

## IF184966 - Big Data

Academic Year: 2022/2023

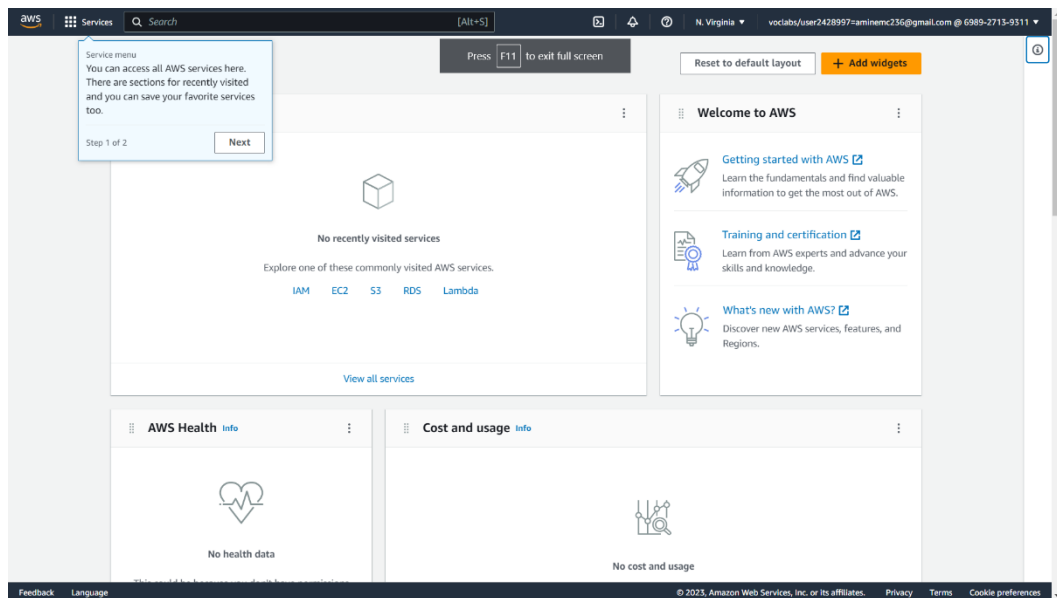
Student's ID : 5025201251

Student's Name : Muhammad Amin

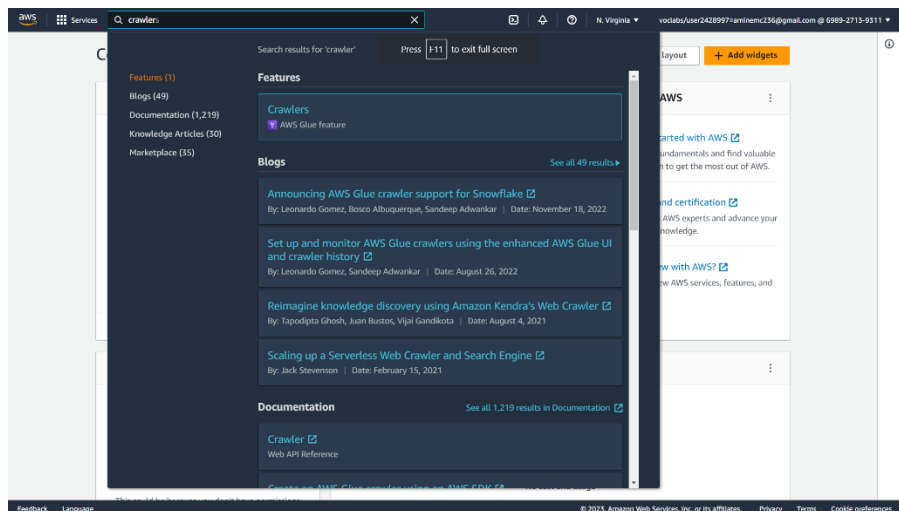
### Task 1

- Do the task on **AWS Academy Data Analytics [39663]: Lab 3 - Query data in Amazon S3 with Amazon Athena and AWS Glue.**
- Make documentation for each step and give a screenshot for each of them. Every screenshot should display your account in the top right corner.

