Glue - lab- Step-4

07:56 AM

Analyze Sales Data with Athena

In previous challenges, you crawled the input CSV file and the output Parquet files,
which populated the Glue Data Catalog with details about those files.

The beauty of the Athena service is that you can easily run SQL queries against files on S3.

You do not need the classic approach of importing those files into a database to run SQL queries.

Still, Athena needs to know details about the files on S3, which the Glue Data Catalog can offer. Since you already populated the Glue Data Catalog in previous challenges, now you can use Athena to run SQL queries against either the input or the output sales data on S3.

1. Use the search bar at the top of the AWS Console to navigate to the **S3** service.
2. Copy the name of the **aws-glue-athena-lab-...** bucket for use later on.

1. Now navigate to the **Athena** service using the search bar at the top of the page.

1. You are now on the **Query editor** page. On the left side of the page, you can see the **sales-database** is already selected under the **Database** field. Also, under the **Tables** field, you can see the **input_data** and **output_data** tables that you created in the previous challenges.

1. The center of the **Query editor** page has a text box in which you can write queries. Above the text box, notice the following message: **Before you run your first query, you need to set up a query result location in Amazon S3**.
2. Click on the **set up a query result location in Amazon S3** text. In the **Settings** dialog that pops up, in the **Query result location** field, do the following. Enter s3:// to show it refers to a S3 bucket. Then, paste the bucket name from your clipboard. Finally, enter /athena-results/ to indicate the folder.
   The resulting path should look something like this:
   s3://aws-glue-athena-lab-bf060el7/athena-results/
3. Click the **Save** button.
4. In the **New query 1** tab of the query editor, enter the following query to look at a few records from the original input CSV file:
   SELECT * FROM input_data LIMIT 10
5. Click the **Run query** button. A few seconds later, you can see ten records with three columns: id, name and amount.
6. In the **New query 1** tab, delete the query so that the editor is empty. Enter the following query: SELECT * FROM output_data LIMIT 10 and click the **Run query** button.
7.
8. Expect to get other records, and the same three columns. This confirms that the output Parquet files have the same schema as the input CSV file.
9.

Congratulations for querying both the input and output sales data with Athena!

You can analyze the data with even more queries. Here is how to get the total sales amount. In the **New query 1** tab of the query editor, delete the query so that the editor is empty, then enter the following query: SELECT SUM(amount) FROM output_data and click the **Run query** button.

Congratulations again for getting the total sales amount (49961659) and finishing this Lab!