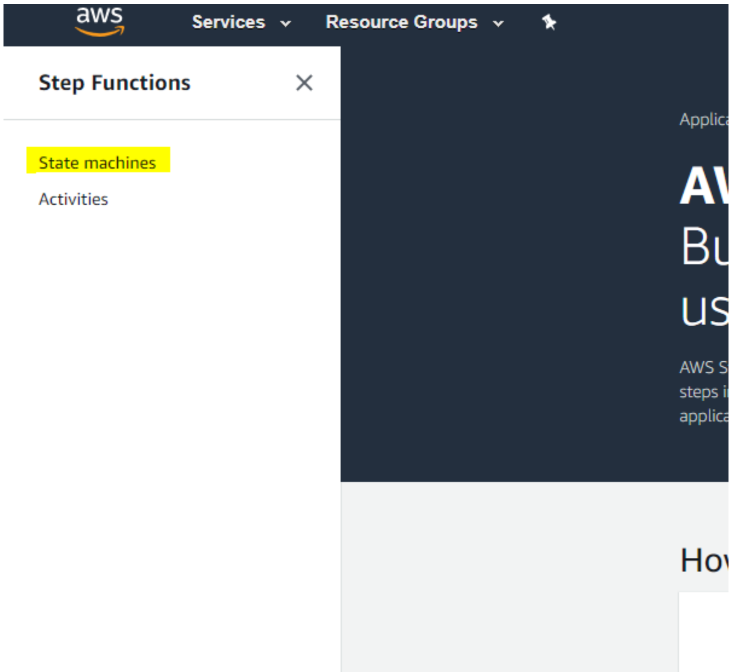
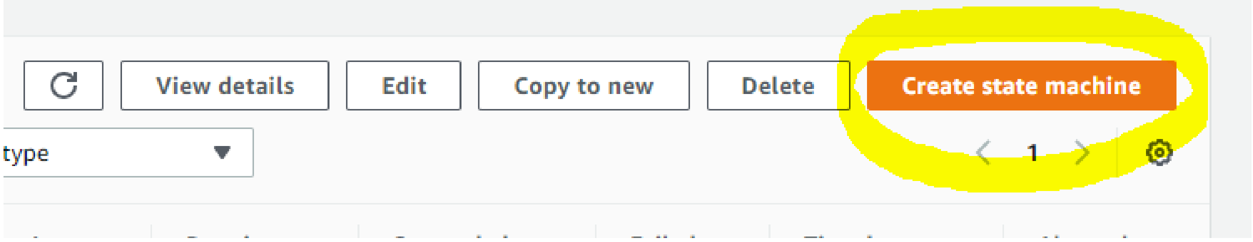
AWS step function

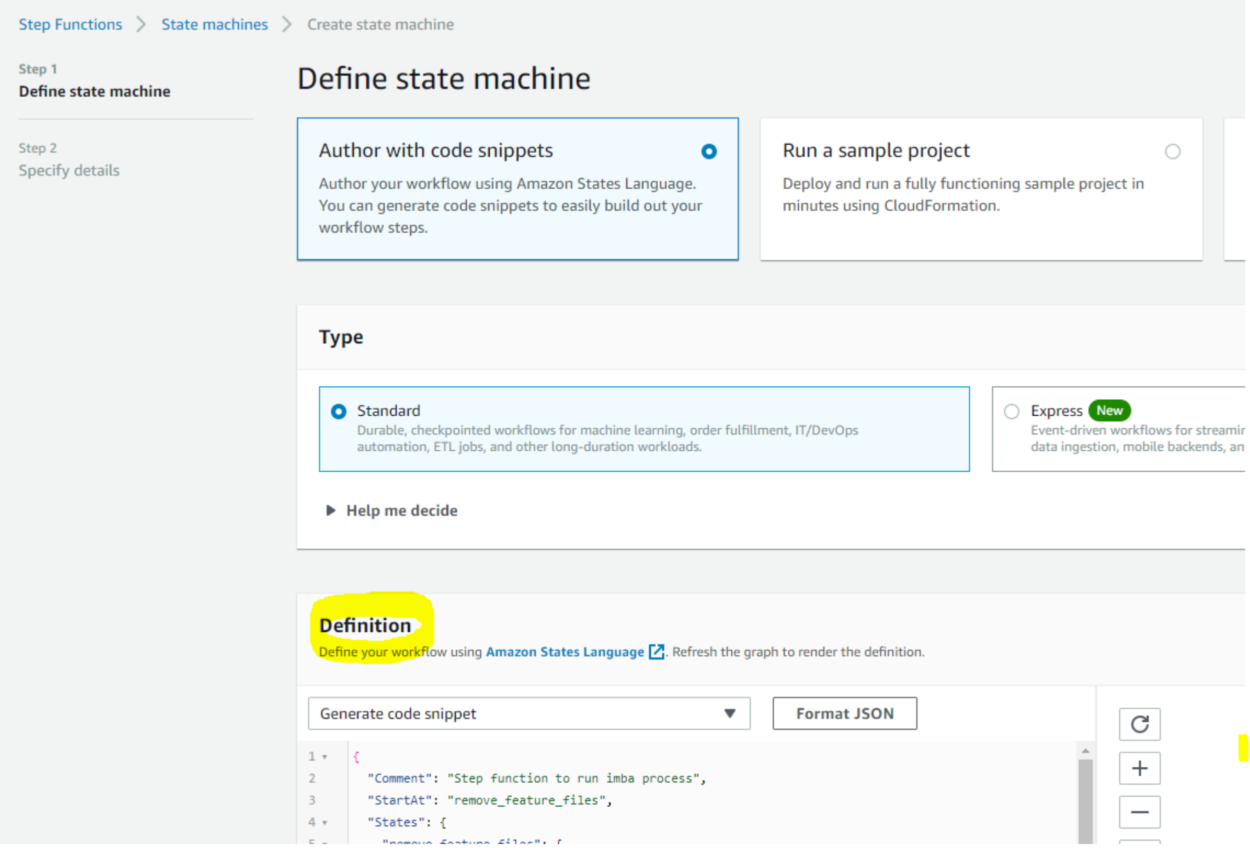
1. Click Step function on AWS console, click State machines on the left side pane, see below:



