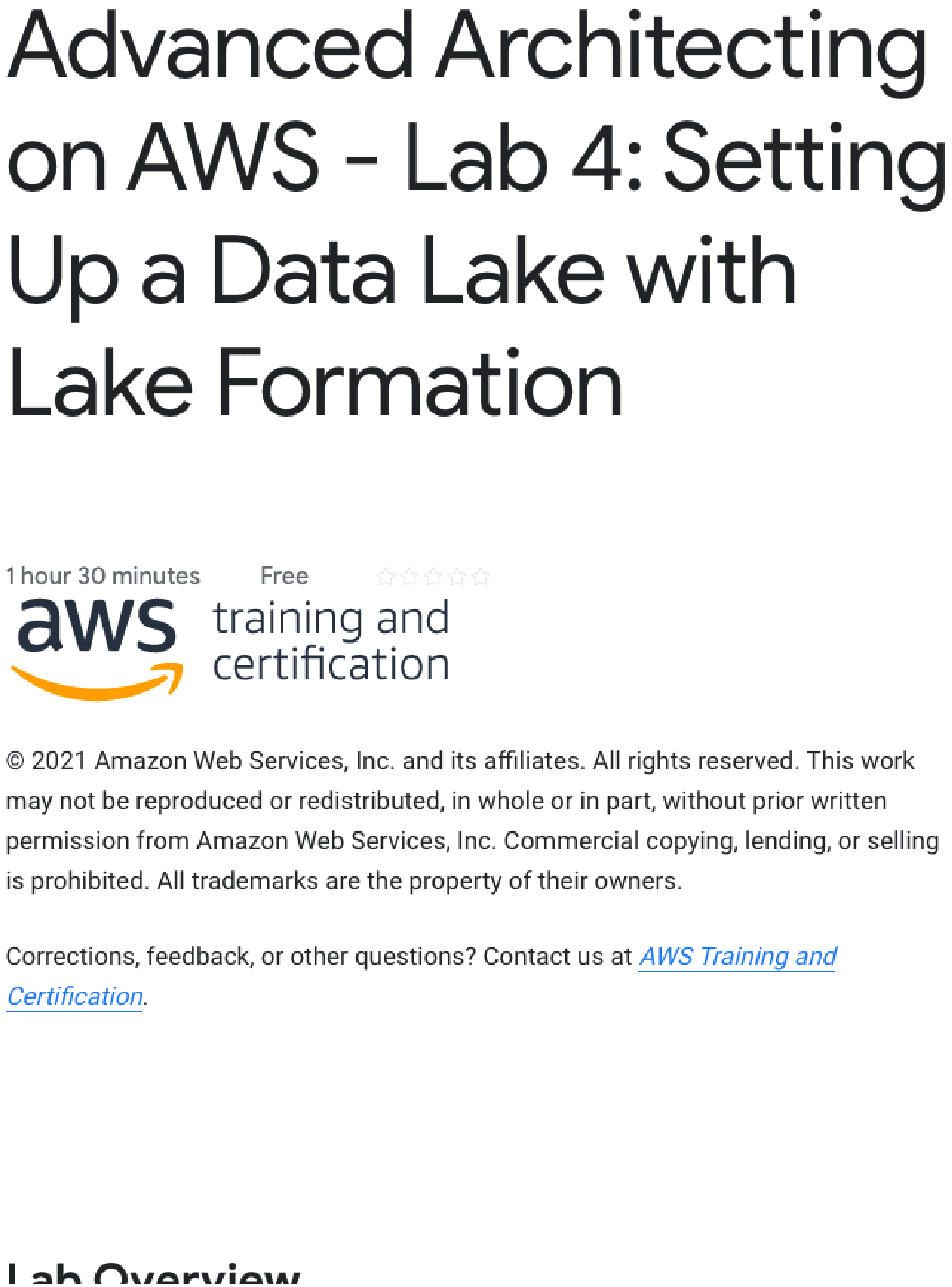
# Lab 4: Setting Up a Data Lake with Lake Formation v3.0.2

