A Day in the Life of a Data Engineer

**SPL-TF-200-ANGLUE-1 - Version 1.0.8**

© 2024 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. All trademarks are the property of their owners.

Note: Do not include any personal, identifying, or confidential information into the lab environment. Information entered may be visible to others.

Corrections, feedback, or other questions? Contact us at [*AWS Training and Certification*](https://support.aws.amazon.com/#/contacts/aws-training).

**Lab overview**

Your colleague is out on vacation, so you’re in charge of your organization’s data engineering practice for the day. Step into their shoes and explore various managed options for data movement on Amazon Web Services (AWS). Consider architecture patterns, performance and cost optimizations, and security best practices—and impress your colleague when they get back to the office!

Your colleague has provided you with a sandbox environment so that you can experiment and test out a proof of concept (POC) while they are away. You have decided to explore data that you believe will help your small company successfully expand into new cities. You will use a National Oceanic and Atmospheric Administration (NOAA) dataset that provides you with historical daily weather data. You believe that this data can help your company determine when it should stock summer, as opposed to winter, items in cities that you are expanding into. For your POC, you will focus on temperature and precipitation metrics.

This lab uses the **NOAA Global Surface Summary of Day (GSOD)** dataset, which was accessed on August 5, 2022, from https://registry.opendata.aws/noaa-gsod. This dataset is a collection of daily weather measurements (temperature, wind speed, humidity, pressure, and more) from 9000+ weather stations around the world. Data was originally collected by the National Climactic Data Center. Because of the lab time limit, the dataset used in this lab has been reduced to include a subset of the years that are available in the full dataset.

OBJECTIVES

By the end of this lab, you should be able to do the following:

* Create an AWS Glue crawler.
* Create and run a job in AWS Glue Studio.
* Explore permissions required to run AWS Glue crawlers and AWS Glue Studio jobs.
* Query the AWS Glue Data Catalog using Amazon Athena.

**Duration**

This lab requires approximately *60* minutes to complete.

ICON KEY

* **Note:** A hint, tip, or important guidance.
* **Learn more:** Where to find more information.
* **Task complete:** A conclusion or summary point in the lab.
* **Refresh:** A time when you might need to refresh a web browser page or list to show new information.

**Start lab**

1. To launch the lab, at the top of the page, choose **Start lab**.

 You must wait for the provisioned AWS services to be ready before you can continue.

1. To open the lab, choose **Open Console**.

You are automatically signed in to the AWS Management Console in a new web browser tab.

**Do not change the Region unless instructed.**

COMMON SIGN-IN ERRORS

**Error: You must first sign out**



If you see the message, **You must first log out before logging into a different AWS account:**

* Choose the **click here** link.
* Close your **Amazon Web Services Sign In** web browser tab and return to your initial lab page.
* Choose **Open Console** again.

**Error: Choosing Start Lab has no effect**

In some cases, certain pop-up or script blocker web browser extensions might prevent the **Start Lab** button from working as intended. If you experience an issue starting the lab:

* Add the lab domain name to your pop-up or script blocker’s allow list or turn it off.
* Refresh the page and try again.

**Task 1: Create and run an AWS Glue crawler**

In this task, you create an AWS Glue crawler to scan the data in an Amazon Simple Storage Service (Amazon S3) bucket, infer the schema of the data, and filter data that will be included in the Data Catalog.

Because you can access the GSOD dataset directly from Amazon S3, you can quickly start exploring it to find data. You do not need to create a relational or NoSQL database to query and prepare the data. Instead, you use AWS Glue.

AWS Glue is a serverless data integration service that you can use to discover, prepare, and combine data for analytics, machine learning (ML), and application development. AWS Glue provides all the capabilities needed for data integration so that you can start analyzing your data and putting it to use in minutes instead of months.

1. In the **AWS Management Console**, use the **AWS search bar** to search for

AWS Glue

 and then, from the list of results, choose the service.

