CASE STUDY

ON AWS GLUE



**AWS Glue**

* **AWS Glue** is a fully managed ETL (extract, transform, and load)

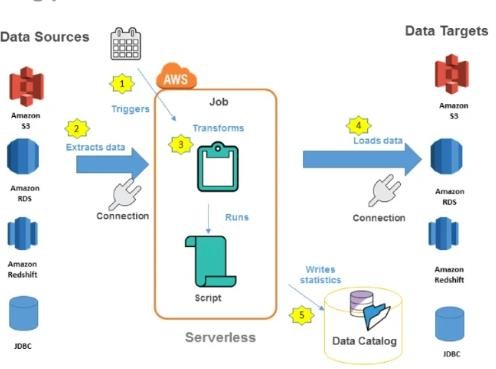
AWS service.

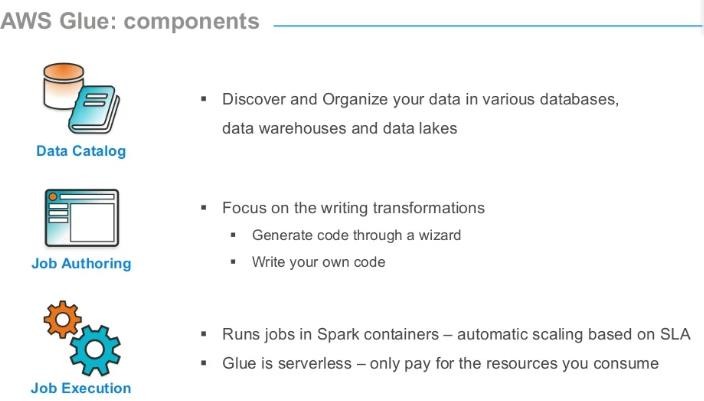
* AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.
* AWS Glue provides all the capabilities needed for data integration so that you can start analyzing your data and putting it to use in minutes instead of months.

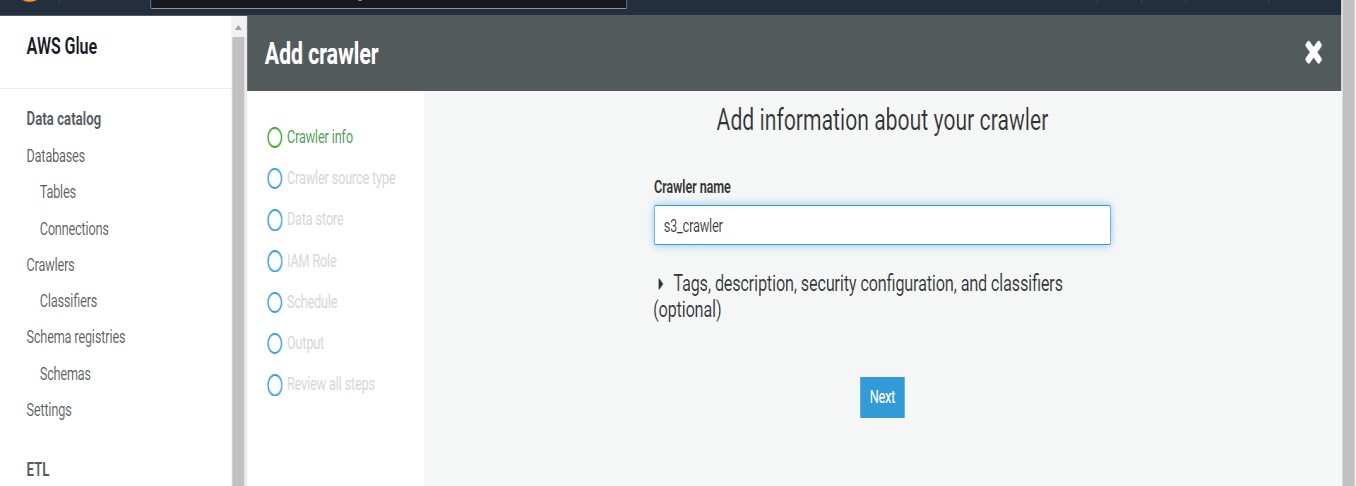
**Benefits:**

* Faster data integration : Different groups across your organization can use AWS Glue to work together on data integration tasks, including extraction, cleaning, normalization, combining, loading, and running scalable ETL workflows. This way, you reduce the time it takes to analyze your data and put it to use from months to minutes.
* Automate your data integration at scale: AWS Glue automates much of the effort required for data integration. AWS Glue crawls your data sources, identifies data formats, and suggests schemas to store your data. It automatically generates the code to run your data transformations and loading processes. You can use AWS Glue to easily run and manage thousands of ETL jobs or to combine and replicate data across multiple data stores using SQL.
* No servers to manage: AWS Glue runs in a serverless environment. There is no infrastructure to manage, and AWS Glue provisions, configures, and scales the resources required to run your data integration jobs. You pay only for the resources your jobs use while running.