This is the dashboard



search crawlers



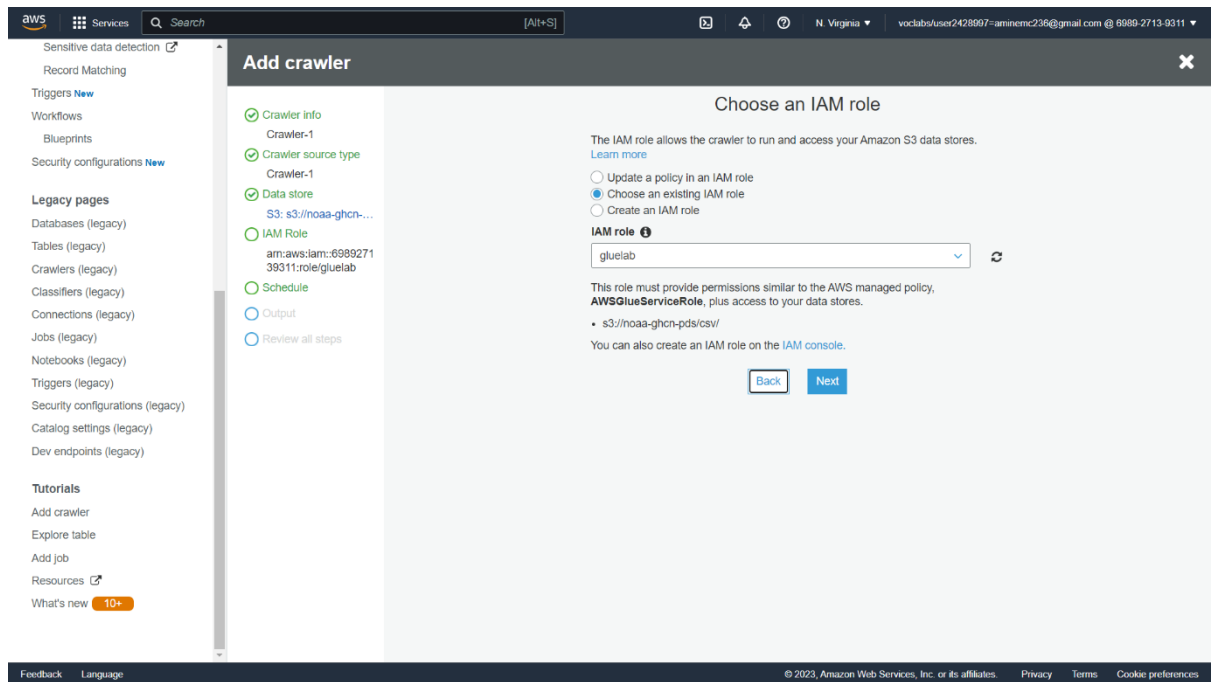
Press the add crawler button to add crawler

The screenshot shows the AWS Crawler console interface. On the left is a navigation menu with categories like 'Sensitive data detection', 'Legacy pages', and 'Tutorials'. The 'Crawlers (legacy)' link is highlighted. The main content area has a header with a notification about a new console experience. Below this, a description of crawlers is provided. A 'User preferences' link is on the right. A toolbar contains an 'Add crawler' button, a 'Run crawler' button, an 'Action' dropdown, and a search filter. Below the toolbar is a table with columns: Name, Schedule, Status, Logs, Last runtime, Median runtime, Tables updated, and Tables added. The table is currently empty and shows a 'Loading' status.

