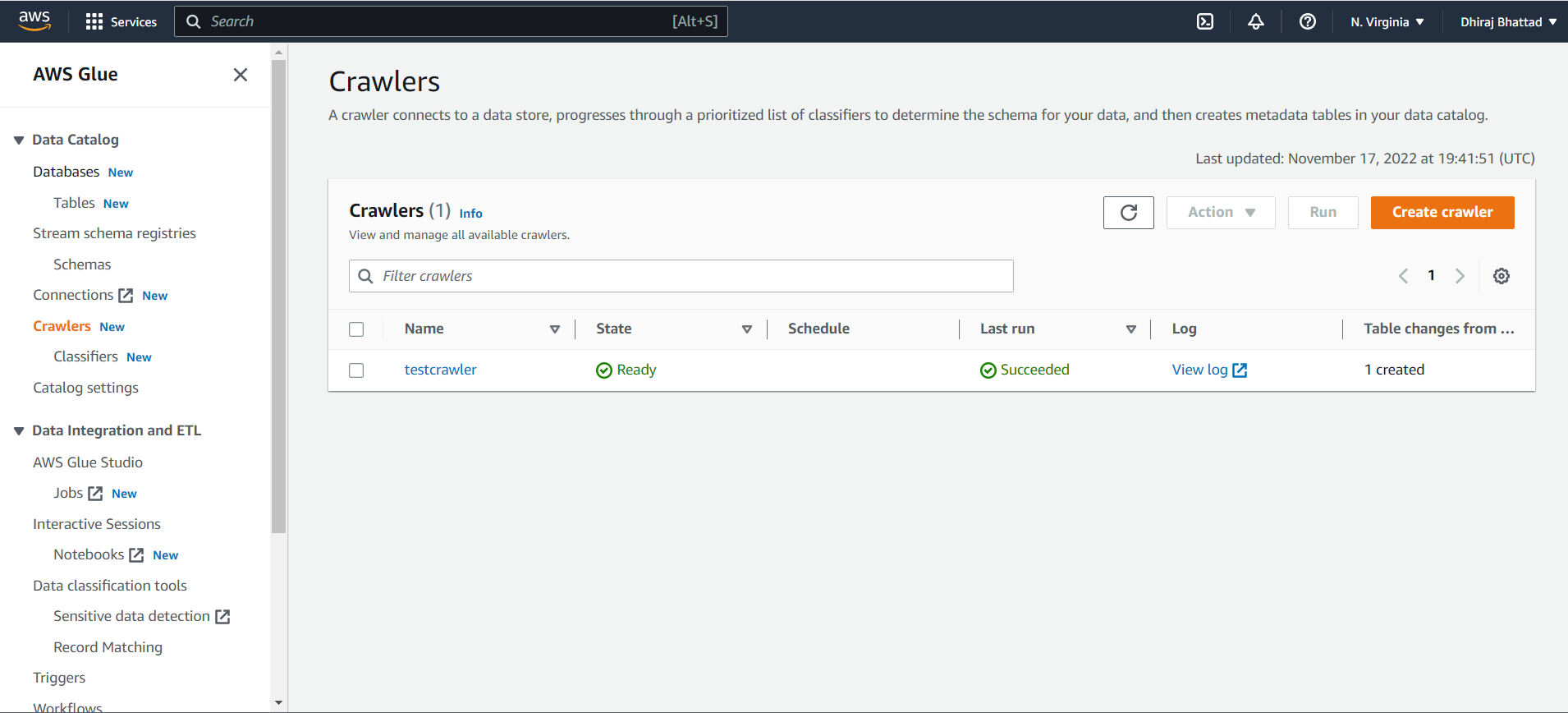
Step 9:- So your database has been created and it is selected by default. Also, next it asks for prefix which is optional. If you want to add the details into it you can or else click on **next** option below.