Wednesday, March 3, 2021 4:04 PM

****Step 52 you might need to remove the filter that the system puts into place.



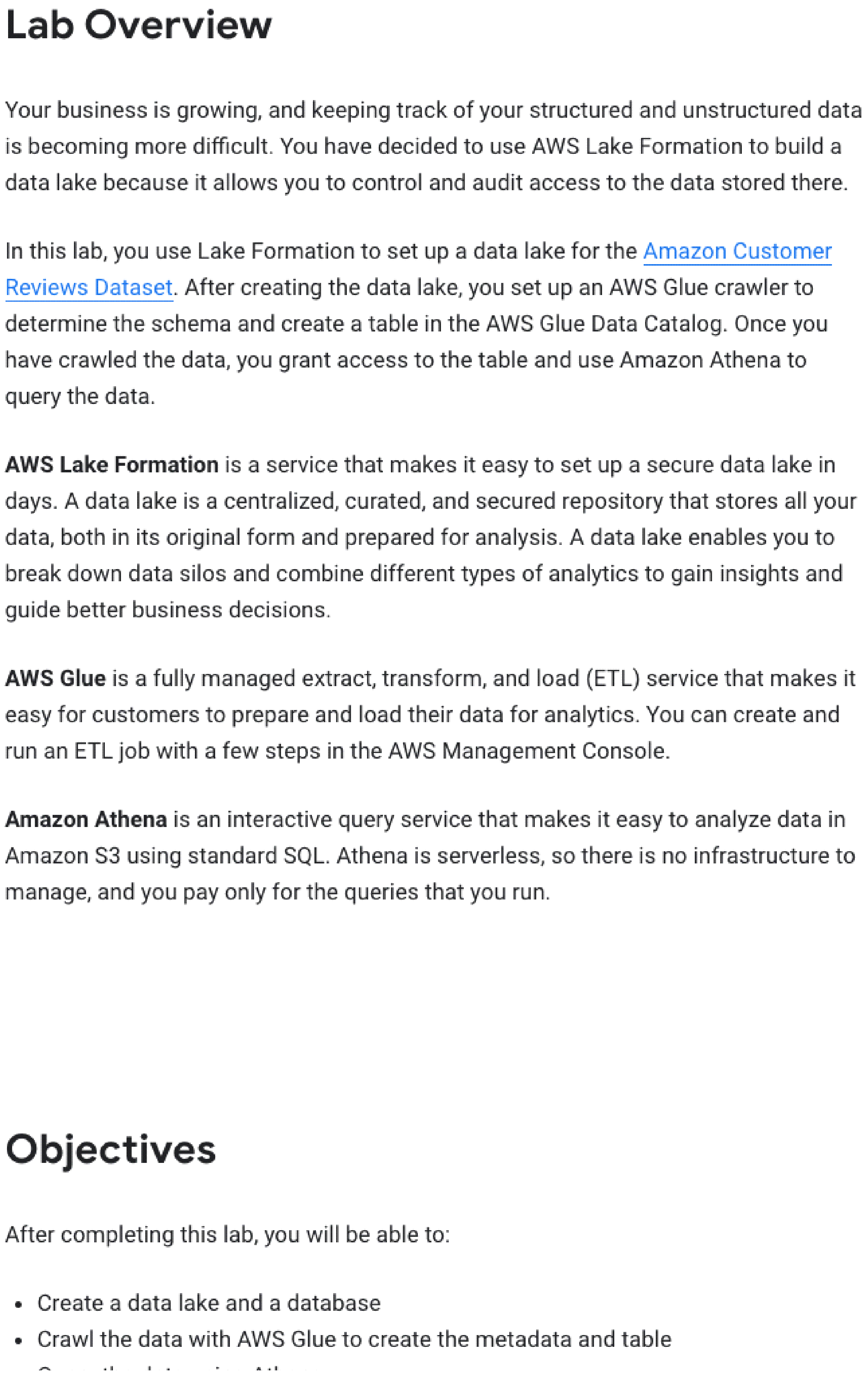
# Advanced Architecting on AWS - Lab 4: Setting Up a Data Lake with Lake Formation

1 hour 30 minutes Free ☆☆☆☆☆

aws training and certification

Corrections, feedback, or other questions? Contact us at *[AWS Training and Certification](#)*.

## Lab Overview



# Lab Overview

Your business is growing, and keeping track of your structured and unstructured data is becoming more difficult. You have decided to use AWS Lake Formation to build a data lake because it allows you to control and audit access to the data stored there.

In this lab, you use Lake Formation to set up a data lake for the Amazon Customer Reviews Dataset. After creating the data lake, you set up an AWS Glue crawler to determine the schema and create a table in the AWS Glue Data Catalog. Once you have crawled the data, you grant access to the table and use Amazon Athena to query the data.

**AWS Lake Formation** is a service that makes it easy to set up a secure data lake in days. A data lake is a centralized, curated, and secured repository that stores all your data, both in its original form and prepared for analysis. A data lake enables you to break down data silos and combine different types of analytics to gain insights and guide better business decisions.

**AWS Glue** is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics. You can create and run an ETL job with a few steps in the AWS Management Console.

**Amazon Athena** is an interactive query service that makes it easy 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.

# Objectives

After completing this lab, you will be able to:

- Create a data lake and a database
- Crawl the data with AWS Glue to create the metadata and table

- Query the data using Athena
- Managing user permissions in Lake Formation

## Prerequisites

This lab requires:

- Access to a notebook computer with Wi-Fi and Microsoft Windows, macOS, or Linux (Ubuntu, SuSE, or Red Hat)
- An internet browser such as Chrome, Firefox, or Microsoft Edge
- A plaintext editor

## Duration

This lab requires approximately **60** minutes to complete.