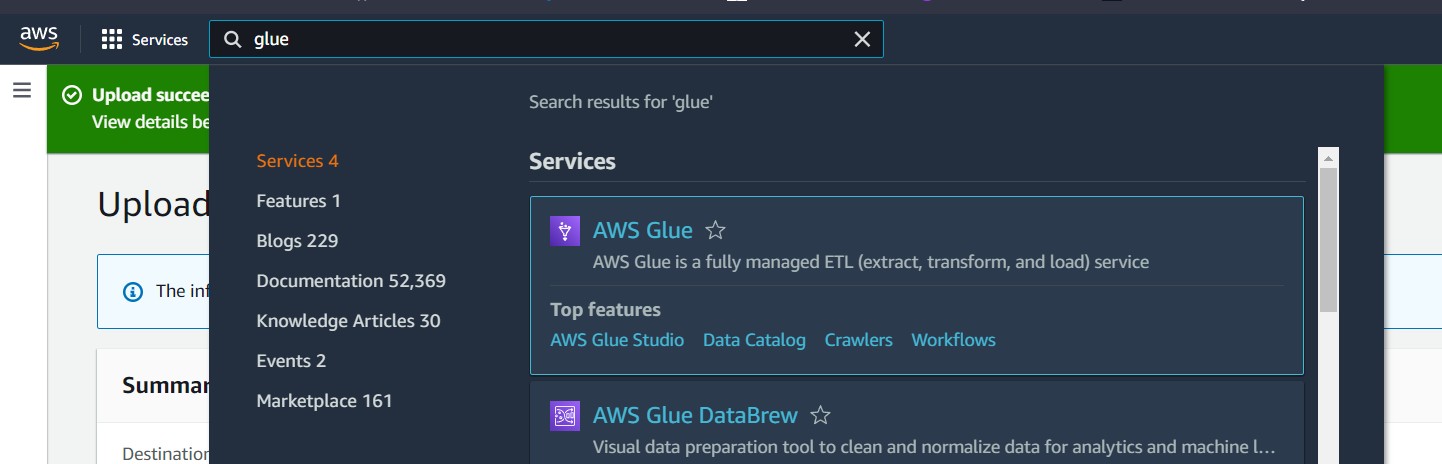
**ETL Pipeline:**



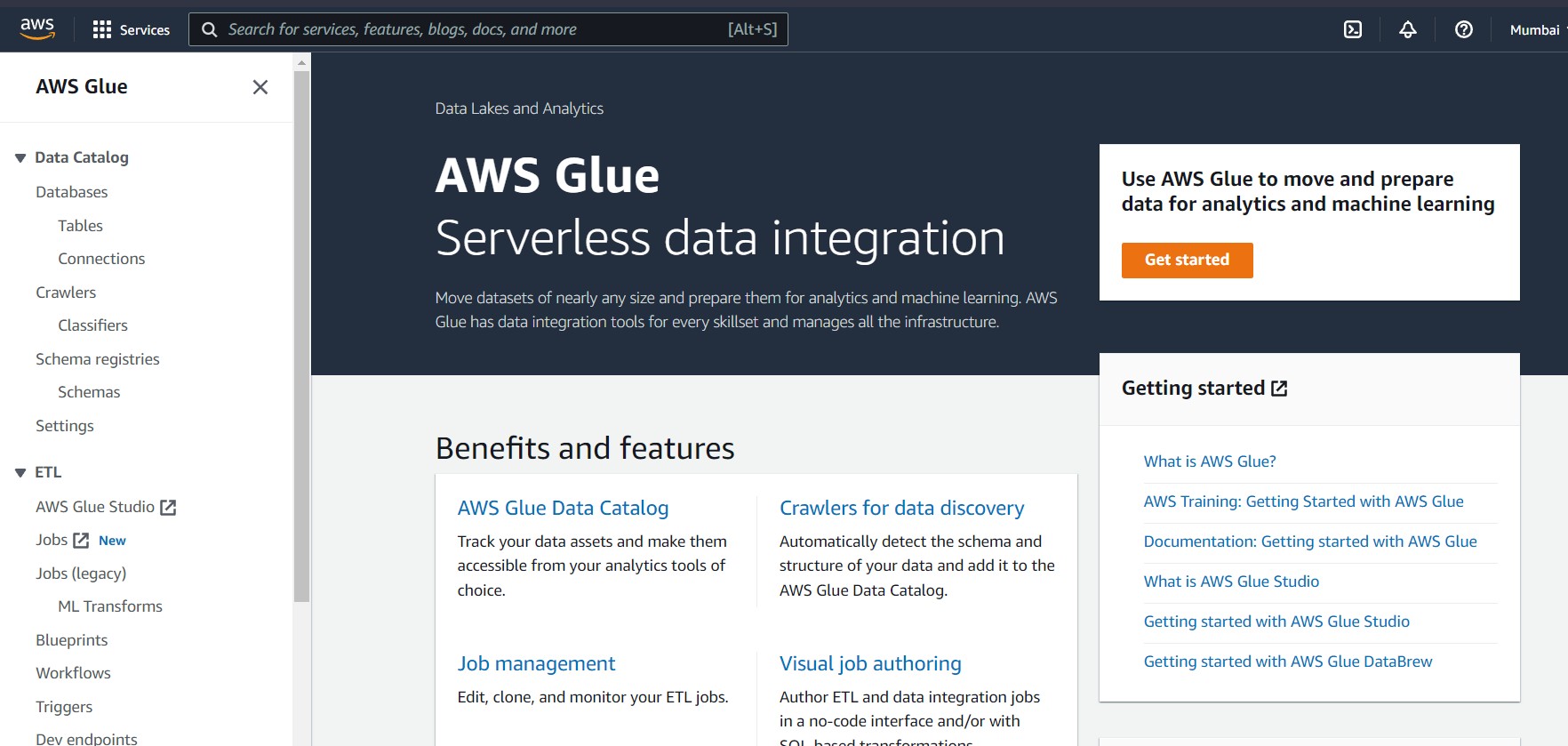


**HANDS ON :**

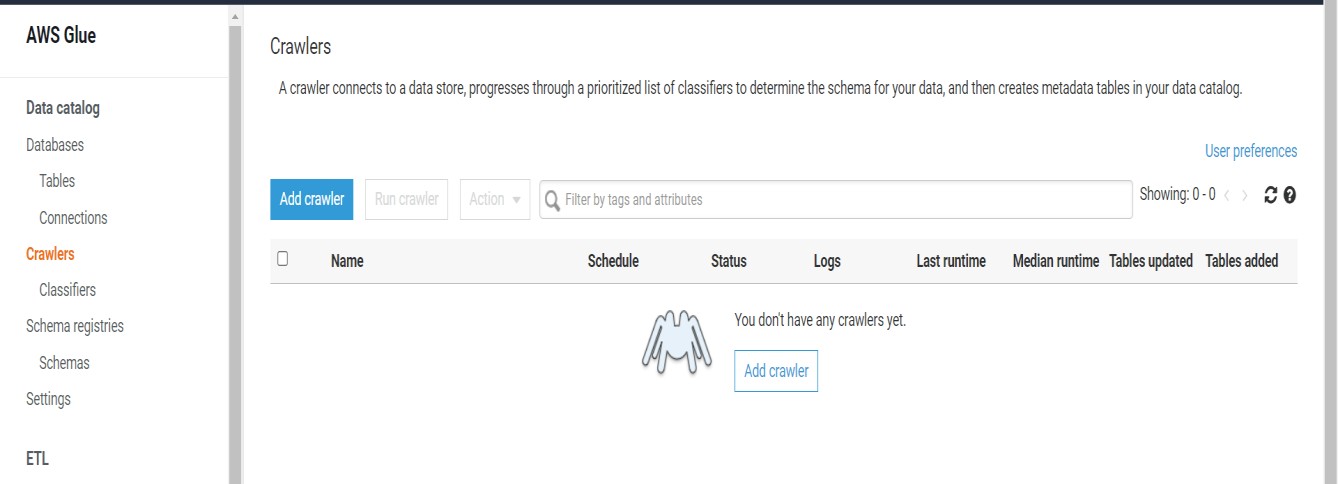
**Step 1:** Search glue on search bar



**Step 2:** Below is the interface of aws glue.



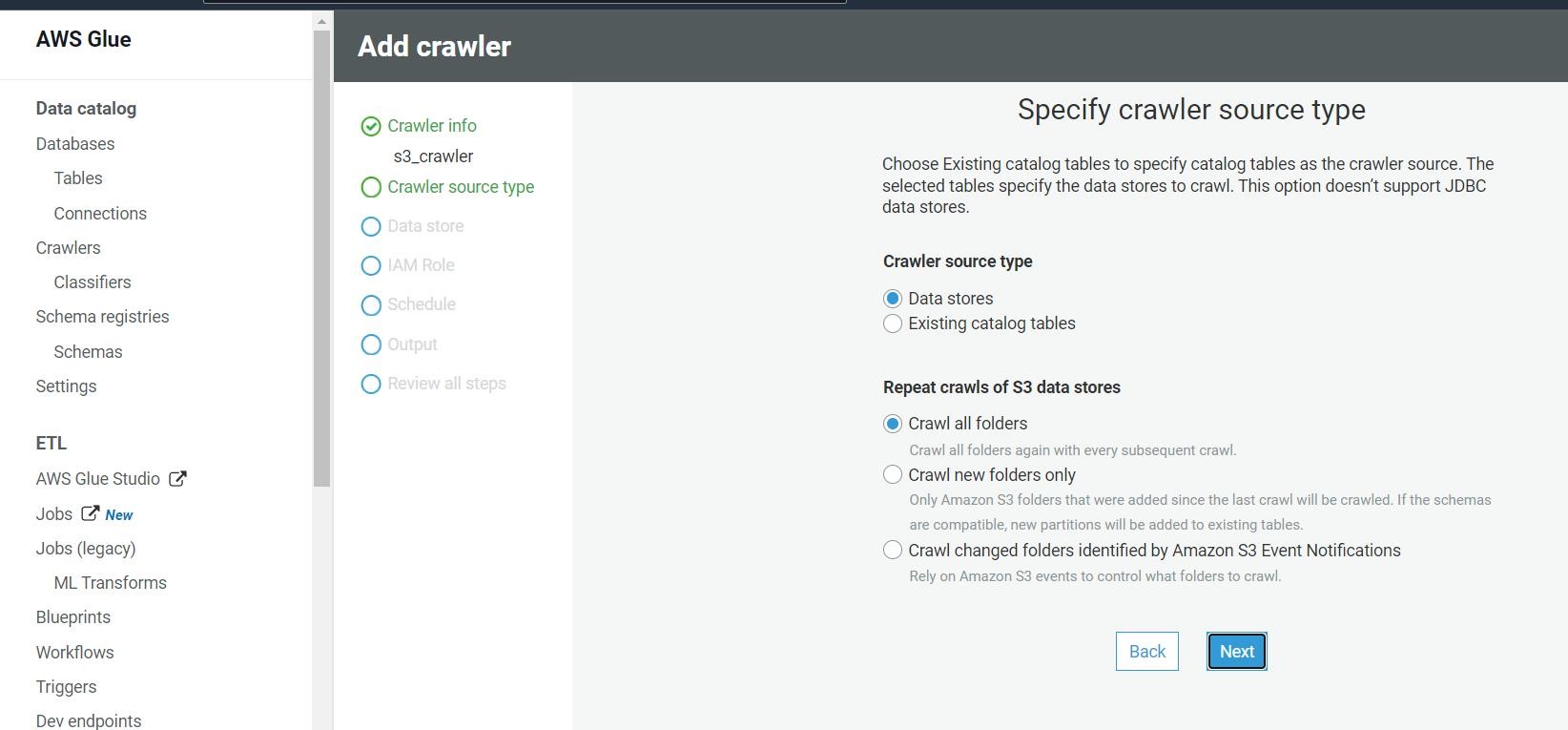
**Step 3:** Click on crawlers to add crawlers



