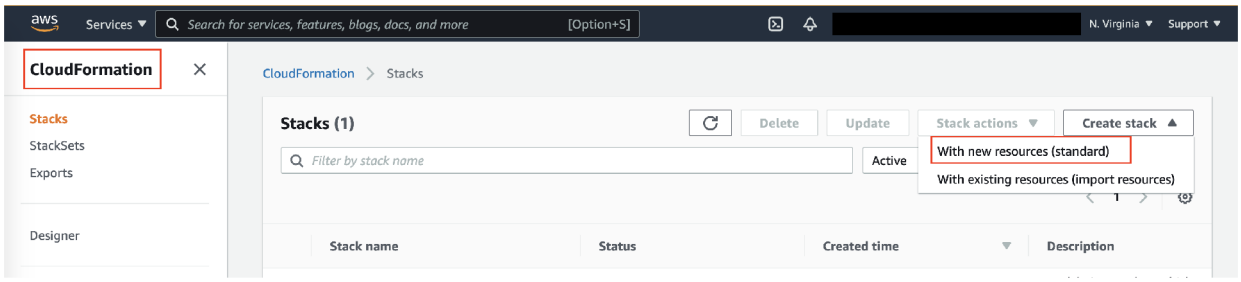
FeatureStore CompactUtil Deployment Guide

Deployment steps:

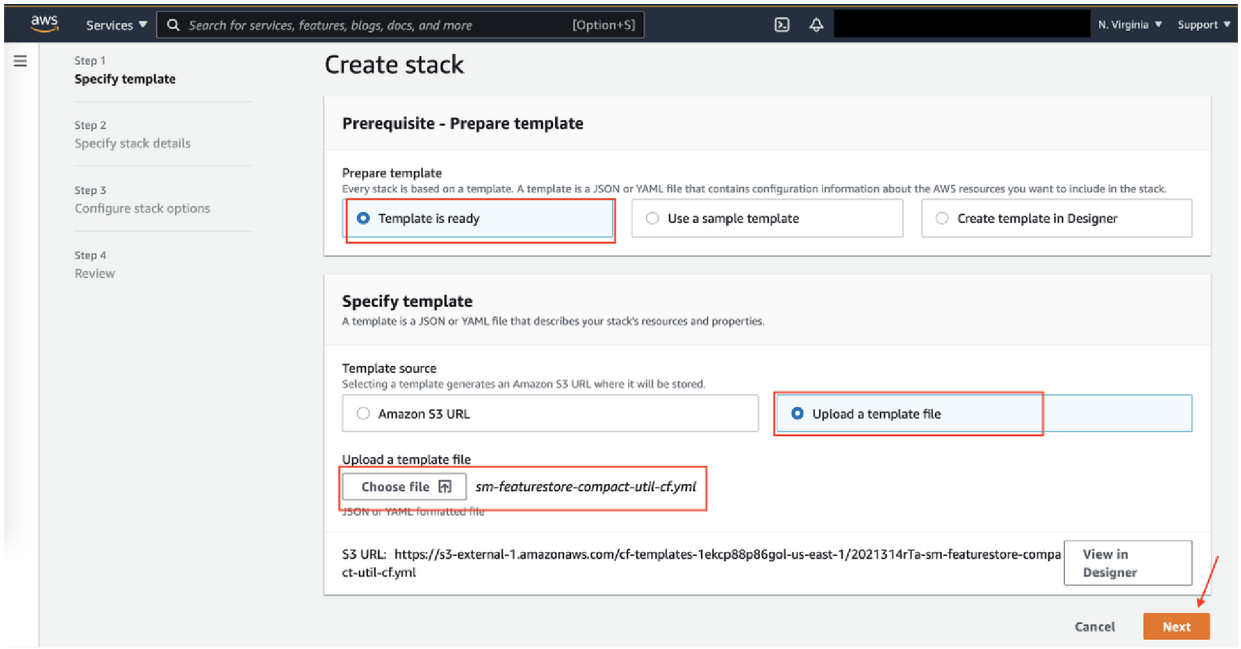
The CloudFormation (CFN) template creates 2 Lambda functions, a Step function, an Event bridge rule and IAM roles necessary for each service to invoke the next service.