2. Click Create state machine on the top right pane:



3. Leave the setting as default, in Definition section(marked as yellow below), copy and paste json below:



Define state machine

● Author with code snippets
Author your workflow using Amazon States Language. You can generate code snippets to easily build out your workflow steps.

○ Run a sample project
Deploy and run a fully functioning sample project in minutes using CloudFormation.

Type

● Standard
Durable, checkpointed workflows for machine learning, order fulfillment, IT/DevOps automation, ETL jobs, and other long-duration workloads.

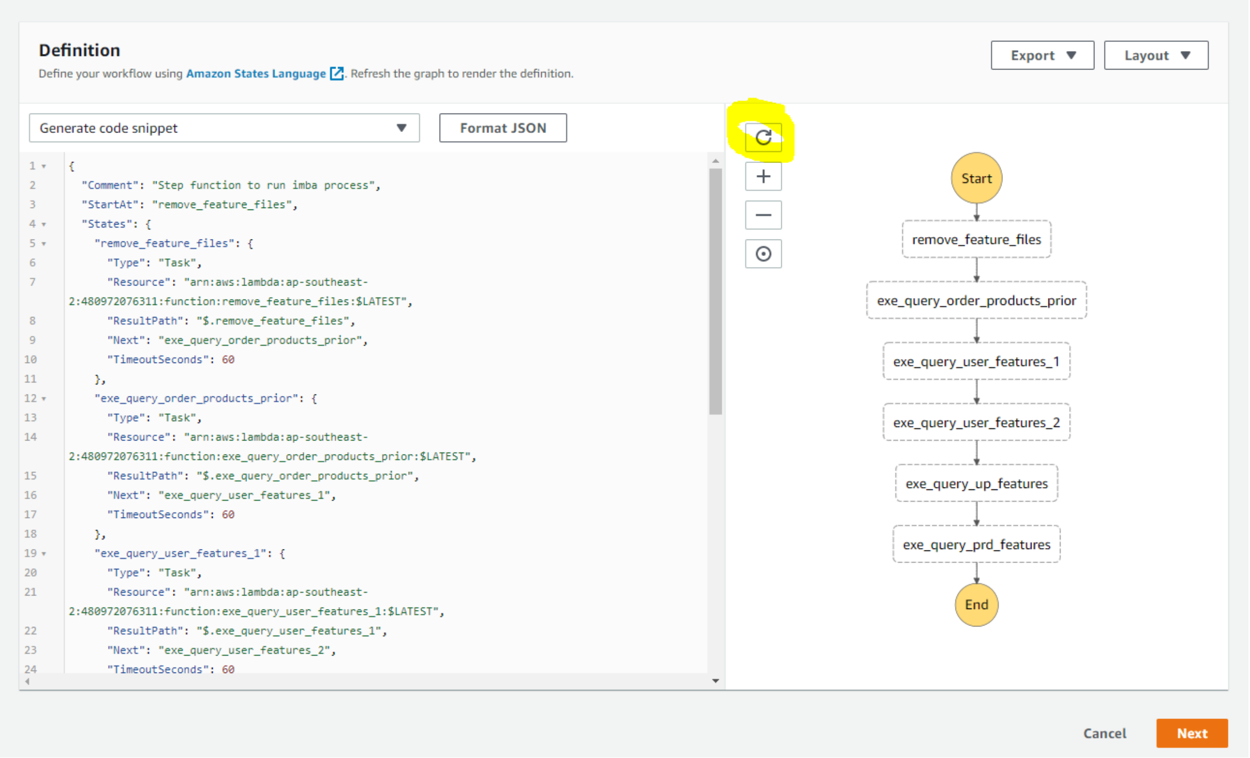
○ Express New
Event-driven workflows for streamir
data ingestion, mobile backends, an