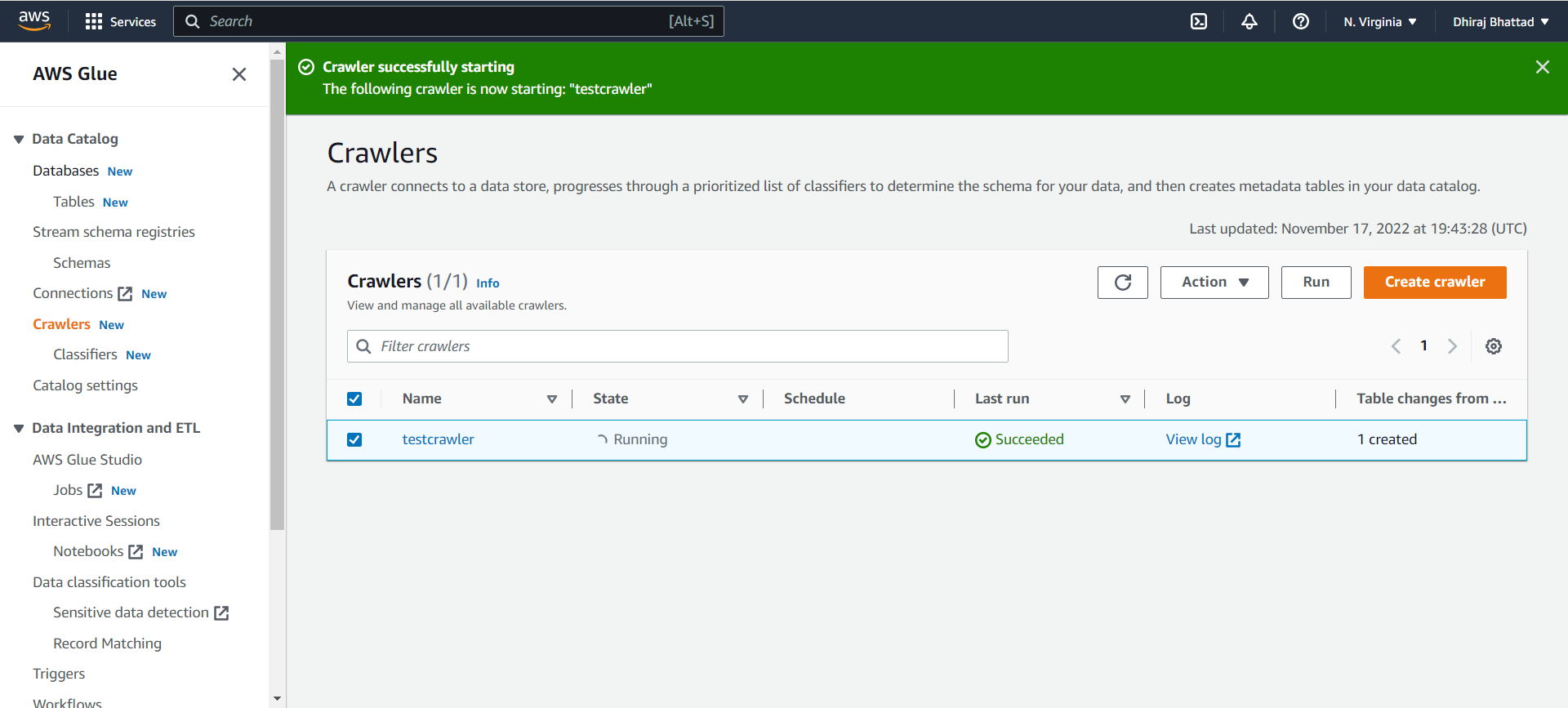
Step-10:- Now you can review all your selections we have done so far in order to create the crawler as shown below after that click on create crawler option.