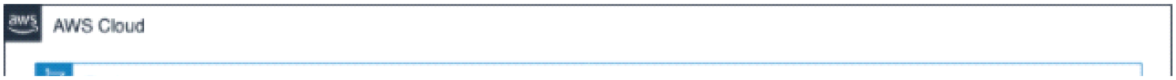
## AWS Services Not Used in This Lab

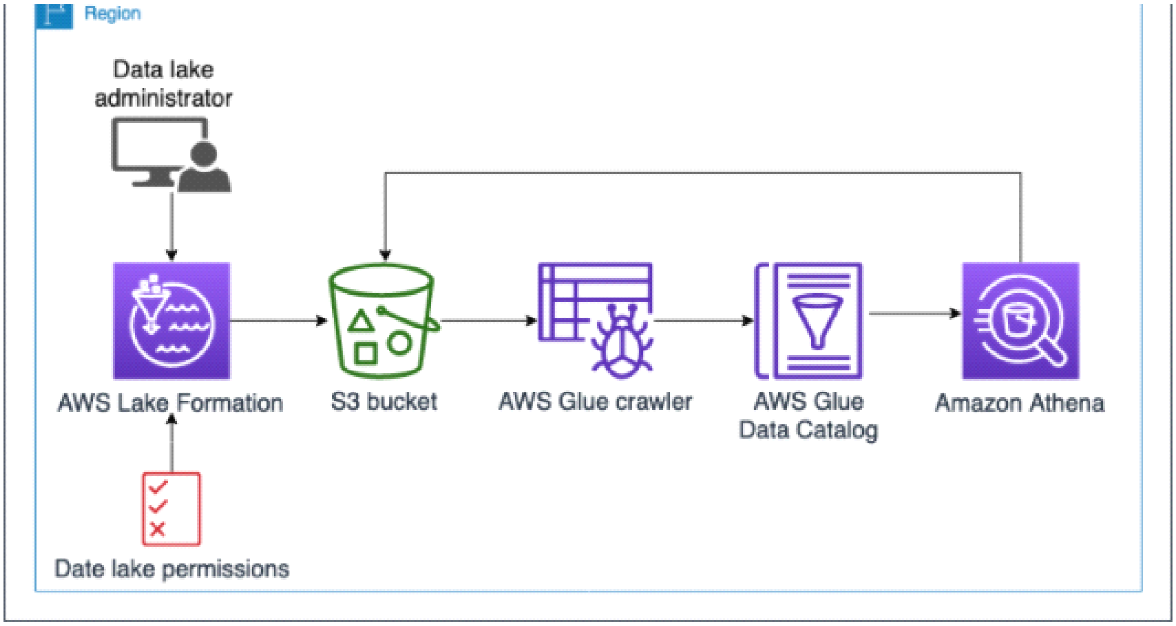
AWS services not used in this lab are disabled in the lab environment. In addition, the capabilities of the services used in this lab are limited to what the lab requires. Expect errors when accessing other services or performing actions beyond those provided in this lab guide.

## Lab Environment

An AWS Cloud9 development environment, AWS Identity and Access Management (IAM) users, and IAM roles are provisioned as part of the lab. You use the AWS Management Console to provision other resources that are required to complete the lab.

The following diagram shows the resources provisioned for this lab and how they will be connected at the end of the lab:





## Start Lab

1. At the top of your screen, launch your lab by choosing **Start Lab**

   This starts the process of provisioning your lab resources. An estimated amount of time to provision your lab resources is displayed. You must wait for your resources to be provisioned before continuing.

   ℹ If you are prompted for a token, use the one distributed to you (or credits you have purchased).

2. Open your lab by choosing **Open Console**

   This opens an AWS Management Console sign-in page.

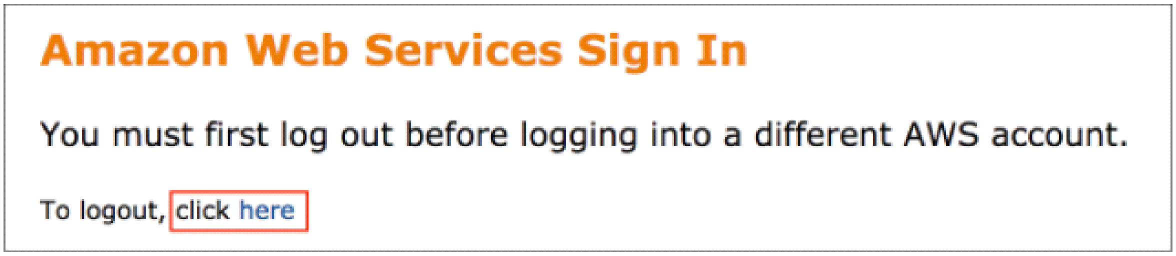
3. On the sign-in page, configure:

   - **IAM user name:** `awsstudent`
   - **Password:** Paste the value of **Password** from the left side of the lab page
   - Choose **Sign In**

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

## Common Login Errors

**Error: You must first log out**



### Amazon Web Services Sign In

You must first log out before logging into a different AWS account.

To logout, click here

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

- Choose **click here**
- Close your browser tab to return to your initial lab window
- Choose Open Console again

# Task 1: Explore the Lab Environment

In this task, you review the account resources created when the lab was started. You then open the AWS Cloud9 development environment and import data from a public Amazon Simple Storage Service (Amazon S3) bucket to your S3 bucket using the AWS Command Line Interface (AWS CLI). You also inspect the data in your AWS Cloud9 environment.

## Create Folders in the S3 Bucket

4. In the AWS Management Console, choose **Services ▾** and select **S3**.

An S3 bucket was created during the lab setup.

5. Choose the name of the **xxxx-databucket-xxxx** bucket.

6. Choose Create folder, and configure the following:

   - **Folder name:** `review`
   - Choose Create Folder

   In a later task, you copy data to this folder in your S3 bucket.

7. Repeat the previous step to create a folder named `results`

   The **results** folder stores the results of your Athena query.

## Load the AWS Cloud9 IDE

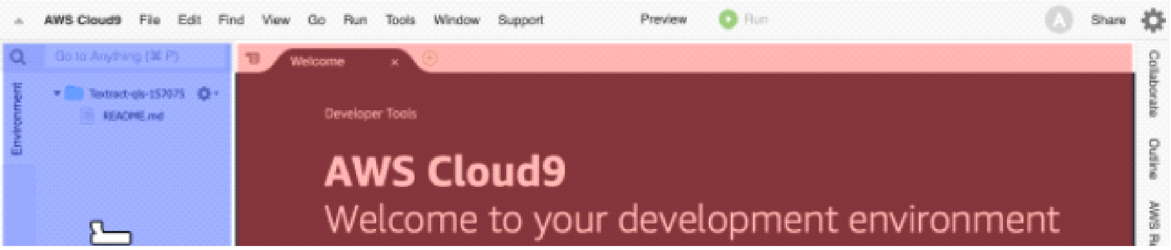
AWS Cloud9 allows you quick access to an integrated development environment (IDE) through your web browser. To access AWS Cloud9, you must first be logged in to the AWS Management Console. Once you are, you can access your development environments directly. A link to launch your AWS Cloud9 environment has been provided.