Definition
Define your workflow using Amazon States Language. Refresh the graph to render the definition.

```
{
  "Comment": "Step function to run imba process",
  "StartAt": "remove_feature_files",
  "States": {
```

```
{
 "Comment": "Step function to run imba process",
 "StartAt": "remove_feature_files",
 "States": {
  "remove_feature_files": {
   "Type": "Task",
   "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:remove_feature_files:$LATEST",
   "ResultPath": "$.remove_feature_files",
   "Next": "exe_query_order_products_prior",
   "TimeoutSeconds": 60
  },
  "exe_query_order_products_prior": {
   "Type": "Task",
   "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:exe_query_order_products_prior:$LATEST",
   "ResultPath": "$.exe_query_order_products_prior",
   "Next": "exe_query_user_features_1",
   "TimeoutSeconds": 60
  },
  "exe_query_user_features_1": {
   "Type": "Task",
   "Resource": "arn:aws:lambda:ap-southeast-
2:480972076311:function:exe_query_user_features_1:$LATEST",
   "ResultPath": "$.exe_query_user_features_1",
   "Next": "exe_query_user_features_2",
   "TimeoutSeconds": 60
  },
  "exe_query_user_features_2": {
   "Type": "Task",
   "Resource": "arn:aws:lambda:ap-southeast-
2:480972076311:function:exe_query_user_features_2:$LATEST",
   "ResultPath": "$.exe_query_user_features_2",
   "Next": "exe_query_up_features",
   "TimeoutSeconds": 60
  },
  "exe_query_up_features": {
   "Type": "Task",
   "Resource": "arn:aws:lambda:ap-southeast-
2:480972076311:function:exe_query_up_features:$LATEST",
   "ResultPath": "$.exe_query_up_features",
   "Next": "exe_query_prd_features",
   "TimeoutSeconds": 60
  },
  "exe_query_prd_features": {
   "Type": "Task",
   "Resource": "arn:aws:lambda:ap-southeast-
2:480972076311:function:exe_query_prd_features:$LATEST",
   "ResultPath": "$.exe_query_up_features",
   "TimeoutSeconds": 60,
   "End": true
  }
 }
}
```

