## Lab 5-02: Creating a Glue Crawler for Athena and S3

### Service Introduction

AWS Glue is a fully managed ETL (Extract, Transform, Load) service that makes it easy to prepare and load data for analytics. Glue Crawler automatically discovers and catalogs data stored in your Amazon S3, making it readily available for querying with Amazon Athena, an interactive query service that makes it easy to analyze data directly in S3 using standard SQL.

### Problem

An organization has a large amount of semi-structured and structured data stored in various formats (CSV, JSON, Parquet, etc.) in Amazon S3. Manually cataloging data for querying in Athena can be time-consuming and prone to errors. Implementing an automated solution to discover and catalog data will streamline your data analytics workflow, ensuring accuracy and efficiency in the process.

### Solution

Deploy an AWS Glue Crawler to automatically discover and catalog your data stored in S3. The Glue Crawler will identify the data schema and store the metadata in the AWS Glue Data Catalog. Once cataloged, Amazon Athena allows you to run SQL queries directly on your S3 data, enabling quick and efficient analysis without manual preparation.

#### Task 1: Creating an S3 Bucket

|  |
| --- |
| 1. In the **AWS console,** navigate to **S3.**      1. In the upper right corner, click **Create bucket.**      1. Under Bucket name, enter a globally unique bucket name.      1. Leave the remaining settings at their defaults, and click **Create bucket**.      1. Select the bucket.      1. Under Files and Folders, upload the unisex\_names.csv file you saved to your machine from the GitHub repo. 2. Click **Upload.**      1. Once you see the file was successfully uploaded, click **Close** in the upper right corner**.** |

#### Task 2: Create a Glue Crawler

|  |
| --- |
| 1. In the top search bar, enter and select **AWS Glue.**      1. In the left navigation menu, under Data Catalog, click **Classifiers.** 2. In the upper right corner, click **Add classifier.**      1. Under Classifier type and properties > Classifier type, select CSV. 2. Under Classifier name, enter data.      1. Leave the other settings at default, and click **Create**      1. In the left navigation menu, under Data Catalog, click **Crawlers.** 2. In the upper right corner, click **Create crawler.**      1. Under **Name,** enter data. 2. Click **Next.**      1. Under Data source configuration, select **Not yet** option. 2. Under Data sources, click **Add a data source**.      1. In the Data source pop-up menu, enter the following values:   Data source: S3.   1. Under S3 path, click **Browse S3.**      1. Select the bucket you created. Click **Choose.**      1. Click **Add an S3 data source.**      1. Under Custom classifiers, click on the dropdown menu and select the data custom classifier. 2. Click **Next.**      1. Under the IAM role, click **Create new IAM role.**      1. In the Create new IAM role pop-up, enter touch at the end of the IAM role name, and click **Create.**      1. Click **Next.**      1. Under Target database, click on the dropdown menu; if a database isn't available, click **Add database.**      1. In the new Databases browser tab, under Name, enter a unique database name. 2. Leave the other settings at default and click **Create database.**      1. Navigate back to the Crawlers browser tab, and under Target database, click on the dropdown menu; you should now see the database you created.   **Note**: You may need to click the Refresh icon a few times for the new database to load.   1. Leave the other settings at default, then click **Next.**      1. Scroll down and click **Create crawler.**      1. Under Crawler runs, click Run crawler. This process may take a few minutes. |

#### Task 3: Check Amazon Athena

|  |
| --- |
| 1. In the search bar at the top, search for **Athena**, right-click on it, and select **Open in new tab** to access the Athena service.      1. In the upper right corner, click **Launch query editor.**      1. In the banner at the top stating before you run your first query, you need to set up a query result location in Amazon S3, click **Edit settings.**      1. Under Location of query result, click **Browse S3** and click on **Save.**      1. Select the bucket that is shown. 2. Click **Choose.** 3. Click the Editor tab. 4. Under Tables and views, you should see the listed table. 5. Click on the listed table to review the columns. 6. Under Query 1, test the table:   select \* from "<TABLE\_NAME>"   1. Click **Run.**      1. Under Results, you should see the names, total number of baby names, and shares between the two. |