1. In the left navigation pane, under **Data Catalog**, choose **Crawlers**.
2. Choose **Create crawler**.
3. Configure **Crawler details**:
   * For **Name**, enter

weather\_crawler

* + For **Description**, enter

Crawler to read data from Amazon S3

1. Choose **Next**.
2. Under **Data source configuration**, for **Is your data already mapped to Glue tables?** choose **Not yet**.
3. Under **Data sources**, choose **Add a data source** and configure the following settings:
   * For **Data source**, select **S3**.
   * For **Network connection**, keep this blank.
   * For **Location of S3 data**, select **In a different account**.
   * For **S3 path**, enter

s3://aws-tc-largeobjects/SPL-TF-200-ANGLUE-1/data/

* + For **Subsequent crawler runs**, select **Crawl all sub-folders**
  + Select **Exclude files matching pattern**

**Note:** You might need to choose the check box more than once before the button named **Add new exclusion pattern** appears.

**Learn more:** Exclude patterns prevent the crawler from reading all data under the target location. Using exclude patterns helps speed up processing when you do not need to ingest all data in the target location. For more information about exclude parameters refer to the [Crawler properties documentation](https://docs.aws.amazon.com/glue/latest/dg/define-crawler.html#crawler-data-stores-exclude).

* + You add four exclusion patterns. Repeat the following process to add each pattern:
    - Choose **Add new exclusion pattern**.
    - **Exclude pattern**:

19\*\*

* + - Additional exclusion patterns:
      * 200\*\*
      * 201[0-2]\*\*
      * index.html
  + After you add all exclusion patterns, choose **Add an S3 data source**.

1. Choose **Next**.
2. Under **IAM role**, for **Existing IAM role**, select **gluelab**.
3. Choose **Next**.
4. Under **Output configuration**, choose **Add a database**.
5. In the **Database details** pane, for **Name**, enter

weather\_data

.

1. Choose **Create database**.
2. Close the **Databases** browser tab to return to the Glue crawler configuration.
3. Under **Output configuration**, next to the **Target database** dropdown list, choose the refresh icon.
4. For **Target database**, select **weather\_data**.
5. For **Crawler schedule**, select **On demand**.
6. Choose **Next**.
7. Review the configuration and then choose **Create crawler**.
8. In the left navigation pane, below **Data Catalog**, choose **Crawlers**.
9. If you don’t see your new crawler, refresh the page.
10. In the **Crawlers** pane, select **weather\_crawler**.
11. Choose **Run crawler**.

The crawler **State** changes from *Ready* to *Running*. The crawler is now reading data from the S3 bucket to build a Data Catalog table. The crawler runs for about 5 minutes.

While the crawler runs, continue to the next task, but be sure to keep the **Crawlers** tab open.

**Task complete:** You have successfully created and ran an AWS Glue crawler.

**Task 2: Review the IAM policies**

In this task, you explore the AWS Identity and Access Management (IAM) permissions that were used to run the AWS Glue processes.

While you wait for the data transformation job to complete, use the IAM console to review the gluelab role. This is the role that you chose when you ran the AWS Glue crawler. You also use this role later, when you run a data transformation job in AWS Glue Studio.

1. In the **AWS Management Console**, use the **AWS search bar** to search for

IAM

, right-click (or an equivalent UI interaction) on IAM, and choose **Open Link in New Tab**.

1. Choose the **IAM Management Console** tab.
2. In the left navigation pane, under **Access management**, choose **Roles**.
3. In the search bar, enter

gluelab

.

1. Under **Role name**, choose **gluelab**.

You find two AWS managed policies that are associated with this role.

1. Choose **gluePolicyRestrictS3**.

A new browser tab will open.

1. In the **Summary** pane, on the Permissions tab, notice the policy definition.

This lab policy provides Amazon S3 with read and write access to the lab environment’s S3 buckets. The AWS Glue job that you created requires these privileges because it reads and writes information to Amazon S3 while processing data.

1. Close the **gluePolicyRestrictS3** browser tab.
2. Choose **AWSGlueServiceRole**.

A new browser tab will open.