8. Copy the **Cloud9url** value from the left side of the lab page.

9. Open a *new* browser tab, paste the **Cloud9url** into the address bar, and press ENTER.

   The AWS Cloud9 IDE loads. The IDE is broken into three main parts, as shown by the colored boxes in the following image:

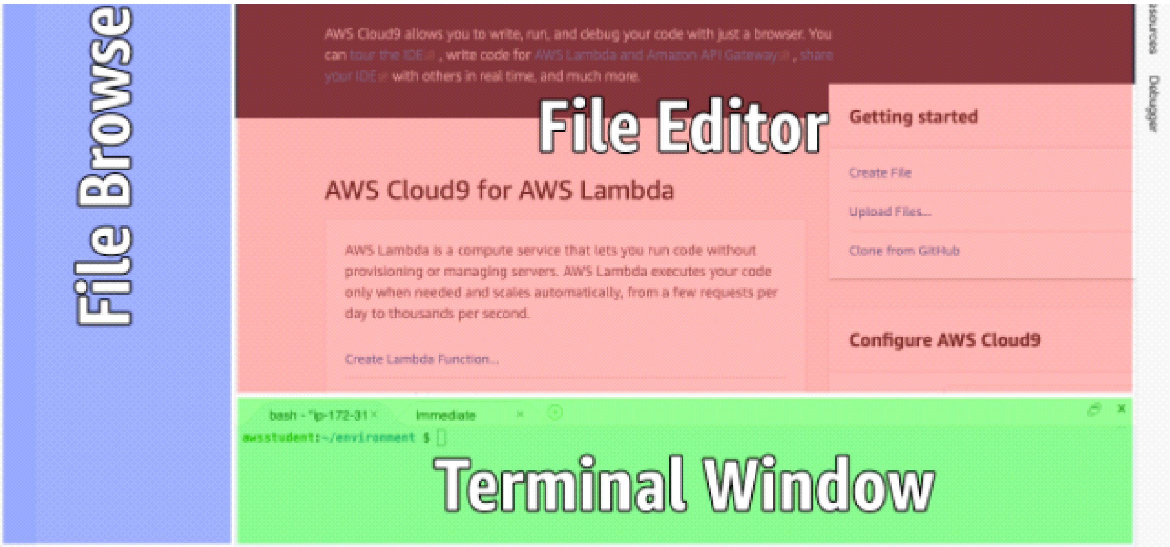
   - File editor, denoted with a red box
   - File browser, denoted with a blue box
   - Terminal window, denoted with a green box



Developer Tools

AWS Cloud9

Welcome to your development environment



AWS Cloud9 allows you to write, run, and debug your code with just a browser. You can tour the IDE, write code for AWS Lambda and Amazon API Gateway, share your IDE with others in real time, and much more.