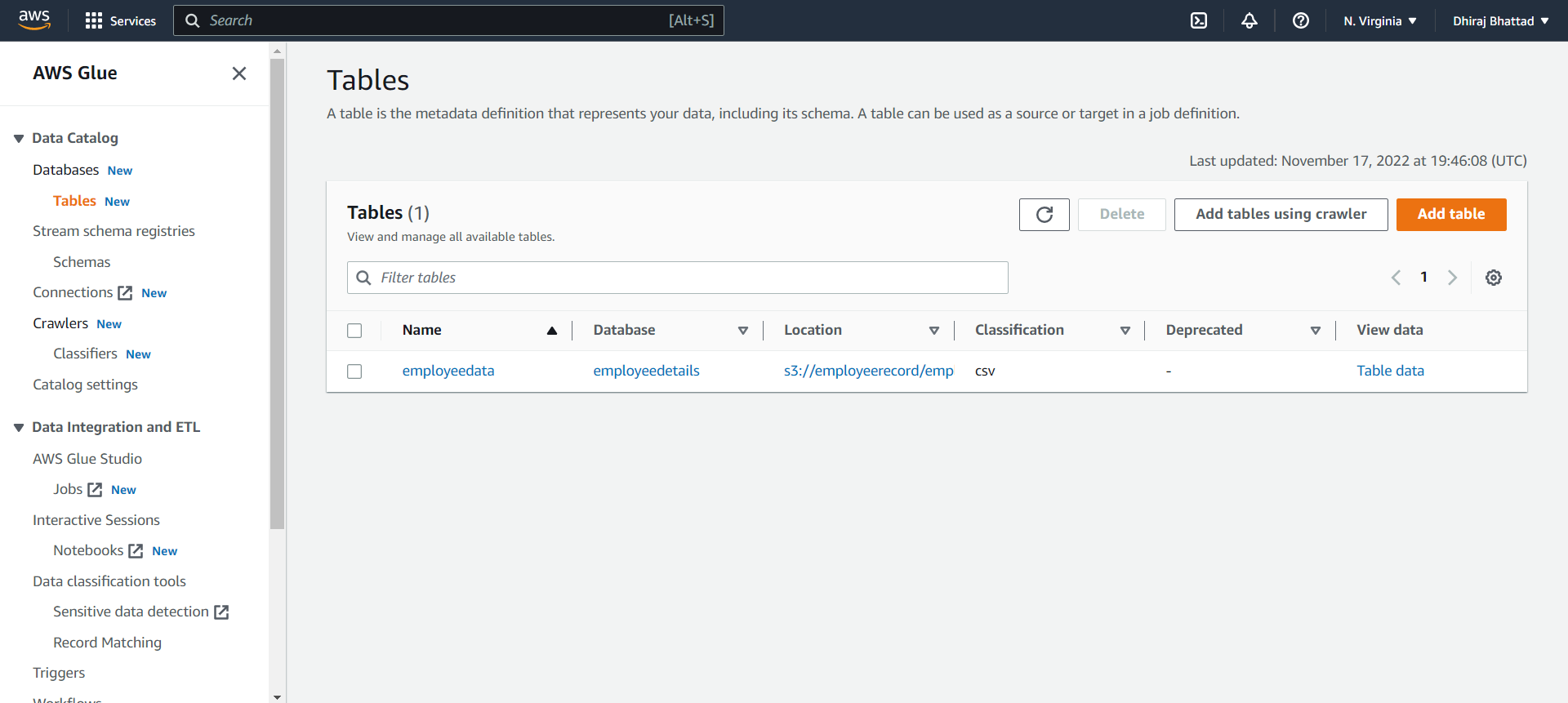
Step-11:- So, now here your crawler has been created successfully. Now we need to run this selected crawler by clicking on the option **Run** crawler.