1. In the **Summary** pane, on the Permissions tab, notice the policy definition.

This lab policy gives AWS Glue access to related services such as Amazon Elastic Compute Cloud (Amazon EC2), Amazon S3, and Amazon CloudWatch Logs. With these privileges, AWS Glue can read, write, and process data. It can also write information to Amazon CloudWatch logs.

1. Close the **AWSGlueServiceRole** browser tab.
2. On the **gluelab** pane, choose the **Trust relationships** tab.

In this trust policy, the AWS Glue service is given access to assume the role. When AWS Glue assumes the role, it receives all privileges that are included in the role’s policies.

1. Close the **IAM Management Console** tab and return to the browser tab where the crawler is running.

You now know the basic permissions that are required to run AWS Glue processes. Nicely done!

1. Verify that the crawler has completed successfully before you continue to the next task.

Wait for the crawler run to complete before proceeding to Task 3. When the job completes, the **State** changes to *Ready*, and the **Last run** status is *Succeeded*. Under **Table changes from last run**, notice that one table was created.

**Task complete:** You have successfully reviewed the IAM permissions that were used to run the AWS Glue processes.

**Task 3: View the table in the Data Catalog**

In this task, you view the table that the AWS Glue crawler added to the Data Catalog. You explore the metadata that was generated by the crawler, and you locate the partitions that are associated with the table.

1. In the navigation pane, under **Data Catalog**, choose **Tables**.

You find an entry for a table named **data**. Notice that the **Classification** for this data is **csv**. This classification was inferred by AWS Glue when it read the data from the source S3 bucket.

1. Choose **data**.
2. From the **Actions** menu, select **View properties**.

From this page, you can view table metadata such as when it was last updated, the input format and delimiter, exclusions, and the table schema.

1. Close the **Properties** window.
2. Choose the **Partitions** tab to view a list of the table’s partitions.

The lab dataset on Amazon S3 includes data from 1972 through 2022. However, you applied a filter to the AWS Glue crawler so your Glue table only contains data from 2013 through 2022.

**Task complete:** You have successfully learned how to locate the metadata for a Data Catalog table and how to find the table’s partition details.

**Task 4: Run a job in AWS Glue Studio to transform the data**

In this task, you create and run an AWS Glue Studio job to clean up the data.

Using an AWS Glue crawler, you created a table in the Data Catalog. However, the data isn’t quite ready to use yet. You discovered that one of the data types needs to change. You also found that many columns in this dataset are empty.

CHOOSE AND CONFIGURE THE DATA SOURCE FOR THE JOB

1. In the navigation pane, under **Data Integration and ETL**, choose **ETL jobs**.

The browser displays the **AWS Glue Studio** page.

1. In the **Create job** pane, select **Visual ETL**.
2. On the **Visual** tab, in the **+ Add nodes** pane, choose the **Sources** tab.
3. From the list of available sources, choose **Amazon S3**.

An **Amazon S3** card is placed on the canvas.

1. Choose the **Amazon S3** card to open the right pane, then configure the following settings under the **Data source properties - S3** tab:
   * For **Name**, enter

Source S3 bucket

* + For **S3 source type**, select **Data Catalog table**.
  + For **Database**, select **weather\_data**.
  + For **Table**, select **data**.

CONFIGURE THE DATA TRANSFORMATION

1. On the top-left side of the canvas, choose the **+** button to open the **+ Add nodes** pane.
2. Choose the **Transforms** tab, then choose **Change Schema**.

A **Change Schema** card is placed on the canvas and connected to the data source.

1. Choose the **Change Schema** card to open the right pane, and then configure the following settings under the **Change Schema (Apply mapping)** section:
   * Change the name of the **date** attribute.
     + Under **Target key**, replace

date

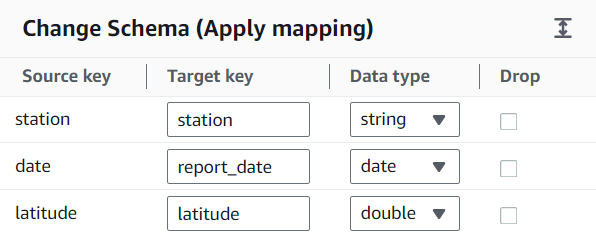
 with

report\_date

.

