

Building Data Lakes on AWS –

Lab 2: Automate Data Lake Creation Using AWS Lake Formation Blueprints

© 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](#).

Lab overview

You are a data engineer at Any Company, a cloud marketing organization. You are asked to populate a data lake with content from AWS CloudTrail so the operational analytics team can efficiently query their data with Amazon Athena.

In this lab, you use a workflow provided as an AWS Lake Formation blueprint to greatly simplify the creation of a data lake and ingestion of data. Lake Formation blueprints are workflows you can apply to an existing Lake Formation data lake. You can also apply them as a task in the setup and creation of a new data lake.

Objectives

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

Create an AWS Glue workflow using a Lake Formation blueprint.

Automate the Lake Formation data lake setup process with an AWS Glue workflow.

Create a custom AWS Glue workflow.

ICON KEY

Various icons are used throughout this lab to call attention to different types of instructions and notes. The following list explains the purpose for each icon:

Note: A hint, tip, or important guidance.

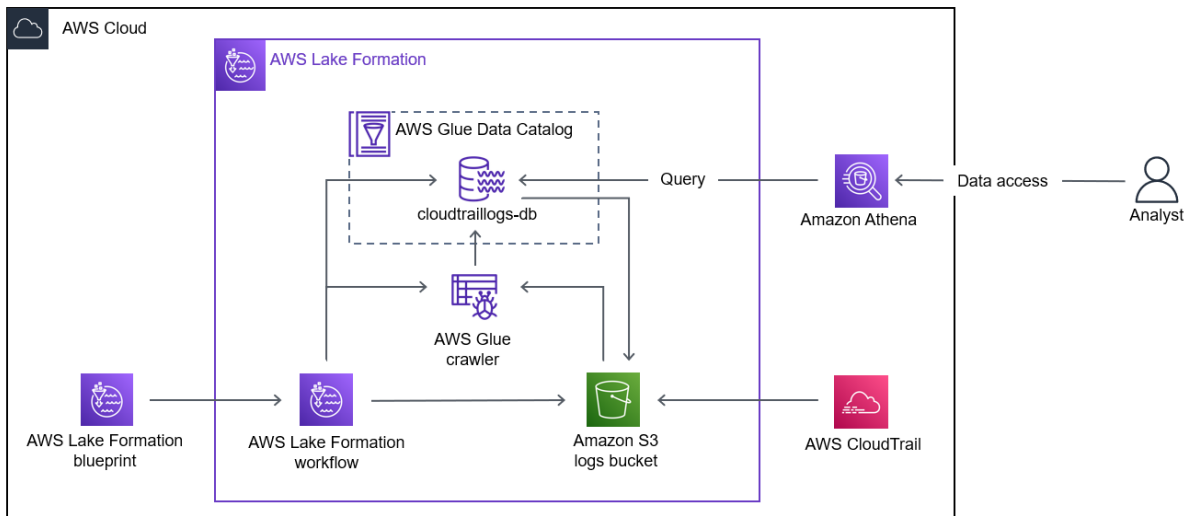
Consider: A moment to pause to consider how you might apply a concept in your own environment or to initiate a conversation about the topic at hand.

Hint: A hint to a question or challenge.

Answer: An answer to a question or challenge.

ENVIRONMENT OVERVIEW

The following diagram shows the basic architecture of the lab environment:



In the preceding diagram, a data lake is created using a Lake Formation blueprint. The blueprint creates a workflow that uses an Amazon Simple Storage Service (Amazon S3) bucket connected to CloudTrail logs, a crawler, and a database. The workflow loads the CloudTrail log data stored in Amazon S3 into the database in the AWS Glue Data Catalog. An analyst queries the Lake Formation database with Athena. Athena uses the AWS Glue Data Catalog and reads data from the database.

Start lab

To launch the lab, at the top of the page, choose [Start lab](#).

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

To open the lab, choose [Open Console](#).

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

WARNING: Do not change the Region unless instructed.

COMMON SIGN-IN ERRORS

Error: You must first sign 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 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: Explore the lab environment

In this task, you review the account resources created before the lab started.

At the top of the AWS Management Console, in the search bar, search for and choose

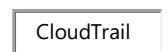
S3.