Follow these steps to deploy the stack:

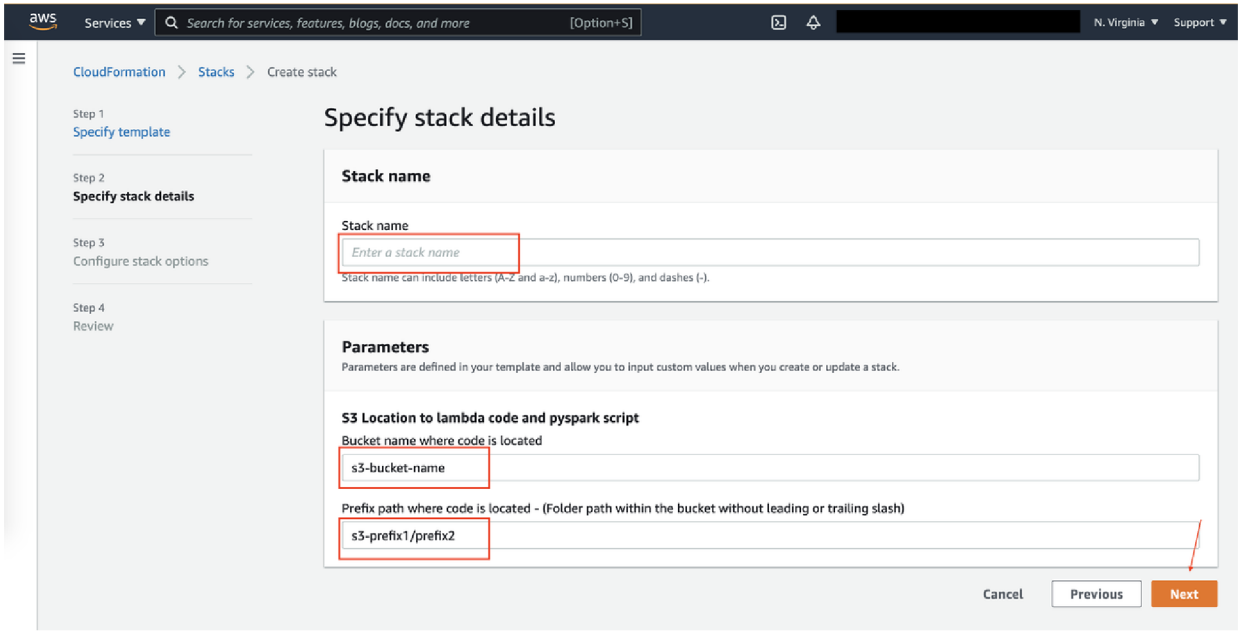
- There are 4 artifacts provided
    - sm-featurestore-offline-compact.zip - Lambda deployment package for compacting function.
    - sm-featurestore-offline-job-monitor.zip - Lambda deployment package for processing job status check.
    - sm-featurestore_offline_compact_spark.py - PySpark script that is fed to sagemaker processing to run compaction.
    - sm-featurestore-compact-util-cf.yml - Cloudformation template that creates all necessary resources.
- Upload the two zip files sm-featurestore-offline-compact.zip and sm-featurestore-offline-job-monitor.zip and the pyspark script sm-featurestore_offline_compact_spark.py to an S3 bucket. The S3 bucket has to be created in the same AWS Region as where Cloudformation stack will be deployed.
- Upload the sm-featurestore-compact-util-cf.yml via AWS CloudFormation console to deploy the stack. Input the S3 bucket name and prefix path as parameters to Cloudformation.
    - Go to CloudFormation console and create stack with new resources



