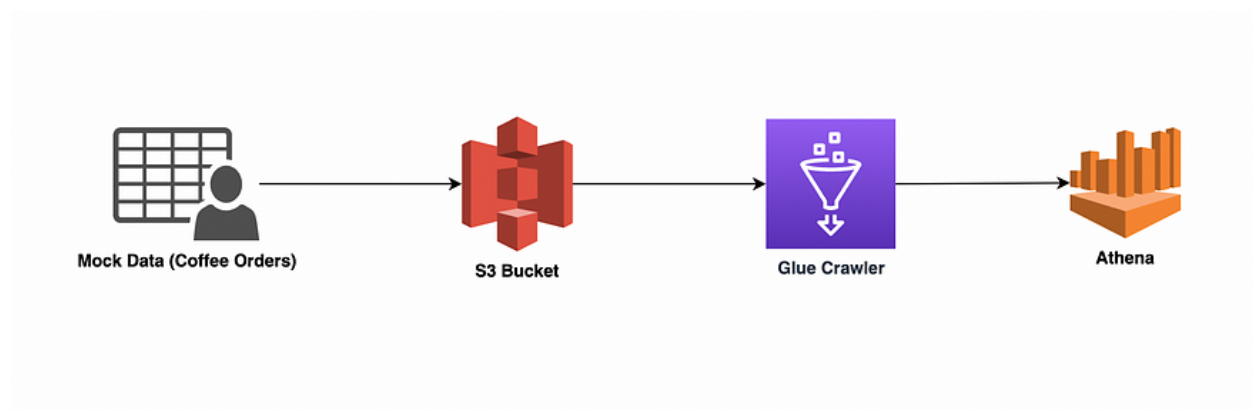# Create an ETL Job using AWS Glue Studio, S3, & Athena

Background

In this blog post, I will walk us through on how to create an ETL job using the following AWS Cloud Services:

- **Amazon S3** *(for intermediary data storage)*
- **AWS Glue Studio** *(main engine of the ETL job)*
- **Amazon Athena** *(query tool)*

ETL stands for **Extract**, **Transform**, and **Load**.

In almost all Enterprise IT settings, it is very common to have huge volumes of Analytical data. In order for that data to be consumable and used by business areas to generate insights and forecast business goals, the data needs to be:

1. Extracted from where it is stored/originated,

2. Transformed to fit a certain format (usually involves cleaning the data, removing duplicates, etc.)
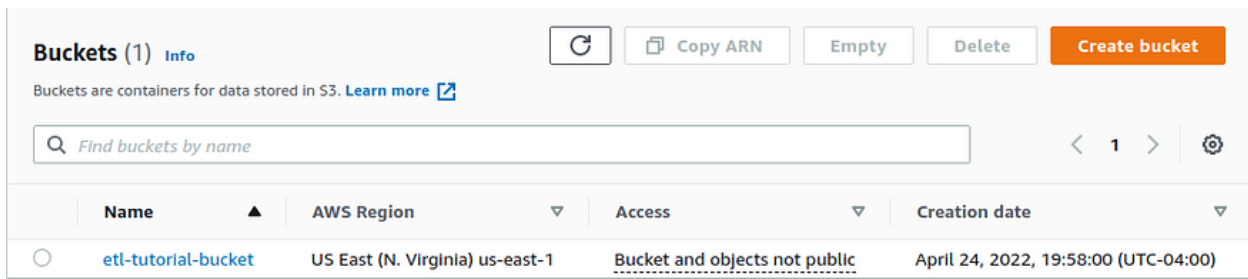
and then,

3. Loaded into the target destination source.

Very often, there is also the presence of a large data warehouse, and

multiple databases/storage layers feeding data into the larger

warehouse for general business consumption.

Without further ado, we will now create an ETL Job using AWS Glue

(an event-driven, serverless computing service).

Step 1 — Create an S3 Bucket

This S3 Bucket will store our data. We're basically just building a

database within the S3 Object storage service.

S3 Bucket created (keep default options when creating it)

Step 2 — Create some Mock Data

For this step, I just created a quick .csv file with data about coffee

orders.

