*Implement a Data Ingestion Solution Using AWS Glue

Obtain Source Data Files
Your company wants to continuously extract semi-structured production data from S3 that is created each day, transform it into a normalized schema, and load it into a downstream relational database for business intelligence reporting and analysis. In preparation for this, you decide to set up a development environment and implement AWS Glue using sample data.

Navigate to https://awsglue-datasets.s3.amazonaws.com/examples/us-legislators/all/areas.json, and save it to your local computer under the same file name, areas.json.

Note: You will use this and the subsequent JSON files later in this lab.

Repeat the first task for the following:

https://awsglue-datasets.s3.amazonaws.com/examples/us-legislators/all/countries.json

https://awsglue-datasets.s3.amazonaws.com/examples/us-legislators/all/events.json

https://awsglue-datasets.s3.amazonaws.com/examples/us-legislators/all/memberships.json

https://awsglue-datasets.s3.amazonaws.com/examples/us-legislators/all/organizations.json

https://awsglue-datasets.s3.amazonaws.com/examples/us-legislators/all/persons.json

You will now have six JSON files saved locally:

areas.json

countries.json

events.json

memberships.json

organizations.json

persons.json

Now that you have all of your data saved locally, move onto the next challenge to create an S3 bucket.

Upload Source Data Files to S3
You will upload the JSON files you downloaded in the first challenge to the bucket you just created.

Under the Name column, click on the link for the bucket you just created

Click Create folder.

For the Folder name, enter inventory, and click Create folder.

At the bottom of the page click the inventory folder link, and then click Upload.

Click Add files, and upload all six JSON files you downloaded in the first challenge.

At the bottom of the page, click Upload.

Once the files are uploaded, in the upper-right click Close.

Under Files and folders you will see the six JSON files:

areas.json

countries.json

events.json

memberships.json

organizations.json

persons.json

Now that you have all of your files uploaded into your inventory folder, move onto the next challenge to create the crawler.

Create the Crawler
Now you will provision a crawler in AWS Glue to populate the AWS Glue Data Catalog every 12 hours with tables from the source S3 data store.

At the top of the page in the search bar, type in and then click on AWS Glue.

Note: Ensure you are using the US West (Oregon) region by clicking the dropdown in the top right of your window next to your pluralsight- username.

In the left-hand menu, click Crawlers, then click Add crawler.

For the Crawler name, enter awsglue-datasets, then click Next.

Click Next.

Note: Leave the Crawler source type values as their defaults.

For the Include path enter s3://awsglue-datasets-<AWS account ID>/inventory, then click Next.

Note:

Choose a data store's default value of S3 can be left as is.

Similar to before, in the include-path replace <AWS account ID> with the AWS account ID you stored earlier.

For Add another data store, click Next.

Note: Leave the default value of No.

For IAM role, enter awsglue-datasets-crawler, then click Next.

Note: The default value of Create an IAM role can be left as is.

Set the Frequency to Custom (you may need to scroll down in the drop-down to see this option), for the Cron expression enter 0 */12 ? * * *, and click Next.

Click Add database, for the Database name enter awsglue-datasets-database then click Create.

Note: Leave all other fields blank.

For Prefix added to tables enter pluralsight_, then click Next.

Click Finish.

At the Crawlers page, you will see an entry for the new crawler, scheduled to run Every 12 hours, and with a Status of Ready.

Manually Run the Crawler
In the last challenge, you finished creating the crawler. Now you can start the crawler manually to check that it performs as expected.

Select the crawler you just created, and click Run crawler, and wait until the Status is Ready.

Note:

This will take about five to ten minutes. It is normal for the Status to be Started, or even Stopped, temporarily. It is not needed, but you can get slightly more up-to-date results by in upper-right of the table clicking the refresh button.

A message including completed and made the following changes will also appear when the Status is Ready.

If the Run crawler button is greyed-out, click the refresh button, then repeat the task.

In the left-hand menu click Tables.

Note: There will be six pluralsight_ prefixed tables, each corresponding to the JSON file names from the first challenge.

Click the pluralsight_countries_json link to verify the crawler worked as expected.

In the table view for pluralsight_countries_json you will see:

A Classification of json,

The Location will be mapped to the countries.json file in the S3 data store,

At the bottom of the page a Schema will appear with five columns named country, code, name, legislatures, and slug.

Congratulations! The crawler performed as expected, and you populated the AWS Glue Data Catalog with tables from the source S3 data store.