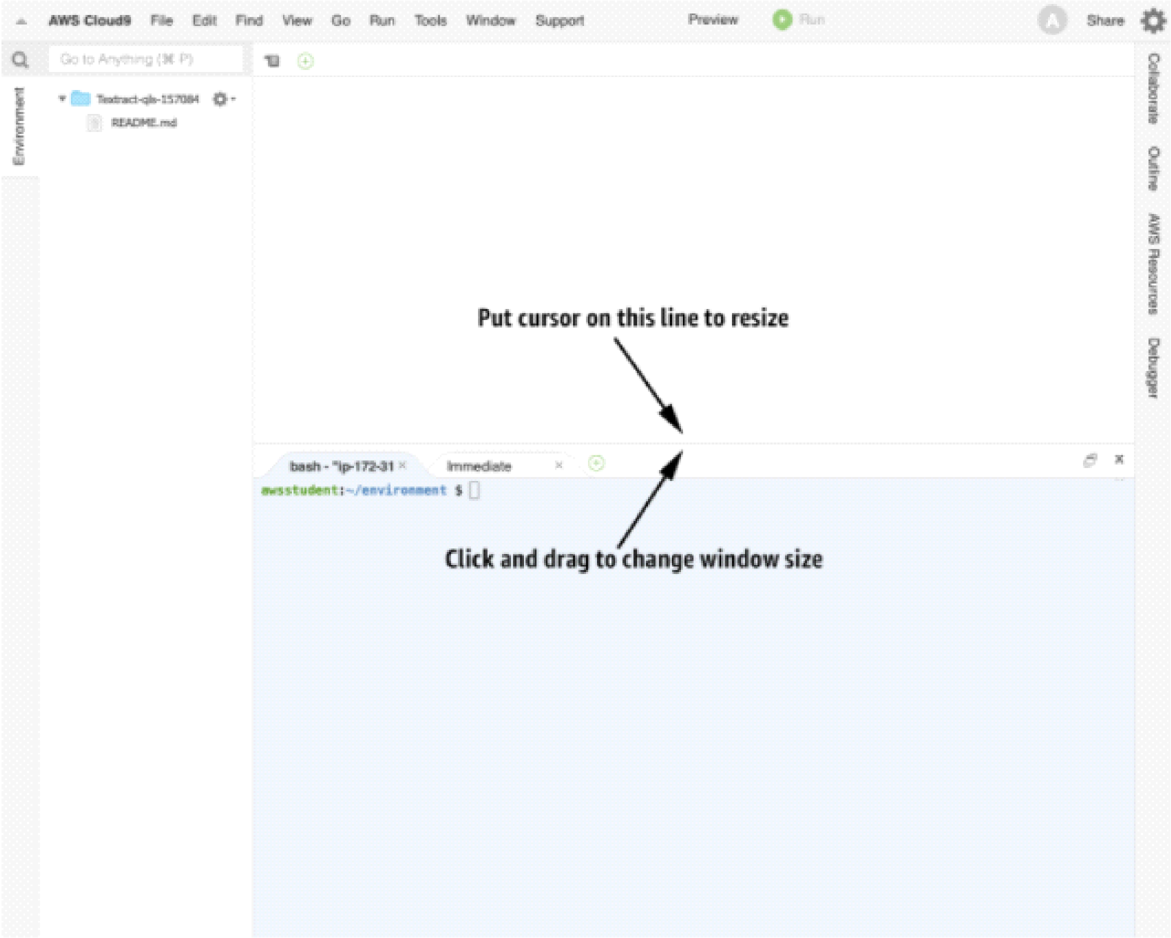
AWS Cloud9 for AWS Lambda

AWS Lambda is a compute service that lets you run code without provisioning or managing servers. AWS Lambda executes your code only when needed and scales automatically, from a few requests per day to thousands per second.

Create Lambda Function...

```
awsstudent:~/environment $
```

You can resize the terminal window and file editor sections by dragging the line between the two sections to reach the desired sizes, as shown in the following image. Depending on what step you are on, you may want to resize multiple times throughout the lab. It can be helpful to make the terminal window section larger to more easily display the command output.



```
awsstudent:~/environment $
```

Put cursor on this line to resize

Click and drag to change window size

## Copy Data to Your S3 Bucket

In this lab, you explore and use the Amazon Customer Reviews Dataset.

Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazon's iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. This makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others.

The dataset contains the following columns:

- **marketplace:** Two-letter country code of the marketplace where the review was written
- **customer_id:** Random identifier that can be used to aggregate reviews written by a single author
- **review_id:** The unique ID of the review
- **product_id:** The unique product ID the review pertains to. In the multilingual dataset, the reviews for the same product in different countries can be grouped by the same product_id.
- **product_parent:** Random identifier that can be used to aggregate reviews for the same product
- **product_title:** Title of the product
- **product_category:** Broad product category that can be used to group reviews (also used to group the dataset into coherent parts)
- **star_rating:** The 1-5 star rating of the review
- **helpful_votes:** Number of helpful votes
- **total_votes:** Number of total votes the review received
- **vine:** Indication of whether the review was written as part of the Vine program
- **verified_purchase:** Indication of whether the review is on a verified purchase
- **review_headline:** The title of the review
- **review_body:** The review text
- **review_date:** The date the review was written

10. In the AWS Cloud9 terminal window, run the following command to list the data in the public bucket:

```
aws s3 ls s3://amazon-reviews-pds/tsv/
```

**Note** To run a command in the terminal window, you may need to press ENTER.

11. In the AWS Cloud9 terminal window, run the following command, replacing *<S3Bucket>* with the **S3Bucket** value from the left side of the lab page:

```
aws s3 cp s3://amazon-reviews-
pds/tsv/amazon_reviews_us_Office_Products_v1_00.tsv.gz \
s3://<S3Bucket>/review/
```

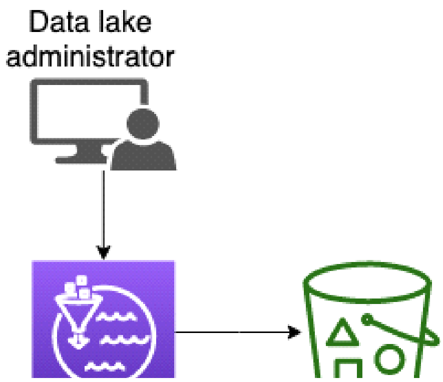
This command copies Amazon customer reviews of U.S. digital eBook purchases to your S3 bucket.

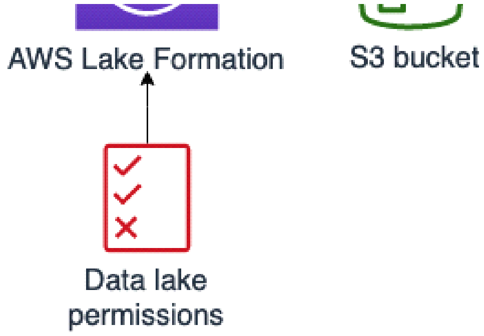
**Learn more** For the AWS Command Line Interface (AWS CLI) command to retrieve the S3 bucket name in AWS Cloud9, refer to **Appendix**.

The AWS Cloud9 instance is not needed for further tasks in this lab. You can safely close the tab and return to the AWS Console.

## Task 2: Set Up AWS Lake Formation

In this task, you use the AWS Management Console to register your data S3 path, create a database, and provide the necessary permissions for the user to access the data lake, as shown in the following diagram:





12. Return to the tab with the AWS Management Console.

13. In the AWS Management Console, choose **Services ⌄** and select **AWS Lake Formation**.

    The first step in creating your data lake in Lake Formation is to define one or more administrators. Administrators have full access to the Lake Formation system, and they control the initial data configuration and access permissions. For this lab, the user '*awsstudent*' is already setup as a data lake administrator.

