# Lab 3 - Orchestrate data processing in Spark using AWS Step Functions

High-level obj - add 2020 data to 2019 data

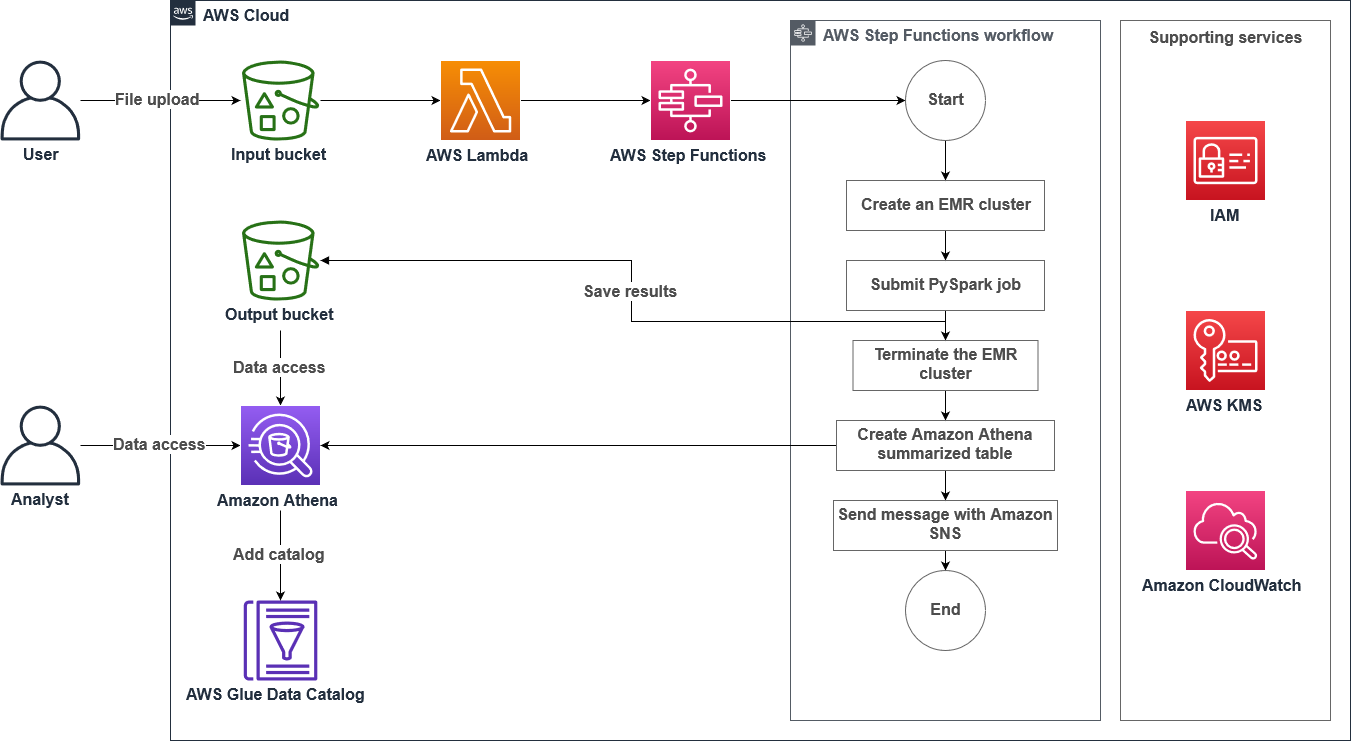
1. Use S3 Event Notifications and AWS Lambda to automate the batch processing of data
2. Use the Step Functions state machine language to:
   1. Create an on-demand Amazon EMR cluster
   2. Add an Apache Spark step job in Amazon EMR and create an Amazon Athena table to query the processed job
   3. Add an Amazon SNS topic to send a notification
3. Validate a Step Functions state machine run

In short - Automate the batch processing of 2020 stock data, append it to 2019 data

AWS Step Functions is a service that orchestrates workflows by coordinating multiple AWS services into serverless workflows.

step functions workflow-

1. Create EMR Cluster: Automatically create an Amazon EMR cluster when needed for processing data.
2. Apache Spark Step Job: Run a Spark job on the EMR cluster to process the 2020 stock data and append it to the 2019 data.
3. Amazon Athena Table: Create a table in Athena to query the processed data.
4. Amazon SNS Topic: Send a notification when the job is completed.



