**Building ML Model to Identify Duplicates using Glue and Athena**

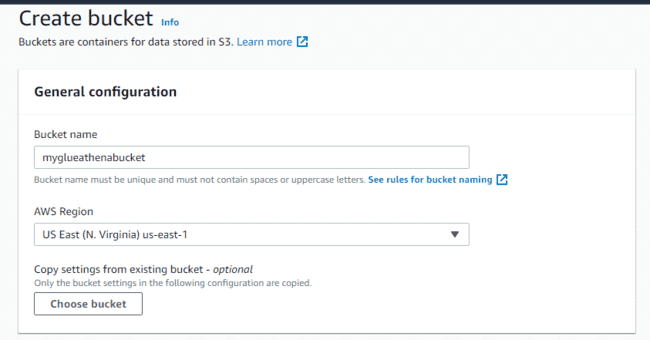
**S3 --> Glue --> Athena**

The raw data is stored on S3 as CSV files, using Glue for ETL, and using Athena to query the data.

**Prerequisite**

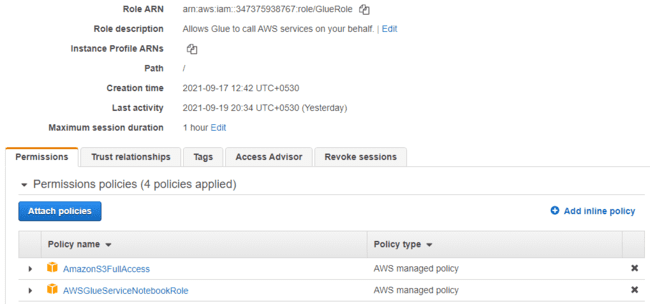
i. Create S3 bucket

* Log in to AWS Account and go to S3
* Create a S3 bucket



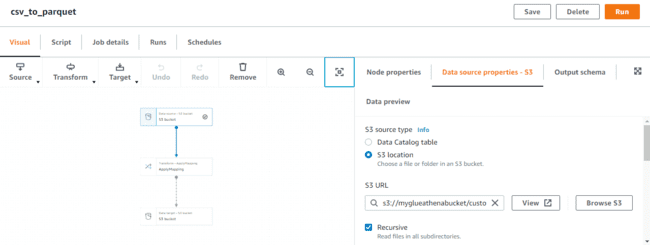
ii. Create IAM Role

* Pick an IAM role that has access to S3 and give the transformation a name.
* Choose three roles: AWSGlueConsoleAccess, AWSGlueServiceRole, AWSGlueServiceNotebook

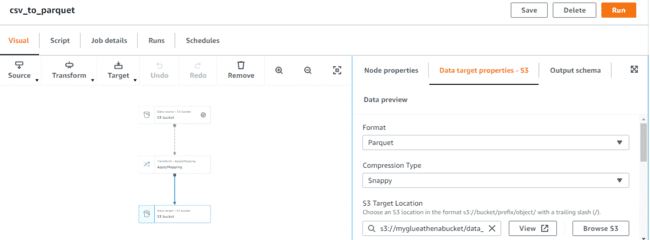


**2. Creating a job, Converting CSV to Parquet, and adding it to database using Glue Studio**

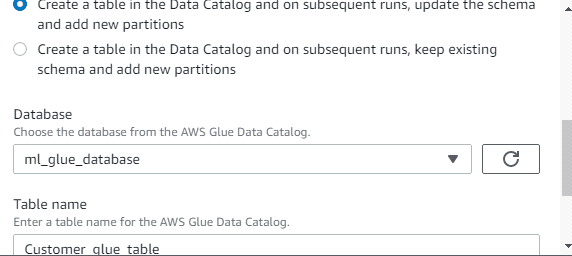
* Open Glue and select Glue Studio
* Click on job, and select Create Jobs
* Choose S3 location



* In the target S3 Bucket choose data format to Parquet.