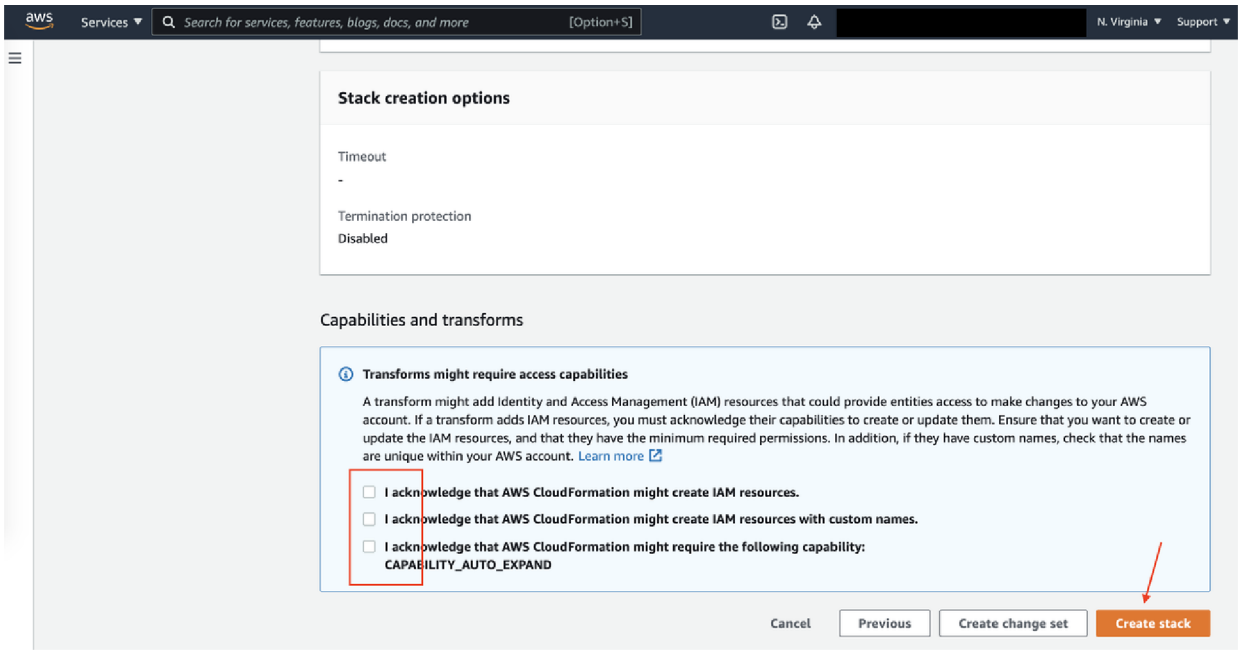
    - Upload the yml file provided and click Next



- Enter a name for the stack, the S3 bucket name where code is uploaded to, and the prefix location if any where code is located. Leave prefix empty if the files are directly uploaded to a bucket and not within folders. Note that prefix should not have leading or trailing slash. Click Next.



- Leave the default settings on the next page and Click Next.
- On the last page, acknowledge the access permissions and click Create Stack.



- SageMaker Processing job is created and run by sm-featurestore-offline-compact Lambda function. The lambda function is created with these environment configuration that can be modified anytime via Lambda console.
    - PYSPARK_SCRIPT_PATH - Pyspark script location in S3
    - SAGEMAKER_INSTANCE_COUNT - Instance count for the job, default is set to 1.
    - SAGEMAKER_INSTANCE_TYPE - Instance type, by default is configured with ml.m5.4xlarge
    - SAGEMAKER_INSTANCE_VOLUME_SIZE - Size of ML storage volume in GB, default is 30.
    - SAGEMAKER_ROLE - IAM role to execute the job, default is what gets created while deploying the stack.
    - SPARK_CONTAINER_IMAGE - Container image that the SageMaker Processing job should use. The location is configured in the Cloudformation template based on the region that the stack gets created in.