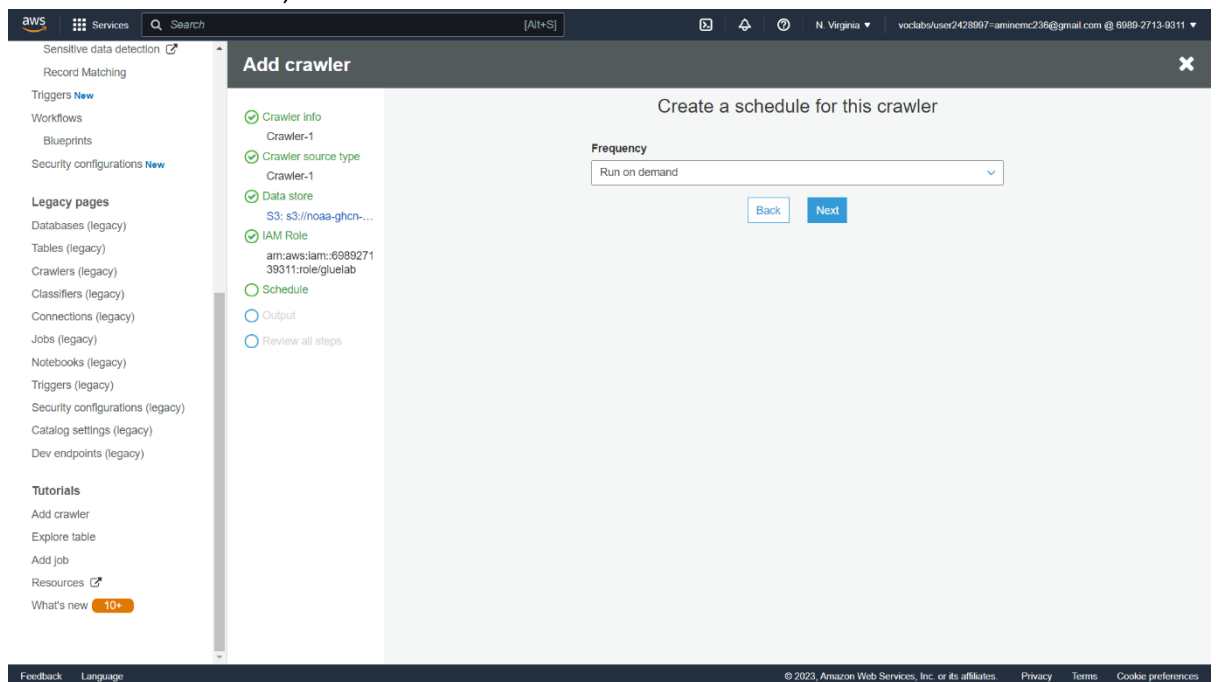
Choose No

The screenshot shows the 'Add crawler' wizard in the AWS Crawler console. The wizard is titled 'Add crawler' and has a close button. It is divided into three main sections. The left section, 'Crawler info', contains a list of steps: 'Crawler info', 'Crawler source type', 'Data store', 'IAM Role', 'Schedule', 'Output', and 'Review all steps'. The 'Crawler info' and 'Crawler source type' steps are marked as complete with green checkmarks. The 'Data store' step is currently active, showing 'S3: s3://noaa-ghcn-...'. The middle section, 'Add another data store', has two radio buttons: 'Yes' and 'No'. The 'No' option is selected. Below the radio buttons are 'Back' and 'Next' buttons. The right section, 'Chosen data stores', shows a list of selected data stores, currently containing 'S3: s3://noaa-ghcn-p...'. The bottom of the screen shows the AWS footer with copyright information and links for Privacy, Terms, and Cookie preferences.

Then choose an existing IAM role and choose gluelab



Just choose the default, Click next



In the configuration, set a database name then click next

The screenshot shows the 'Add crawler' configuration page in the AWS Glue console. The left sidebar contains navigation links for Sensitive data detection, Record Matching, Triggers, Workflows, Blueprints, Security configurations, Legacy pages, Databases, Tables, Crawlers, Classifiers, Connections, Jobs, Notebooks, Triggers, Security configurations, Catalog settings, Dev endpoints, Tutorials, Add crawler, Explore table, Add job, Resources, and What's new. The main content area is titled 'Add crawler' and 'Configure the crawler's output'. It includes a progress bar with steps: Crawler info, Crawler source type, Data store, IAM Role, Schedule, Output, and Review all steps. The 'Output' step is currently active. The configuration fields are: Database (dropdown menu with 'crawler-1-database' selected), Prefix added to tables (optional) (text input), Table threshold (optional) (text input), and a 'Table threshold' description. There are also expandable sections for 'Grouping behavior for S3 data (optional)' and 'Configuration options (optional)'. At the bottom are 'Back' and 'Next' buttons.

**Add crawler**

**Configure the crawler's output**

**Database**

crawler-1-database

**Prefix added to tables (optional)**

Type a prefix added to table names

**Table threshold (optional)**

Enter a number greater than 0

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.

► Grouping behavior for S3 data (optional)

► Configuration options (optional)

Back Next

Review the crawler information

The screenshot shows the 'Add crawler' review page in the AWS Glue console. The left sidebar is the same as the previous screenshot. The main content area is titled 'Add crawler' and 'Review the crawler information'. It displays a summary of the crawler configuration in a card-based layout. The cards are: Crawler info (Name: Crawler-1, Tags: -), Data stores (Data store: S3, Include path: s3://noaa-ghcn-pds/csv/, Connection: -, Exclude patterns: -), IAM role (IAM role: arn:aws:iam::698927139311:role/gluelab), Schedule (Schedule: Run on demand), and Output (Database: crawler-1-database, Prefix added to tables (optional): -, Table threshold (optional): false, Create a single schema for each S3 path: false, Table level (optional): false). At the bottom is a 'Configuration options' link.