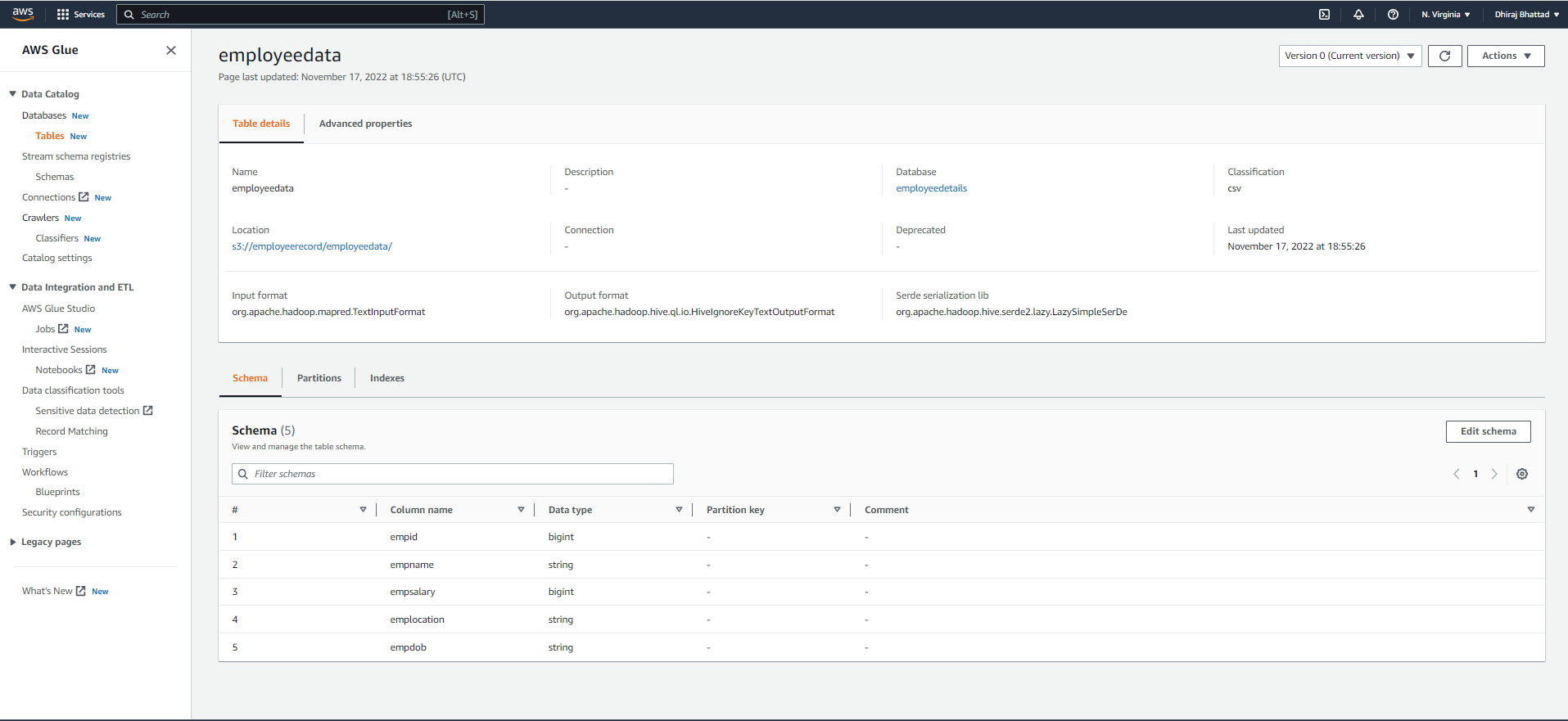
Step-12:-So, as you crawler is running and in the backend crawler will go to storage path of s3 and it will bring the data set and it will also automatically create schema and store it’s metadata in that database as we just created for this crawler.



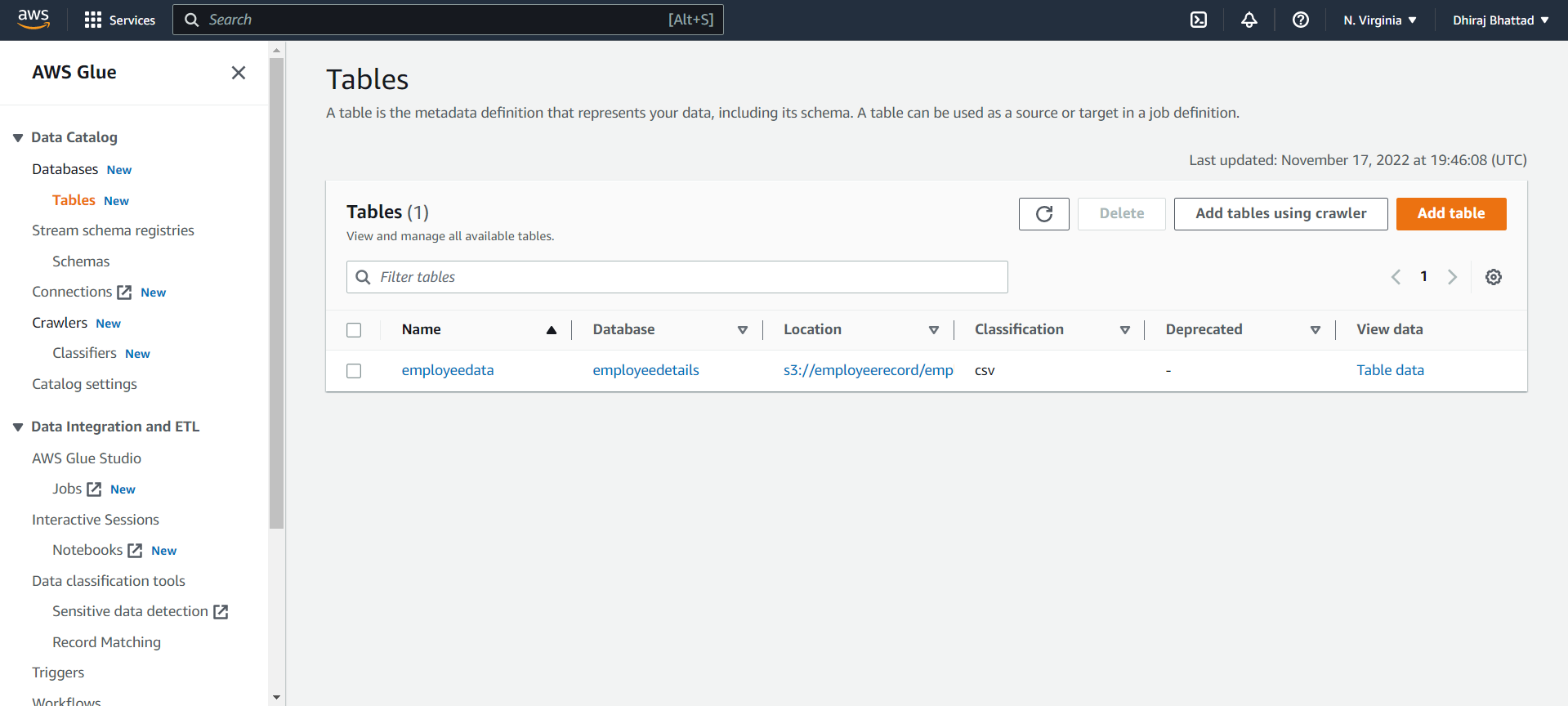
Step 13:-Now our table is created and it will show in database as shown below for that you need to go into table in left side of option under **AWS glue**.