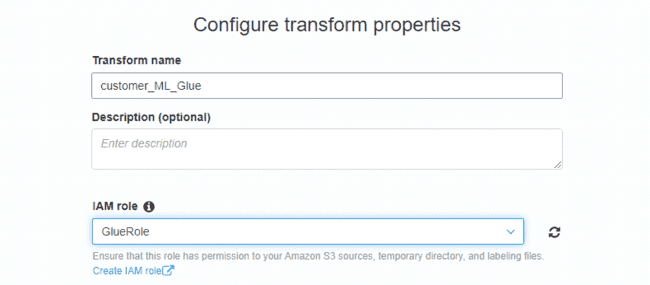


* Create a database and add a table name

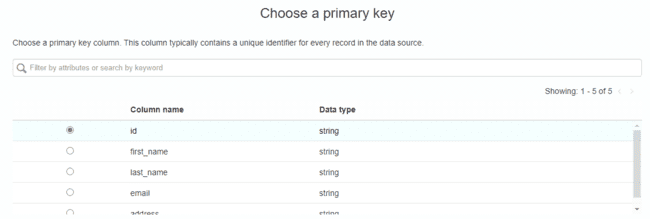


**3. Model Transformation**

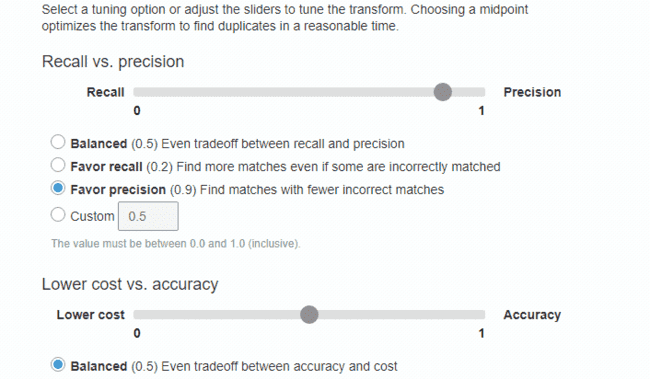
* Enter Transform Name
* Select the IAM Role created



* Select a Primary Key for the algorithm to do its matching logic.



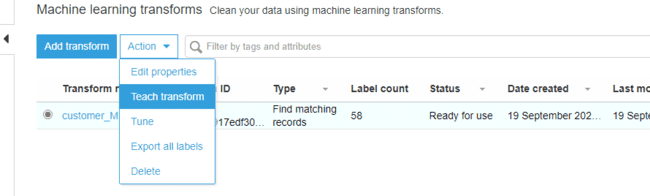
* Then it asks you to tune the transformation. These are tradeoffs between **cost vs accuracy** and **recall vs Precision,** usually precision is set to 0.9



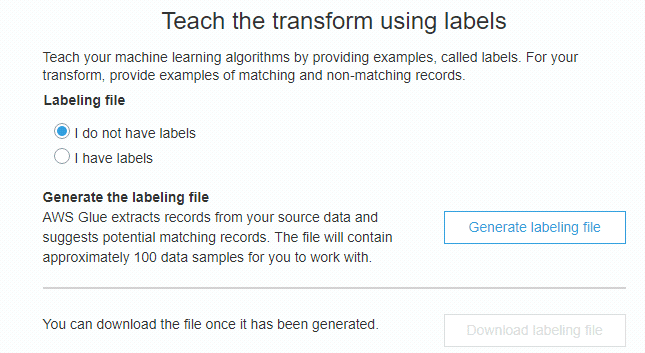
* You can't use a machine language transform in an extract, transform, and load (ETL) job until its status is **Ready for use**.

**4. Teach the model**

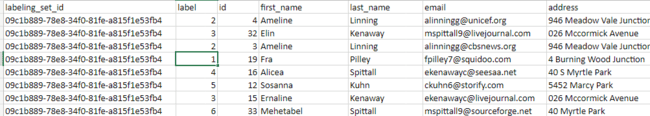
* To get your transform ready, you must teach it how to identify matching and nonmatching records by providing examples of matching and nonmatching records.
* Go to ML Transform and select the transform, from action select Teach transform



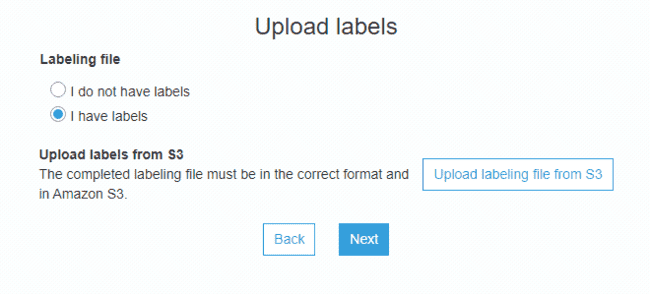
* To teach your transform, you can **Generate a label file.**



* **Add labels using Excel or spreadsheet** In column labels, mark the same number to all detected duplicates. All items not labelled as duplicate should be labelled with unique number, empty value is not allowed.