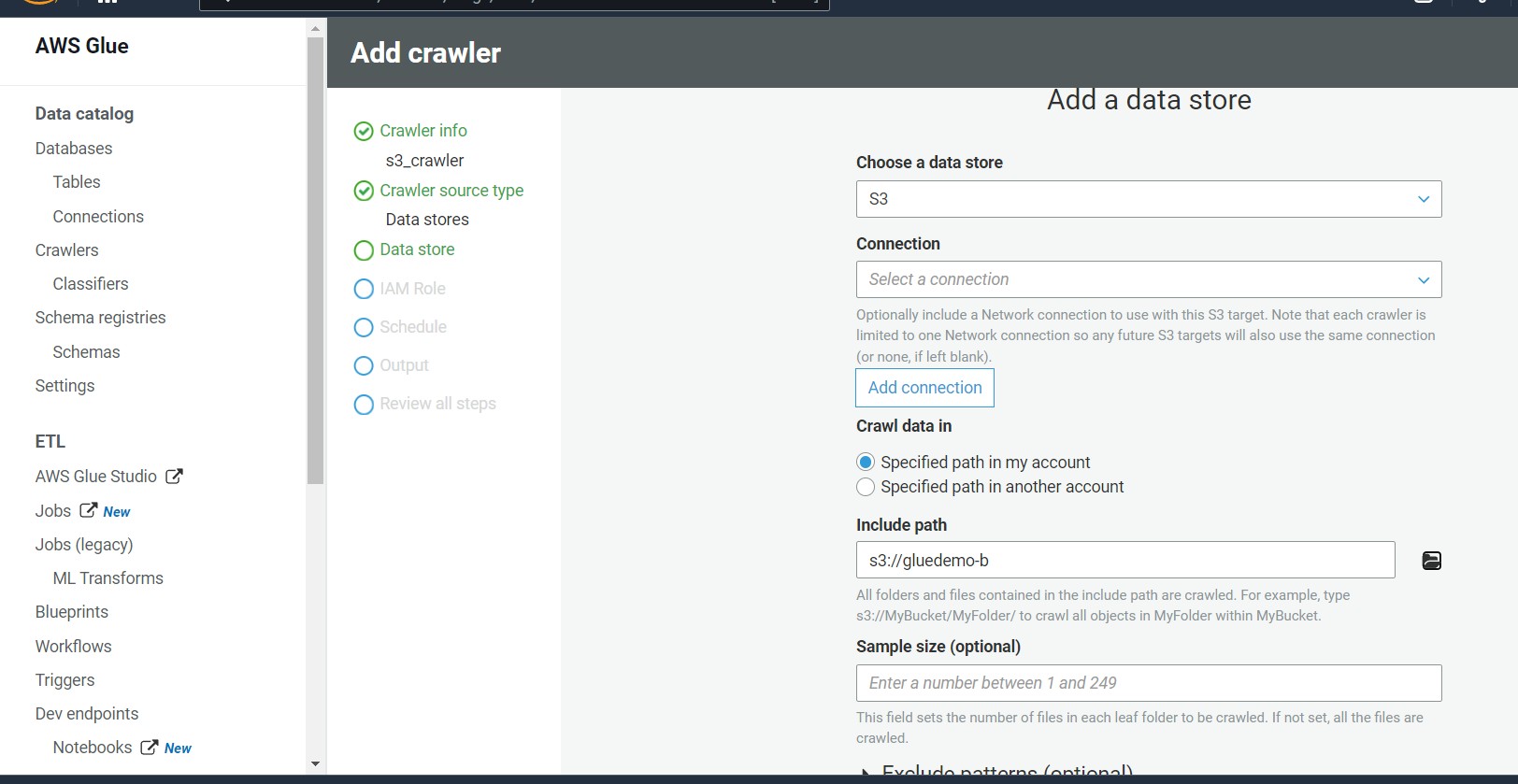
**Step 4:** Give crawler name and click

next.