Choose the link for the bucket name that starts with **cloudtraildatabucket**.

Note: The **data/** folder will contain your dataset. The **results/** folder stores the results of your Athena queries. You will specify the **results/** folder as the query results location in Athena later in this lab.

Choose the **data/** folder. You will specify the **data/** folder as your designated storage location for your data lake later in this lab.

At the top of the AWS Management Console, in the search bar, search for and choose

CloudTrail.

In the **Trails** section, choose **LabCloudTrail**.

Note: If you receive an error related to “Access Denied” or “You do not have permission to perform this action” on cloudtrail console in rest of lab, then please ignore and proceed through next steps.

Consider: Which bucket are the logs stored in?

You use this trail as part of your Lake Formation blueprint.

Congratulations! You successfully explored the lab environment.

Task 2: Set up Lake Formation

In this task, you register your Amazon S3 data storage and create a database.

TASK 2.1: REGISTER YOUR AMAZON S3 STORAGE

Lake Formation manages access to designated storage locations in Amazon S3. Register the storage locations you want to be part of the data lake.

At the top of the AWS Management Console, in the search bar, search for and choose

AWS Lake Formation

In the **Welcome to Lake Formation** popup window, make sure **Add myself** is selected and then choose **Get started**.

Note: If you see **Read Only Admin is not supported**.

Go to settings under **Data catalog** and choose **Get started**.

choose **Save**.

In the left navigation pane, in the **Administration** section, choose **Data lake locations**.

Choose **Register location**.

On the **Register location** page, in the **Amazon S3 location** section:

For **Amazon S3 path**, copy and paste the **SourceDataLocation** value that is listed to the left of these instructions.

For **IAM role**, select **LakeFormationServiceRole**.

For **Permission mode**, select **Lake Formation**.

Choose **Register location**.

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

Choose **Grant**.

On the **Grant permissions** page, configure the following:

For **IAM users and roles**, select **LakeFormationWorkflowRole**.

For **Storage locations**, copy and paste the **SourceDataLocation** value that is listed to the left of these instructions.

Choose **Grant**.

Your Amazon S3 bucket **data/** folder is now registered as the storage location for your data lake.

TASK 2.2: CREATE A DATABASE

Lake Formation organizes data into a catalog of logical databases and tables. Create a database in the AWS Glue Data Catalog.

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

Choose **Create database**.

On the **Create database** page, in the **Database details** section:

For **Name**, enter

cloudtraillogs-db

For **Location - optional**, copy and paste the **SourceDataLocation** value that is listed to the left of these instructions.

Choose **Create database**.

You created a database in the AWS Glue Data Catalog using Lake Formation.

Congratulations! You successfully registered your Amazon S3 data storage and created a database.

Task 3: Use a Lake Formation blueprint to create an AWS Glue workflow

In this task, you use a Lake Formation blueprint to create an AWS Glue workflow that will automatically add new content to your data lake.

A workflow encapsulates a complex multi-job extract, transform, and load (ETL) activity. Workflows generate AWS Glue crawlers, jobs, and triggers to orchestrate the loading and updating of data. Lake Formation runs and tracks a workflow as a single entity. You can configure a workflow to run on-demand or on a schedule.

Workflows you create in Lake Formation are visible in the AWS Glue console as a directed acyclic graph (DAG). Each DAG node is a job, crawler, or trigger. To monitor progress and troubleshoot, you can track the status of each node in the workflow.

In the left navigation pane, in the **Ingestion** section, choose **Blueprints**.

Choose **Use blueprint**.

On the **Use a blueprint** page:

For **Blueprint type**, choose **AWS CloudTrail**.

For **CloudTrail name**, choose **LabCloudTrail**.

For **Start date**, choose today's date.

For **Target database**, choose **cloudtraillogs-db**.

For **Target storage location**, add the **S3 bucket** from the left of instructions.

For **Data format**, choose **Parquet**.

For **Frequency**, choose **Run on demand**.

For **Workflow name**, enter

If-cloudtrail-workflow

For **IAM role**, choose **LakeFormationWorkflowRole**.