4. Click the refresh button(marked as yellow below), you should see the execution graph:

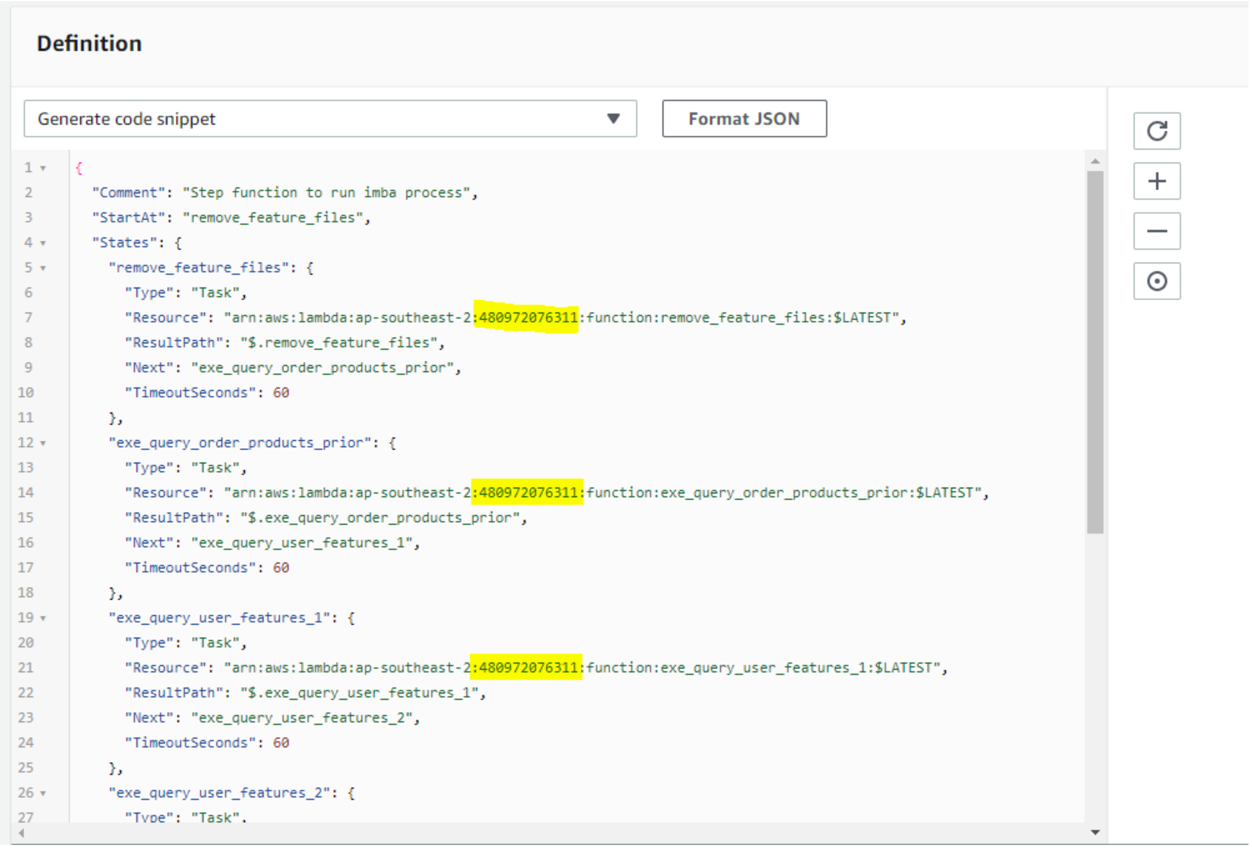


Definition

Define your workflow using Amazon States Language. Refresh the graph to render the definition.

```
{
  "Comment": "Step function to run imba process",
  "StartAt": "remove_feature_files",
  "States": {
    "remove_feature_files": {
      "Type": "Task",
      "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:remove_feature_files:$LATEST",
      "ResultPath": "$.remove_feature_files",
      "Next": "exe_query_order_products_prior",
      "TimeoutSeconds": 60
    },
    "exe_query_order_products_prior": {
      "Type": "Task",
      "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:exe_query_order_products_prior:$LATEST",
      "ResultPath": "$.exe_query_order_products_prior",
      "Next": "exe_query_user_features_1",
      "TimeoutSeconds": 60
    },
    "exe_query_user_features_1": {
      "Type": "Task",
      "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:exe_query_user_features_1:$LATEST",
      "ResultPath": "$.exe_query_user_features_1",
      "Next": "exe_query_user_features_2",
      "TimeoutSeconds": 60
```

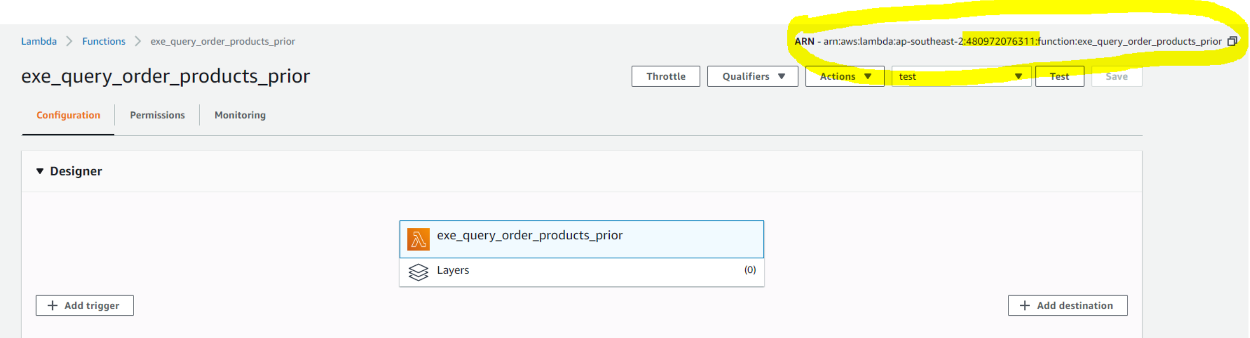
5. In the json object, you need to replace all the account_id in the five lambda ARN to your account id, see below:



Definition

```
{
  "Comment": "Step function to run imba process",
  "StartAt": "remove_feature_files",
  "States": {
    "remove_feature_files": {
      "Type": "Task",
      "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:remove_feature_files:$LATEST",
      "ResultPath": "$.remove_feature_files",
      "Next": "exe_query_order_products_prior",
      "TimeoutSeconds": 60
    },
    "exe_query_order_products_prior": {
      "Type": "Task",
      "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:exe_query_order_products_prior:$LATEST",
      "ResultPath": "$.exe_query_order_products_prior",
      "Next": "exe_query_user_features_1",
      "TimeoutSeconds": 60
    },
    "exe_query_user_features_1": {
      "Type": "Task",
      "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:exe_query_user_features_1:$LATEST",
      "ResultPath": "$.exe_query_user_features_1",
      "Next": "exe_query_user_features_2",
      "TimeoutSeconds": 60
    },
    "exe_query_user_features_2": {
      "Type": "Task",
```

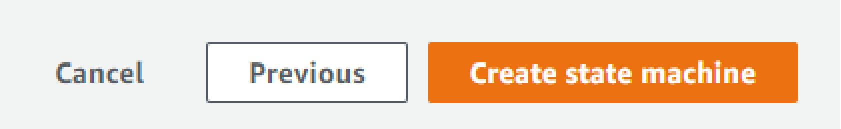
The account id or the lambda arn is located in your lambda function details, see below for one example:



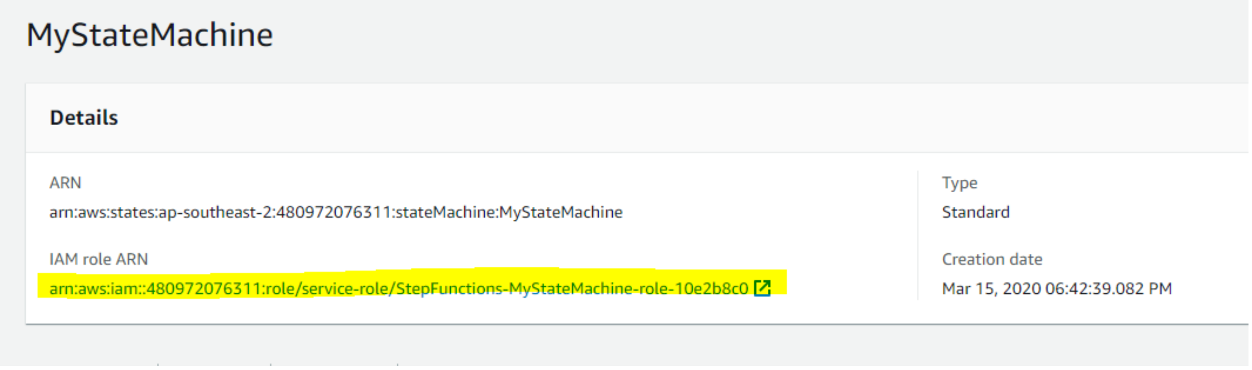
exe_query_order_products_prior

ARN - arn:aws:lambda:ap-southeast-2:480972076311:function:exe_query_order_products_prior

6. Click next, give your state machine a name and then click Create state machine in the bottom:



7. Give your step function the permission to invoke lambda function by clicking the IAM role ARN:



MyStateMachine

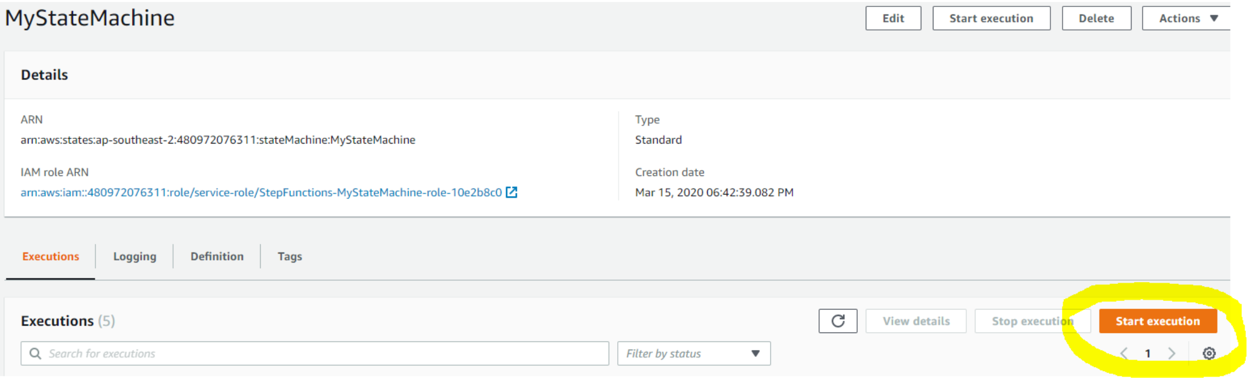
ARN
arn:aws:states:ap-southeast-2:480972076311:stateMachine:MyStateMachine

IAM role ARN
arn:aws:iam::480972076311:role/service-role/StepFunctions-MyStateMachine-role-10e2b8c0

Type
Standard

Creation date
Mar 15, 2020 06:42:39.082 PM

8. Attach AWSLambdaFullAcess permission to this role.
9. Go back to the state machine you just created, click Start execution:



MyStateMachine

ARN
arn:aws:states:ap-southeast-2:480972076311:stateMachine:MyStateMachine

IAM role ARN
arn:aws:iam::480972076311:role/service-role/StepFunctions-MyStateMachine-role-10e2b8c0