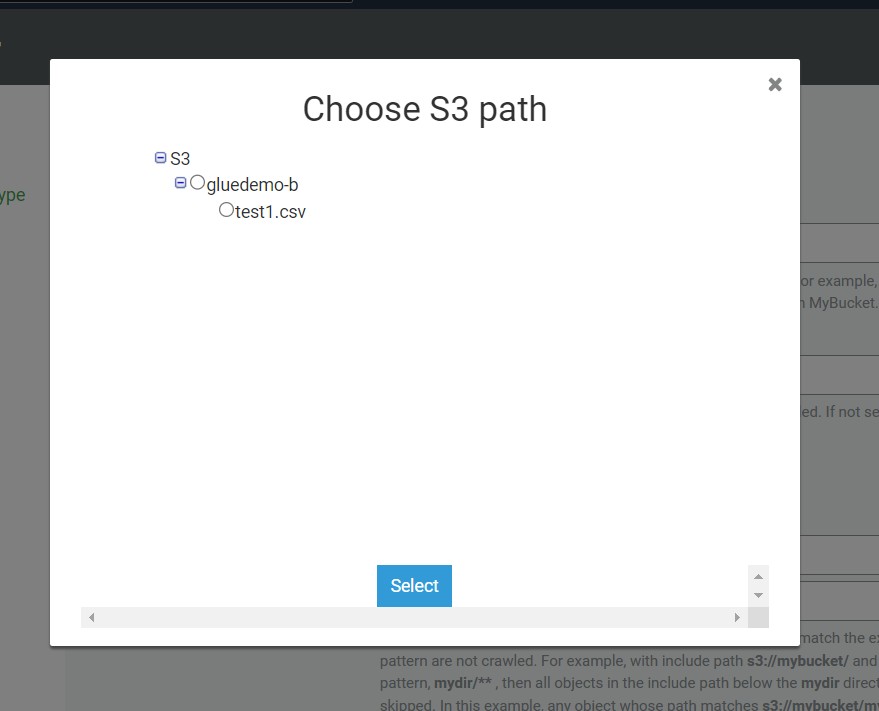
**Step 5:** Specify crawler source type



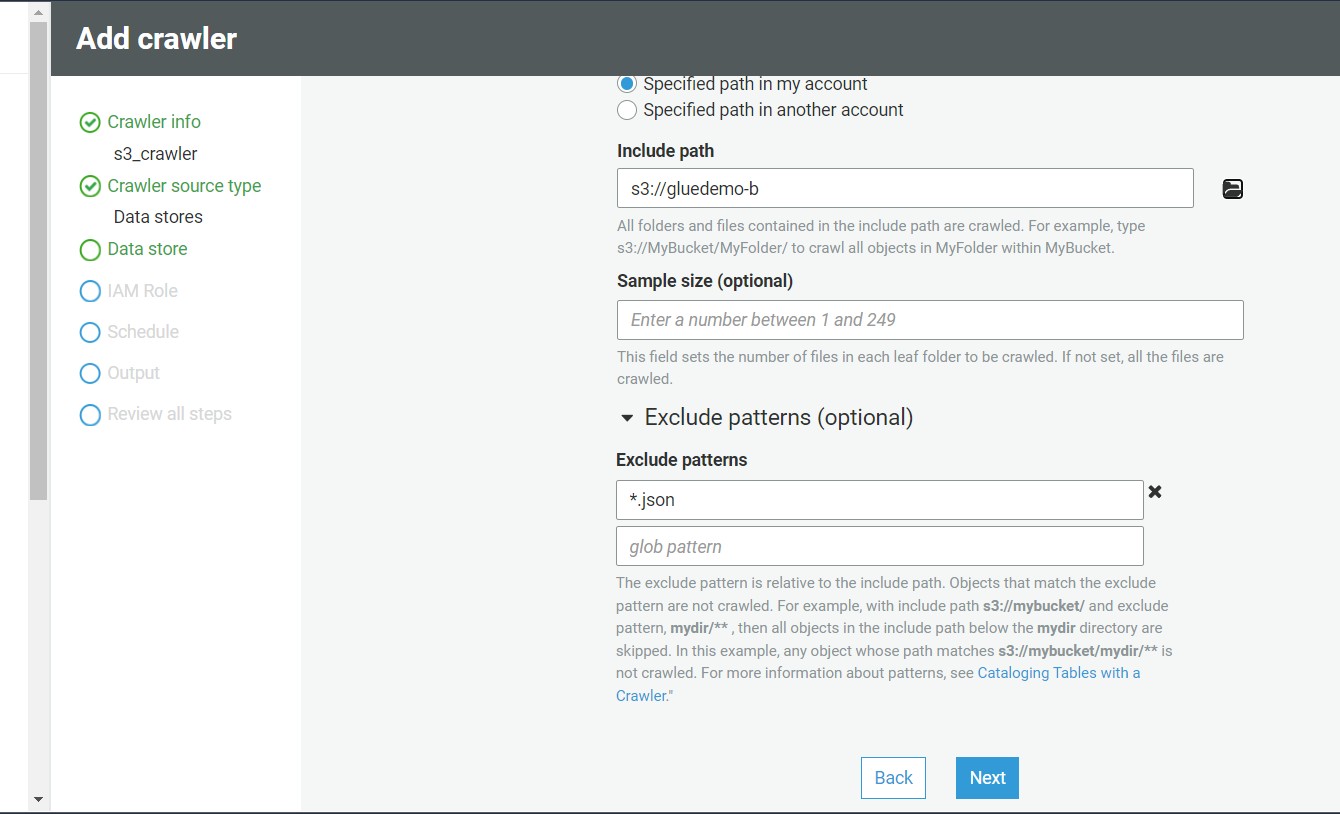
**Step 6:** Add a data source



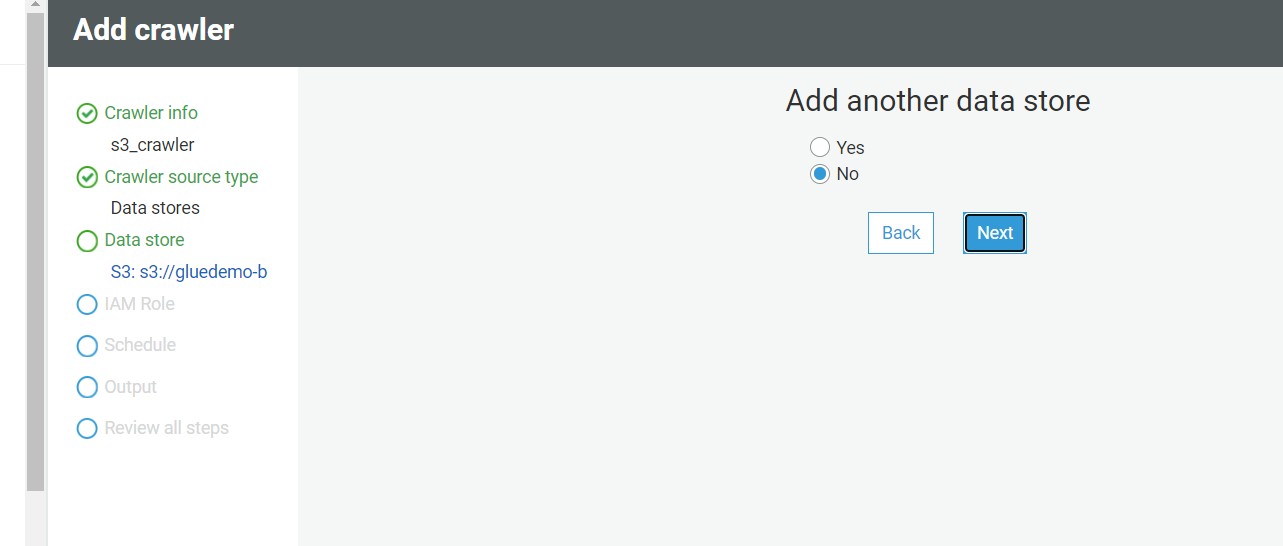
**Step 7:** Adding path of S3 data source