* + Change the **date** attribute **Data type**:
    - From the **Data type** drop-down list, select **date**.

**Note:** The values in the row that begins with **date** should match the example in the following image:



* + Remove all attributes that have a **long** data type.
    - For each column with a **long** data type, under **Drop**, select the check box.

**Note:** These are the columns that should be removed from the dataset:

* + temp\_attributes
  + dewp\_attributes
  + slp\_attributes
  + stp\_attributes
  + visib\_attributes
  + wdsp\_attributes
  + frshtt

1. Review attributes and their data types. Ensure the **date** data type was updated and renamed. Verify that all long columns have been excluded from the data transformation output.

DEFINE THE JOB TARGET AND THE OUTPUT DETAILS FOR THE JOB

1. On the top-left side of the canvas, choose the **+** button to open the **+ Add nodes** pane.
2. Choose the **Targets** tab, then choose **Amazon S3**.

An **Amazon S3** card is placed on the canvas and connected to the **Change Schema** card.

1. Choose the newly created **Amazon S3** card to open the right pane, then configure the following settings under the **Data target properties - S3** tab:
   * For **Name**, enter

Target S3 bucket

.

* + For **Format**, select **Parquet**.
  + For **Compression type**, select **Uncompressed**.
  + For **S3 Target Location**, choose **Browse S3**, then select the bucket that begins with **glue-bucket-**, then select **Choose**.
  + For **Data Catalog update options**, select **Create a table in the Data Catalog and on subsequent runs, update the schema and add new partitions**.
  + For **Database**, select **weather\_data**.
  + For **Table name**, enter

data\_parquet

.

* + Choose **Add a partition key**, then for **Partition (0)**, select **report\_date**.

COMPLETE THE JOB CONFIGURATION AND SAVE THE JOB

1. Above the canvas, choose the **Job details** tab and configure the following settings:
   * For **Name**, enter

DataPreparation

* + For **IAM role**, select **gluelab**.
  + Under **Advanced properties**:
    - For **Script filename**, enter

data\_prep.py

1. Above the canvas, choose the **Script** tab.

Take a moment to review the Python code that was automatically generated using the visual editor. This method for creating the job provides a helpful jump start for creating extract, transform, and load (ETL) code. You can update the Python script, if needed, before you save and run your ETL job. For this job, you do not need to make any changes, so you will only save the job and continue to the next step.

1. In the upper-right corner, choose **Save**.

A banner with the message **Successfully updated job - Successfully updated job data. To run the job choose the Run Job button.** will appear at the top of the page.

START THE JOB

1. Choose **Run** to start your job.
2. Above the canvas, choose the **Runs** tab.

Notice that your job’s **Run status** is *Running*.

This job will run for 15-20 minutes.

You now know how to cleanse and transform data using an AWS Glue Studio job. Great work!

While the **DataPreparation** job is running, review the following helpful tips for managing AWS Glue crawlers and ETL jobs:

* [Why is the AWS Glue Crawler Running for a Long Time?](https://aws.amazon.com/premiumsupport/knowledge-center/long-running-glue-crawler/)
* [Best Practices to Optimize Cost and Performance for AWS Glue Streaming ETL Jobs](https://aws.amazon.com/blogs/big-data/best-practices-to-optimize-cost-and-performance-for-aws-glue-streaming-etl-jobs/)

1. Verify that the job completed successfully before you continue to the next task.

**Refresh:** You may need to refresh the jobs under the **Runs** tab to check the status of your job.

**Task complete:** You have successfully created and ran an AWS Glue Studio job to clean up the data.

**Task 5: Query the data\_parquet table in Amazon Athena**

In this task, you query the data using Amazon Athena.