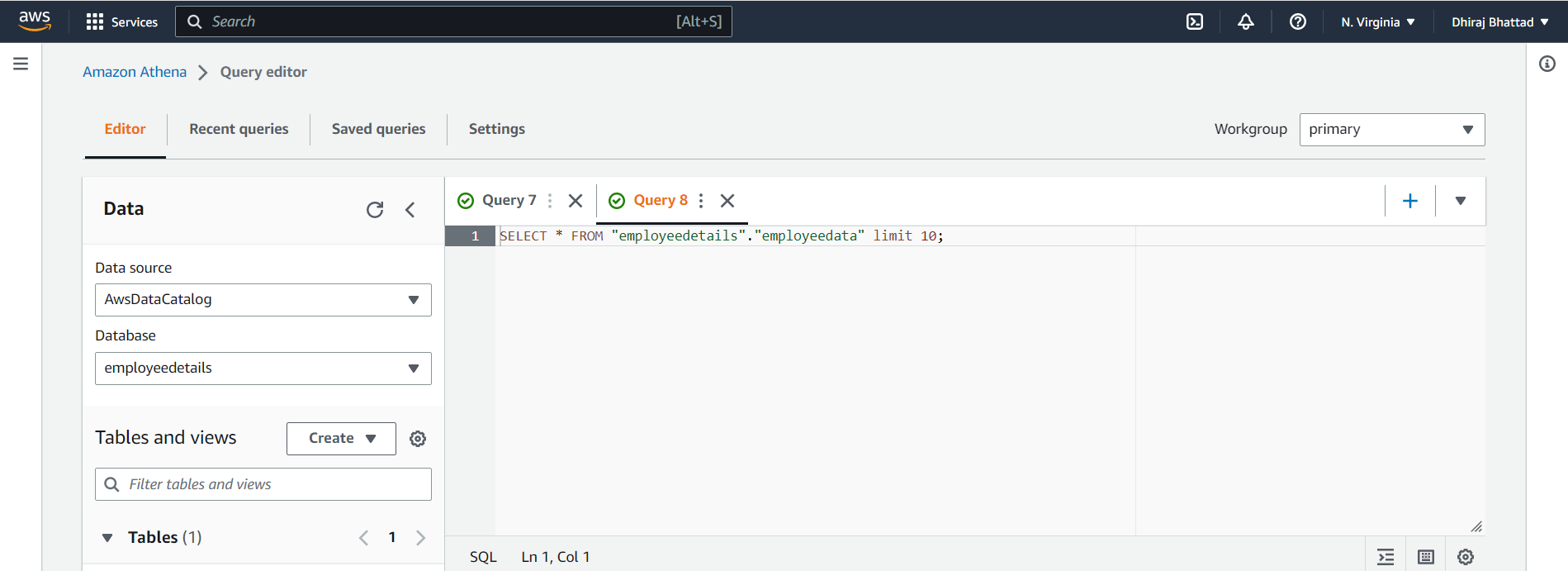
Step 14:- By clicking on the created table, you will see the details of the same along with schema created by this **crawler**.