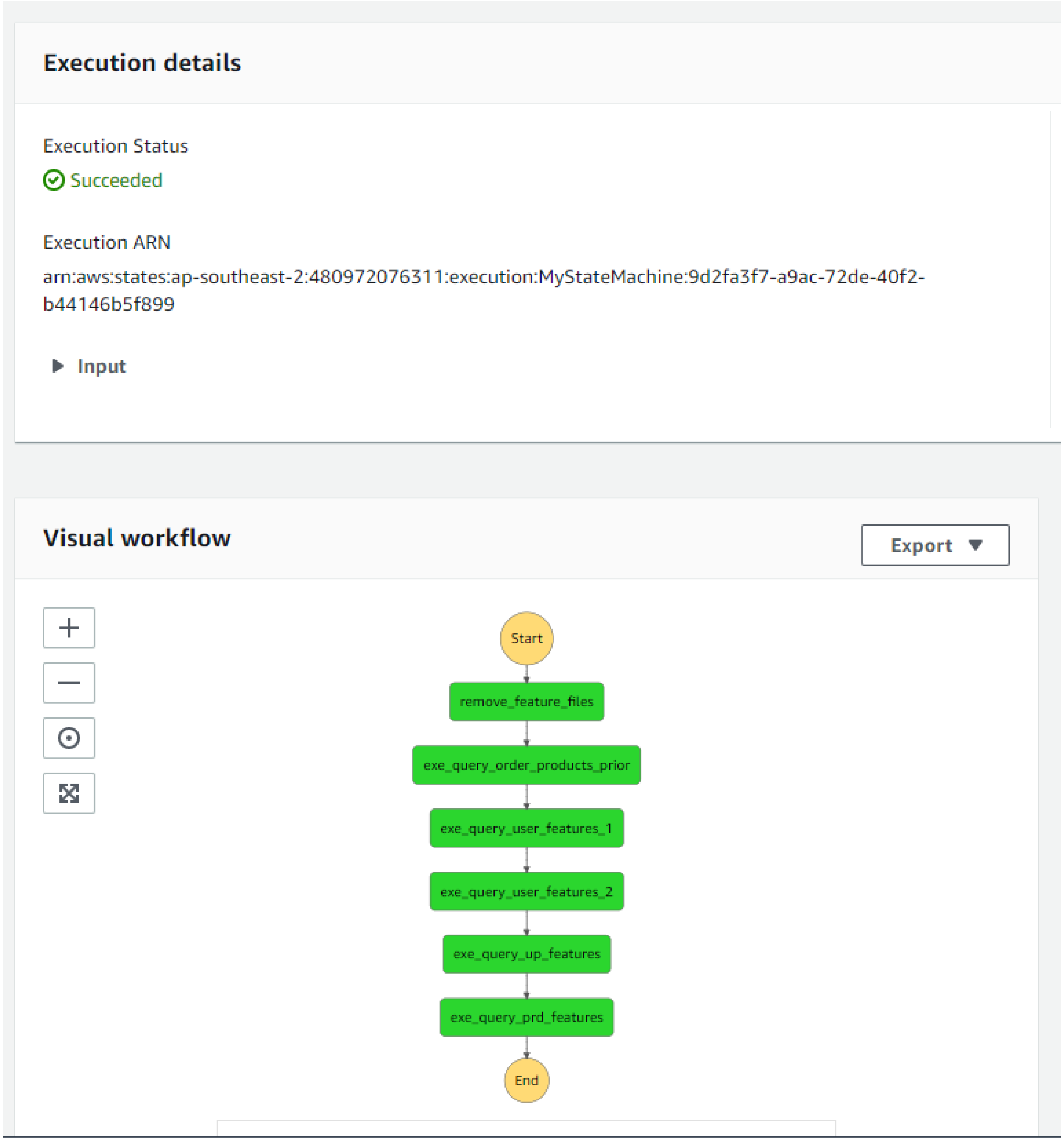
Type
Standard

Creation date
Mar 15, 2020 06:42:39.082 PM

Put below json as input and click Start execution:

```
{
 "bucket": "<your s3 bucket>",
 "prefix": "features/",
 "database": "prd",
 "query_output": "s3://<your s3 bucket>/query_results/"
}
```

10. You should see result similar to below if everything is configured correctly:



Execution details

Execution Status

✓ Succeeded

Execution ARN

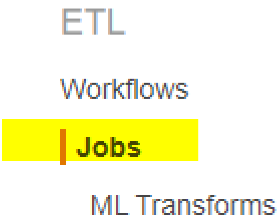
arn:aws:states:ap-southeast-2:480972076311:execution:MyStateMachine:9d2fa3f7-a9ac-72de-40f2-b44146b5f899