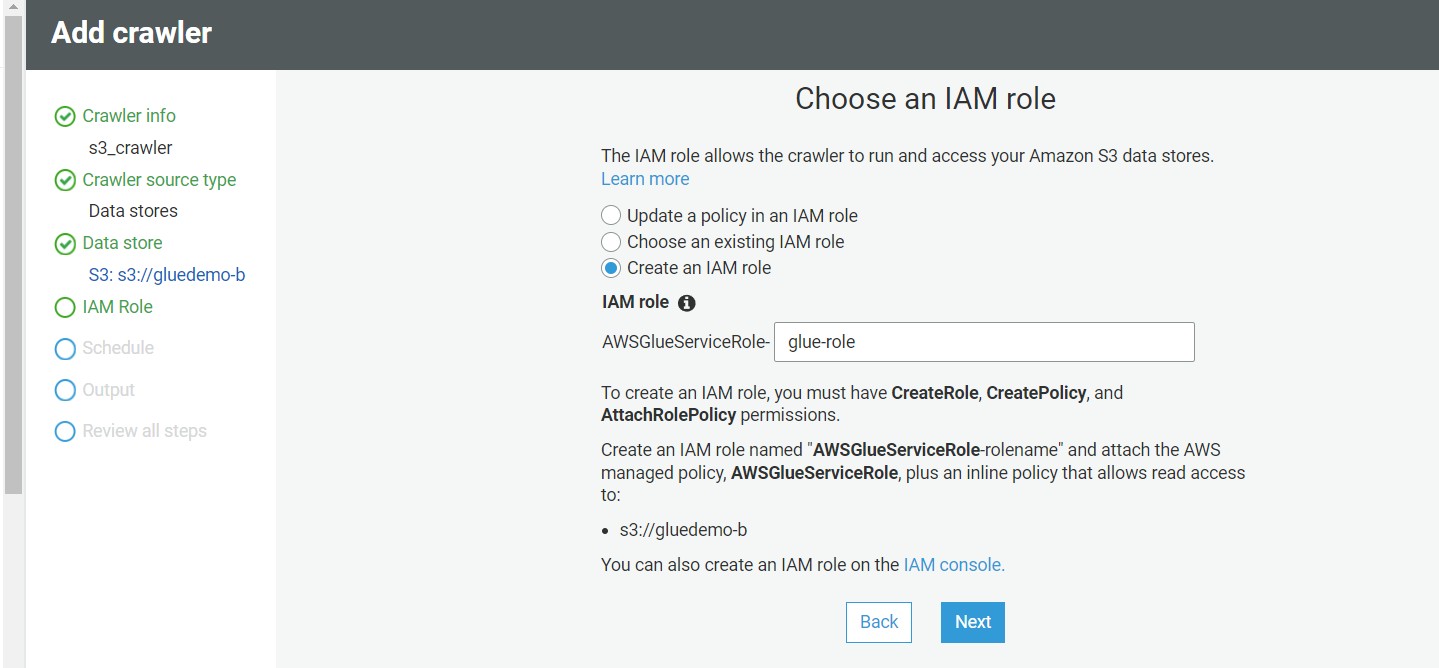


**Step 8:** Click on Next

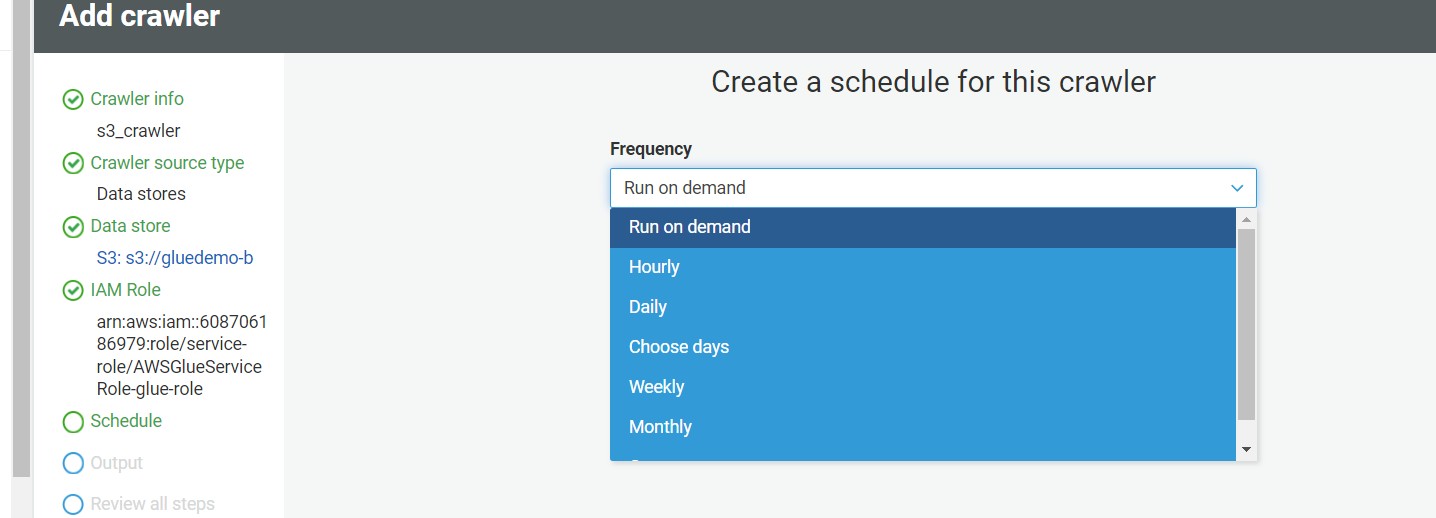


**Step 9:** If you wanna add another data source then click yes or else No

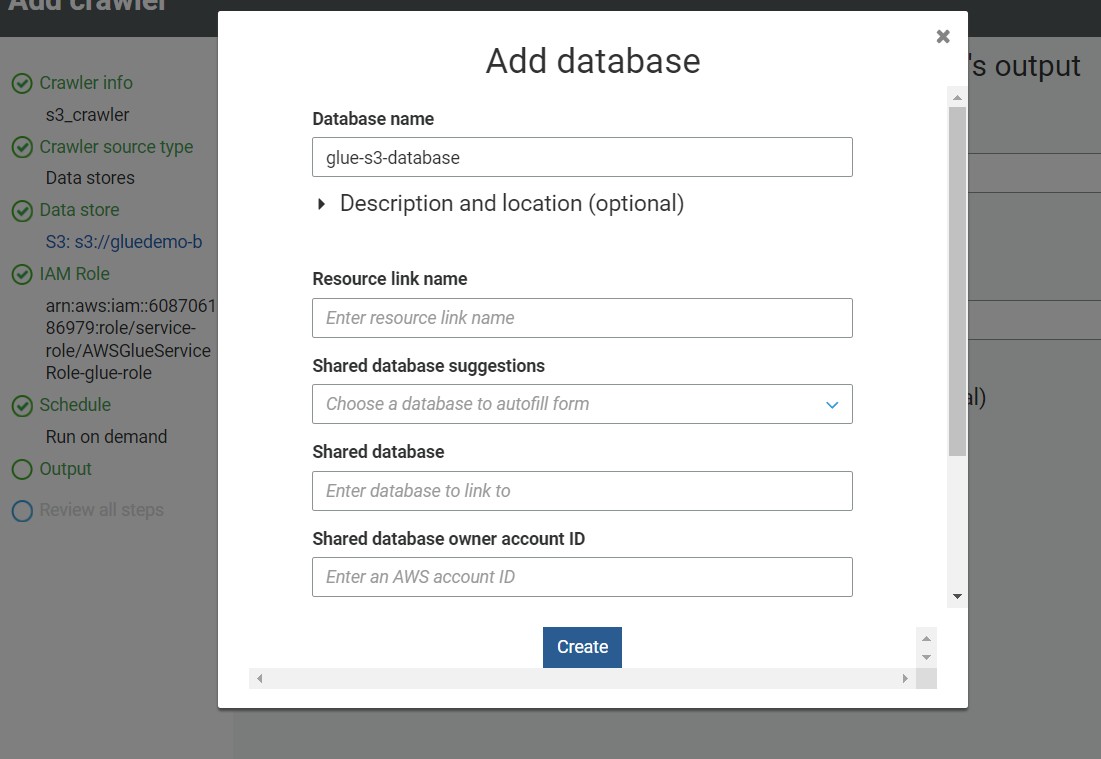


**Step 10**: Now choose IAM role , click on create an IAM role And select the aws service role for glue.

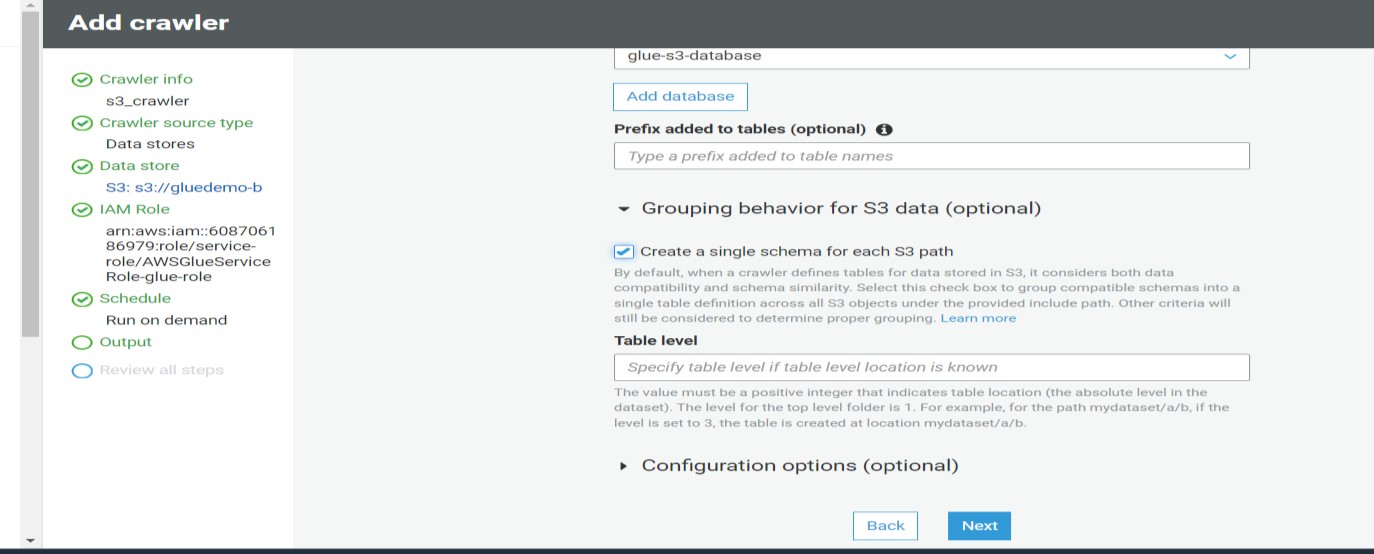
**Step 11:** Now click on Run on demand



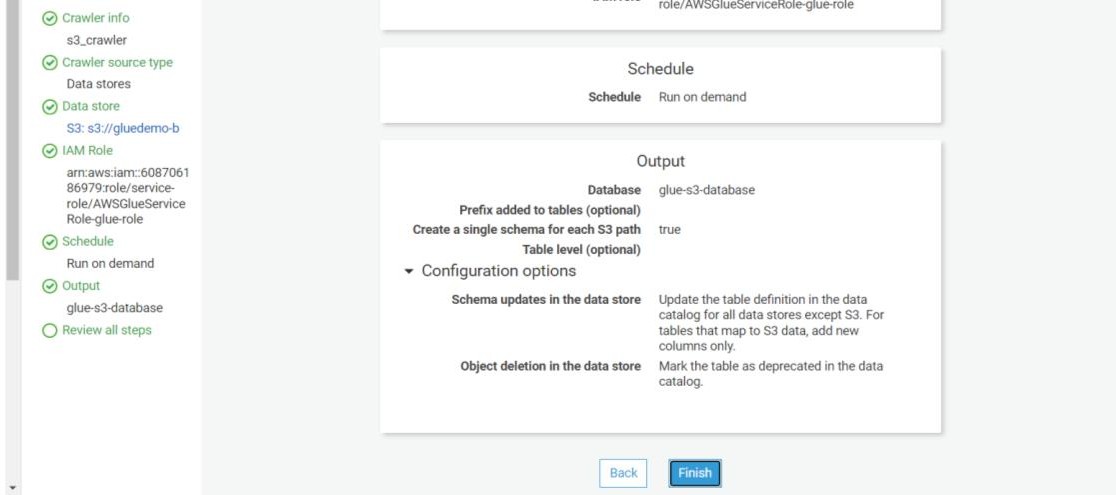
**Step 12:** Add a database as shown below