**Add crawler**

**Review the crawler information**

**Crawler info**

Name Crawler-1

Tags -

**Data stores**

Data store S3

Include path s3://noaa-ghcn-pds/csv/

Connection -

Exclude patterns -

**IAM role**

IAM role arn:aws:iam::698927139311:role/gluelab

**Schedule**

Schedule Run on demand

**Output**

Database crawler-1-database

Prefix added to tables (optional) -

Table threshold (optional) false

Create a single schema for each S3 path false

Table level (optional) false

► Configuration options

The crawler is ready to use

**AWS Glue**

**Data Catalog**

- Databases
- Tables
- Stream schema registries
- Schemas
- Connections [↗](#)
- Crawlers
- Classifiers
- Catalog settings **New**

**Data Integration and ETL**

- AWS Glue Studio
- Jobs [↗](#)
- Interactive Sessions
- Notebooks [↗](#)
- Data classification tools
- Sensitive data detection [↗](#)
- Record Matching
- Triggers **New**
- Workflows
- Blueprints
- Security configurations **New**

**Legacy pages**

- Databases (legacy)
- Tables (legacy)

[Feedback](#) [Language](#)

**New console experience for AWS Crawlers available!**  
We've redesigned the AWS Crawlers console to make it easier to use. [Switch to the new console.](#)

**Crawlers** A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

[Add crawler](#) [Run crawler](#) [Action](#)  Showing: 1 - 1 [User preferences](#)

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	<a href="#">Crawler-1</a>		Ready		0 secs	0 secs	0	0

© 2023, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

## Run the Crawler

**AWS Glue**

**Data Catalog**

- Databases
- Tables
- Stream schema registries
- Schemas
- Connections [↗](#)
- Crawlers
- Classifiers
- Catalog settings **New**

**Data Integration and ETL**

- AWS Glue Studio
- Jobs [↗](#)
- Interactive Sessions
- Notebooks [↗](#)
- Data classification tools
- Sensitive data detection [↗](#)
- Record Matching
- Triggers **New**
- Workflows
- Blueprints
- Security configurations **New**

**Legacy pages**

- Databases (legacy)
- Tables (legacy)

[Feedback](#) [Language](#)

**New console experience for AWS Crawlers available!**  
We've redesigned the AWS Crawlers console to make it easier to use. [Switch to the new console.](#)

**Crawlers** A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

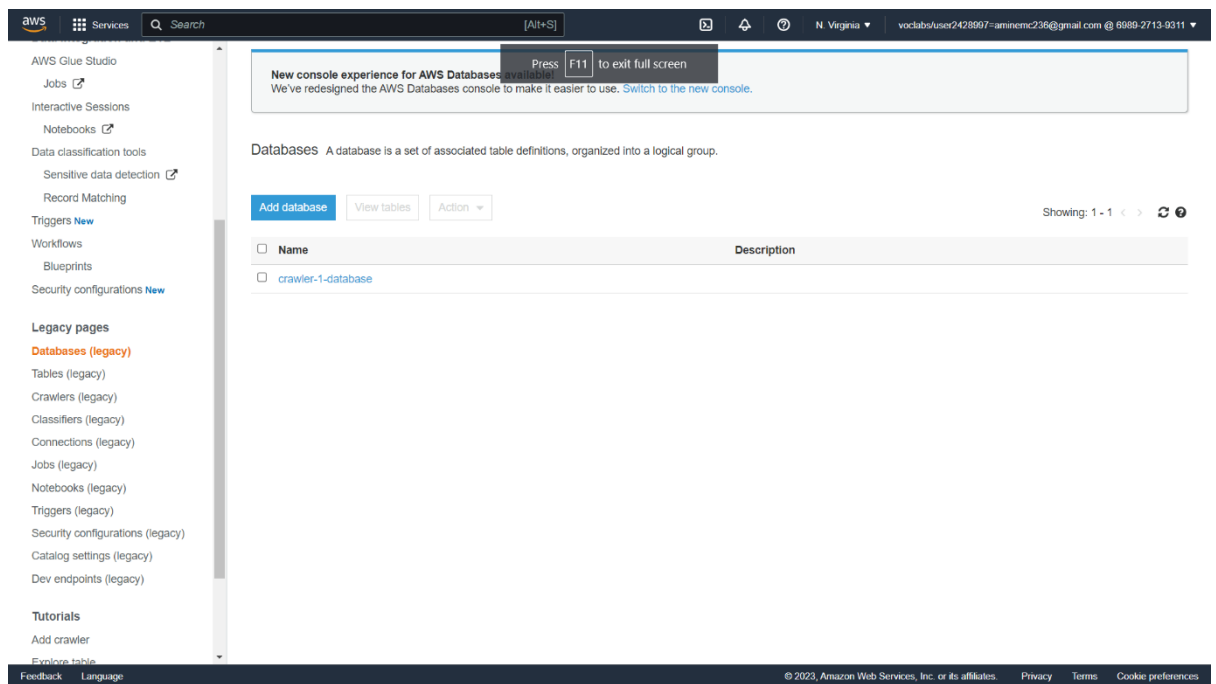
Crawler "Crawler-1" is now running.