Step 15:-Now our table is ready so now we can query it directly by going into **AWS Athena** service.



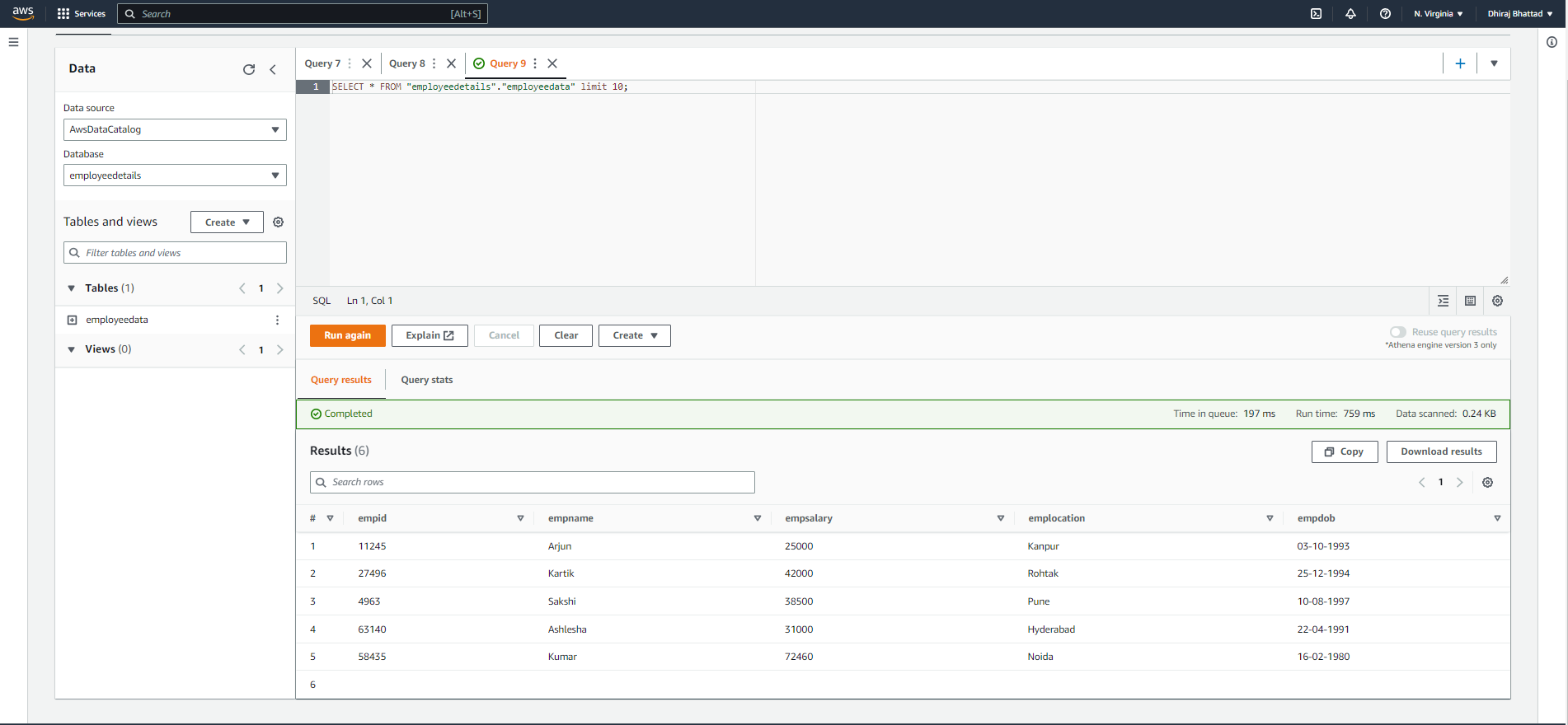
Step 16:- Now go to AWS Athena dashboard.



So, you can directly go to preview table from left side of table created in **Athena** service

You can see the query result below.

**Output**



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_