AWS Glue

1. Open glue_job.py, change <your s3 bucket> to your s3 bucket name, check to make sure all the s3 path is correct, save this file and upload it to s3://<your s3 bucket>/scripts/.
2. Open Glue console in AWS, click Jobs on the left pane:

3. Click Add job and you should fill in the details similar to below, name the job to "imba-glue", create a new IAM role or re-use an existing one (you just need to make sure AmazonS3FullAccess and AWSGlueServiceRole is attached). Make sure you select "An existing script that you provide" for "this job runs". Specify the s3 path where your script is stored: s3://<your s3 bucket>/scripts/glue_job.py and Temporary directory: s3://<your s3 bucket>/root. Leave everything else as default and click next.

Configure the job properties

Name

imba-glue

IAM role ⓘ

AWSGlueServiceRole-imba   ⌄   ⟳

Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job. Create IAM role.

Type

Spark   ⌄

Glue version

Spark 2.4, Python 3 (Glue Version 1.0)   ⌄

This job runs

◉ A proposed script generated by AWS Glue ⓘ

◯ An existing script that you provide

◯ A new script to be authored by you

Script file name

imba-gluejob

S3 path where the script is stored

s3://<your s3 bucket>/scripts/glue_job.py   📁

Temporary directory ⓘ

s3://<your s3 bucket>/root   📁

▸ Advanced properties

▸ Monitoring options

4. Click Save job and edit script:

| Back | Save job and edit script |

5. Have a look at the script and close it by clicking the top right X button:

Insert template at cursor ⓘ | Source | Target | Target Location | Transform | Spigot | ❓ | ✖

6. Select the job you created and click Run job from Action drop down menu:

Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are

New in AWS Glue
Streaming ETL in AWS Glue (preview): Process streaming data and make it available for analysis in sec
Reduced start times for AWS Glue Spark jobs (preview): Glue Spark jobs will start in under a minute. Le

Add job	Action ▼	🔍 Filter by tags and attributes

	Run job
☑ **Name**	
☑ imba-g	Stop job run
	Choose job triggers
	Delete
	Edit job
	Edit script
	Reset job bookmark
	Create development endpoint

History	Details	Script	Metrics

View run metrics	Rewind job bookmark

Run ID	Retry attempt	Run status	Error	Logs	Error logs

7. Be patient, it should take around 10 minutes or so to finish, once done you should see an csv file is created in s3://<your s3 bucket>/output/part-xxxxxxxxxxx.csv
8. Download this file to your local desktop and rename it as "data.csv"