Justin Slade

G00409678

CS 435

Fall 2024

Big Data & Cloud Security

AWS Lab Report Capstone

Date: 12/03/2024

Grambling State University

Big Data and Cloud Security

**Purpose or Objective:** Learn how to utilize everything that was learned prior to completing this scenario for this assignment.

**Procedure/s:**

* Create infrastructure to host info on fish
* Create AWS Glue for database and tables
* Store data on Amazon S3
* Create an Athena to interact with cloud9 IDE
* Be able to query said information from S3 buckets

**Task 1.**

I will first to create an environment for my capstone. This will require an IAM and an EC2 minimum and roles to go along with it.

A screenshot of a computer

Description automatically generated

This is me creating an environment with Cloud9 which evokes an EC2 instance to be created. Doing this all within the VPC.

I also went and named 2 buckets named query-results-23451 and data-source-2222 respectively.  
A screenshot of a computer

Description automatically generated

Then I need to download the necessary files where im going to be taking data from.

A screenshot of a computer

Description automatically generated

SAU-GLOBAL-1-v48-0.csv

SAU-HighSeas-71-v48-0.csv

SAU-EEZ-242-v48-0.csv

These are the files I’ve downloaded the following is from SAU-GLOBAL data set

A screen shot of a computer

Description automatically generated

Showing the year, Who did it, Which sector, IF reported and ect.  
Lastly I needed to upload this data to my S3 bucket that I made earlier this task.



This is effectively saying I want to tell AWS to copy what I just downloaded on cloud9 and paste it within the specified bucket.

**Task 2.**

Now I need to set up the automation of AWS Glue Crawler and Querying multiple files with Athena.

First, I needed to see if I could see this information.

A computer screen with white text

Description automatically generated

This command I ran allowed me to see 5 different elements within this table. I can change the amount of elements provided by changed the -5 to any value -0.

A computer screen with white text

Description automatically generated

I also needed to convert the *SAU-HighSeas-71-v48-0.csv* file into a parquet format and upload this to the S3 bucket I made prior. It took me a while to realize I was just repeating the previous steps since this was a different file an all, but I was able to figure out it was just the same.

Now I wanted to create a database and a Crawler using AWS Glue, so I can create my tables properly.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Then I needed to configure Athena to take my findings and put in into my 2nd bucket I made.

A screenshot of a computer error

Description automatically generated

Now that I’ve done this, I ran a simple query and got this to pop up.

A screenshot of a computer

Description automatically generated  
Now they want my to query the data base to get information on Fiji and there revenue on years since 2001. So using the code provided and filling in the necessary information I was able to get these results.

A screenshot of a computer code

Description automatically generated

This was the code needed to produce the following. (needed to change From command and the specified year to 2000)

A screenshot of a computer

Description automatically generated

This gives me the year, country, and Value not in USD. So, I need to prompt further to get a USD conversion.  
Changing the Value to ValueAllHighSeasCatch and area\_name to “is null”

A screenshot of a computer

Description automatically generated

We get the following results.

A screenshot of a computer

Description automatically generated  
This change changed the value from whatever it was to USD.

**Task 3.**

Add a new file (table) with different names for the same values. Transform said file into proper notation and add it to the data set.   
Ok so I couldn’t go step by step with my writing and process but I’m able to consolidate the steps since it did get a bit involved. Here is the following code I needed to up in. (most was given to me from the assignment, but I needed to fill in the blanks in some areas)

A computer screen with white text

Description automatically generated

This starts off with me having to go back to cloud9 IDE and following some code. I need to get the last data set I haven’t used since the start of the capstone. I print the head to see anything different. I don’t see anything different but the table that was given to me by the lab showed that two names in the columns of those tables a labeled incorrectly. This was why this line was important

df.rename(columns = {"fish\_name": "common\_name", "country": "fishing\_entity"}, inplace = True)  
This line renames the two columns to names that are already established within my other two tables. After I’ve done this I go back into the process of converting this file into a parquet file to be used in further querying.

Now I need to upload this to my bucket that contains my other tables. Not the results bucket.

A screen shot of a computer

Description automatically generated

Showing some messed up code because I can’t spell “environment” correctly… but I was able to upload it to my s3 bucket.

A screenshot of a computer

Description automatically generated

After this I now need to get my AWS Glue crawler up and running again to parse the new information with the table.

A screenshot of a computer

Description automatically generated

Now that this is done, I’m now able to use Athena to run a query to test if I did things properly.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThis Is good because it has given me 3 results instead of 2, so I can be certain the last dataset is now accessible via Athena.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThese 3 pictures prior are just the result of the 3 queries I put in. Effectively what is happening here is that I have the first two images be two separate entities with their own values. The third photo is just the combination of the two prior photos

Now that these datasets are compatible with each other I can now use them in tandem with each other. Which give a bunch of information about mackerels and the amount of them caught measured by the ton from each country. Fun fact USA caught 800 tons of it in 2015A screenshot of a computer

Description automatically generated

This is the last thing I needed to do for the capstone and is the end of this lab.

**Conclusions:**

This lab was actually very involved and required much more time than I was ready to allocate to such an assignment. This was more up my alley as I am a bit more acquainted with databases, but it was fun utilizing Cloud9 and Crawler more extensively than I have previously used in the past. It was fun for me at least seeing everything just work when I wasn’t I figured out the problems that the capstone brought upon me. This is probably the most well-made assignment on AWS academy because they gave me all the access I needed to do a bit more unconventional methods to get the job done.   
A screenshot of a score

Description automatically generated

**Presentation. 30mins**

[**https://youtu.be/nDlyIna-5j8**](https://youtu.be/nDlyIna-5j8)