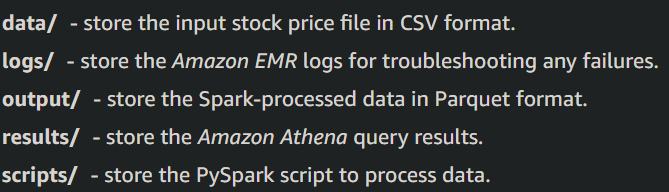
Arch workflow -

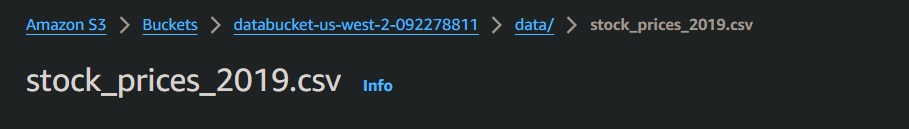
1. Amazon S3: Stores the 2020 stock data.
2. AWS Lambda: Responds to new data uploads and initiates processing.
3. Step Functions: Manages the workflow of creating the EMR cluster, running the Spark job, and terminating the cluster.
4. Amazon EMR: Processes the stock data using Apache Spark.
5. Amazon Athena: Allows querying of the processed data.
6. Amazon SNS: Sends notifications about the job status.
7. AWS Glue: Catalogs the processed data for easy querying with Athena.

**Task 1: Explore the lab environment**

1.1 view 2019 data in S3

File in bucket



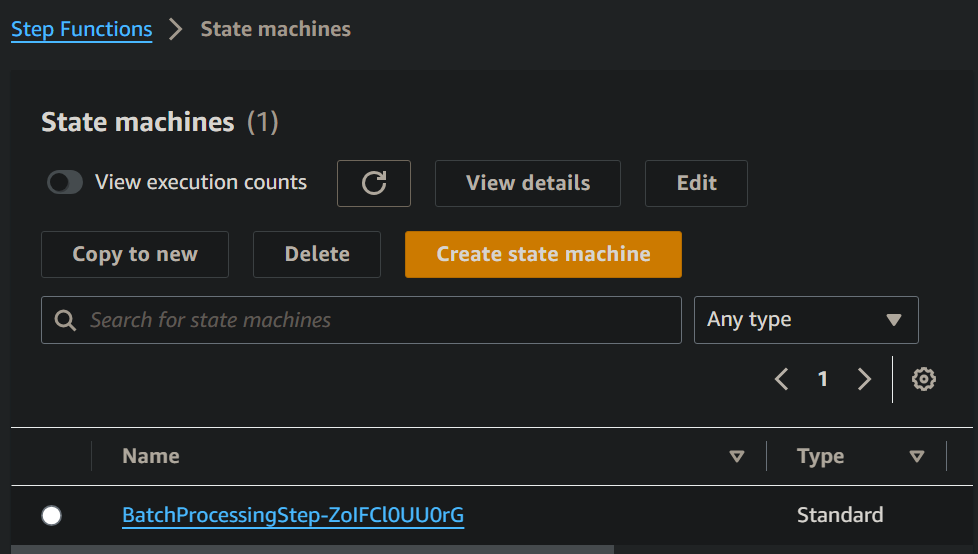


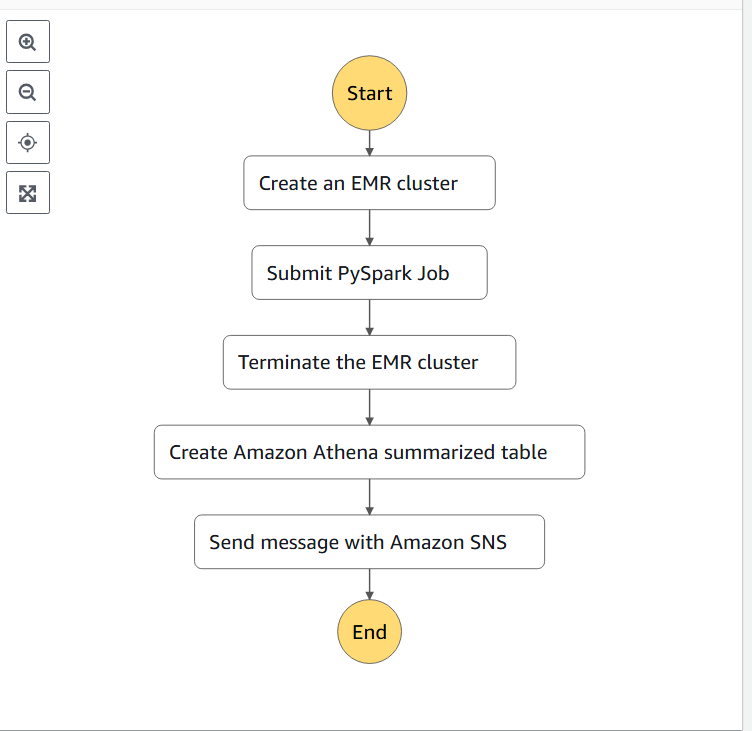
1.2 lambda function - The Lambda function initiates the Step Functions run



