AWS Glue

It is a system for building table definitions and doing ETL.

Glue is a serverless system. It will automatically handle discovery and definition of table definitions and schemas. It's main use is to serve as a central metadata repository for your Data lake. Will discover schemas out of S3 and publish table definitions for use for analytics tools like Redshift and other SQL databases.

So if we have unstructured data lying around in S3 it can help us query it using various analytical tools for which it would create a schema / publish table definitions.

It also does custom ETL jobs. Can be trigger driver / on schedule / on demand. ETL jobs use Apache Spark under the hood.

We don't have to manage the spark cluster but have the entire power of the Spark cluster with Glue.

Glue Crawler:

Scans data in S3. Can schedule it to run categorically.

Upon doing this, it would store table definitions in Glue Data Catalog. It stores only the table definitions. The original data stays in S3.

Once it is catalogued, can treat data as structured data with the help of – Redshift, Athena, EMR, Quicksight etc

Glue and S3 Partitions:

Glue will extract parititons based on how your S3 data is organized. Think up front about how you will be querying your Data lake in S3.

For example if we have device sensor data –

- If we are querying by time ranges then it would be good to organize by yyyy/mm/dd/device
- If we are querying by the devices then it would be more sensible to organize by device/yyyy/mm/dd

Glue + Hive:

Hive would let your run SQL like queries from EMR. The glue data catalog can server as a Hive "metastore" . Can import hive metastore into Glue

Glue ETL :

Can be used to Transform your data, clean your data, enriching your data. Will generate the etl code in python or scala and we can modify the code once it has been generated. Jobs run on a serverless Spark platform.

Can automatically generate code for your for transforming your data. Can do this in a graphical manner from the Amazon console. Can add your own code in Scala or Python.

Can be event driven, can have ETL process kick off if it sees a new job.

Can provision addition Data processing units (DPUs) to increase performance of the underlying spark jobs.

How to go about applying the ETL transforms in Glue? Dynamic Frame

The main data structure that we would be interacting with is called a *dynamic frame*. It is very much like a spark dataframe but has more ETL tools available to it. Like a dataframe is a collection of rows, a dynamic frame is a collection of dynamic records.

Glue ETL – Transformations

- Bundled transformations :
  1. DropFields , DropNullFields – removes empty fiels
  2. Filter – to filter out records
  3. Joins – to enrich data
  4. Map – to add fields, delete fields, perform external lookups
- Machine Learning Transformations :
  1. FindMatches ML : identify duplicate or matching records in your dataset even when records do not have a common unique identifier or no fields match exactly.
- Format conversions : CSV, JSON, Parquet
- Apache Spark Transformations

Glue ETL – ResolveChoice

Deals with ambiguities in a DyanmicFrame and returns a new one. Multiple ways to solve ambiguity.

- make_cols :It can create a new column if two column names are the same – Price : 100 and Price : "$100" . Then it can create new columns Price_double and Price_string to solve the problem.
- cast : casts all values to specified type
- make_struct : creates a structure that contains each data type

Glue Development Endpoints:

These allow you to develop your ETL code using a notebook. We can connect via various methods like Apache Zeppelin on your local machine.

