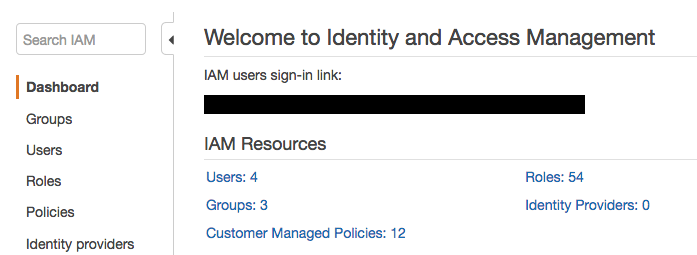
# Exercise 4 – Use AWS Glue to Crawl SQL Server Baseball DB on S3 and setup ETL job to convert CSV tables to Parquet

In this exercise, you will use AWS Glue to convert 4 CSV tables in the Baseball database on S3 to Parquet format. Parquet is a compressed data format that is optimized for storing large data sets. It is a machine readable format and is much faster for computers to analyze as compared to text based formats, such as CSV. You will create a Glue Data Catalog to house the metadata for this operation. You will create 4 corresponding ETL jobs that will persist the parquet files to S3.

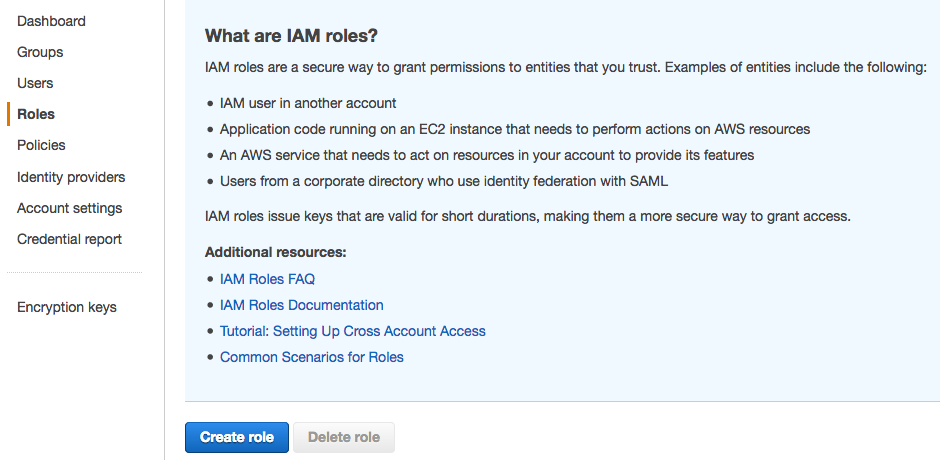
AWS Glue is a fully managed data catalog and ETL (extract, transform, and load) service that simplifies and automates the difficult and time-consuming tasks of data discovery, conversion, and job scheduling. AWS Glue crawls your data sources and constructs a data catalog using pre-built classifiers for popular data formats and data types, including CSV, Apache Parquet, JSON, and more. To learn more about AWS Glue click here: https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html

**Step 1:**

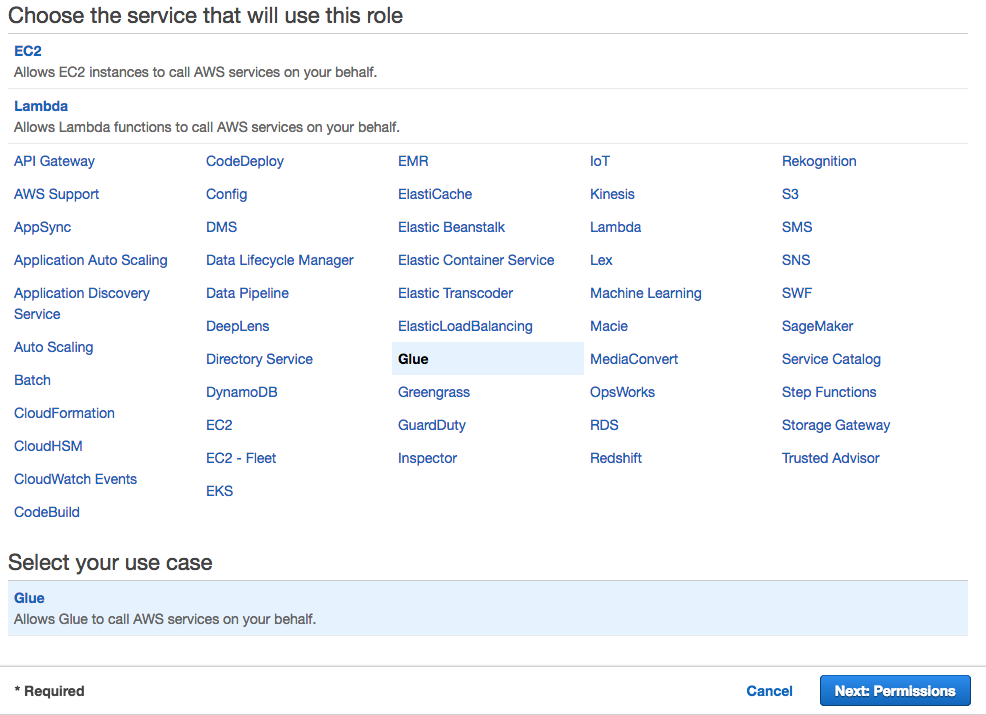
* On the AWS Console, under the drop-down Services, select IAM listed under Security, Identity & Compliance section.
* Select “Roles” in the left margin:



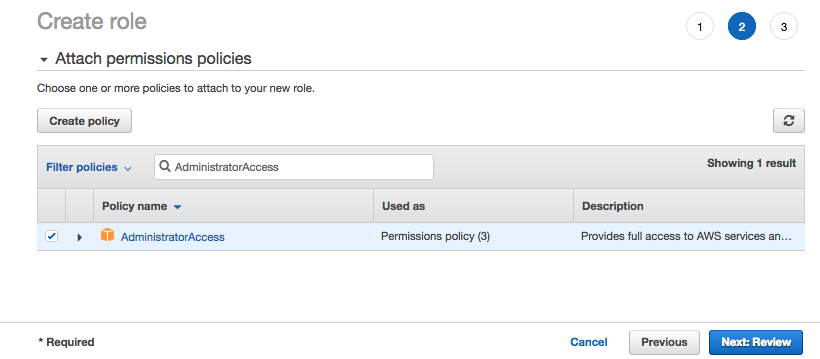
* Then Select **“Create role”**



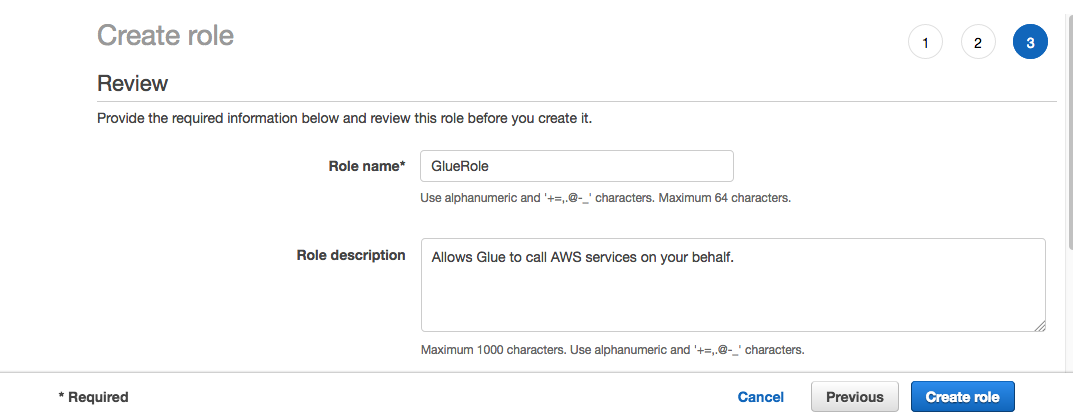
* Select “Glue” from the list of choices and then click “Next:Permissions”



* Search for the “AdministratorAccess” policy, select the checkbox next to it and then click on “Next:Review”:

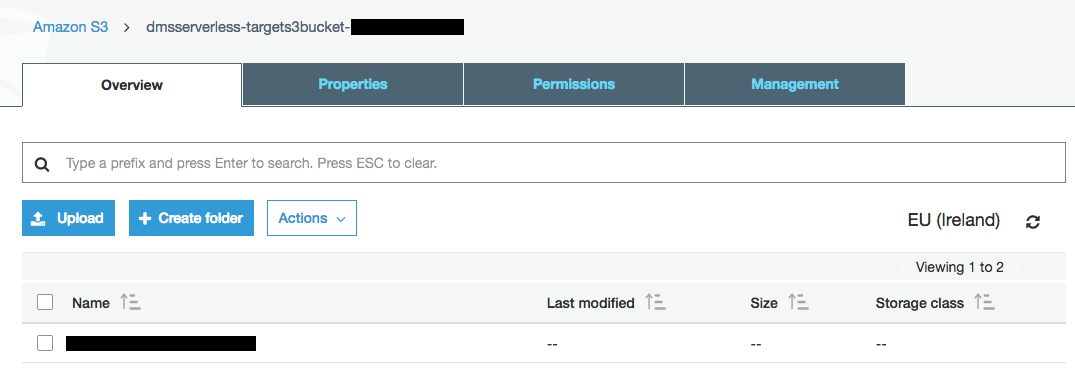


* Type “GlueRole” for the Role name and then click on “Create role”

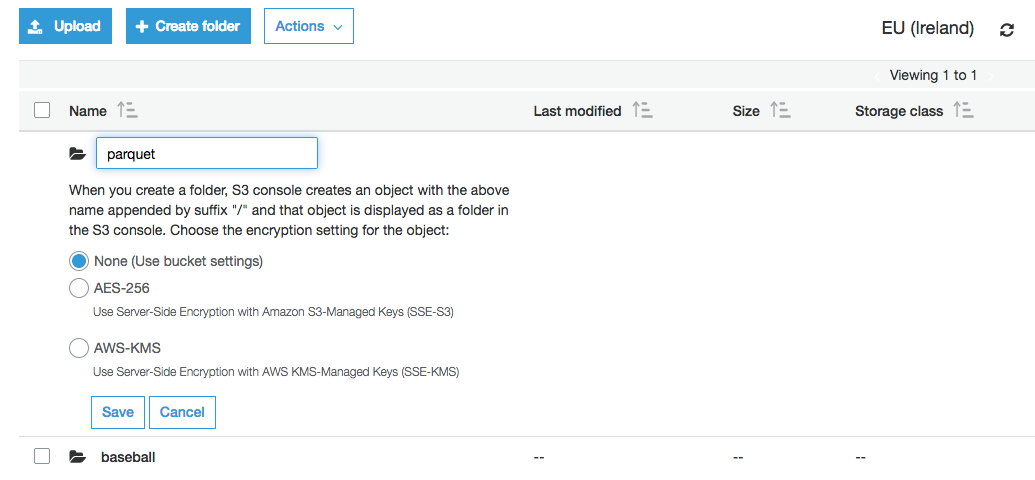


**Step 2:**

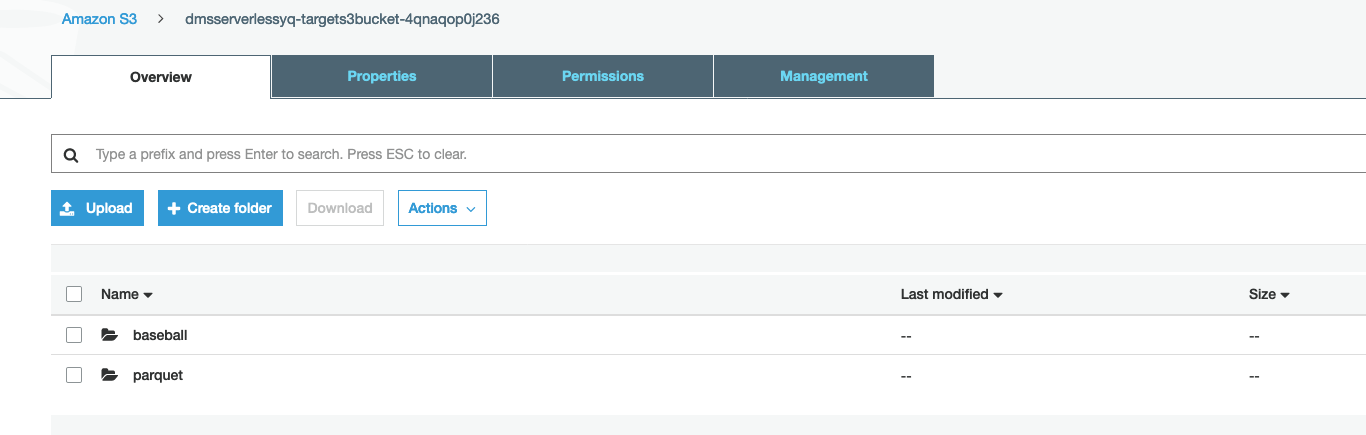
* On the AWS Console, under the drop-down Services, select S3 listed under Storage section or type in “S3” in the search bar.
* Navigate back to your S3BucketName dmsserverless-targets3bucket-:



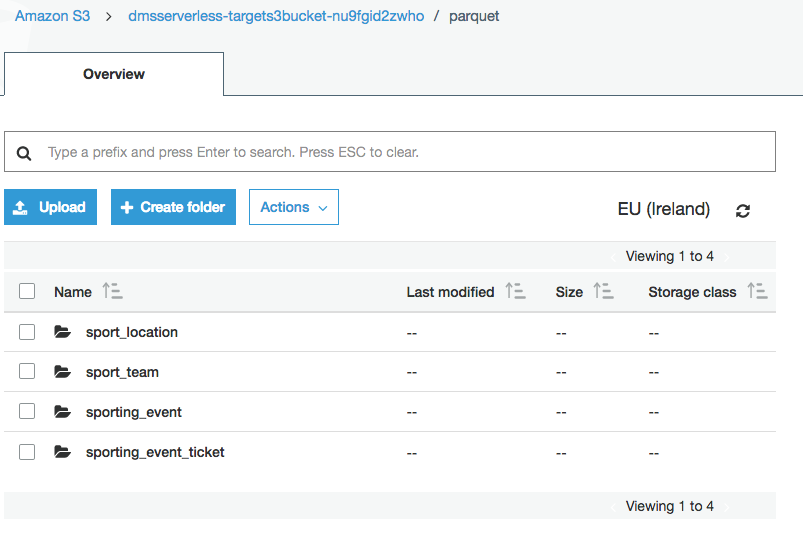
* Select **“Create Folder”** and enter **“parquet”** for the folder name and then click “Save”.



* You will now see two folders listed – baseball and the newly created parquet folder:

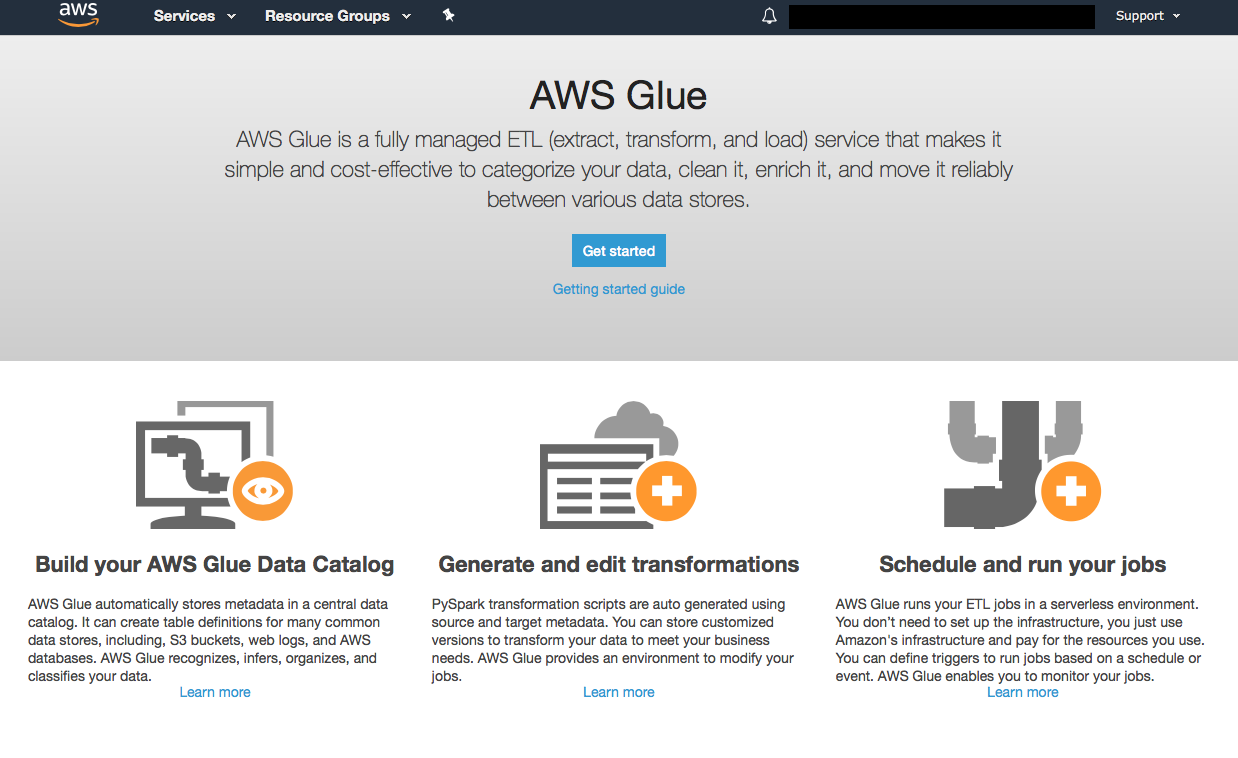


* Open the parquet folder and create four additional folders as depicted in the following image. You will use similar steps as you did when you created the parquet folder earlier. The difference here is these four folders will be ‘subfolders’ of the parquet folder:

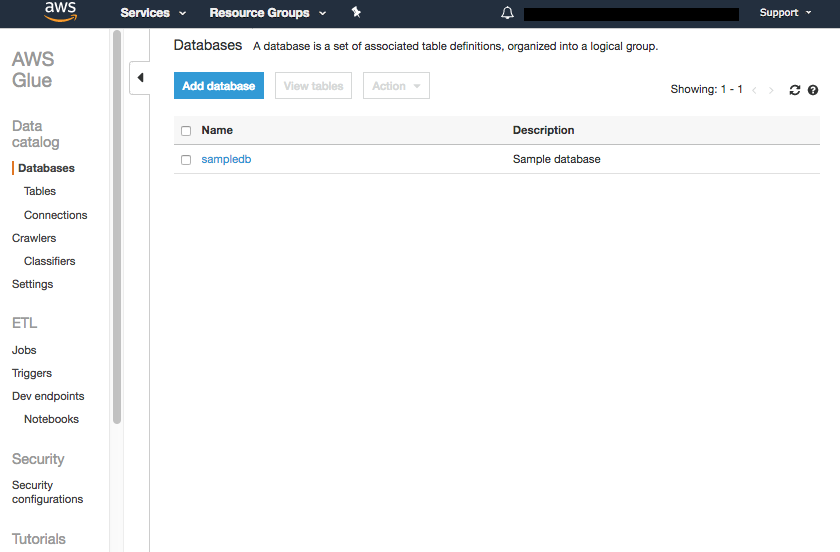


**Step 3:**

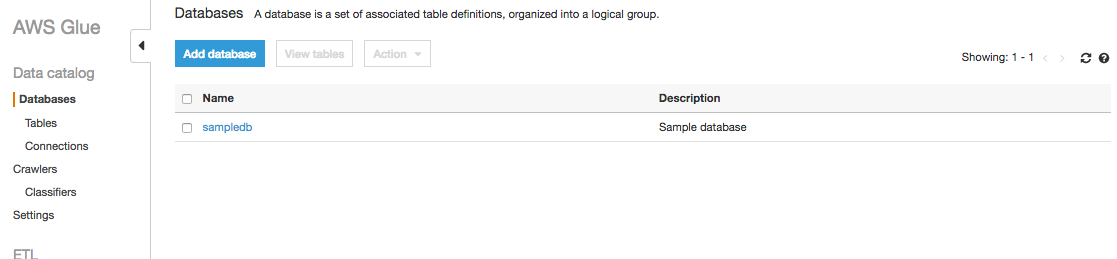
* On the AWS Console, under the drop-down Services, select AWS Glue under Analytics section (or type in “Glue” in the search bar)
* You will be taken to the Glue console as shown below:



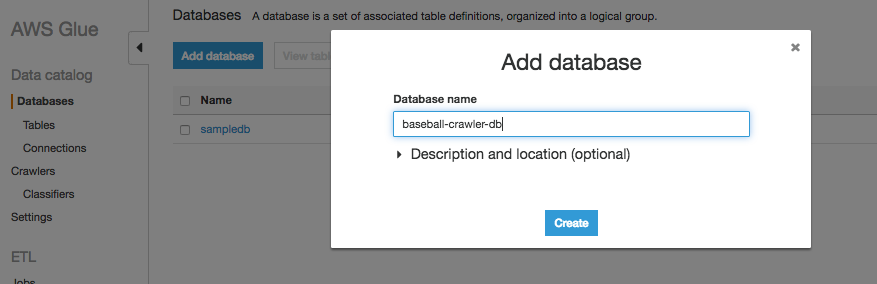
* Click on “Get Started”
* You should now see the following window:

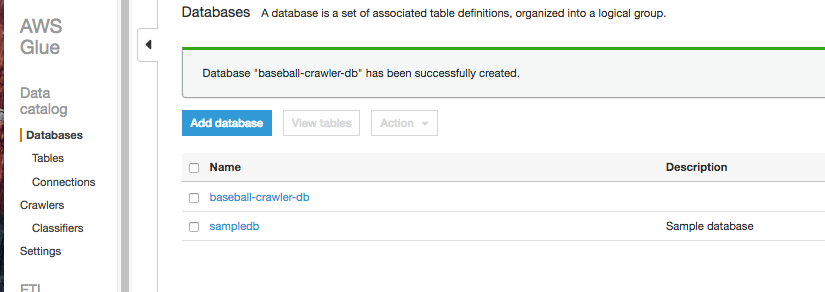


* Click on “Databases” under the “Data catalog” section in the left margin of the window and then click on “Add database”:

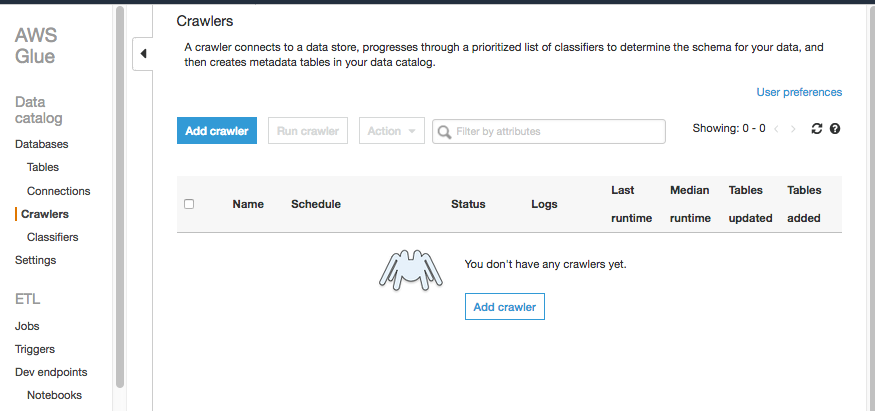


* In the “Database name” text field enter “baseball-crawler-db” and click “Create”:

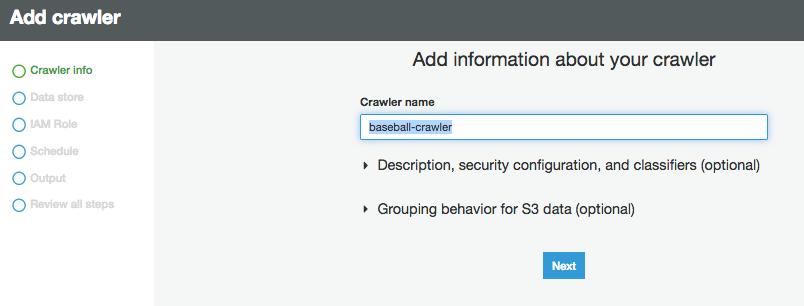




* Next click on “Crawlers” under the “Data Catalog” section. A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.



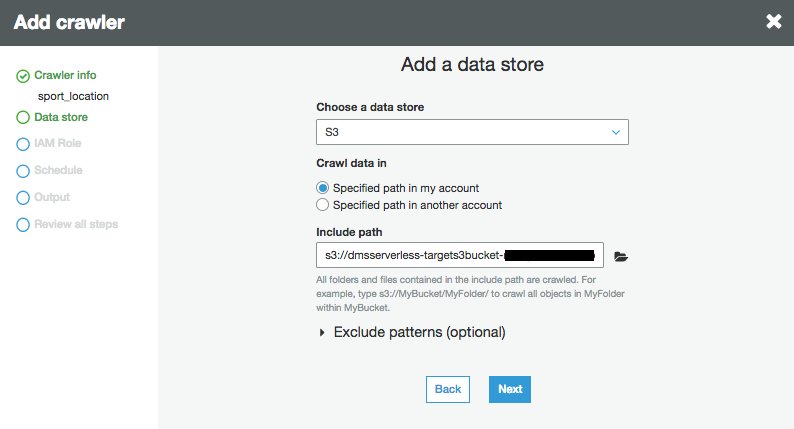
* Select “Add Crawler” and enter “baseball-crawler” for the crawler name and click “next”:



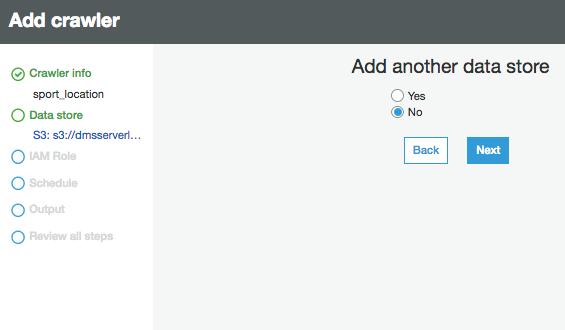
* For the “include path” text field type. Make sure to replace the highlighted portion with your bucket name from CloudFormation outputs.:

s3://dmsserverless-targets3bucket-XXXXXX/baseball/dbo

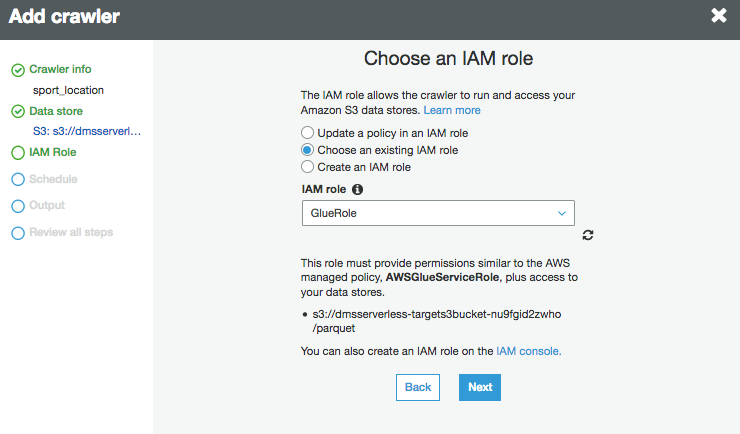
and click “Next”.



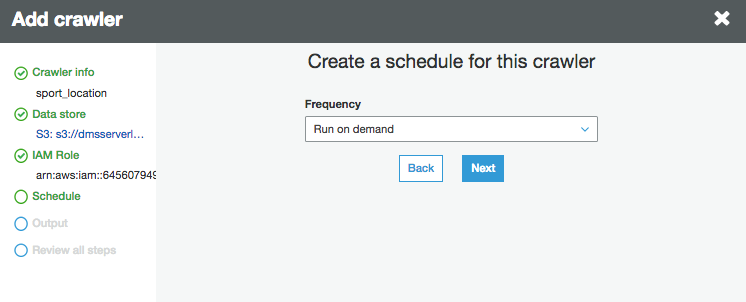
* Keep the default of “No” for “Add another data store” and click “Next”



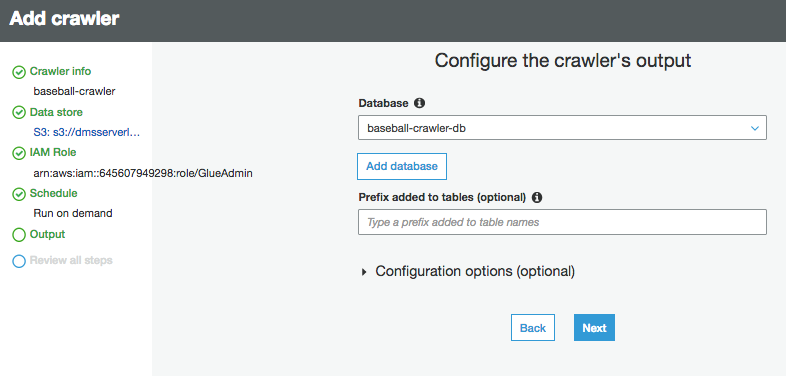
* Choose an existing IAM role and select the “GlueRole” IAM role you created earlier and click “Next”:



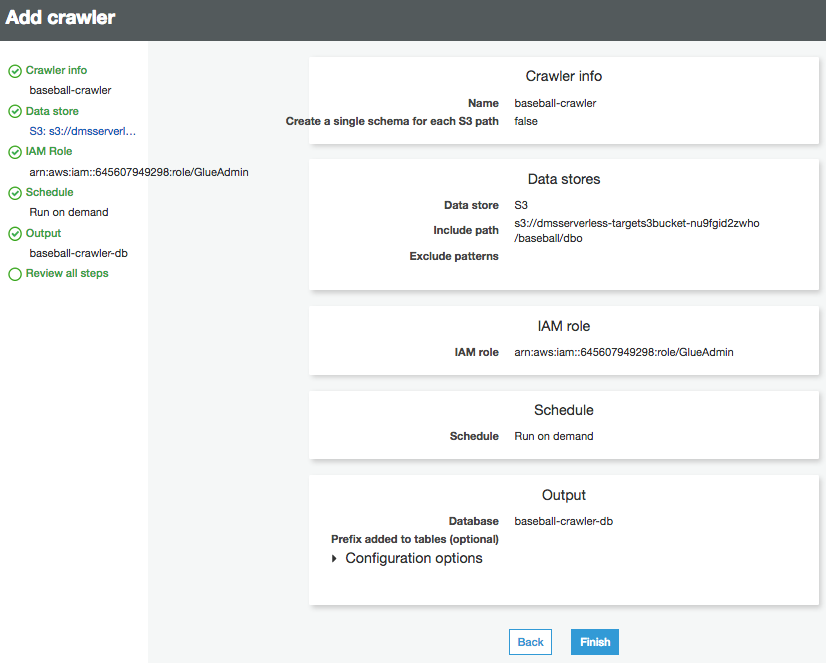
* Leave the default of “Run on demand” and click “Next”



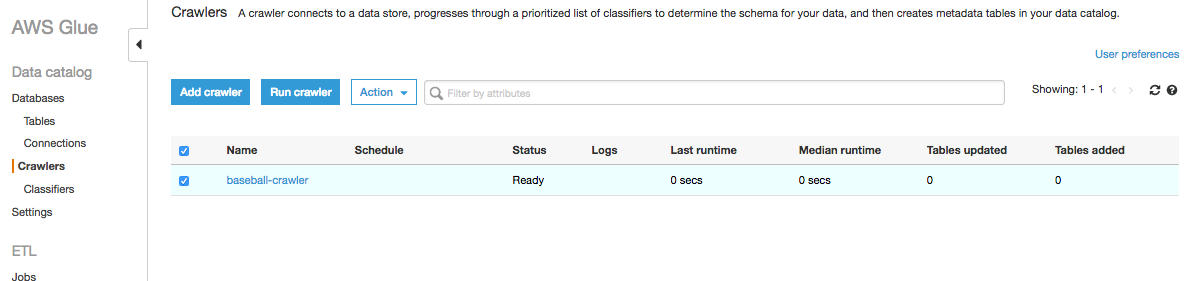
* Select the baseball-crawler-db database from the Database dropdown menu:



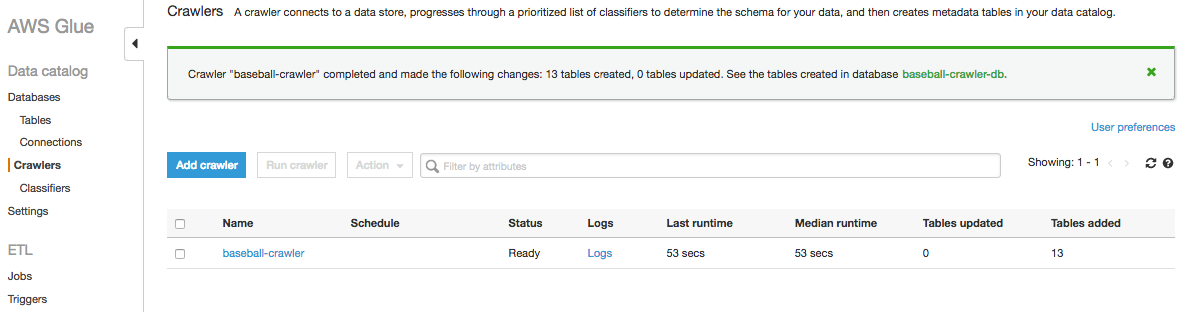
* Review details and then click “Finish”:



* Select baseball-crawler and click on “Run crawler”:

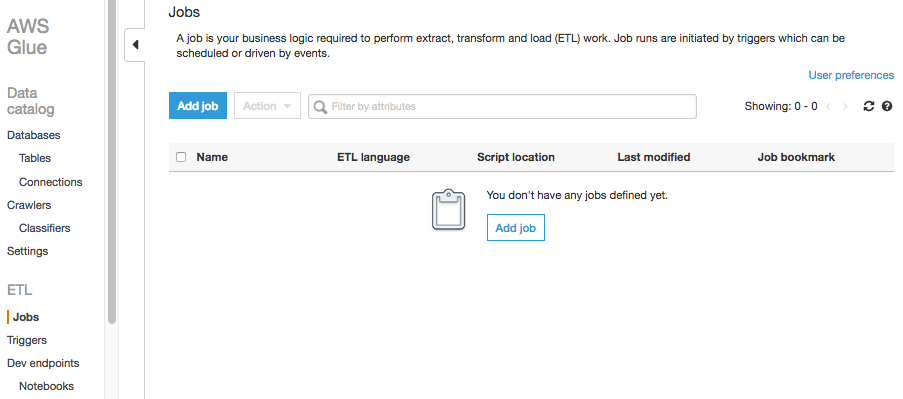


* Baseball-crawler will run for a few minutes and will eventually add 13 tables:

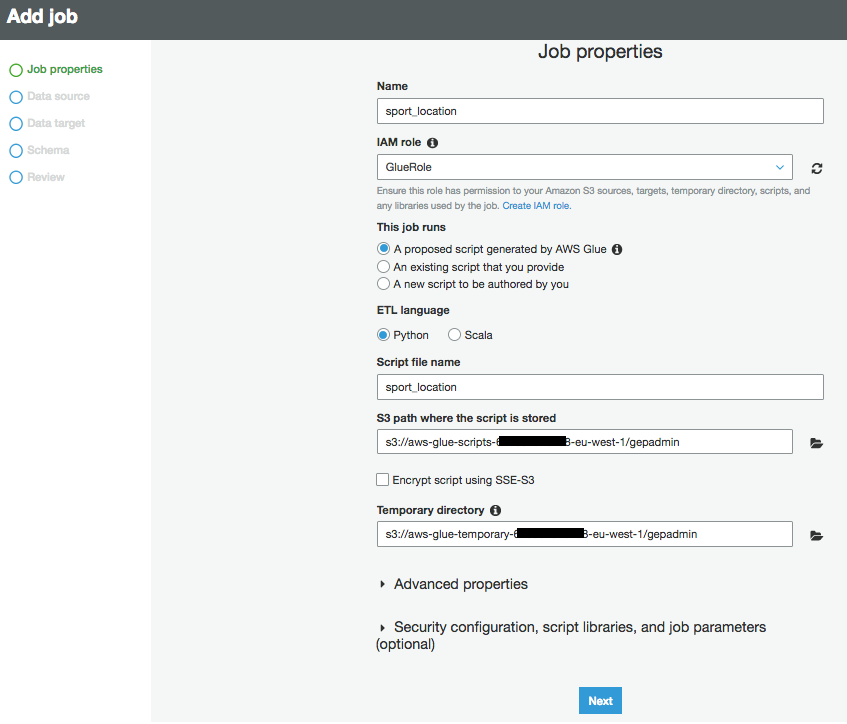


**Step 4:**

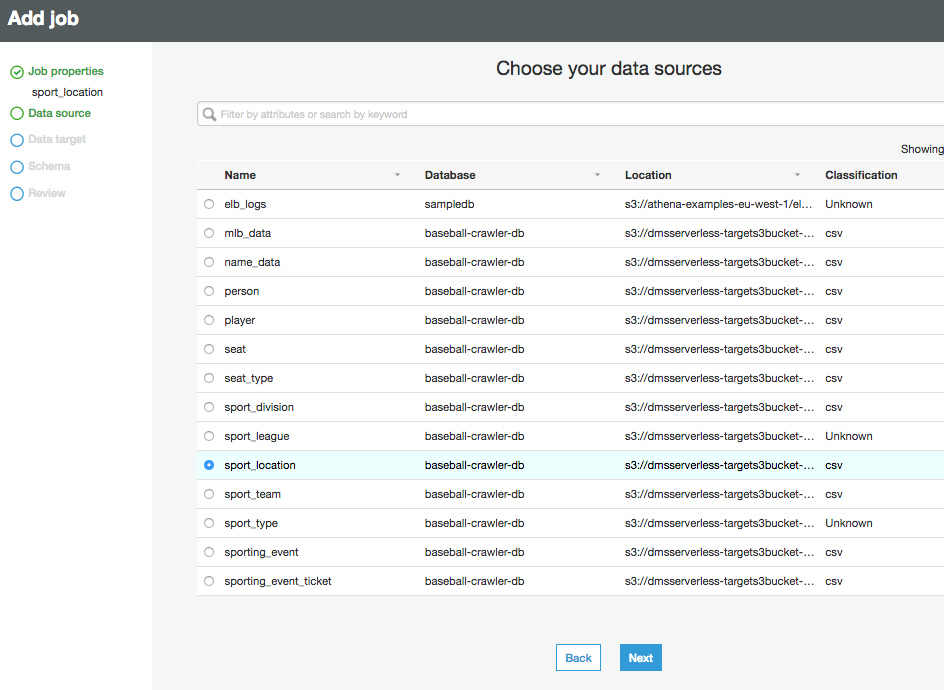
* Click on “Jobs” under “ETL”:



* Click on “Add Job”. Populate the “Name” text field with “sport\_location”. Select “GlueRole” for the IAM role. Leave defaults for the remaining items and click “Next”.