In this code, the Lambda function starts the Step Functions step if the event contains databucket as a source bucket and the event name contains ObjectCreated:Put.

1.3 view the step functions (open Step Functions )



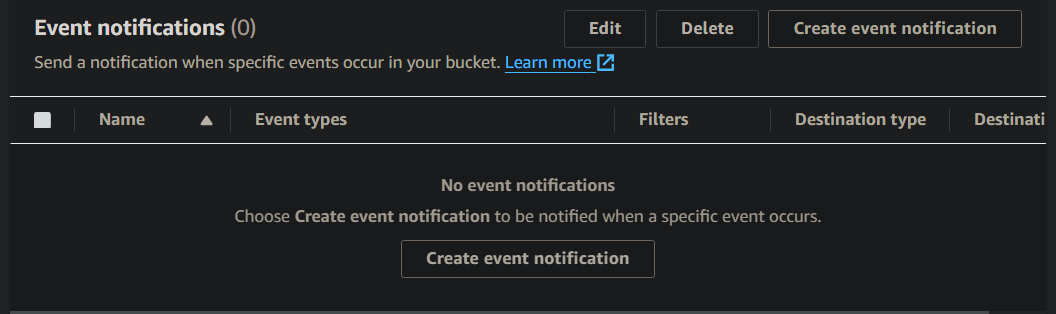


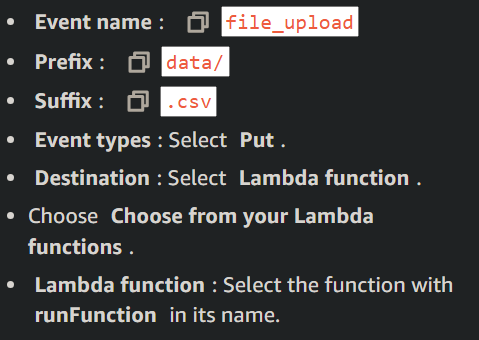
**Task 2: Run the Step Functions state machine task**

2A - Configure S3 Event Notifications

use the Amazon S3 Event Notifications feature to run the Lambda function when a new file is uploaded to the S3 folder.

2.1 open bucket, choose



2.2

2B - Update and upload the PySpark script in the Amazon S3 dataBucket

2.3 save this script, locally



This script is made to only process high volume stocks

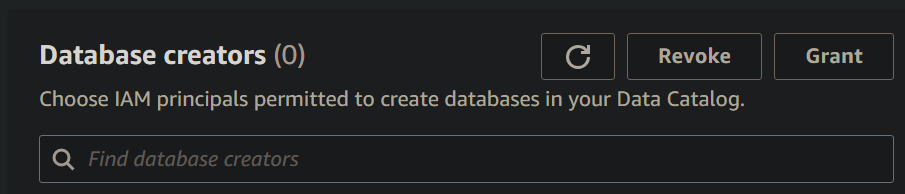
Save this script to script folder in S3, which stores pyspark code on how to process data

The o/p of this is stored to output folder, which stores pyspark processed data

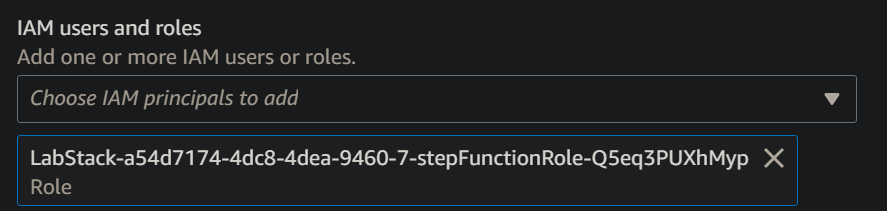
2C - grant the Step Functions state machine permission to create a database. (A DB is created in AWS Glue Data Catalog.)