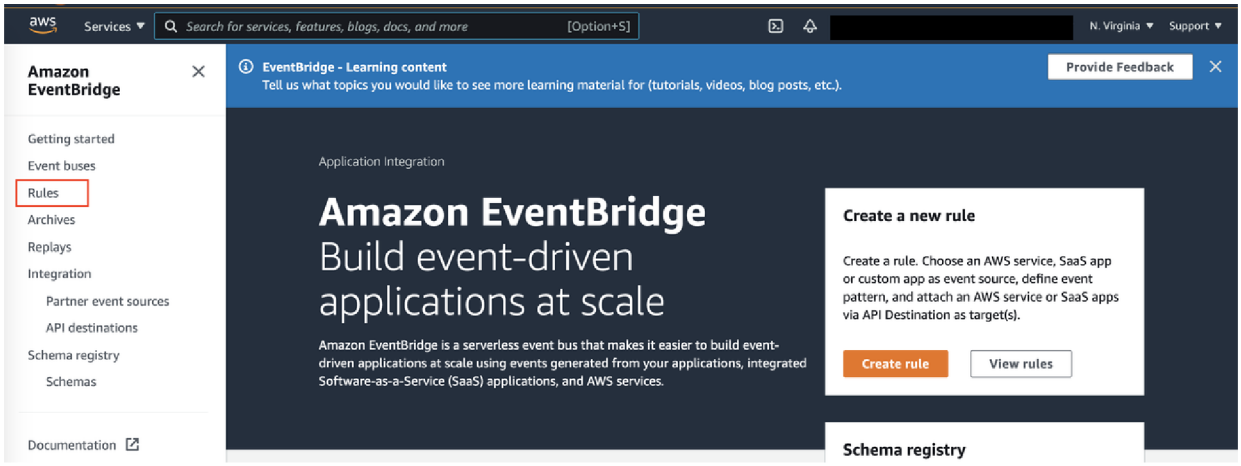
Using the util:

Refer to FeatureStore-CompactUtil-Solution.pdf for detailed architecture and input JSON structure.
The Cloudformation creates an event bridge rule disabled by default and with a default schedule to run every 24 hours. The target for the rule is the step function.
Go to EventBridge console and edit the event rule to

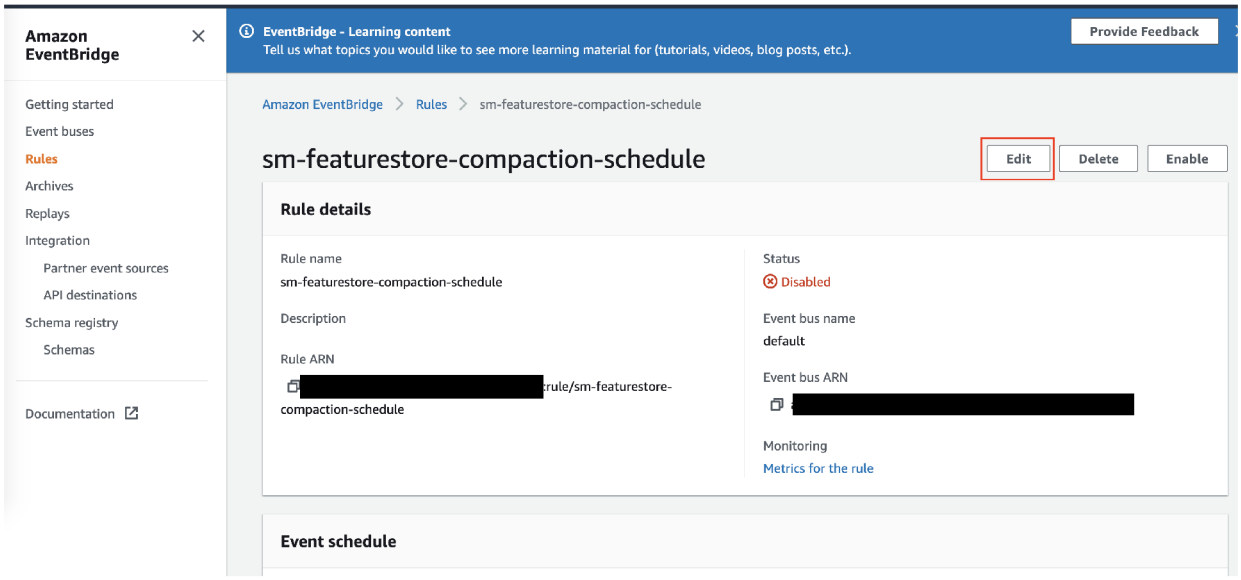
1. Run the util in full mode the first time
2. Run the util in incremental mode after full mode

For #1 Run the util in full mode the first time
:

- **Navigate to EventBridge console and click Rules.**



- **Click sm-featurestore-compaction-schedule rule and click Edit.**

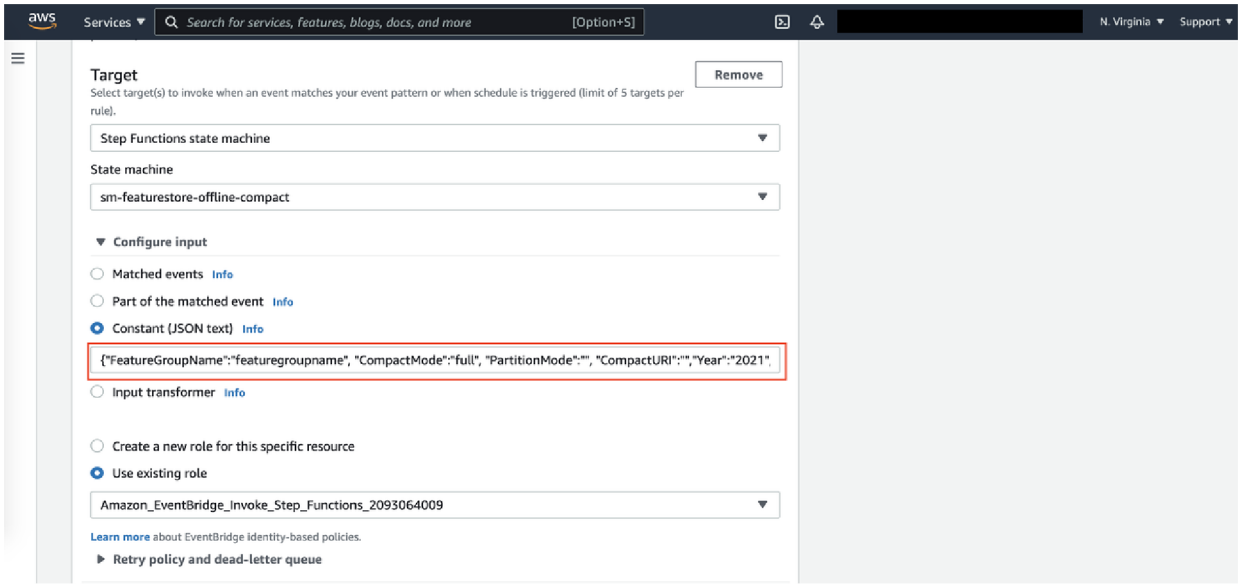


- **Modify the input JSON with the feature group name, compact_mode as “full”, partiton_mode as “hour” or “day” based on needs, “compact_uri” if compated files need to go into user specified location and click Update.**

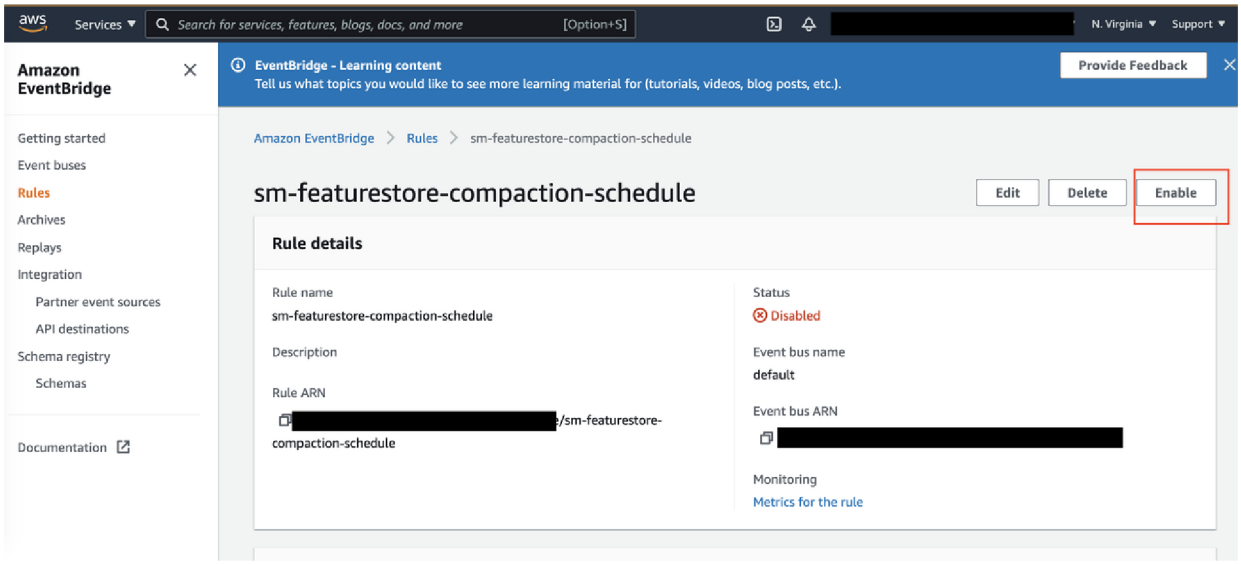
Use the event JSON structure shown here. All keys are mandatory but optional values can be blank.

```
{
    "feature_group_name":"REPLACE_WITH_FEATURE_GROUP_NAME",
    "compact_mode":"full",
    "partition_mode":"hour",
    "compact_uri":"",
    "year":"",
```

```
    "month":"",
    "day":""
}
```