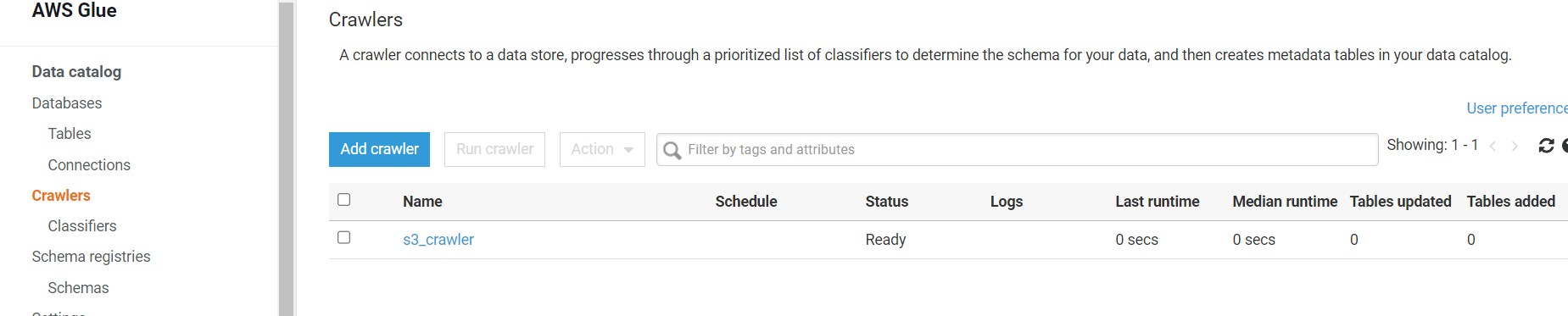


**Step 13:** after adding database click on next.

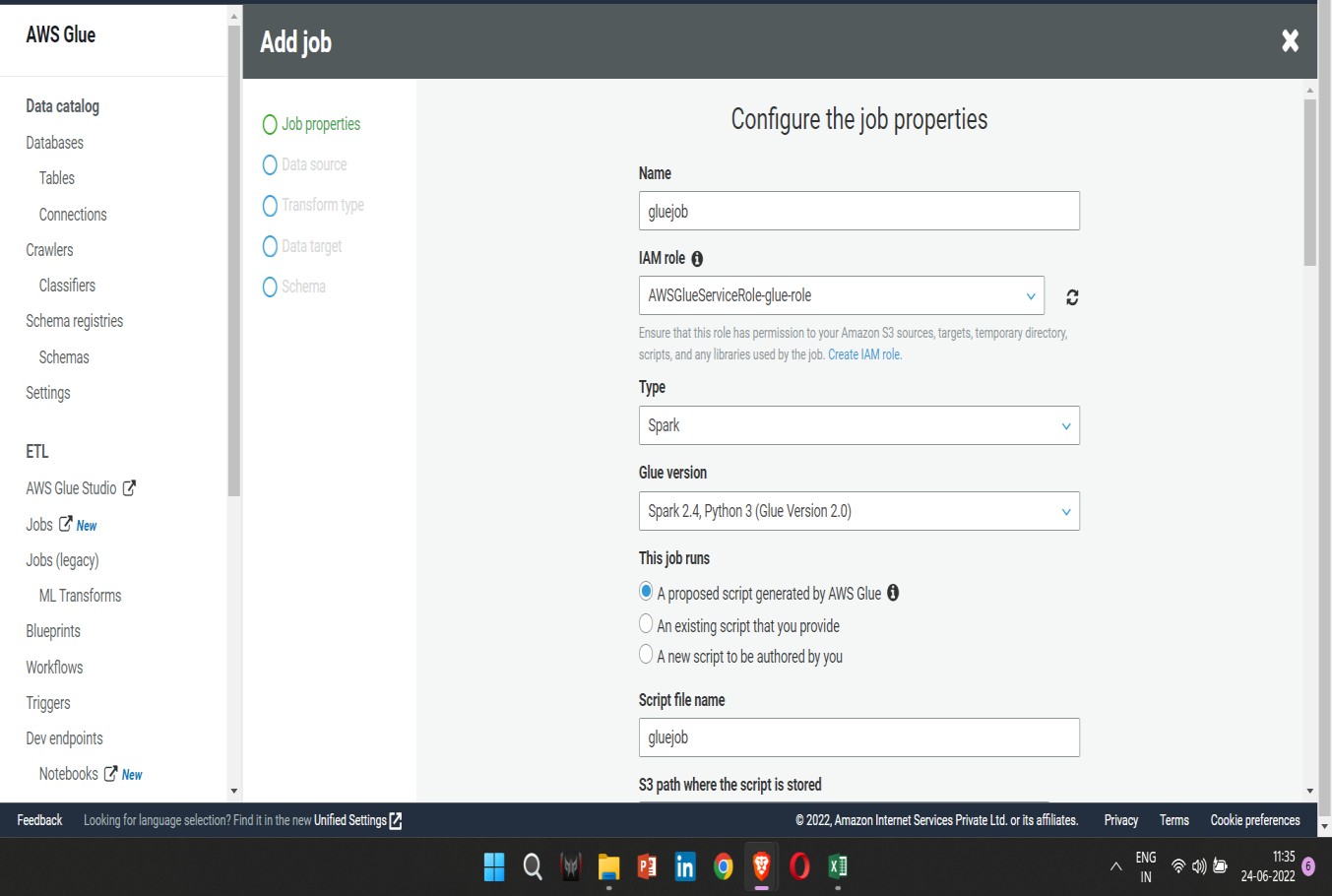


**Step 14:** Final Review and click on finish

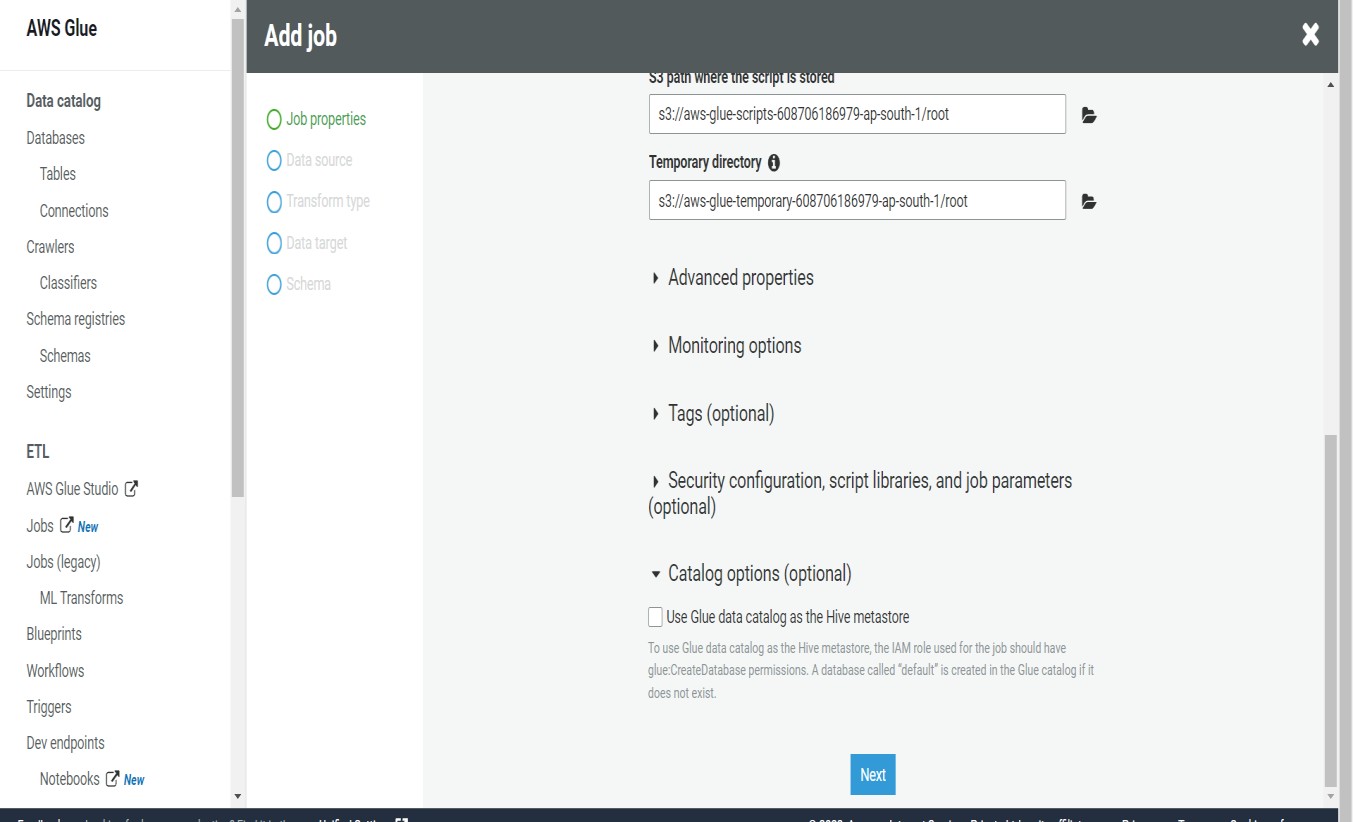


**Step 15:** After creating crawlers below interference will appear. Then select the created crawler and click on Run crawler.