For **Table prefix**, enter

lab

Choose **Create**

Choose the refresh icon in the **Workflows** section until you see your new workflow.

Congratulations! You successfully used a Lake Formation blueprint to create an AWS Glue workflow that will automatically add new content to your data lake.

Task 4: Run and monitor the workflow

In this task, you run an AWS Glue workflow. While it is running, explore AWS Glue workflows and monitor your workflow until it is complete.

TASK 4.1: RUN THE WORKFLOW

Run a workflow and view the DAG in AWS Glue.

Choose the link for **If-cloudtrail-workflow**.

Choose **Start**.

Refresh your browser tab.

In the **Workflow runs** section, a new run appears.

At the top of the AWS Management Console, in the search bar, search for and choose

AWS Glue

In the left navigation pane, in the **Data Integration and ETL** section, choose **Workflows**.

Note: If the **Workflows** page does not load, choose **Workflows** again.

The **Last run status** for the **If-cloudtrail-workflow** is **Running**.

Choose **If-cloudtrail-workflow**.

Choose the **History** tab.

Choose the run that starts with **wr_**.

Choose **View run details**.

The workflow DAG appears.

The workflow takes 15-20 minutes to run. You can view the workflow progress using the DAG in the **Graph** section.

Consider: Take a moment to view the DAG. Which nodes have already completed?

While you wait for this workflow to complete, continue with the next task.

TASK 4.2: EXPLORE AWS GLUE WORKFLOWS

While your workflow is running, build a sample AWS Glue workflow and explore the DAG.

In the left navigation pane, in the **Data Integration and ETL** section, choose **Workflows**.

Choose **Add workflow**.

For **Workflow name**, enter

sample-workflow

Choose **Create workflow**.

Choose **sample-workflow**.

An empty graph editor appears.

In an AWS Glue workflow, you have four node types to choose from:

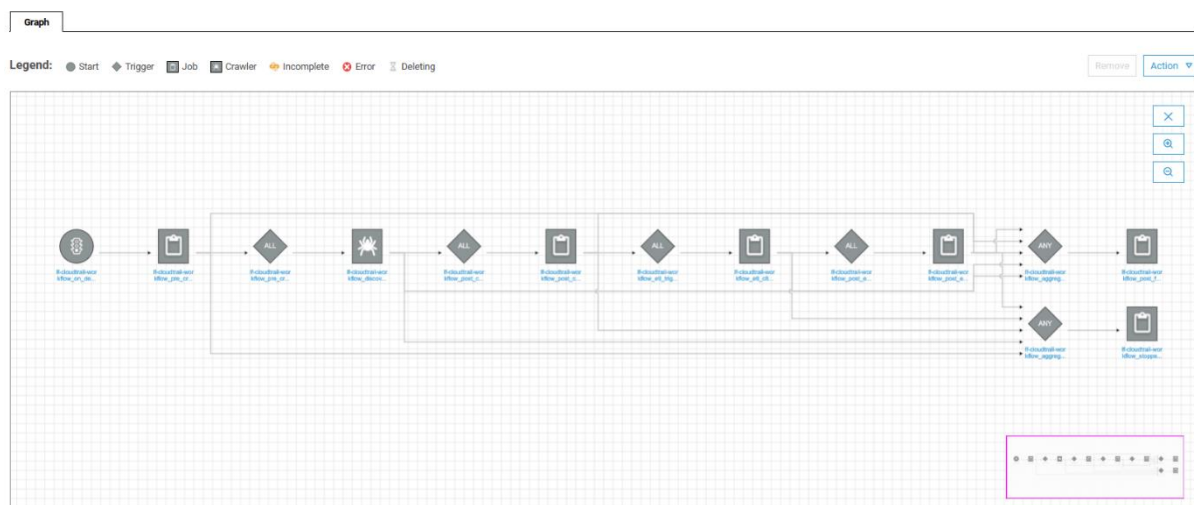
Start: A node that starts the workflow.

Trigger: A node that activates a job based on an event.

Job: A node that completes work.

Crawler: A node that discovers a source schema.

The Lake Formation blueprint created a workflow with several nodes, all connected and displayed in a DAG.



Note: If you cannot see the full trigger names, expand the **Name** section with the bar | icon next to **Name** until the full name of each trigger is displayed.