[Add crawler](#) [Run crawler](#) [Action](#)  Showing: 1 - 1 [User preferences](#)

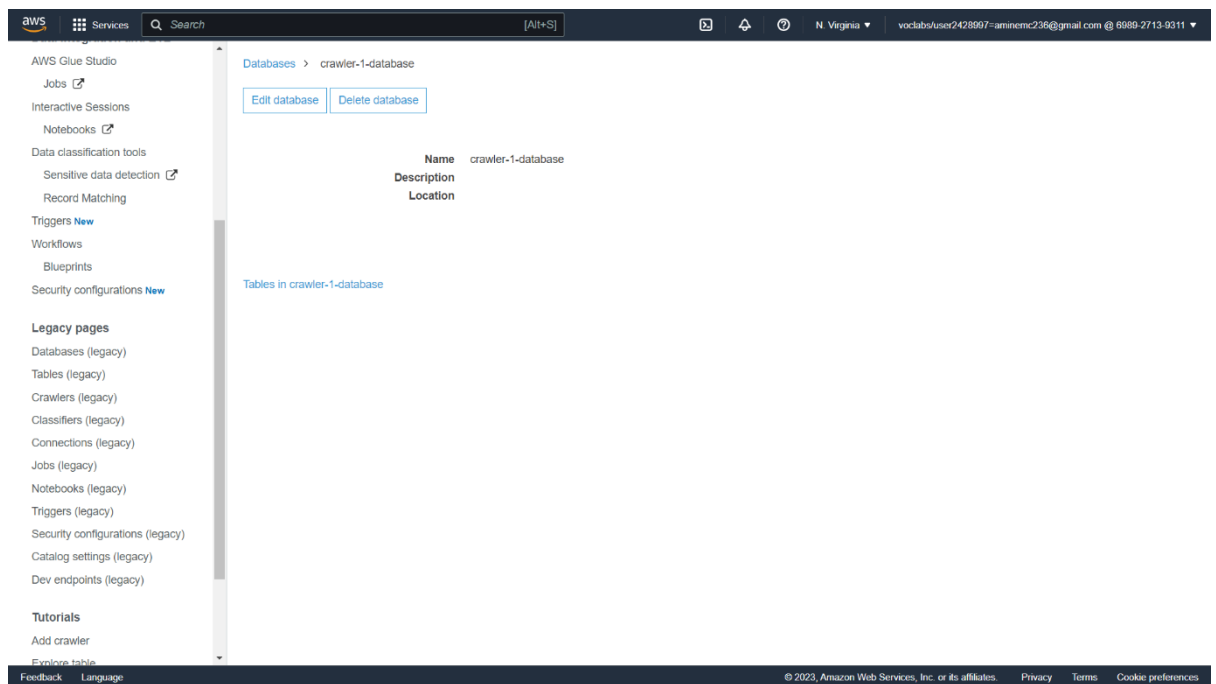
<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
<input type="checkbox"/>	<a href="#">Crawler-1</a>		<a href="#">↻</a> Starting		0 secs	0 secs	0	0

© 2023, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

Go to database (legacy) to see the database



After that try to edit the database



## Choose csv tables

New console experience for AWS Tables available!  
We've redesigned the AWS Tables console to make it easier to use. [Switch to the new console.](#)

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.

Add tables Action Database: crawler-1-database Filter or search for tables. Save view Showing: 1 - 1

Name	Database	Location	Classification	Last updated	Deprecated
csv	crawler-1-database	s3://noaa-ghcn-pds/csv/	csv	23 February 2023 11:24 AM U...	