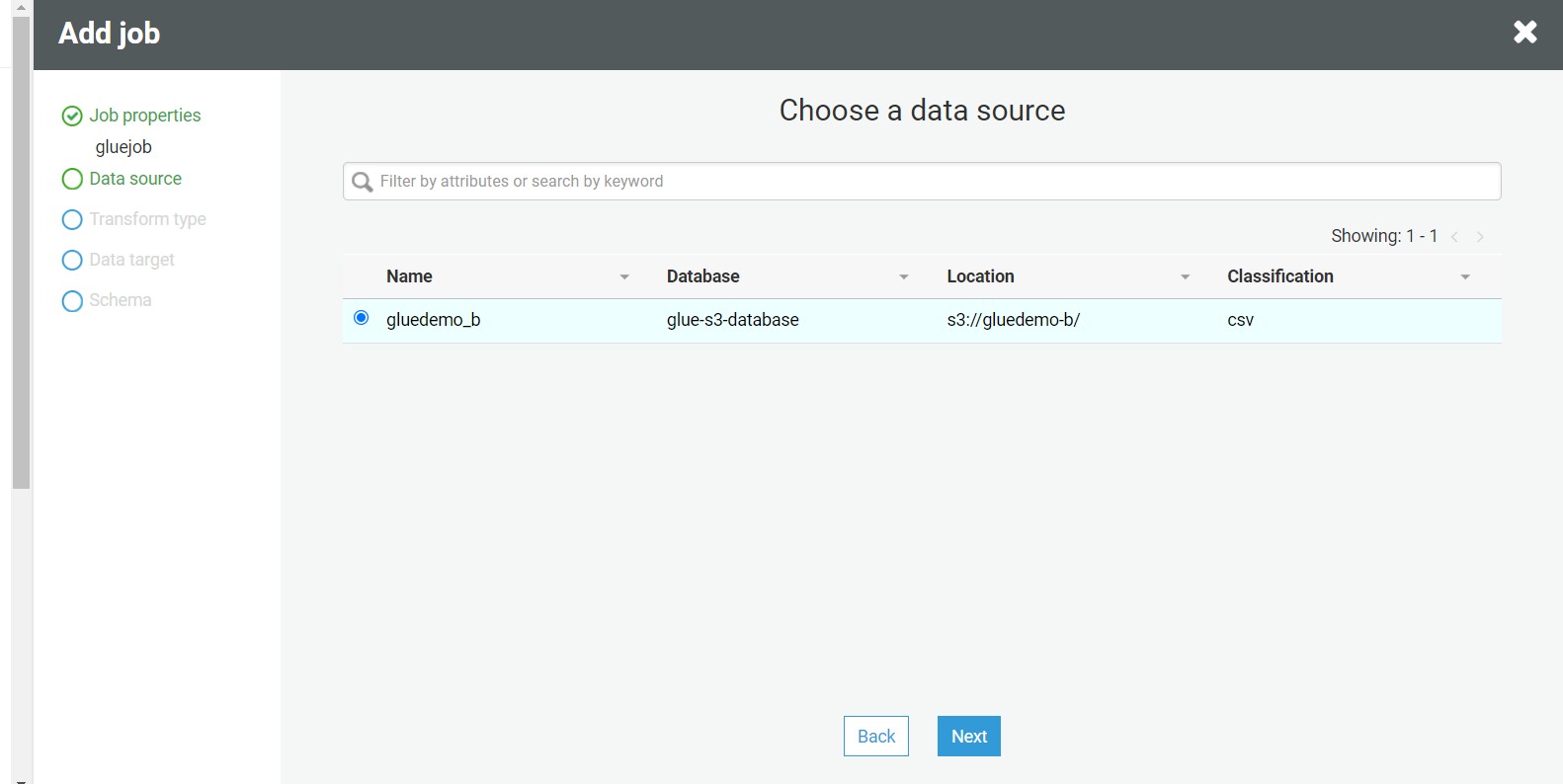
**Step 16:** Now we have to create Job so, click on Job (legecy)



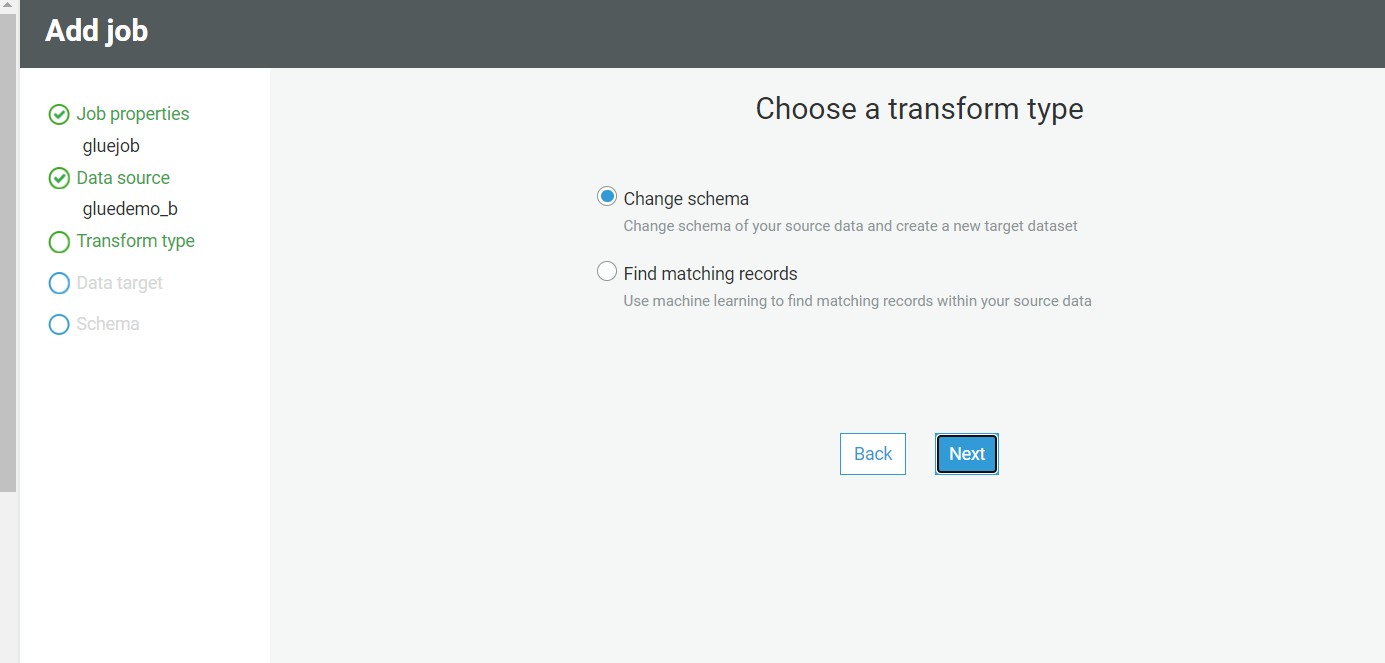
**Step 17:** Keep the default properties