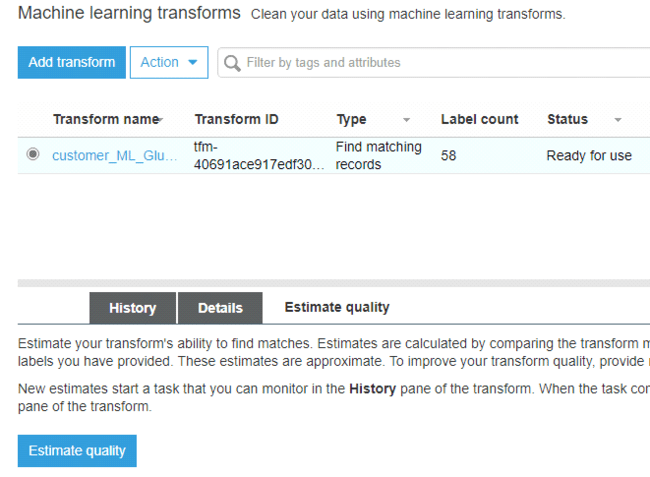


* Upload label file in S3 bucket, and then upload it to the transform to teach it



**5. Estimate the Quality of Your Machine Learning Transform**

* On the AWS Glue console, in the navigation pane, choose **ML Transforms**.
* Choose **Estimate quality** to start a task to estimate the quality of the transform. The accuracy of the quality estimate is based on the labeling of the source data.

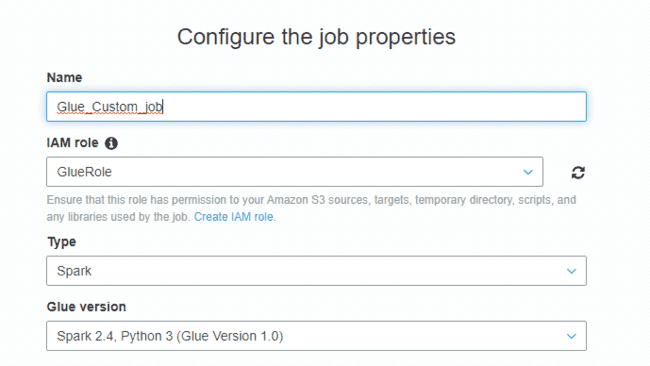


**6. Add and Run a Job with Your Machine Learning Transform**

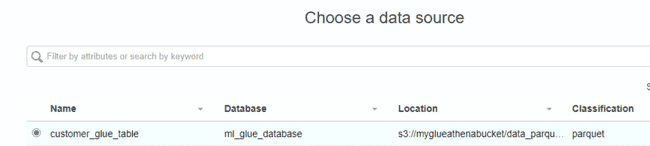
A job is the business logic that performs the extract, transform, and load (ETL) work in AWS Glue. When you start a job, AWS Glue runs a script that extracts data from sources, transforms the data, and loads it into targets.

* On the AWS Glue console, in the navigation pane, choose **Jobs**.
* Choose **Add job**, and follow the steps in the wizard to create an ETL Spark job with a generated script. Choose the following property values for your transform:
* For **Name**, choose a name for your job.
* For **IAM role**, choose an IAM role with permission to the Amazon S3 source data, labeling file, and AWS Glue API operations.
* For **Type**, choose Spark.
* For **Glue Version**, choose Spark 2.4, Python 3 (Glue Version 1.0).

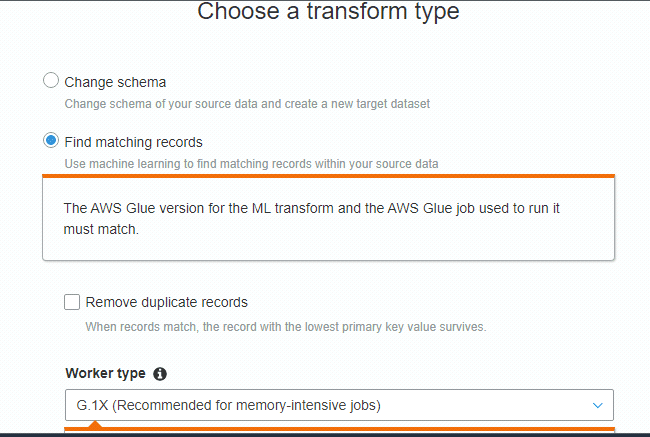
**#Spark version when creating the job and when transforming the job should match**



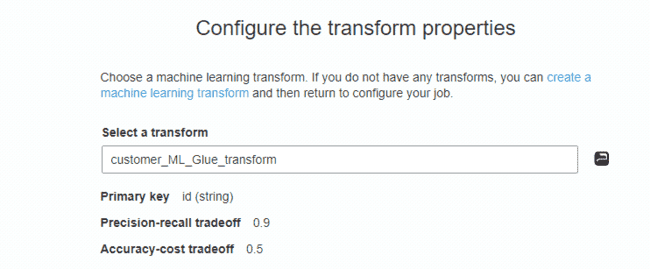
* For **Data source**, choose the data that was created by Glue Studio



