[version_1.0]

Note

The exercises in this course will have an associated charge in your AWS account. In this exercise, you will create the following resources:

- AWS Glue crawlers
- Amazon Simple Storage Service (Amazon S3) bucket
- Amazon Athena query
- AWS Identity and Access Management (IAM) roles (created by AWS CloudFormation)

The final exercise task includes instructions to delete all the resources that you create for this exercise.

Familiarize yourself with **AWS Glue**, **Amazon S3 pricing**, **Amazon Athena**, and the **AWS Free Tier**.

Exercise 3: Processing data in a data lake

In this exercise, you define a database and configure a crawler to explore data in an Amazon S3 bucket. Next, you create a table. You then transform the comma-separated values (CSV) file into Parquet and create a table for the Parquet data. Finally, you query the data with Amazon Athena.

Setting up

This exercise requires an IAM role and an Amazon S3 bucket. You will create these resources by using the provided CloudFormation template.

1. Download the following CloudFormation template: exercise-3-processing.yml. This template will set up backend resources that you need to complete the exercise.

   Note: If you have an existing virtual private cloud (VPC) with the Classless Inter-Domain Routing (CIDR) block *10.16.0.0/16*, you must edit the template and change its CIDR block.

2. Sign in to the AWS Management Console as a user that has the necessary permissions to create an IAM role and CloudFormation stack. You have already created the CloudFormation role in exercise 2 of this course. If you don't have a user or role with permissions to create a stack in AWS CloudFormation, you must create the user or role before you proceed to the next step. If you are unsure how to create a role, see the step-by-step instructions in the **Setting up** section in exercise 2.

3. After you create the user or role to have permissions to work with AWS CloudFormation, open the **CloudFormation** console. Make sure that you are in the **US East (N. Virginia)** Region.

4. Choose **Create stack**.

5. In the **Specify template** section, choose **Upload a template file**.
6. Select **Choose file**, browse to where you downloaded the `exercise-3-processing` template, and select the template.
7. Choose **Next**.
8. For **Stack name**, enter `exercise-3-processing`.
9. Choose **Next**, and then choose **Next** again.
10. Select the acknowledgement, and choose **Create stack**.
11. After the stack is created, choose the **Outputs** tab and copy the name of the S3 bucket.

Task 1: Discovering the data

In this task, you first create the AWS Glue database. With AWS Glue, you can discover and connect diverse data sources and manage your data in a centralized data catalog. After you create the database, you add a crawler to extract, transform, and load (ETL) data into the database tables by using a source comma-separated values (CSV) file.

1. Choose **Services**, and search for and open **AWS Glue**.
2. In the navigation pane, in the **Data Catalog** section, choose **Databases**.
3. Choose **Add database**.
4. For **Database name**, enter `nycitytaxi`
5. Choose **Create database**.
6. In the navigation pane, choose **Tables**.

   You can add a table manually or by using a crawler. A crawler is a program that connects to a data store and progresses through a prioritized list of classifiers to determine the schema for your data. AWS Glue provides classifiers for common file types, such as CSV, JavaScript Object Notation (JSON), Apache Avro, and others. You can also write your own classifier by using a grok pattern.
7. Choose **Add tables using a crawler**.
8. For **Crawler name**, enter `nytaxicrawler` and choose **Next**.
9. For **Data source configuration**, under **Is your data already mapped to Glue tables?**, keep **Not yet** selected.
10. Under **Data sources**, choose **Add a data source**.
11. For **Data source**, keep **S3**.
12. For **S3 path**, paste the following path:

    ```
    s3://aws-tc-largeobjects/DEV-AWS-MO-Designing_DataLakes/week3/
    ```

    This S3 bucket contains the data file, which includes data for all rides from the green taxis in the month of January 2020.

13. Keep all other settings for this page at their default values. Then choose **Add an S3 data source**.

14. Choose **Next**.

15. On the **Configure Security Settings** page, under **Existing IAM role**, choose **AWSGlueServiceRoleDefault**, and then choose **Next**.

16. On the **Set output and scheduling** page, under **Target database**, choose **nycitytaxi**.

17. For **Frequency**, keep **Run on demand** selected and choose **Next**.

18. On the **Review all steps** page, choose **Create crawler**.

19. In the **Crawlers** pane, select **nytaxicrawler** and choose **Run**.

    When the crawler finishes running, one table is added to the database. After the job stops, you should see that the **Tables added** column now shows *1*.

20. In the navigation pane, choose **Tables**.

21. In the **Tables** pane, choose the **week3** link.

    This screen describes the table, including its schema, properties, and other information. If you want to look at the schema information, you can choose **Edit schema**.

Task 2: Transforming the data from CSV to Apache Parquet

In this task, you transform the data from CSV into Apache Parquet format. Apache Parquet organizes data in columns. This format type is more lightweight and brings efficiency compared to row-based files, such as CSV.

1. In the navigation pane of AWS Glue, in the **ETL** section, choose **Jobs**.

   This action opens a new browser tab in AWS Glue Studio.

2. In the **Create job** section, keep all the default settings and choose **Create**.

3. In the job diagram, choose the **Data source - S3 bucket** tile.

4. In the right pane, configure the following settings:
   - **S3 source type**: Keep *Data Catalog table* selected
   - **Database**: *nycitytaxi*
   - **Table**: *week3*
5. In the job diagram, choose the **Data target - S3 bucket** tile and configure the following settings:
   - **Format**: *Parquet*
   - **S3 Target location**: Paste `s3://<FMI>/data/` and replace the FMI with your bucket name. When you replace the FMI with your own value, make sure that you also delete the angle brackets (<>).