Feedback Language © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

## Review the metadata of the table

Name csv

Description

Database crawler-1-database

Classification csv

Location s3://noaa-ghcn-pds/csv/

Connection

Deprecated No

Last updated Thu Feb 23 23:24:15 GMT+700 2023

Input format org.apache.hadoop.mapred.TextInputFormat

Output format org.apache.hadoop.hive.q1.io.HiveIgnoreKeyTextOutputFormat

Serde serialization lib org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe

Serde parameters field.delim , skip.header.line.count 1 sizeKey 210967072764 objectCount 123913 UPDATED\_BY\_CRAWLER Crawler-1

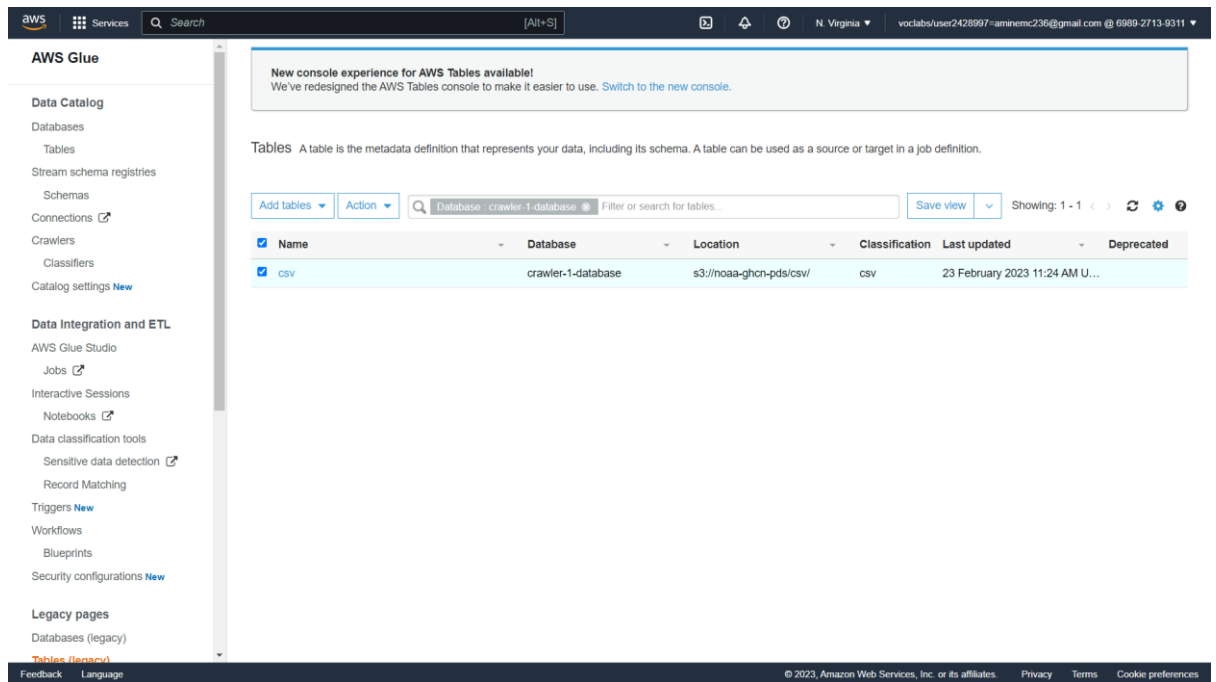
Table properties CrawlerSchemaSerializerVersion 1.0 recordCount 5156847131 averageRecordSize 40 CrawlerSchemaDeserializerVersion 1.0 compressionType none columnsOrdered true areColumnsQuoted false delimiter , typeOfData file