▶ Input

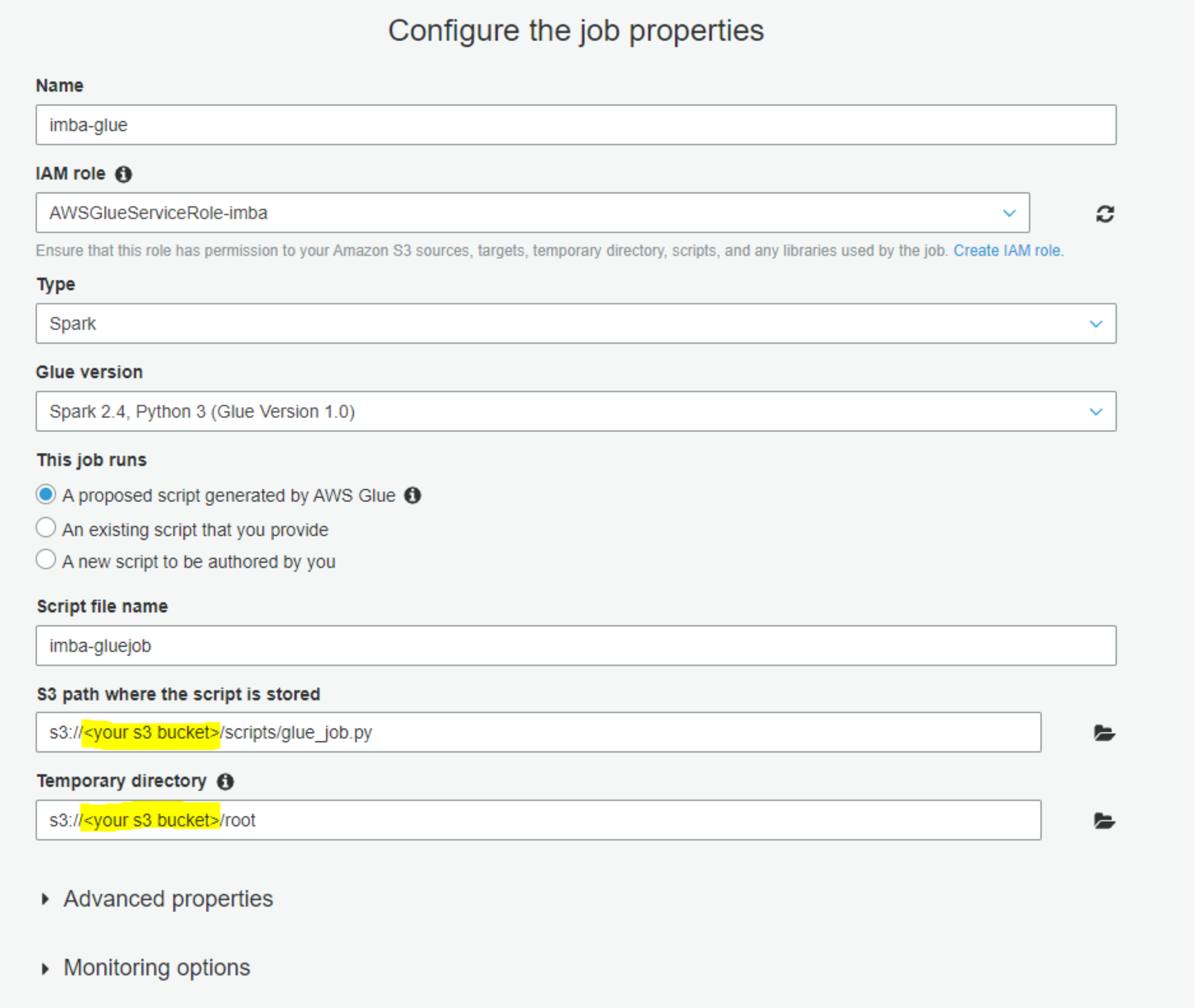
Visual workflow

AWS Glue

1. Open glue_job.py, change <your s3 bucket> to your s3 bucket name, save this file and upload it to s3://<your s3 bucket>/scripts/.
2. Open Glue console in AWS, click Jobs on the left pane:



3. Click Add job and you should fill in the details similar to below, name the job to "imba-glue", create a new IAM role or re-use an existing one (you just need to make sure AmazonS3FullAccess and AWSGlueServiceRole is attached). Make sure you select "An existing script that you provide" for "this job runs". Specify the s3 path where your script is stored: s3://<your s3 bucket>/scripts/glue_job.py and Temporary directory: s3://<your s3 bucket>/root. Leave everything else as default and click next.



Configure the job properties

Name

imba-glue

IAM role ℹ

AWSGlueServiceRole-imba

Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job. Create IAM role.

Type

Spark

Glue version

Spark 2.4, Python 3 (Glue Version 1.0)

This job runs

● A proposed script generated by AWS Glue ℹ
○ An existing script that you provide
○ A new script to be authored by you