2.4 go here

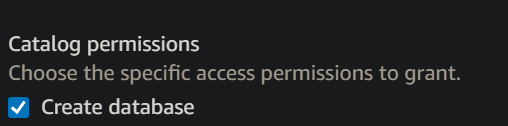


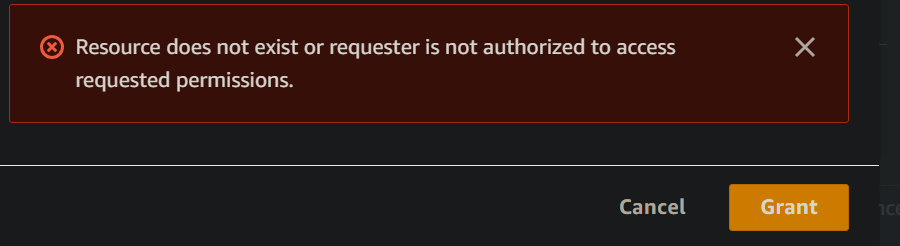
2.5 choose grant

2.6 grant permission to this role



2.7 give permission

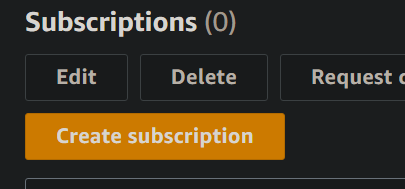


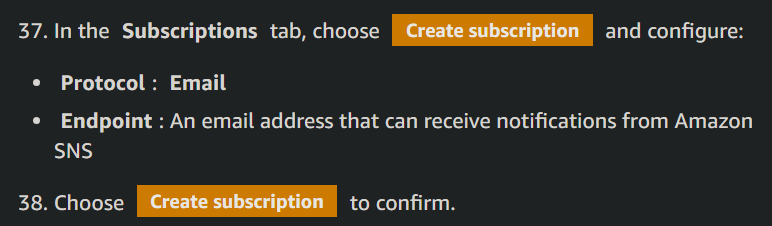
Issue = permission not granted, let’s see how this affects

2D - Subscribe to an Amazon SNS topic

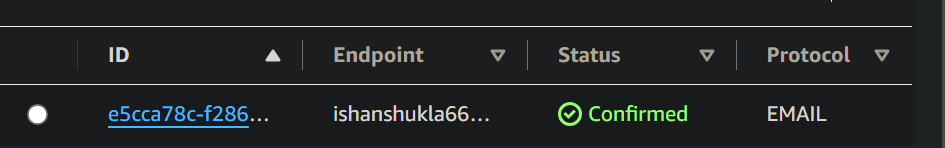
2.8 open SNS, topic has been pre-created and create a subscription