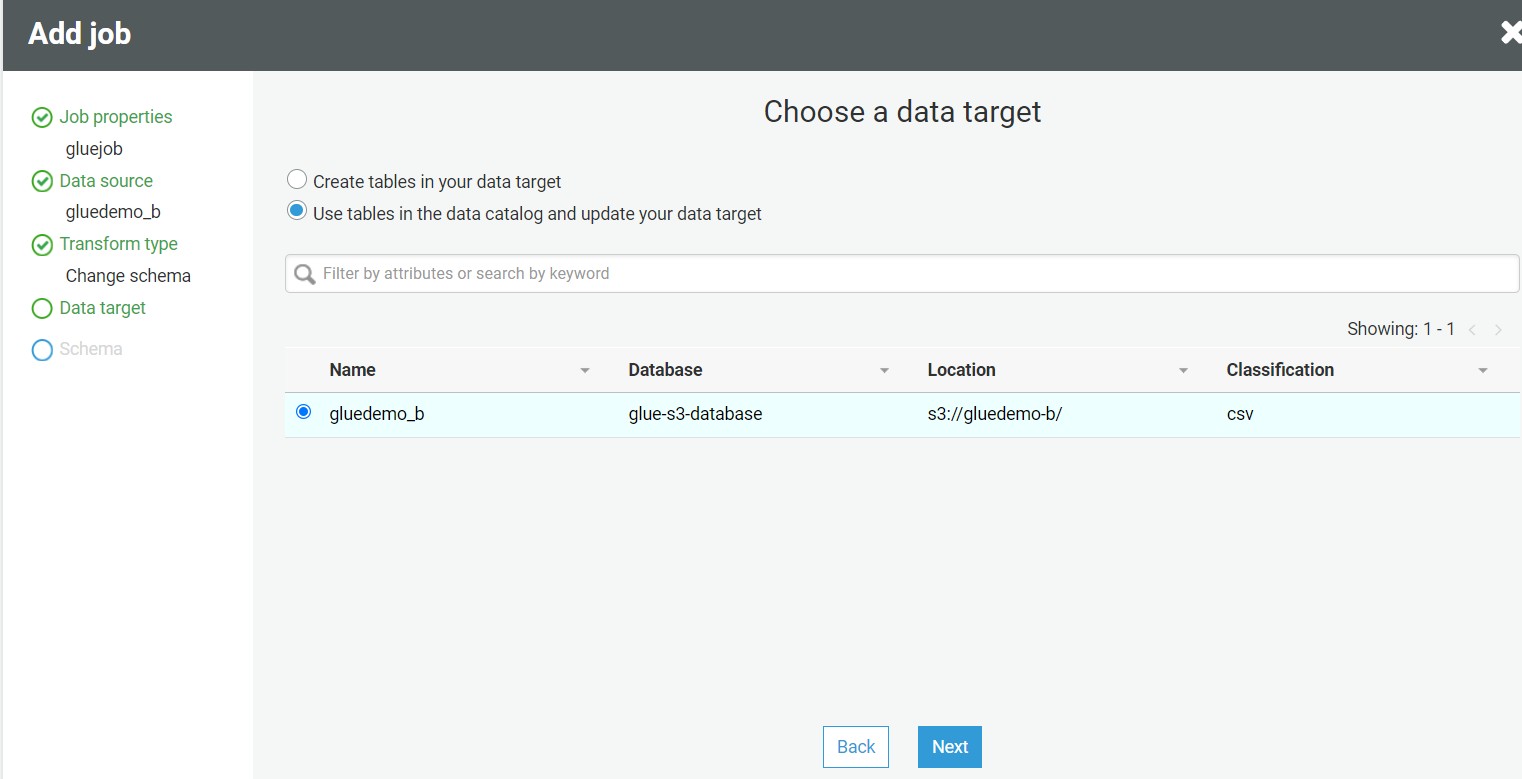
**Step 18:** Add data source

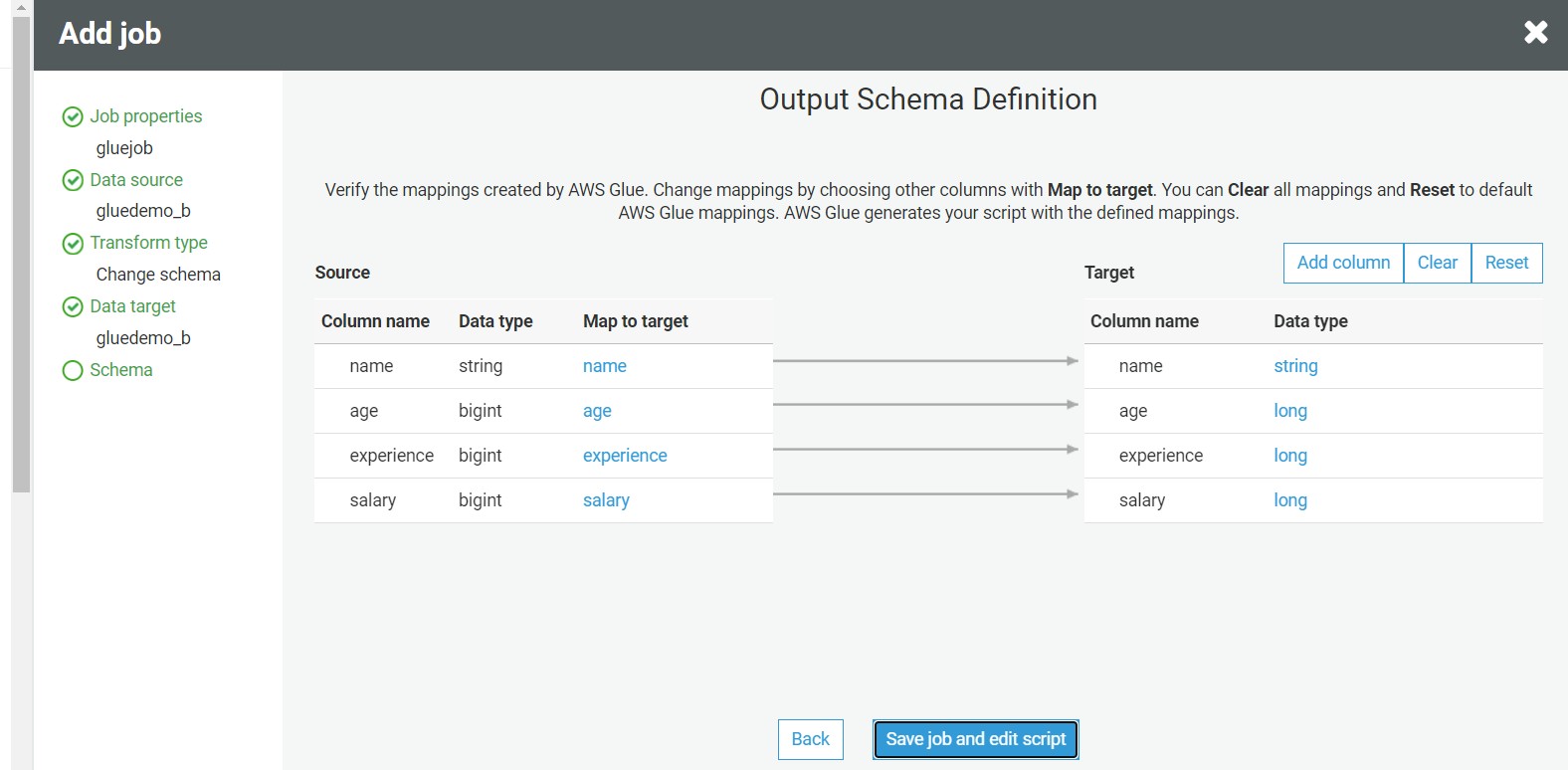


**Step 19:** choose a transform type

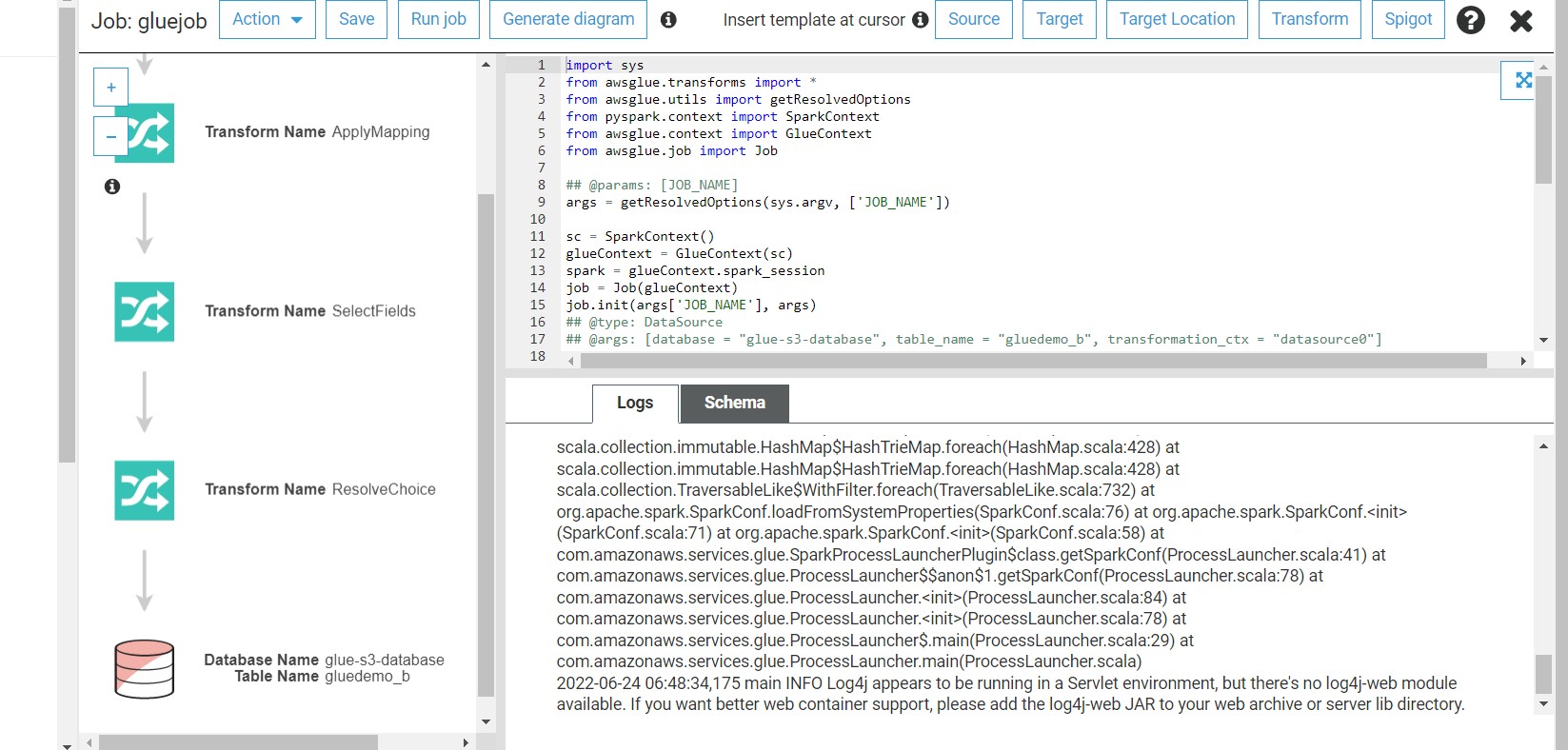


**Step 20**: Choose a data target



**Step 21:** This is the output schema. You can change it by clicking on edit

**Step 22:** now click on Run Job



**Step 23:** Job run successful .