Schema Showing: 1 - 9 of 9

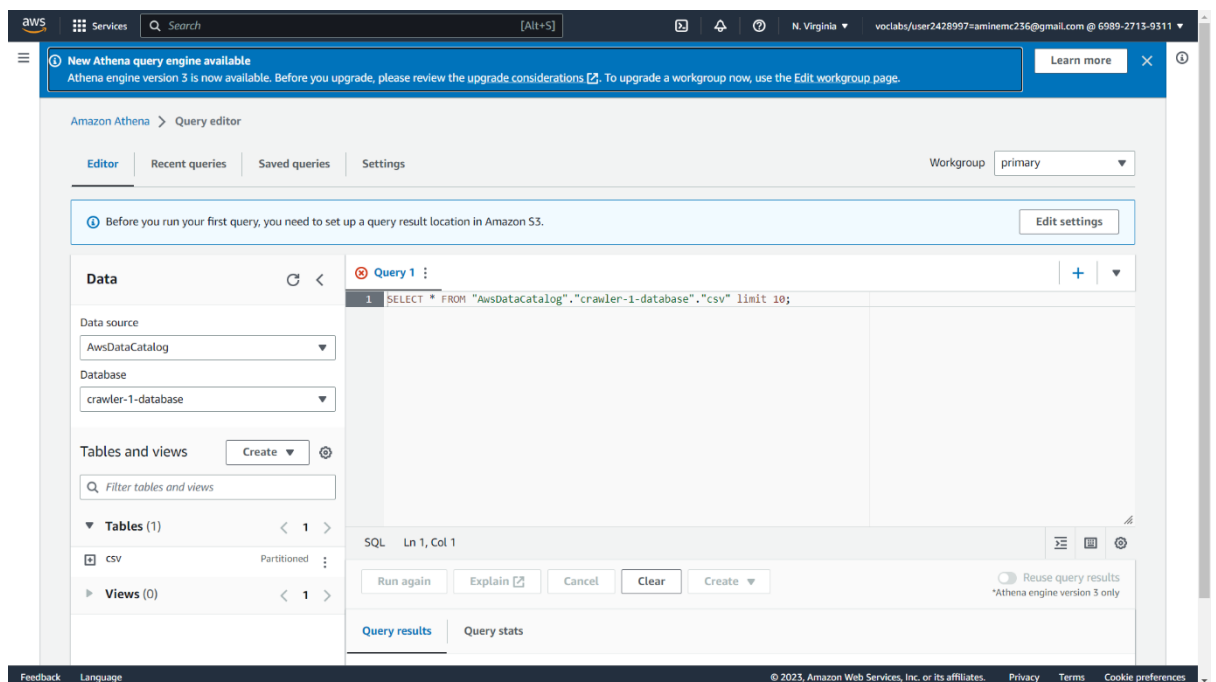
	Column name	Data type	Partition key	Comment
1	id	string		
2	date	bigint		
3	element	string		
4	data_value	bigint		
5	m_flag	string		
6	q_flag	string		