Simple Notification Service





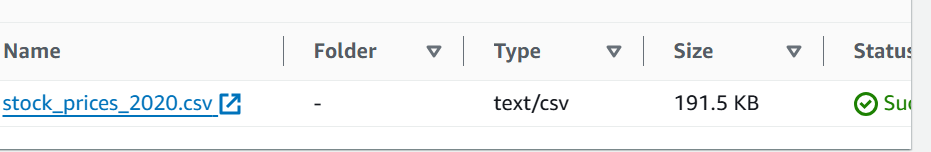
Confirm on email



2E - Upload a file to the Amazon S3 bucket

2.9 

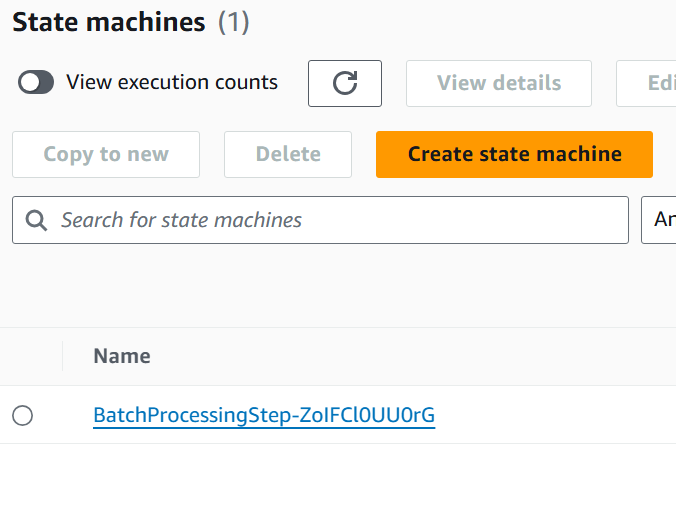
2.10 upload in data folder



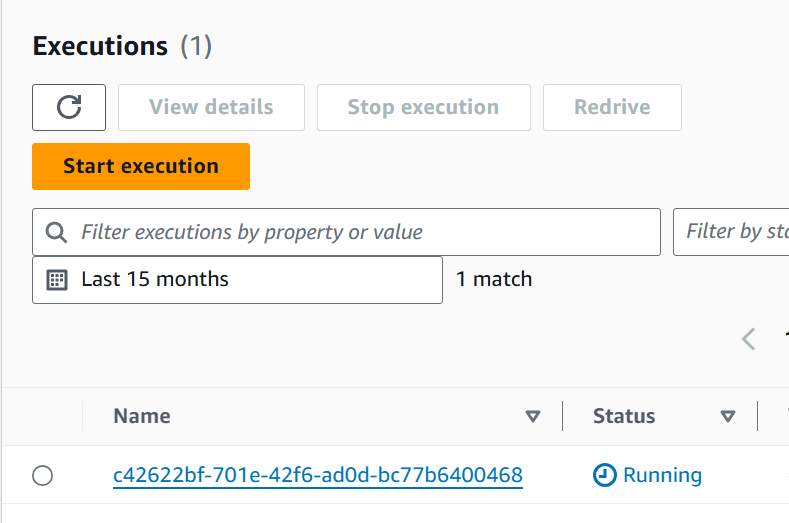
We uploaded a new file,now the lambda function will start step functions.

**Task 3: Validate the Step Functions run**

3.1 Open step functions in console



3.2 select the name of the currently executing state machine



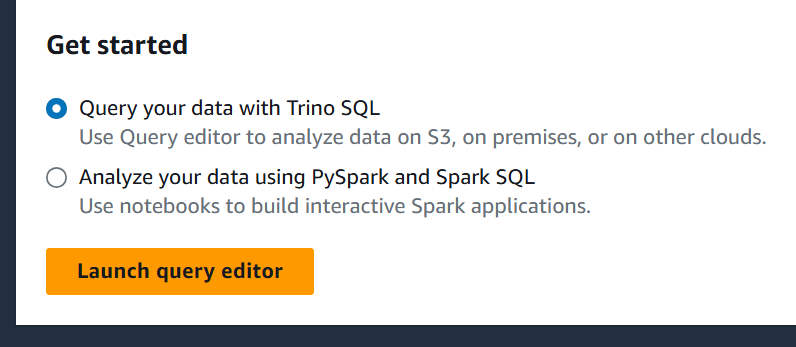
3.3

|  |  |
| --- | --- |

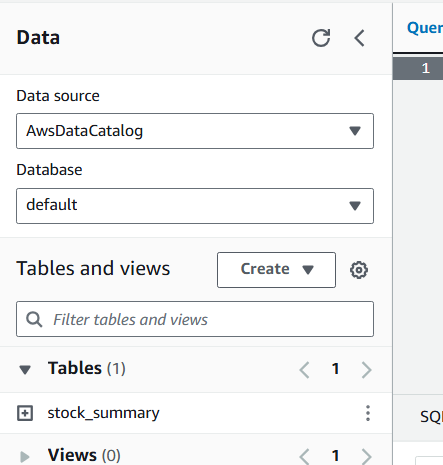
Mail is sent

3.4 **run a query on Athena to validate the data**

3.5 open athena, and

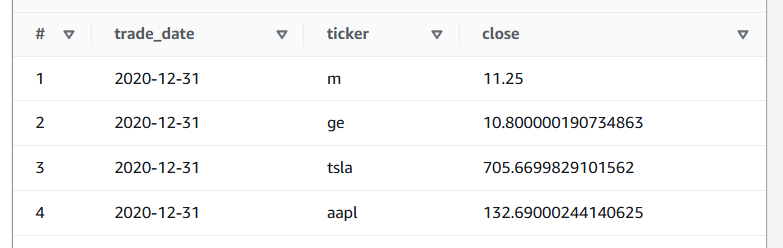


3.6 the data source



3.7 run this

SELECT \* FROM "default"."stock\_summary" ORDER BY trade\_date DESC LIMIT 10;



This shows that 2020 data has been appended to 2019 data