   Example:

   ```
   s3://glue-934169e0/data/
   ```

6. At the top of the pane, choose the **Job details** tab and configure the following settings:
   - **Name**: `nytaxiparquet`
   - **IAM role**: *AWSGlueServiceRoleDefault*

   Note: This role grants access to resources that AWS Glue needs to automatically generate the *nytaxi-csv-parquet* script.

7. To verify the script, choose the **Script** tab. Feel free to review the script.

8. Choose **Save** to save the job and then choose **Run**.

   Wait for the job to complete. You can view the status by choosing the **Run details** link (in the system message at the top of the pane) or by choosing the **Runs** tab.

9. Return to the main **AWS Glue** console.

10. In the navigation pane, in the **Data catalog** section, choose **Crawlers**.

11. Choose **Add crawler**.

12. For **Crawler name**, paste `nytaxiparquet` and choose **Next**.

13. For **Data source configuration**, keep **Not yet** selected.

14. For **Data sources**, choose **Add a data source**.

15. For **S3 path**, paste `s3://<FMI>/data/`. Replace the FMI with the name of the bucket where the Parquet file is located.

    Example:

    ```
    s3://glue-934169e0/data/
    ```

16. Choose **Add a data source**. Then choose **Next**.

17. On the **Configure security settings** page, under **Existing IAM role**, choose **AWSGlueServiceRoleDefault**. Then choose **Next**.

18. For **Target database**, select **nycitytaxi**.

19. For **Frequency**, keep **Run on demand** selected and choose **Next**.

20. On the **Review and create** page, choose **Create crawler**.

21. Select the **nytaxiparquet** crawler and choose **Run crawler**. Wait for the job to finish.

22. In the navigation pane, in the **Data catalog** section, choose **Tables**.

    You should see two tables:
    - **week3** – The original CSV version from the source bucket
    - **data** – The Parquet table in your S3 bucket

Task 3: Analyzing the data with Amazon Athena

Athena is an interactive query service that you can use to analyze data in Amazon S3 with standard SQL. Athena can query CSV data. However, the Parquet file format significantly reduces the time and cost of querying the data.

In this task, you use Athena to analyze the data in the S3 bucket.

1. Choose **Services**, and search for and open **Athena**.
2. If you are a new user, choose **Explore the query editor**. For existing users, the Query editor may open automatically.
3. On the **Data** tile, under **Database**, choose **nycitytaxi**.
4. In the query editor box, paste the following:

   ```
   Select * From "nycitytaxi"."week3" limit 10;
   ```

5. Choose the **Save** menu and select **Save as**.
6. For both **Query name** and **Query description**, paste `taxidata` and choose **Save query**.
7. At the top of the query editor pane, choose the **Settings** tab and then choose **Manage**.
8. For **Location of query result**, paste the following path for your bucket and replace the FMI with your bucket name.

   ```
   s3://<FMI>/sql/
   ```

   Example:

   ```
   s3://glue-934169e0/sql/
   ```

9. Choose **Save**.
10. Choose the **Editor** tab and choose **Run**.

You can now browse the results and see information such as the *passenger_count*, *trip_distance*, and *tip_amount*.

After you query the data, you can optionally connect Amazon Athena with Amazon QuickSight to visualize data through dashboards.

Cleaning up

Delete the AWS resources that you created for this exercise by completing the following steps.

1. Delete the Athena query.
   - Open the **Amazon Athena** dashboard.
   - In the navigation pane, choose **Query editor** and then choose the **Saved queries** tab.
   - Delete the **taxidata** query and confirm the deletion.
2. Delete the AWS Glue resources.
   - Open the **AWS Glue** dashboard.
   - In the navigation pane, choose **Crawlers**.
   - Delete the following crawlers, and confirm their deletion:
     - **nytaxicrawler**

        - **nytaxiparquet**
    - In the navigation pane, choose **Tables**.
    - Delete the table and confirm the deletion.
    - In the navigation pane, choose **Databases**.
    - Delete the **nycitytaxi** database and confirm the deletion.
    - In the navigation pane, choose **Jobs** (in the **ETL** section).
    - Delete **nytaxiparquet**, and confirm the deletion.
3. Delete the S3 buckets.
    - Open the **Amazon S3** dashboard.
    - Empty and delete the following buckets, and confirm their deletion:
        - **glue-** bucket
        - **aws-glue-assets-** bucket
        - **cf-templates-** bucket
4. Delete the CloudFormation stack.
    - Open the **AWS CloudFormation** dashboard.
    - Delete the stack and confirm the deletion.
5. Optionally, delete the IAM role. IAM users, roles, and policies don't have an associated charge in your AWS account.
    - Open the **IAM** dashboard.
    - In the navigation pane, choose **Roles**.
    - Delete the IAM role for CloudFormation, and confirm the deletion.

Congratulations! You successfully completed the final exercise this course. In this exercise, you gained a deeper understanding of how to transform and process data in a data lake. You defined a database, configured a crawler, and created a table in AWS Glue. You then transformed the CSV file into Parquet to save data processing costs, and repeated the steps to create a table for the Parquet data. Finally, you queried the data with Amazon Athena.