Choose **Add**.

A start node and a job node are created. These are the first two nodes shown in your lf-cloudtrail-workflow.

Next, create the trigger that starts an AWS Glue crawler.

Choose the job node, and then choose **Add trigger**.

For **Name**, enter

pre_crawl_trigger

For **Trigger logic**, choose **Start after ALL watched event**.

Choose **Add**.

After the **pre_crawl_trigger** node, choose **Add node**.

Choose the **Crawlers** tab.

Choose the crawler that contains **discoverer** in its name.

Choose **Add**.

A trigger and a crawler are added. There are now four nodes in your workflow.

Next, create a trigger that waits for the crawler to finish.

Choose the crawler node, and then choose **Add trigger**.

For **Name**, enter

post_crawl_trigger

For **Trigger logic**, choose **Start after ALL watched event**.

Choose **Add**.

After the **post_crawl_trigger** node, choose **Add node**.

Choose the **Jobs** tab.

Choose the job that contains **post_crawl** in its name.

Choose **Add**.

A trigger and a job node are added. There are now six nodes in your workflow.

At this point, your workflow starts a job that triggers a crawler. Then, a trigger notifies the next job when the crawler is done.

To finish your workflow, complete the challenge task to add pre-ETL and post-ETL triggers and jobs.

CHALLENGE A: ADD MORE NODES TO THE AWS GLUE WORKFLOW

To finish your sample workflow, add the following nodes:

An **etl** trigger

An **etl** job

A **post_etl** trigger

A **post_etl** job

Hint: Start by adding the **etl** trigger after the **post_crawl** job node. When you add a trigger, you are prompted to add a job node.

Answer: Navigate [here](#) for a solution.

TASK 4.3: MONITOR THE WORKFLOW

Now that you have explored AWS Glue workflows, return to the workflow you created with the Lake Formation blueprint and view its progress.

In the left navigation pane, in the **Data Integration and ETL** section, choose **Workflows**.

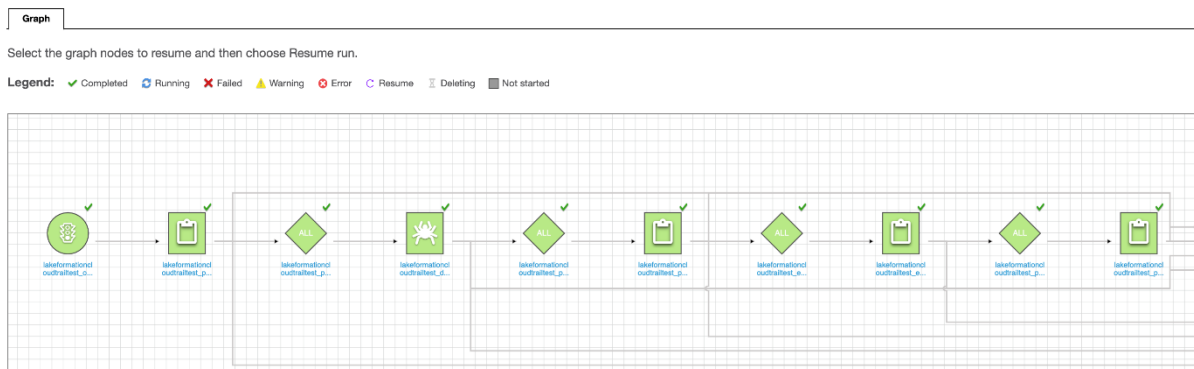
Choose **lf-cloudtrail-workflow**.

Choose **History**.

Choose the run that starts with **wr_**.

Choose **View run details**.

The workflow DAG appears.



In the preceding image, all of the nodes in the main line of the workflow graph have a Completed status.

A **Run status** of **Completed** indicates that your new CloudTrail logs database was successfully created.

Note: If your workflow run is not finished, wait until the **Run status** is **Completed**.

Congratulations! You have successfully run an AWS Glue workflow, created a custom workflow, and monitor the status of your workflow.

Task 5: Validate the data lake setup results

In this task, you validate the data lake setup results and use Athena to view a sample of the table.