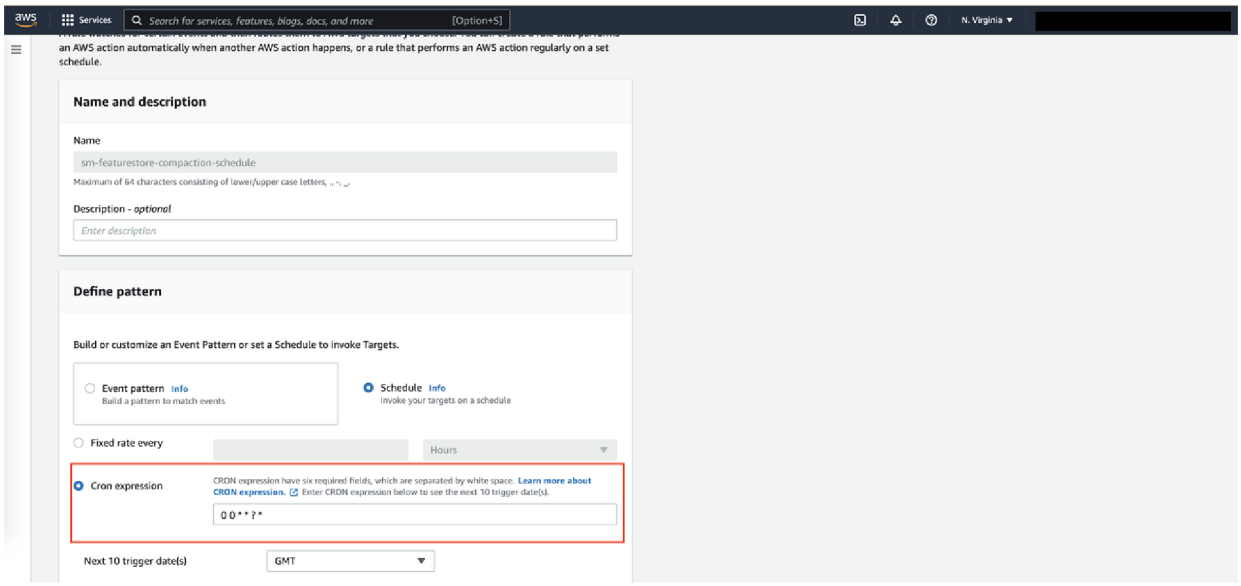
- **Enable the event**



- **The event will fire in 30-60 seconds. Navigate to Step functions and check state machine sm-featurestore-offline-compact if execution has begun (Refer to monitoring section below).**

For #2 Run the util in incremental mode after full mode

- **Once the util runs in "full" mode, it is time to setup the util to run in incremental manner. Navigate to EventBridge and edit the schedule to set schedule to run at midnight UTC everyday. The util will run every night at 12 AM UTC and compacts previous day's files.**



Change the Schedule from Fixed rate every to Cron expression and set an expression to fire event at 12 AM UTC.