    A '*Welcome to Lake Formation*' message box is displayed.

14. Ensure the checkbox for ☑ *Add myself* is filled.

15. Choose **Get Started**.

    The '*Admins and database creators*' page is displayed.

    **Note** If the '*Welcome to Lake Formation*' message box was not displayed, follow these steps to confirm if *awsstudent* is already an administrator of the data lake:

    - Expand the navigation menu on the left side, if necessary, by choosing the menu icon ☰.
    - Choose **Admins and database creators** from the navigation menu. It is located under the **Permissions** section
    - Locate the **Data lake administrators** section
    - Confirm the IAM User *awsstudent* is listed

## Register Your Amazon S3 Storage

Lake Formation manages access to designated storage locations within Amazon S3.

Register the storage locations that you want to be part of the data lake.

16. Expand the navigation menu on the left side, if necessary, by choosing the menu icon ≡.

17. In the left navigation pane, choose **Data lake locations** from the **Register and ingest** section.

18. Choose **Register location** to include your S3 storage location a part of the data lake.

    The '*Register location*' page is displayed.

19. Configure the following:

    **Note** Replace *<S3Bucket>* with the **S3Bucket** value from the left side of the lab page.

    - *Amazon S3 path:* `s3://<S3Bucket>/review/`
    - *IAM role:* Select the **xxxx-LakeFormationServiceRole-xxxx** role
    - Choose **Register location**

    The '*Data lake locations*' page is displayed.

## Update Permissions

Lake Formation manages access for IAM users, roles, and Active Directory users and groups via flexible database, table, and column permissions. Grant permissions to one or more resources for your selected users.

20. In the left navigation pane, choose **Data locations** from the **Permissions** section.

21. Choose **Grant**

22. Configure the following:

    - *IAM users and roles:* **awsstudent**

    **Note** Replace *<S3Bucket>* with the **S3Bucket** value from the left side of the lab page.

- **Storage locations:** `s3://<S3Bucket>/review/`
- Choose **Grant**

The '*Data locations*' page is displayed.

## Validate Permissions for Databases and Tables

23. In the left navigation pane, choose **Settings** from the **Data catalog** section.

    Configure the following:

    - Select ☑ **Use only IAM access control for new databases**
    - Select ☑ **Use only IAM access control for new tables in new databases**
    - Choose **Save**

## Create a Database

24. In the left navigation pane, choose **Databases** from the **Data catalog** section.

    The '*Databases*' page is displayed.

25. Choose **Create database**

    The '*Create database*' page is displayed.

    Lake Formation organizes data into a catalog of logical databases and tables. It creates one or more databases, and then automatically generate tables during data ingestion for common workflows.

26. Configure the following:

    - **Name:** `review-db`

    **Note** Replace *<S3Bucket>* with the **S3Bucket** value from the left side of the lab page.

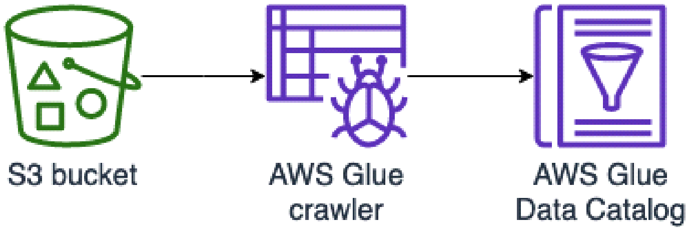
    - **Location:** `s3://<S3Bucket>/review/`

- Select ☑ **Use only IAM access control for new tables in this database**
- Choose **Create database**

The '*Databases*' page is displayed.

## Task 3: Crawl Review Data with AWS Glue

In this task, you use an AWS Glue crawler to create a **review** table for the database you created previously.



### Use a Crawler to Add a Table

A *crawler* connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

27. In the AWS Management Console, choose **Services ⌄** and select **AWS Glue**.

28. If prompted, choose **Get Started**

29. In the left navigation pane, choose **Tables** from the **Data catalog** section.

30. Choose **Add tables**, and select **Add tables using a crawler** from the drop-down menu.

31. Configure the following:

    - Crawler name:

- **Crawler name:** `review-tb`
- Choose **Next**
- *Crawler source type:* **Data stores**
- Choose **Next**
- *Choose a data store:* **S3**
- *Crawl data in:* **Specified path in my account**

🗒 **Note** Replace *<S3Bucket>* with the **S3Bucket** value from the left side of the lab page.

- *Include path:* `s3://<S3bucket>/review/`
- Choose **Next**
- *Add another data store:* **No**
- Choose **Next**
- Select **Choose an existing IAM role**
- *IAM role:* **xxxx-AdminGlueServiceRole-xxxx**
- Choose **Next**
- *Frequency:* **Run on demand**
- Choose **Next**
- *Database:* **review-db**
- Choose **Next**

32. On the review page, choose **Finish** to add the crawler.

## Run the Crawler to Add Data to the Table

33. Select the checkbox next to the ☑ **review-tb** crawler.

34. Choose **Run crawler**

The crawling task can take a few minutes to complete. Choose the refresh ↻ icon at the top of the page to get the current status of the task. The status should change from **Starting** >> **Stopping** >> **Ready**.

## Task Validation

Once the task completes, the **Tables added** value changes to **1**.

35. Choose **Logs** to review the Amazon CloudWatch Logs, including logs related to data classification and table creation.