At the top of the AWS Management Console, in the search bar, search for and choose

AWS Lake Formation

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

You will see two tables as follows:

_lab_cloudtrail

lab_cloudtrail

The **_lab_cloudtrail** table contains the CloudTrail data before it was transformed to the Parquet format. The **lab_cloudtrail** table contains the transformed data.

On the **Tables** page, choose the **lab_cloudtrail** table to see its details.

The table has a Parquet data format and contains 20 columns.

At the top of the AWS Management Console, in the search bar, search for and choose

Athena

Note: If the Athena **Get Started** menu appears, choose **Query your data with Trino SQL** option and then choose **Launch query editor**.

In the **Workgroup primary settings** window, choose **Acknowledge**.

In the **Data** section, the **AWSDataCatalog** data source and **cloudtraillogs-db** database are automatically selected. In the **Tables** section, you see two tables listed.

Choose the plus sign + to create a new query.

In the query editor, enter the following:

```
SELECT * FROM "cloudtraillogs-db"."lab_cloudtrail" limit 10;
```

Choose **Run**.

In the **Results** section, you see 10 records from the **lab_cloudtrail** table.

Consider: Take a moment to explore the results. Which columns are included in the **lab_cloudtrail** table?

CHALLENGE B: FIND ALL THE ERRORS IN THE CLOUDTRAIL LOGS

Now that you have set up the CloudTrail data lake, AnyCompany wants you to find all the errors captured in today's logs. Run a new query to find all the logs that contain an error code.

Hint: Find all the error messages by using

WHERE NOT errorcod

in your query.

Answer: Navigate [here](#) for a solution.

Congratulations! You successfully verified the data lake setup results in Athena.

Conclusion

Congratulations! You now have successfully:

Created an AWS Glue workflow using a Lake Formation blueprint

Automated the Lake Formation data lake setup process with an AWS Glue workflow

Created a custom AWS Glue workflow

End lab

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

Return to the **AWS Management Console**.

At the upper-right corner of the page, choose **AWSLabsUser**, and then choose **Sign out**.

Choose **End lab** and then confirm that you want to end your lab.

Appendix

CHALLENGE A SOLUTION

To finish your sample workflow, add the following nodes:

An **etl** trigger

An **etl** job

A **post_etl** trigger

A **post_etl** job

Create a trigger that waits for the **post_crawl** job to finish.

Choose the **post_crawl** job node, and then choose **Add trigger**.

For **Name**, enter

workflow_etl_trigger

For **Trigger logic**, choose **Start after ALL watched event**.

Choose **Add**.

After the **workflow_etl_trigger** node, choose **Add node**.

Choose the **Jobs** tab if it is not already selected.

Choose the job that contains **workflow_etl** in its name.

Choose **Add**.

A trigger and a workflow_etl job node are added. There are now eight nodes in your workflow.

Create a trigger that waits for the ETL job to finish.

Choose the crawler node, and then choose **Add trigger**.

For **Name**, enter

post_etl_trigger

For **Trigger logic**, choose **Start after ALL watched event**.

Choose **Add**.

After the **post_etl_trigger** node, choose **Add node**.

Choose the **Jobs** tab if it is not already selected.

Choose the job that contains **post_etl** in its name.

Choose **Add**.

A trigger and a post_etl job node are added. There are now 10 nodes in your workflow.

You successfully replicated the main line of nodes created by your Lake Formation blueprint.

Consider: Take a moment to explore other options available to you in the AWS Glue workflows graph editor. You can view the trigger and job details for each node, connect nodes together using **Action**, and add additional nodes to your workflow.

Note: You can run the workflow if you want to. It completes the same tasks as the workflow created by the Lake Formation blueprint.

To continue this lab, go to [Task 4.3](#).

CHALLENGE B SOLUTION

Query your table for all the CloudTrail logs with error code.

Choose the plus sign + to create a new query.

In the query editor, enter the following:

```
SELECT * FROM "cloudtraillogs-db"."lab_cloudtrail" WHERE NOT errorcode=";
```

Choose **Run**.

In the **Results** section, you see all the CloudTrail logs that contain an error code.

You successfully queried all the CloudTrail logs that contain an error code.