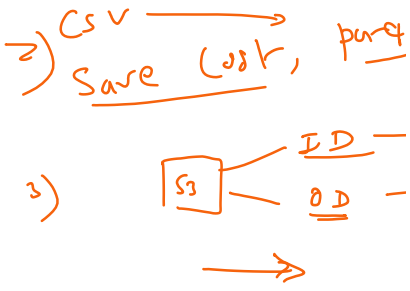
*Glue - lab - step 3

07:54 AM



Crawl Output Sales Data on S3

To analyze the data in the Parquet files, you need to crawl the data and save its schema to the Glue Data Catalog. This is quite similar to the first challenge, in which you crawled the input CSV file.

1. Navigate back to the **AWS Glue** service if you're not already there.
2. On the left-hand **AWS Glue** column menu under the **Data catalog** section, click on the **Crawlers** item.
3. Click on the **Add crawler** button to start the crawler creation wizard.
4. On the **Add information about your crawler** page, under **Crawler name**, enter output-data-crawler, and click on the **Next** button.
5. On the **Specify crawler source type** page, keep defaults and click on the **Next** button.
6. On the **Add a data store** page, under the **Include path** label, click on the small folder icon at the right of the text box. A **Choose S3 path** dialog appears. Click on the **+** sign next to the **aws-glue-athena-lab-...** bucket, then select the **output-data** folder, and click on the **Select** button.
7. The S3 path to the **output-data** folder is now entered in the **Include path** field. Click on the **Next** button.
8. On the **Add another data store** page, click **Next**.
9. On the **Choose an IAM role** page, click to select the **Choose an existing IAM role** option. Under the **IAM role** label, click on the **Choose an IAM role** drop-down, and select the **AWSGlueServiceRole-SalesData** role. Click on the **Next** button.
10. On the **Create a schedule for this crawler** page, leave the default value and click on the **Next** button.
11. On the **Configure the crawler's output** page, under the **Database** label, click on the **Choose a database...** drop-down, and select **sales-database**. Click on the **Next** button.
12. On the review page, scroll down and click on the **Finish** button.
13. The new crawler is created and you are now back to the **Crawlers** page. Click the checkbox next to the **output-data-crawler** crawler, then press the above **Run crawler** button.
14. Watch the **Status** field of **output-data-crawler** as it changes to **Starting**. Wait until it changes to **Stopping**, and then back to **Ready**. The last column on the right is named **Tables added** and it indicates that one table is now added to the Glue Data Catalog by **output-data-crawler**.

Congratulations for successfully adding another table to the Glue Data Catalog!

To check the schema of the new table, on the left-hand **AWS Glue** column menu, under **Databases**, click on **Tables** and you can see the **input_data** and **output_data** tables (if you don't see **output_data**, you may need to click the refresh button on the top-right of the page. Click on the **output_data** table and scroll down to the **Schema** section. Notice the three column names (id, name, amount) and their data types. Just like in the first challenge, you did not specify them manually, instead the Glue Crawler recognized them automatically for you by crawling the Parquet files.