*Glue-lab -step2

07:51 AM

Convert Sales Data with a Glue Job

The original raw input CSV file is about 1.8MB in size. However, the IT policy of Globomantics states that the Parquet format is preferred, to better support complex data processing across all departments, including saving on storage to lower costs. You need a new Glue Job to convert the CSV sales data into the Parquet format.

1. On the left-hand **AWS Glue** menu, under the **ETL** section, click on **Jobs**.
2. On the **Jobs** page, click on the **Add job** button to start the job creation wizard.
3. On the **Configure the job properties** page, under **Name**, type convert-to-parquet. Under the **IAM role** label, click on the drop-down and select **AWSGlueServiceRole-SalesData**, which you created previously. Scroll down and click on the **Next** button.
4. On the **Choose a data source** page, select the **input_data** table that you created in the previous challenge, and click on the **Next** button.
5. On the **Choose a transform type** page, leave the **Change schema** option selected and click on the **Next** button.
6. On the **Choose a data target** page, click on the **Create tables in your data target** option. Under **Data store**, click on the drop-down and select **Amazon S3**. Under **Format**, click on the dropdown and select **Parquet**.
7. Under **Target path**, click on the small folder icon to the right of the text box. In the **Choose S3 path** dialog, click on the **+** sign next to the **aws-glue-athena-lab-...** bucket, select the **output-data** folder and click on the **Select** button. The **Target path** is now set to the **output-data** S3 path, so click on the **Next** button.
8. On the **Output Schema Definition** page, you have the option to delete some columns from the target or re-arrange target columns. For now, leave defaults and click on the **Save job and edit script** button.
9. Press **Esc** or click on the small top-right **x** to dismiss the **Script editor tips** dialog.
10. On the script editor page, click on the **Save** button at the top of the page.
    On the left of the page, you can see a diagram with boxes and arrows. From top to bottom, you can see the **input_data** table from the **sales-database** database, three transformation steps, and finally the **output-data** path on S3. The diagram summarizes the code in the editor, which can be used to add custom logic.
11. Keep the current code and close the editor by clicking on the top-right **X** label of the script editor, next to the label with a question mark.
12. You can now see the list of Glue Jobs. Select the **convert-to-parquet** job, then click on the **Action** button with a drop-down, and click the **Run job** action. A dialog with **Parameters (optional)** appears, click the **Run job** button in the dialog.
13. Again, you can see the list of Glue Jobs. Select again the **convert-to-parquet** job. More details appear at the bottom of the page with details about the job. The **History** tab shows a table that starts with **Run ID**. Wait until the **Run status** column changes from the **Running** to the **Succeeded** status.

Congratulations for successfully running the conversion Glue job!
Here is how you can check the output of this Glue job. On the top-left of the page, click on **Services** to see all services, then click on **S3** under the **Storage** section. Next, click on the **aws-glue-athena-lab-...** bucket, then click on the **output-data** folder.
Look at the sizes of the Parquet files: in total they occupy around 600KB. This is about 3 times less storage than the input file, so congratulations again for lowering storage costs!