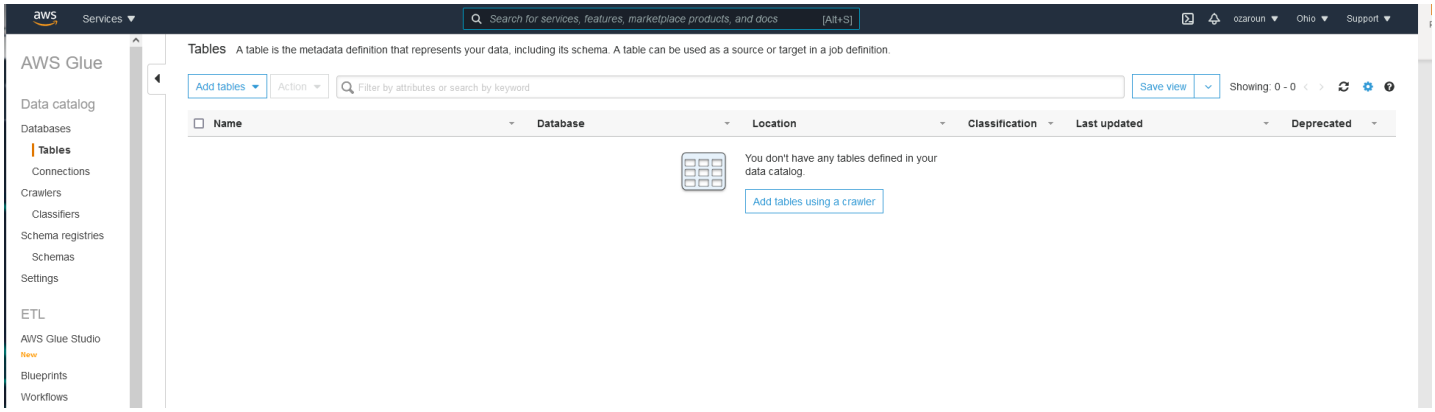
Running Glue jobs:

- Can be a time based schedule like (Cron jobs)
- Job bookmarks:
  1. Persists state from the job run
  2. Prevents reprocessing of old data
- CloudWatch events compatible with Glue

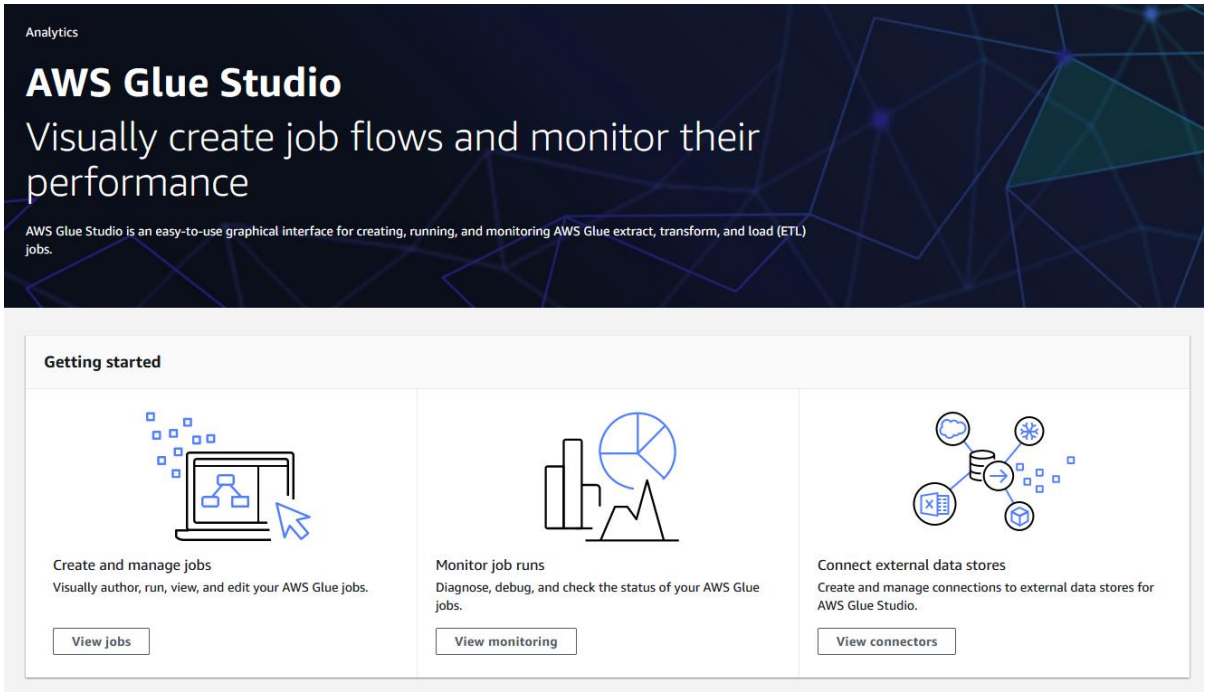
AWS Glue Studio:

Is a visual interface for defining and creating ETL workflows. Visual job editor allows your to create DASs for complex workflows. Uses sources such as S3, Kinesis, etc. Transforms / joins data if needed. Targets can be S3 or the Glue Data Catalog. Supports partitioning.

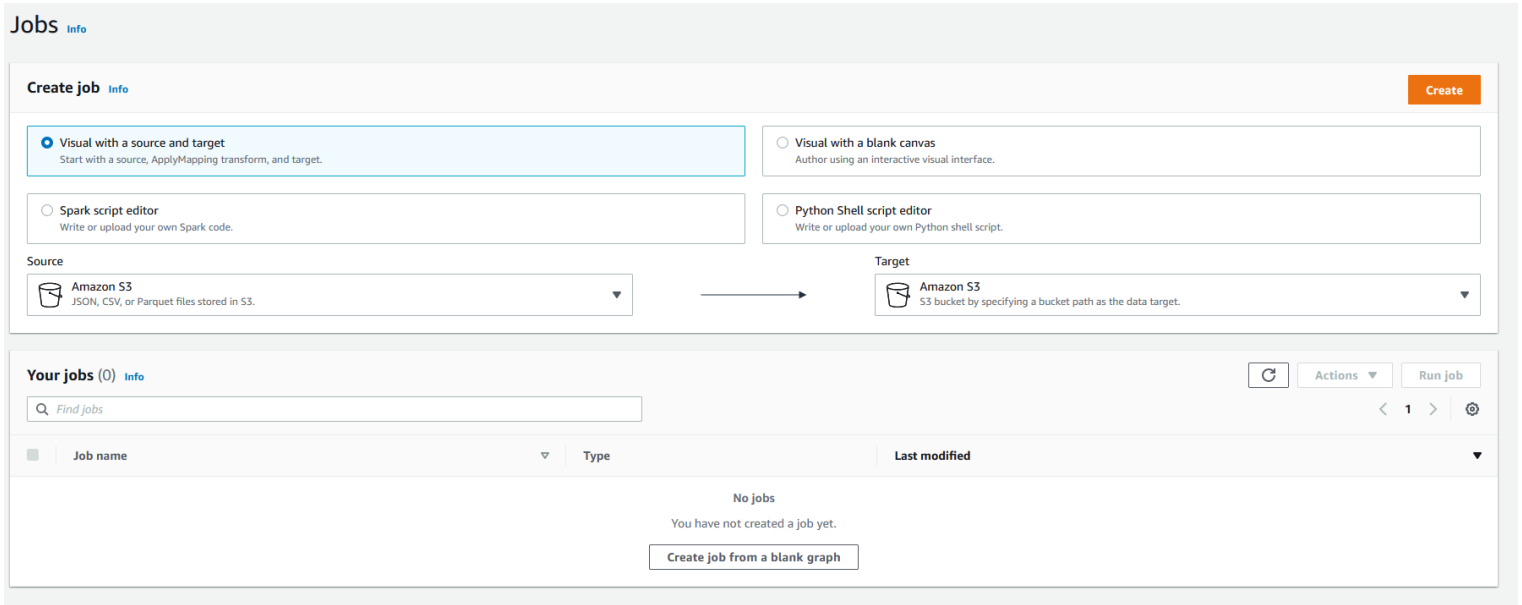
Overview:



Click on the AWS Glue Studio to get start making an ETL job

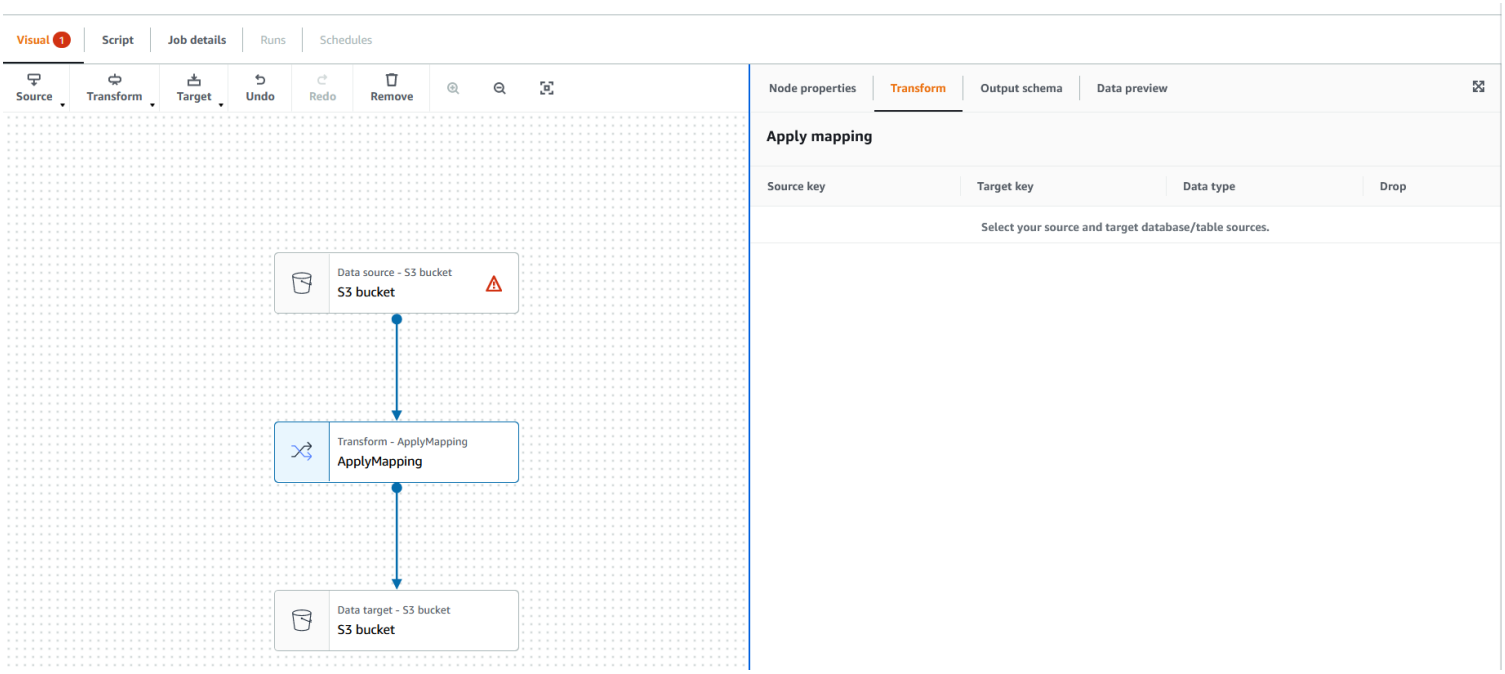


Click on view jobs to create one.



Select source and destination and then click create.

Click on the source and the target to set source and targets. Click on the transform to do transformations if required. Can also do multiple other transformations present on the top left “Tranform” button.



AWS Glue DataBrew:

A visual data preparation tool meant for doing graphically complex transformation on large datasets coming from S3.

It is a bit different than glue studio. Glue studio lets you build these complicated transforms in top of Glue for transforming data in complex workflows. DataBrew is much more simple. You take a data source and build up recipes of transformations that you want to apply and just put the output in some S3 bucket somewhere.

It has more than 250 ready made transformations ready to use.

So it is purpose-built for very quickly transforming the data in some way.

UI :