Script file name

imba-gluejob

S3 path where the script is stored

s3://<your s3 bucket>/scripts/glue_job.py

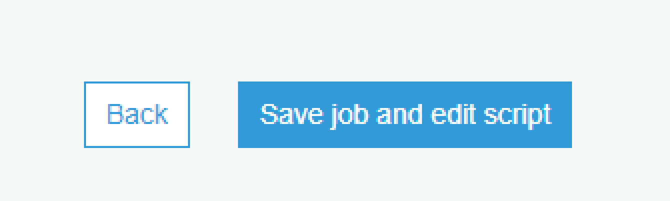
Temporary directory ℹ

s3://<your s3 bucket>/root

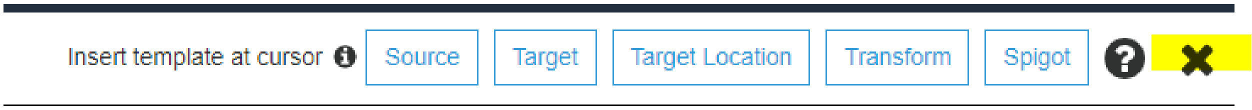
▸ Advanced properties

▸ Monitoring options

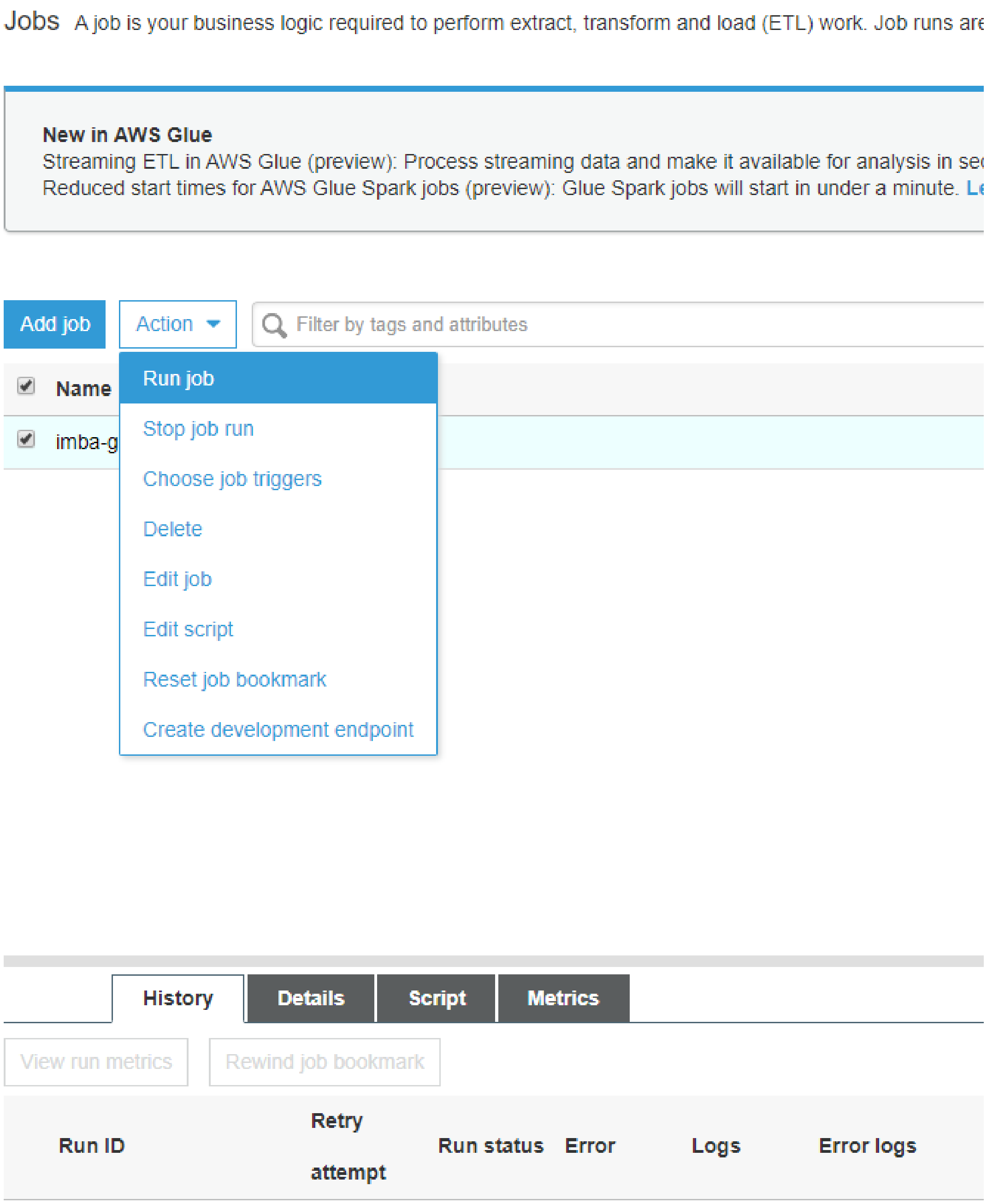
4. Click Save job and edit script:



5. Have a look at the script and close it by clicking the top right X button:



6. Select the job you created and click Run job from Action drop down menu:



Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are

New in AWS Glue

Streaming ETL in AWS Glue (preview): Process streaming data and make it available for analysis in sec

Reduced start times for AWS Glue Spark jobs (preview): Glue Spark jobs will start in under a minute. Le

Run ID	Retry attempt	Run status	Error	Logs	Error logs

7. Be patient, it should take around 10 minutes or so to finish, once done you should see an csv file is created in s3://<your s3 bucket>/output/part-xxxxxxxxxxx.csv
8. Download this file to your local desktop and rename it as “data.csv”