36. Return to the AWS Glue console.

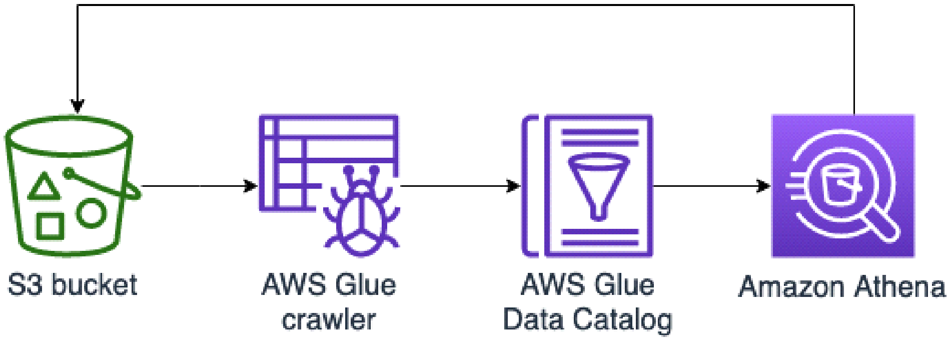
37. In the left navigation pane, in the **Data catalog** section, choose **Tables**.

38. Choose the name of the **review** table to display its schema.

    **Note** To show the table in the list, you may need to refresh the page or clear the filter area above the list.

## Task 4: Use Athena to Query Data

In this task, you use the Athena Query editor to review the data in your table.



39. In the AWS Management Console, choose **Services ⌄** and select **Athena**.

    ⚠ **Note** If there is a banner to upgrade the Athena Data Catalog, **DO NOT** choose to upgrade.

40. If prompted, choose **Get Started**

### Update the Query Results Location

Before you run your first query, you need to set up a query result location in Amazon S3.

41. At the top-right of the page, choose **Settings** and configure the following:

    **Note** Replace *<S3Bucket>* with the **S3Bucket** value from the left side of the lab page.

    - **Query result location:** `s3://<S3bucket>/results/`

      **Note Query result location:** can be set by choosing **Select** icon and then selecting `s3://<S3bucket>/results/` and clicking **Select** button.
      **Note:** If you are unable to update the **Query results location** setting in the Amazon Athena console, refer to the Troubleshooting section.

    - Choose **Save**

## Run a Query

⚠ **Note** If there is a banner to upgrade the Athena Data Catalog, choose **here** and select **Upgrade**.

42. Choose the **Query editor** tab at the top of the page and configure the following:

    - *Data source:* **AwsDataCatalog**
    - *Database:* **review-db**

    Under **Tables**, the **review** table is listed.

43. Copy and paste the following command in the **New query 1** tab:

    ```
    SELECT * FROM review LIMIT 10;
    ```

44. To run the command, choose **Run query**

    The first 10 records from the **review** table appear in results window.

⚠ **Note** If the results do not appear, wait for 5 minutes, and re-run the query.

# Task 5: Manage Users with AWS Lake Formation Policies

To maintain backward compatibility with AWS Glue, Lake Formation has the following initial security settings:

- The *Super* permission is granted to the group **IAMAllowedPrincipals** on all existing Data Catalog resources.
- *Use-only* IAM access control settings are enabled for new Data Catalog resources.

These settings effectively cause access to Data Catalog resources and Amazon S3 locations to be controlled solely by IAM policies. Individual Lake Formation permissions are not in effect.

To use the Lake Formation permissions, revoke permission for **IAMAllowedPrincipals**.

45. In the AWS Management Console, choose **Services ⌄** and select **AWS Lake Formation**.

46. In the left navigation pane, choose **Settings** from the **Data catalog** section.

47. Clear both checkboxes for **Use only IAM access control....**

48. Choose **Save**

49. In the left navigation pane, choose **Admins and database creators** from the **Permissions** section.

50. In the **Database creators** section, select **IAMAllowedPrincipals**, and choose **Revoke**.

    The **Revoke permissions** dialog box appears, showing that **IAMAllowedPrincipals** has the *Create database* permission.

51. Choose **Revoke**.

52. In the left navigation pane, choose **Data permissions** from the **Permissions** section.

53. Select the **IAMAllowedPrincipals** principal with the **Database** resource type and **review-db** as resource.

54. Choose **Revoke**.

    The **Revoke permissions** dialog box appears, showing that **IAMAllowedPrincipals** has *Super* database permissions.

55. Choose **Revoke**.

56. Repeat the previous steps to revoke *Super* database permissions for the **IAMAllowedPrincipals** principal with the **Table** resource type.

    **Note** To show the full list of principals, you may need to refresh the page or clear the filter area above the list.

    **Question** Run a query in the Athena Query Editor. Are any results returned? What could be the reason for failure?

## Grant the User Access to the Table

57. In the left navigation pane, choose **Data permissions** from the **Permissions** section.

58. Choose **Grant** at the top-right of the page and configure the following:

    - *IAM users and roles:* **awsstudent**
    - *Database:* **review-db**
    - *Table:* **review**
    - *Table permissions:* Choose only the checkbox for **Select**
    - *Grantable permissions:* Uncheck all options except the **Select** permission; be sure to uncheck **Super**

    **Note** With these selections, you are allowing the **awsstudent** user to perform *select* operations on the **review** table. The user can also grant *select* permissions to other users.

59. Choose **Grant**.

60. Run a query in the Athena Query Editor. The query should be successful.

## Challenge Task: Add a User with Restrictive Permissions to Access Data

Add permissions to **testuser** to limit the user's access to the *product_title* and *star_rating* columns.

**Note** After adding the permissions to the user, log in to the AWS Management Console using the **testuser** credentials on the left side of the lab page and validate access to the table columns.