This data file will be uploaded to the etl-tutorial-bucket from the

previous step.

	A	B	C	D	E	F
1	order_id	coffee	cust_name	drive_thru	walk_in	
2	1	Cappuccino	Adriana	N	Y	
3	2	Latte	Riya	Y	N	
4	3	Espresso	Wahab	Y	N	
5	4	Cocoa	Hardy	N	Y	
6	5	Frappe	Molly	N	Y	
7	6	Plain	Ryan	N	Y	
8						

coffee.csv

Step 3 — Upload the coffee data file to the S3 Bucket

coffee.csv uploaded to S3

Step 4 — Create this IAM Role for Glue with the Required permissions

Step 5— Define the Glue Data Crawler in AWS Glue

AWS Glue was launched in 2017. It is a fully-managed, serverless, and AWS Cloud-optimized ETL Service offering.
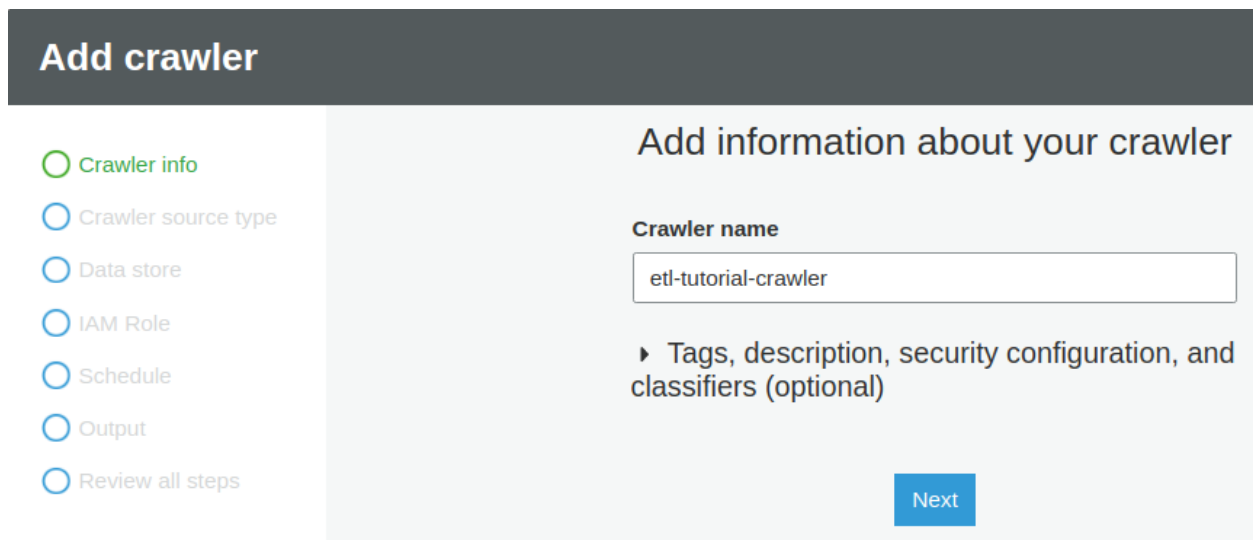
Using Glue, there are numerous ways we can connect to our data store in S3. However, for this tutorial, I am going to demo this using Glue Crawler.

Glue Crawler is one of the most widely-used method among Data Engineers.

Basically, this Crawler will "crawl" the data file in our S3 bucket, and using the data available in the file in S3 will create a table schema in Glue.

To create and add the Crawler to our S3 data store:

5a. Gave my crawler the name `etl-tutorial-crawler` then click **Next**.

5b. You can keep the default settings as is for the Crawler. Click **Next**.

5c. Select **S3** for the **data store**, and add the name of your S3 bucket.

The click **Next**.

5d. Select the **Choose an existing IAM Role** radio button, and

select the IAM Role we created in Step 4. Then click **Next**.

5e. Set the frequency to **Run on Demand**. Then click **Next**. (Note: if required, you can always set customized scheduling for your Crawler to run on an hourly, daily, weekly, o monthly cadence).