Feedback Language © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

In table list, choose csv and click action to view data

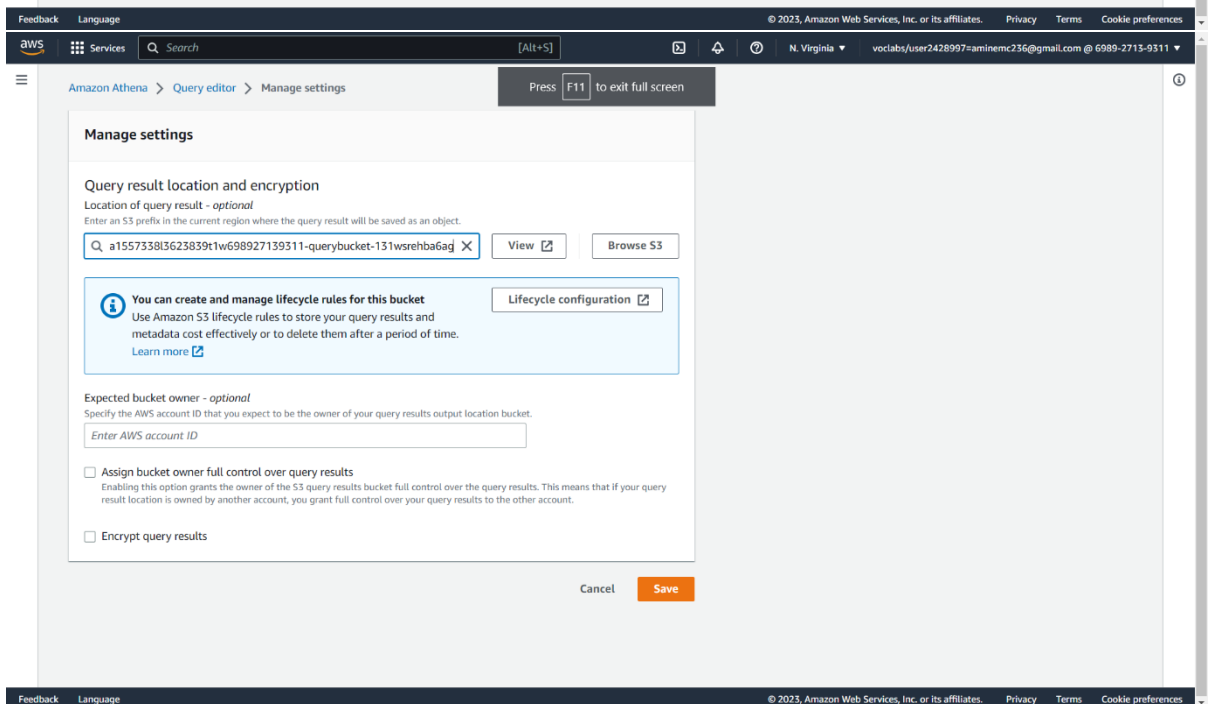
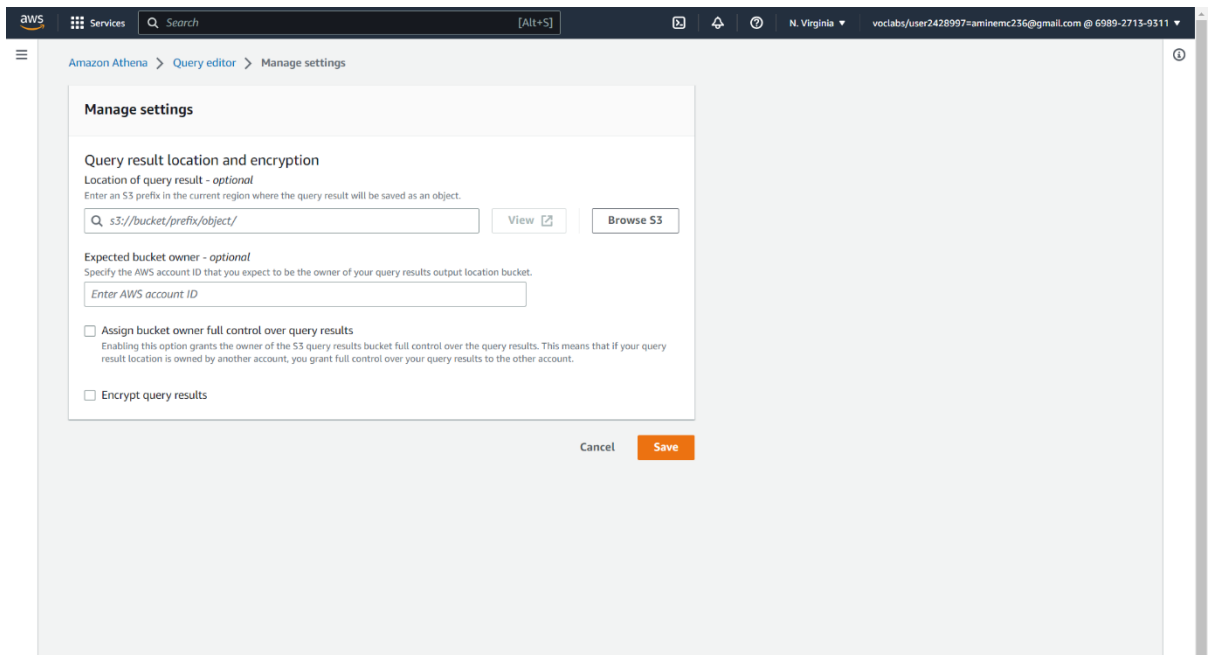


Now we are inside of the athena, before running a query, we need to set where to save the result (called bucket) just click edit settings



Choose a bucket





Run the query

Query results

Completed Time in queue: 188 ms Run time: 1.385 sec Data scanned: 369.81 KB

Results (10)

Search rows

#	id	date	element	data_value	m_flag	q_flag	s_flag	obs_time	partition
1	EZE00100082	17750101	TMAX	-48			E		by_year
2	EZE00100082	17750101	TMIN	-101			E		by_year
3	ITE00100554	17750101	TMAX	-26			E		by_year
4	ITE00100554	17750101	TMIN	-46			E		by_year
5	EZE00100082	17750102	TMAX	-14			E		by_year
6	EZE00100082	17750102	TMIN	-48			E		by_year
7	ITE00100554	17750102	TMAX	-13			E		by_year
8	ITE00100554	17750102	TMIN	-43			E		by_year
9	EZE00100082	17750103	TMAX	6			E		by_year
10	EZE00100082	17750103	TMIN	-22			E		by_year

## Task 2 (Map Reduce using Python)

- Redo the programming example from here:
  - [https://icaml.org/canon/basics/mapreduce\\_wordcount\\_python.html](https://icaml.org/canon/basics/mapreduce_wordcount_python.html)
  - [https://colab.research.google.com/drive/1blwHxoV55wHo11bqj2VTOJkZy5lyz\\_el?usp=sharing](https://colab.research.google.com/drive/1blwHxoV55wHo11bqj2VTOJkZy5lyz_el?usp=sharing)
- Upload/commit your code on your GitHub.
- Paste the GitHub link here: ...

<https://github.com/Aminemcc/Big-Data/tree/master/Week%202>

Collect this template as PDF file.