Amazon Athena is an interactive query service that you can use to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.

1. From the **AWS Management Console**, use the **AWS search bar** to search for

Athena

, and then choose the service from the list of results.

1. In the left navigation pane, choose **Query editor**.
2. Choose the **Settings** tab.
3. Choose **Manage**.
4. Under **Query result location and encryption**, choose **Browse S3**.
5. Select the bucket the begins with **query-result-**.
6. Choose **Choose**.

**Note:** Before you run a query in Amazon Athena, you must first specify the query result output location.

1. Choose **Save**.
2. Choose the **Editor** tab.
3. In the **Data** pane, configure the following settings:
   * For **Data source**, select **AWSDataCatalog**.
   * For **Database**, select **weather\_data**.

Under **Tables and views**, you find two tables: **data**, and **data\_parquet**. You will query the **data\_parquet** table.

**Note:** If you do not find a table named **data\_parquet**, ensure that the AWS Glue job has completed. If the job has completed, refresh the browser page and the table should appear.

1. Next to the table named **data\_parquet**, choose the ellipsis.
2. Under **Run Query**, choose **Preview Table**.

A new query tab will open. The following query will run automatically:

SELECT \* FROM "weather\_data"."data\_parquet" limit 10;

In the lower pane, notice a preview of the first 10 rows that the query retrieved.

1. To the right of the query tab, choose the **+** symbol, then enter the following SELECT statement:

SELECT report\_date, latitude, longitude, temp, prcp

FROM "weather\_data"."data\_parquet"

WHERE report\_date between CAST('2022-01-01' AS DATE) and CAST('2022-06-30' AS DATE)

ORDER BY report\_date

1. Choose **Run**.

This query returns the location, temperature, and precipitation data that was reported between January and June of 2022.

1. Replace the previous SELECT statement with the following SELECT statement:

SELECT report\_date, avg(temp) as "Avg Temp", avg(prcp ) as "Avg Prcp"

FROM "weather\_data"."data\_parquet"

WHERE report\_date between CAST('2022-01-01' AS DATE) and CAST('2022-09-30' AS DATE)

Group by report\_date

Order by 3 desc

1. Choose **Run**.

This query shows the average temperature, and average precipitation for each report\_date between January 1, 2022 and September 30, 2022. The data returned is ordered so the wettest day during this time period appears first. Which report\_date did your query return first?

As next steps, you could continue to explore the dataset, or you could hand access to your data science team so they can discover trends in the data. You could use this dataset to enrich other datasets that are already in use. You might also consider including partitions that you originally filtered out during your POC to provide more historic data to your data science team.

**Task complete:** You have successfully ran queries to the data using Amazon Athena.

**Conclusion**

You now have successfully done the following:

* Created an AWS Glue crawler.
* Created and ran a job in AWS Glue Studio.
* Explored permissions required to run AWS Glue crawlers and AWS Glue Studio jobs.
* Queried the AWS Glue Data Catalog using Amazon Athena.

In a very short time, you were able to access, prepare, and explore the NOAA GSOD dataset.

**End lab**

Follow these steps to close the console and end your lab.

1. Return to the **AWS Management Console**.
2. At the upper-right corner of the page, choose **AWSLabsUser**, and then choose **Sign out**.
3. Choose **End lab** and then confirm that you want to end your lab.

**Additional Resources**

* For more information about how to use AWS Glue, see [AWS Glue Documentation](https://docs.aws.amazon.com/glue/index.html).
* For more information about how to use Amazon Athena, see [Amazon Athena Documentation](https://docs.aws.amazon.com/athena/index.html).

For more information about AWS Training and Certification, see [*https://aws.amazon.com/training/*](https://aws.amazon.com/training/).

*Your feedback is welcome and appreciated.*  
If you would like to share any feedback, suggestions, or corrections, please provide the details in our [*AWS Training and Certification Contact Form*](https://support.aws.amazon.com/#/contacts/aws-training).