If you have trouble, refer to the Challenge Solution section.

## Conclusion

👍 Congratulations! You now have successfully:

- Created a data lake and a database
- Crawled the data with AWS Glue to create the metadata and table
- Queried the data using Amazon Athena
- Managed user permissions in Lake Formation

# End Lab

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

61. Return to the AWS Management Console.

62. On the navigation bar, choose **awsstudent@<AccountNumber>**, and then choose **Sign Out**.

63. Choose **End Lab**

64. Choose **OK**

65. (Optional):

    - Select the applicable number of stars ☆
    - Type a comment
    - Choose **Submit**

        - 1 star = Very dissatisfied
        - 2 stars = Dissatisfied
        - 3 stars = Neutral
        - 4 stars = Satisfied
        - 5 stars = Very satisfied

You may close the window if you don't want to provide feedback.

For more information about AWS Training and Certification, see *http://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*.

# Appendix

- In the AWS Cloud9 terminal window, run the following AWS CLI command to retrieve the **S3 bucket name**:

```
dataBucket=$(aws s3api list-buckets --query "Buckets[?contains(Name,
'databucket')].Name" --output text)

echo "S3Bucket=$dataBucket"
```

💡 Click here to go to the next step in the lab.

# Challenge Solution

66. In the AWS Management Console, choose **Services⌄** and select **AWS Lake Formation**.

67. In the left navigation pane, in the **Permissions** section, choose **Data permissions**.

68. Choose **Grant** at the top-right of the page and configure the following:

    - *IAM users and roles:* **testuser**
    - *Database:** **review-db**
    - *Table:* **review**
    - *Columns:* **Include columns**
    - *Include columns:* Choose **product_title** and **star_rating**
    - *Table permissions:* Choose **Select**
    - *Grantable permissions:* Choose **Select**

69. Choose **Grant**.

70. Sign out of the console.

71. Sign back in to the console with the following credentials:

    - **IAM user name:** `testuser`
    - **Password:** Paste the **testuserPassword** value from the left side of the lab page

72. Choose Sign in

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

## Update the Query Result Location

Before you run a query, you need to set up a query result location in Amazon S3.

73. In the AWS Management Console, choose **Services ⌄** and select **Athena**.

74. At the top-right of the page, choose **Settings** and configure the following:

    - **Query result location:** `s3://<S3bucket>/results/`

⚠ **Note:** Replace *<S3Bucket>* with the **S3Bucket** value from the left side of the lab page.

```
<i class="fas fa-exclamation-triangle" style="color:orange"></i>
**Note:** If you are unable to update the **Query results location**
setting in the Amazon Athena console, refer to the <a
href="#Trobuleshooting">**Troubleshooting**</a> section.
```

- Choose **Save**

## Run a Query

75. Choose the **Query editor** tab at the top of the page and configure the following:

- *Data source:* **AwsDataCatalog**
- *Database:* **review-db**

Under **Tables**, the **review** table is listed.

76. Copy and paste the following command in the **New query 1** tab:

```
SELECT * FROM review LIMIT 10;
```

77. To run the command, choose **Run query**

In the results, notice that only data for the *product_title* and *star_rating* columns was returned.

# Troubleshooting

## Unable to Set Query Results Location in Amazon Athena

In the AWS Management Console, choose Services and select **Athena**

Choose **Get Started**

Choose **Workgroup: primary** at the top of the page

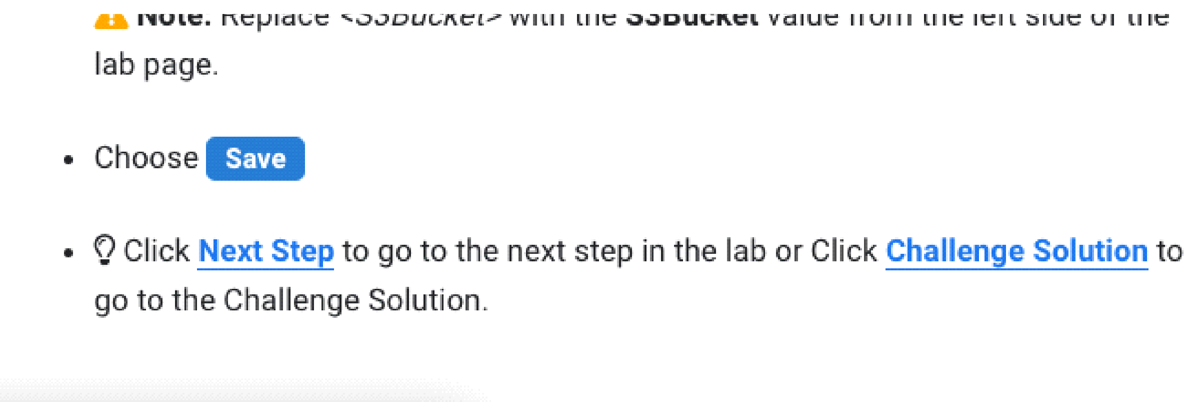
Select the **primary** workgroup.

Choose **View details**.

Choose **Edit workgroup** and configure the following:

- **Query result location:** `s3://<S3bucket>/results/`

⚠ **Note:** Replace *<S3Bucket>* with the **S3Bucket** value from the left side of the



⚠ **Note:** Replace *<S3Bucket>* with the **S3Bucket** value from the left side of the lab page.

- Choose **Save**

- 💡 Click **Next Step** to go to the next step in the lab or Click **Challenge Solution** to go to the Challenge Solution.