5f. For storing the output from the Crawler, I created a database `coffee-database` and then clicked **Next**.

Add database

Database name

coffee-database

▸ Description and location (optional)

Resource link name

Enter resource link name

Shared database suggestions

Choose a database to autofill form ⌄

Shared database

Enter database to link to

Shared database owner account ID

Enter an AWS account ID

Create

5g. Lastly, review all the settings that we've configured and click

Finish to create our Crawler.

5h. Once the Crawler has been created, we should see this:

Try running the Crawler. If the job status updates from starting to

stopping, and then to ready, that tells that the Crawler job was

successful.

Step 6 — Defining our Glue Job

Now that we have finished setting up the Glue Crawler to point and

recognize the data file in S3, we will define our Glue job to perform the

ETL portion of this tutorial.

6a. From the AWS Console, go to AWS Glue Studio:
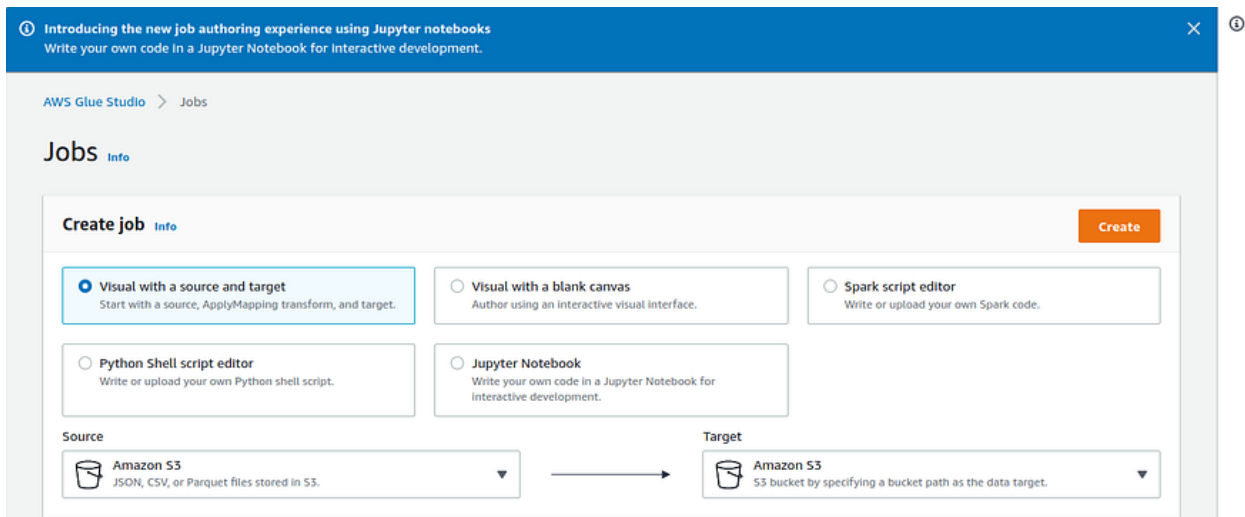
You will want to select Amazon S3 as both the Source and Target. Click

Create.

Set up your Glue Job to look like this:

6b. Run the job.

While it's running, you should see an output like this:

After the Glue job is finished running, your output should look like

this:

Step 7 — Querying our data using Athena

AWS Athena is a query service offered by AWS that analyzes data in S3

using SQL. For this tutorial, I will run Athena to query the ETL data

and run some SQL operations.

Go to the AWS console, and search for Athena. Then run this query as

shown below:

After running the query, we should see our results. :)

Overall summary

When it comes to data engineering, in general, ETL operations and pipelines are a huge part of it. Apart from Glue, there are a multitude of other ETL services (i.e. Cloud Air Fusion by Google Cloud, Data Factory offerred by Azure, and many others).

Aside from whichever one you pick, the basic idea is that it should

assist with reduction of engineering efforts required by supplying a

base (source and target locations and configurations).

We have just completed ETL using AWS Glue. :)

My hope is that this walkthrough helps you with your projects!

In case I missed anything, or if you have any questions/thoughts,

please submit them in the comments section below.