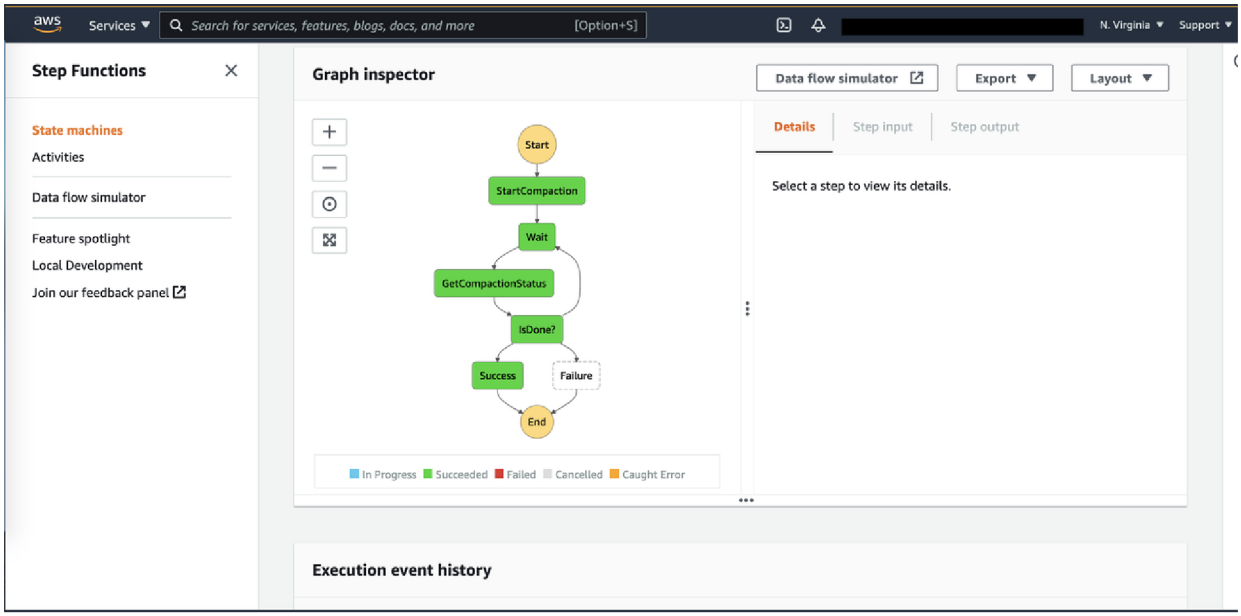
```
0 0 * * ? *
```

Set the input (Constant JSON text) as":

```
{
    "feature_group_name":"REPLACE_WITH_FEATURE_GROUP_NAME",
    "compact_mode":"incremental",
    "partition_mode":"",
    "compact_uri":"",
    "year":"",
    "month":"",
    "day":""
}
```

Monitoring:

- Lambda and SageMaker processing jobs will create logs in Cloudwatch.
- Step functions can be monitored to check if all steps executed without errors. Navigate to Step Functions via AWS Console → StateMachines → Click on sm-featurestore-offline-compact → Click on the latest execution. A graph should be displayed to show the status of the execution. Each step in the graph shows details of input, output, execution success/failure and errors if any.



Validation:

- Check the S3 compact URI location for the compacted files. They will be partitioned by year, month day (and hour if partition_mode is hour). If compact_uri was not specified as an input, the files will be in the same bucket as the offline store. The S3 location will be in the below format
    - `s3://<bucket-name>/<customer-prefix>/<account-id>/sagemaker/<aws-region>/**compact-offline-store**/<feature-group-name>-<feature-group-creation-time>/data/year=<event-time-year>/ month=<event-time-month>/day=<event-time-day>/hour=<event-time-hour>/`
- Create glue crawler and tables for this location and run queries to validate the data.

Helpful resources:
EventBridge rules - https://docs.aws.amazon.com/eventbridge/latest/userguide/eb-create-rule-schedule.html
EventBridge schedule expressions
- https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/ScheduledEvents.html
EventBridge targets - https://docs.aws.amazon.com/eventbridge/latest/userguide/eb-targets.html
SageMaker Processing job - https://docs.aws.amazon.com/sagemaker/latest/dg/processing-job.html
Step Functions invoking Lambda - https://docs.aws.amazon.com/step-functions/latest/dg/connect-lambda.html
Lambda - https://docs.aws.amazon.com/lambda/latest/dg/welcome.html