

Glue lab - step1

07:34 AM

ETL

Transform Data Using AWS Glue and Amazon Athena

you'll practice transforming data with ~~AWS Glue~~ and then analyzing the result with ~~AWS Athena~~. When you're finished with this lab, you'll have a solution to transform and query data.

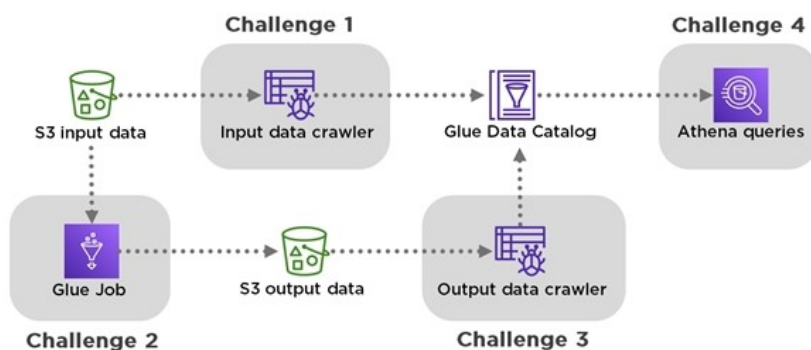
1. **Crawl Input Sales Data on S3**
2. **Convert Sales Data with a Glue Job**
3. **Crawl Output Sales Data on S3**
4. **Analyze Sales Data with Athena**

Crawl Input Sales Data on S3

You received a CSV file with 100k sales records from the Globomantics Corporation. You need to achieve two objectives. First,

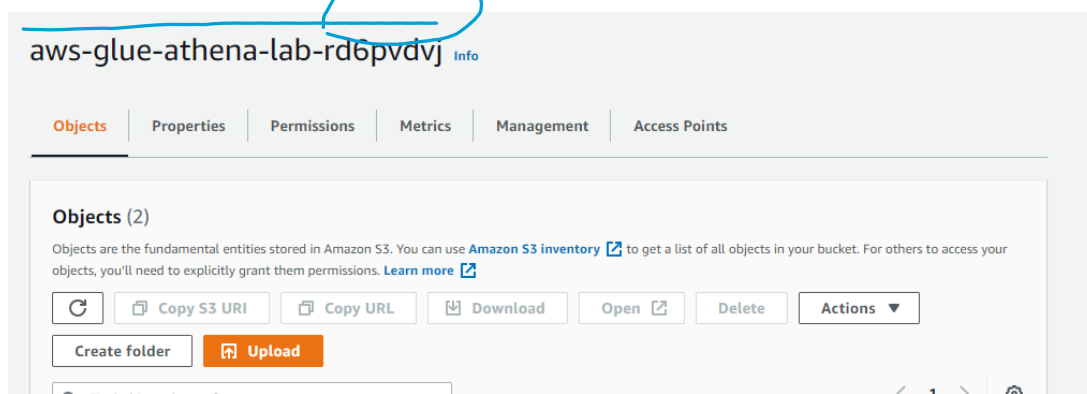
transform the file into a Parquet file compressed with Snappy, so that it takes less storage and can be used by other Globomantics departments. Second, analyze the data and return the total sales amount.

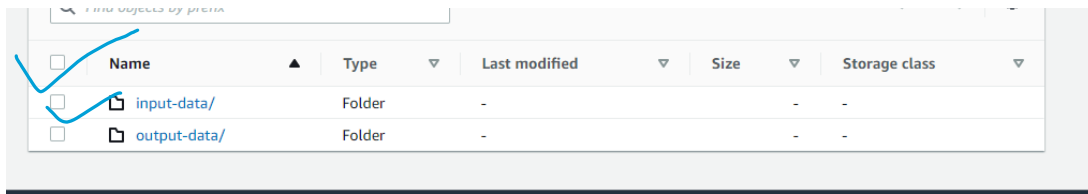
This diagram shows the high-level pictures of how the challenges fit together. The S3 input data has 100k lines with sales details. These challenges will help you analyze the input sales data using a high-performance and low-cost approach.



Note: We will create s3 bucket so that we can use for input and output part.

aws-glue-athena-lab-<random number>





Name	Type	Last modified	Size	Storage class
input-data/	Folder	-	-	-
output-data/	Folder	-	-	-



sales

~~Upload in input data folder~~

The first challenge is about creating a Glue crawler that crawls the sales input data on S3 and writes details about it into the Glue Data Catalog.

Here is what the input sales data looks like:

The first challenge is about creating a Glue crawler that crawls the sales input data on S3 and writes details about it into the Glue Data Catalog.

Here is what the input sales data looks like:

id,name,amount

1,Diego,671

2,Cristiano,405

...

100000,Roddy,652

Click the Open AWS console button to the right of these instructions, then use the provided credentials to log in.

Make sure you are in the proper region – the region selector is in the top right hand corner and it should display Oregon; if not, set the region to US West (Oregon).

Use the search box at the top of the page to navigate to the AWS Glue service.

From the left-hand menu underneath Data catalog, click on Crawlers.

On the Crawlers page, click on Add crawler.

Under Crawler name, enter **input-data-crawler** and click Next.

Leave defaults on the Specify crawler source type page and click Next.

On the Add a data store page, notice that S3 is already selected as a data store and we need to specify the Include path, as follows:

- Click on the folder icon at the right of the textbox that displays `s3://bucket/prefix/object`. After clicking, a dialog with Choose S3 path appears.
- Click on the + sign next to the **aws-glue-athena-lab-...** bucket.
- Click on input-data to select it.

- Click on the Select button. This will select the S3 path to the input sales data and close the dialog.

- Back in the Add a data store page, click the Next button.

On the Add another data store page, leave the default No selected and click Next.

On the Choose an IAM role page, notice Create an IAM role is already selected and it needs a name. In the text box displaying Type a name..., type SalesData and click Next.

On the Create a schedule for this crawler page, leave Run on demand selected and click Next.

On the Configure the crawler's output page, you define a database in the Glue Data Catalog to store the crawler's output. Click on the Add database button, under the Database name type sales-database, and click Create. The sales-database is now selected in the Database field, so click Next.

In the page with an overview of all steps, scroll down and click on the Finish button.

On the Crawlers page you can see the input-data-crawler and its Ready status. Click on the checkbox next to input-data-crawler, then click on the Run crawler button above it.

Watch the Status field of input-data-crawler as it changes to Starting. After about a minute, it changes to Stopping. Wait about a minute more until it changes 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.

Congratulations for successfully adding a table to the Glue Data Catalog! To look at the new table, on the left-hand AWS Glue column menu, under Databases, click on Tables and you can see the input_data table. Click on the input_data table and scroll down to the Schema section. Notice the three column names (id, name, amount) and their data types. You did not specify them manually, instead the Glue Crawler recognized them automatically for you by crawling the input sales data on S3.