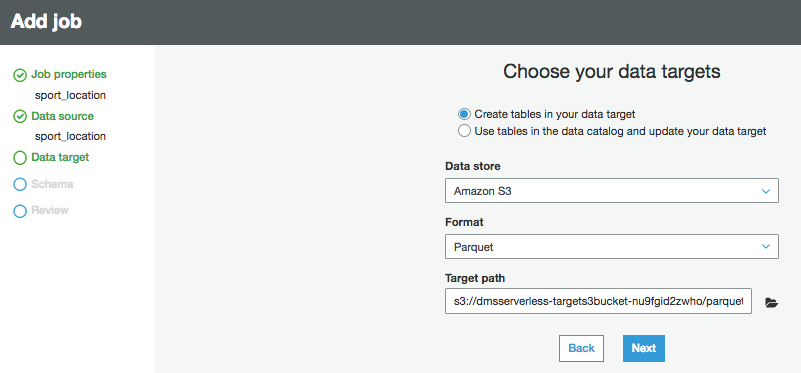
* For “Choose your data sources” select “sport\_location” and click “Next”:



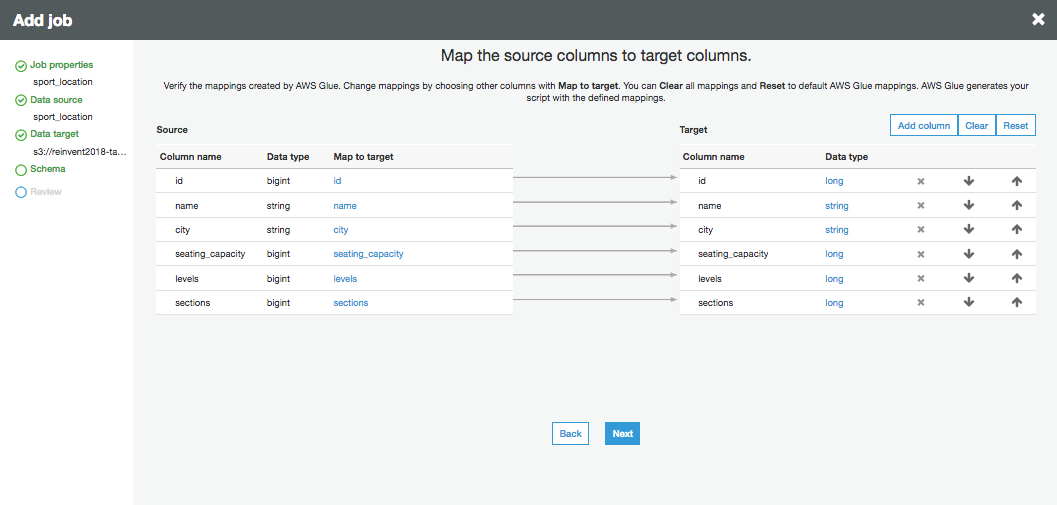
* For “Choose your data targets” select “Create tables in your data target”, select Amazon S3 as the Data Store, chose Parquet as the format, and for the target path select your path to “sport\_location” folder in S3 you created earlier in this Exercise:

s3://dmsserverless-targets3bucket-XXXXXXXX/parquet/sport\_location

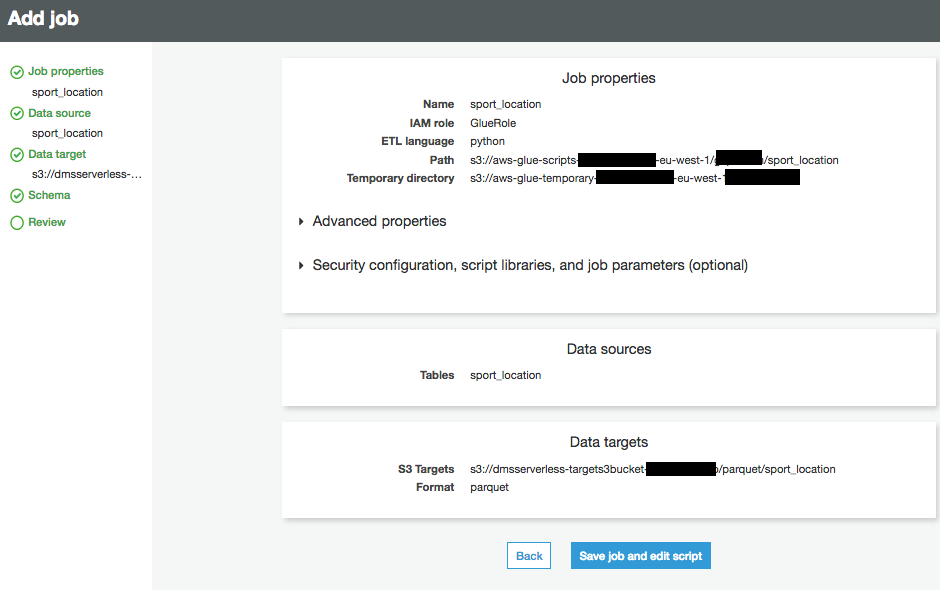
and then click “Next”



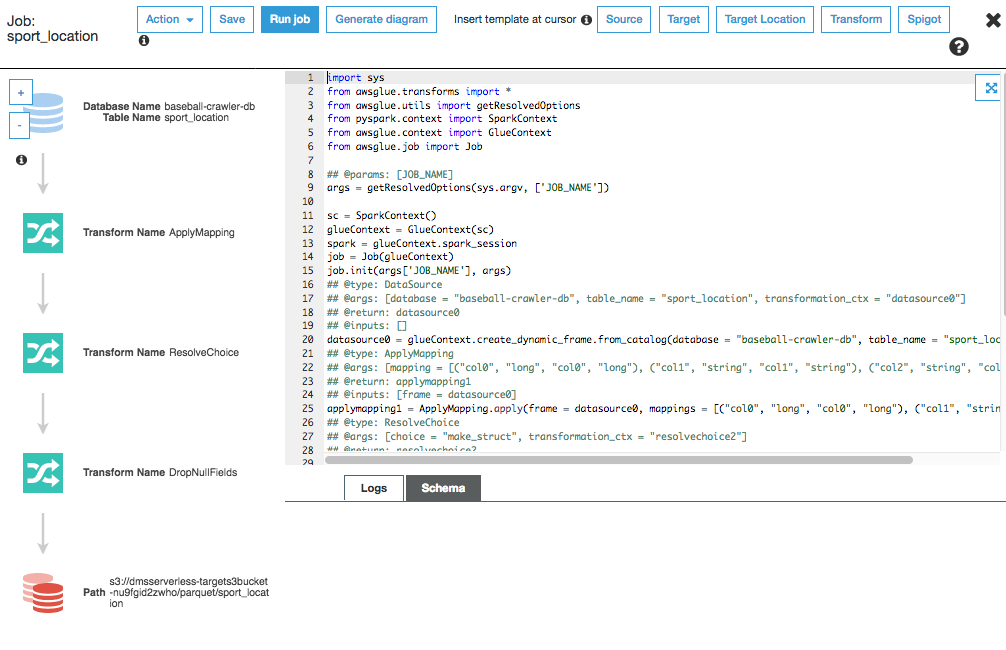
* Confirm source and target schema match and then click “Next”:



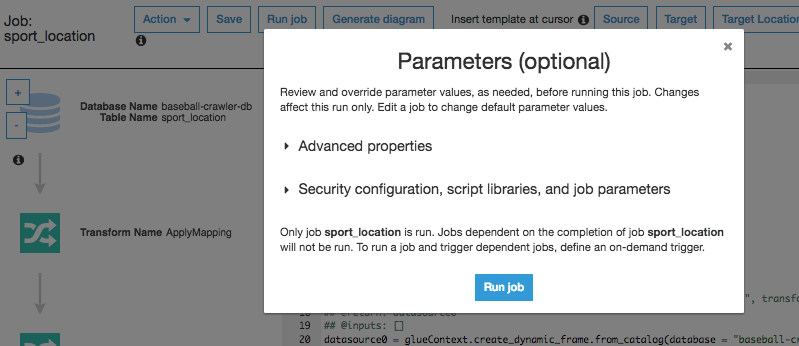
* Confirm job properties and then click “Save job and edit script”



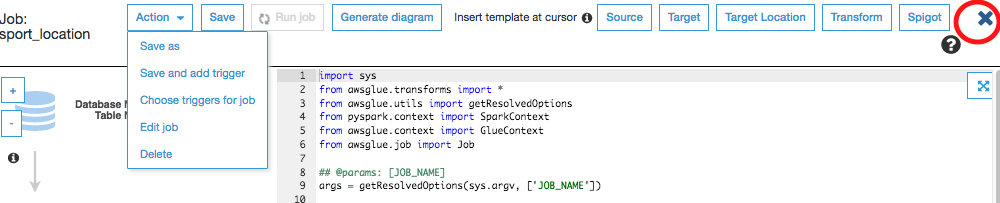
* Click “Run Job”:



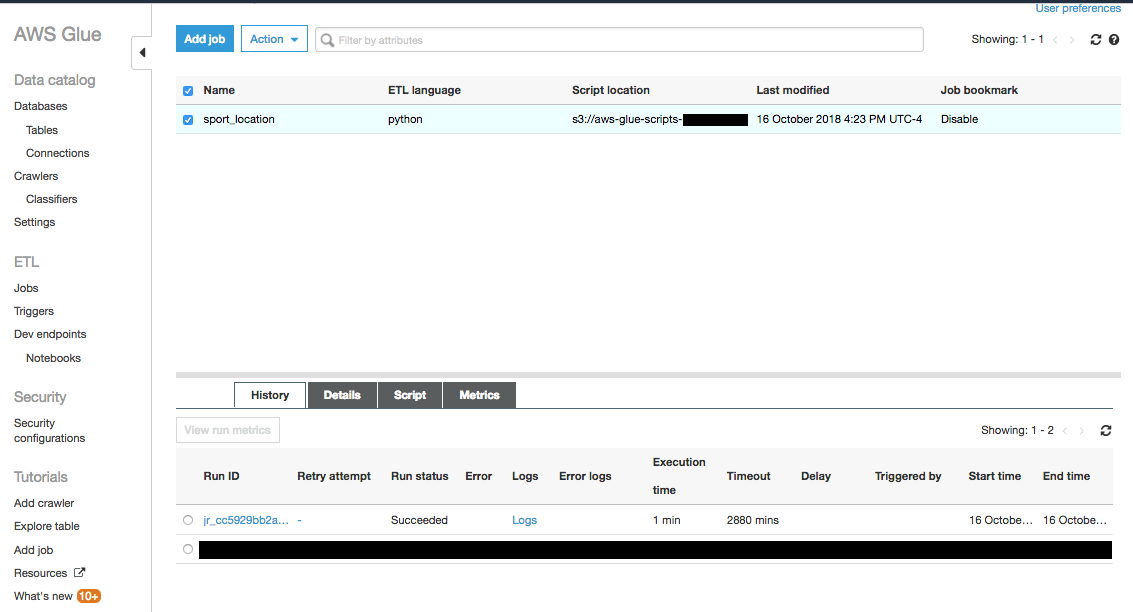
* Click “Run Job” again on “Parameters (optional)” window:



* Click the “X” to return to the ETL Jobs console:



* After a few minutes, confirm your “Run status” for the ETL job for “sport\_location” reads “Succeeded”



* Create 3 additional ETL Jobs one each for:

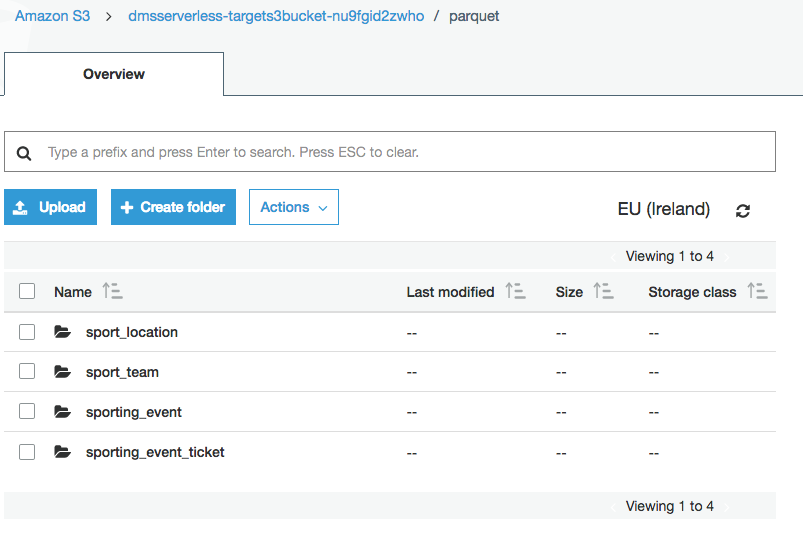
**sporting\_event\_ticket**

**sport\_team**

**sporting\_event**

Repeating the steps as performed for the “sport\_location” crawler.

\*Remember adjust the target path for the parquet output to their corresponding S3 buckets. This will be where all the corresponding ETL Parquet files will persist:



Exercise recap –

In this exercise, we learnt to do the following:

1. Use AWS Glue to crawl & build data catalog

2. Transform data stored on Amazon S3 from CSV to Parquet format