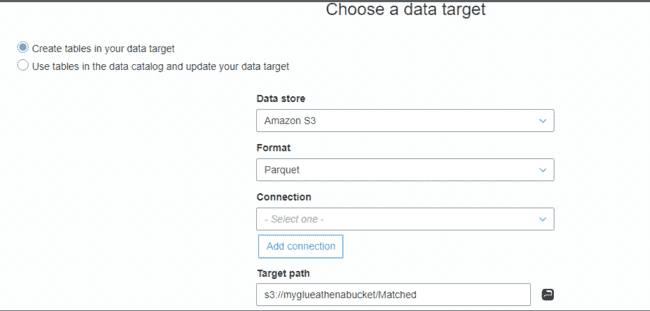
* For **Transform type**, choose **Find matching records**.



* For **Transform**, choose a transform file that was created by machine learning transform.



* For **Create tables in your data target**, choose to create tables with the following properties:
* **Data store type** —S3
* **Format** — Parquet
* **Target path** — The Amazon S3 path where the output of the job is written.

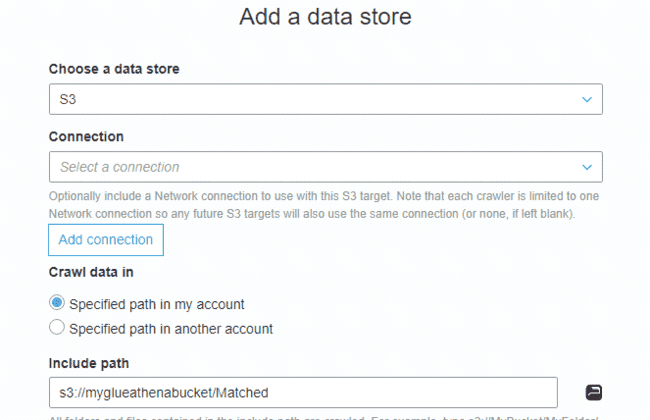


* Choose **Save job and edit script** to display the script editor page.

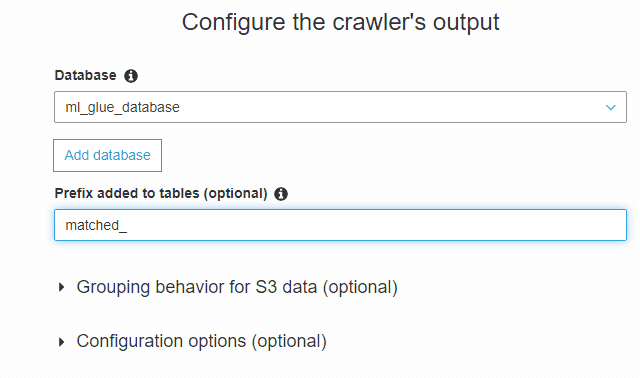
**7. Crawl the data source**

A crawler accesses your S3 bucket, extracts metadata, and creates table definitions in the AWS Glue Data Catalog. The Crawler pane in the AWS Glue console lists all the crawlers that you create.

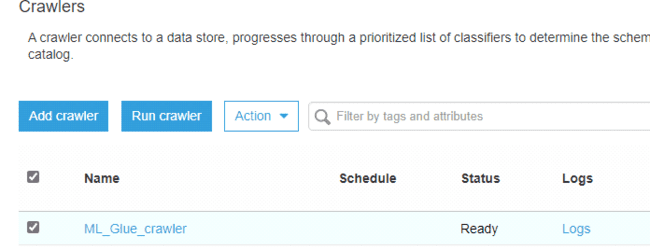
* Choose **Crawlers** in the navigation pane.
* Choose **Add crawler**, and select a name for crawler.
* Add data Store



* Select IAM role.
* In configure the crawler output select the database that was created using Glue Studio



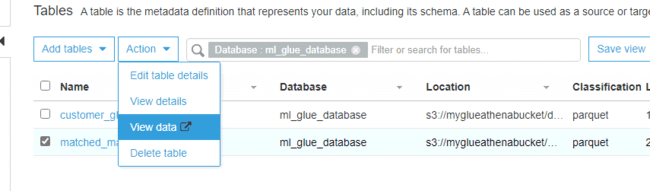
* Run the Crawler



* Choose **Tables** in the navigation pane to see the tables that were created by your crawler in the database that you specified.

**8. Preview Data using Athena**

* Click on Tables and choose the table that is created by the **crawler**
* From **Actions** choose View data
* Select **Preview Data**



* Athena Console will open
* Click on s**et up a query result** and select the folder
* From Tables choose the table and **Preview Table**
* From the Query remove limit 10 and add **ORDER BY** **match\_id** clause and Run Query
* Records that are duplicate have the same match\_id.

