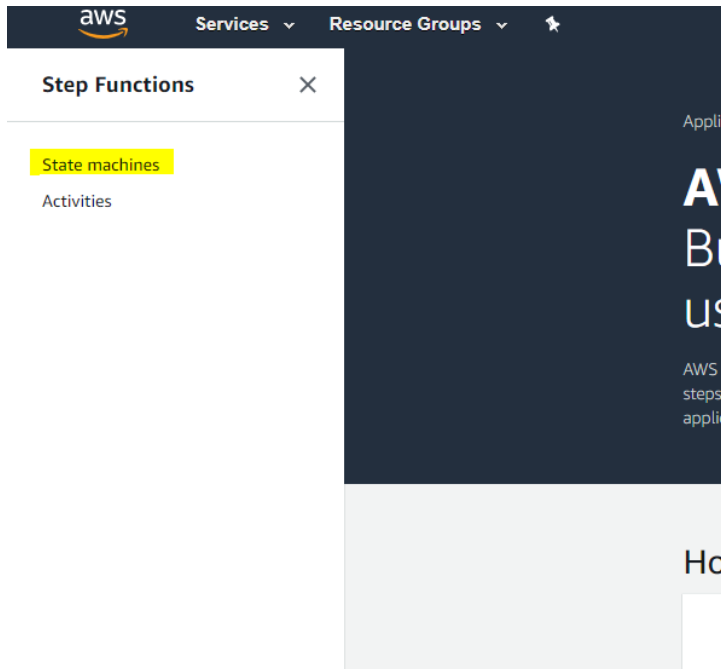
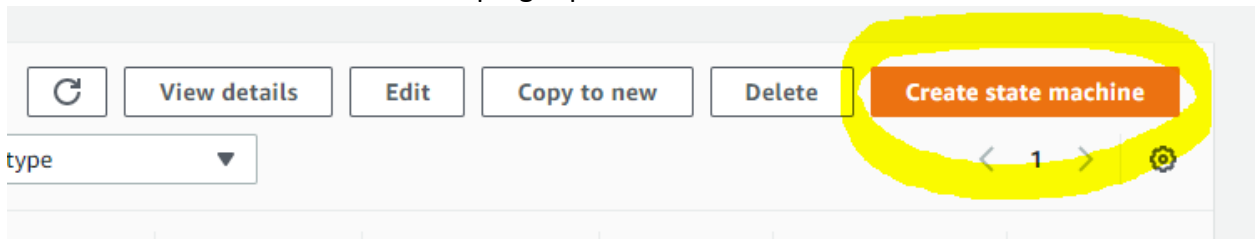


AWS step function

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:

Step Functions > State machines > Create state machine

Step 1
Define state machine

Step 2
Specify details

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 streaming data ingestion, mobile backends, and more.

► Help me decide

Definition

Define your workflow using [Amazon States Language](#). Refresh the graph to render the definition.

Generate code snippet ▼

Format JSON

```

1 {
2   "Comment": "Step function to run imba process",
3   "StartAt": "remove_feature_files",
4   "States": {
5     "remove_feature_files": {
6       "Type": "Task",
7       "Resource": "arn:aws:lambda:ap-southeast-

```

```

{
  "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.

Export

Layout

Generate code snippet

Format JSON



+

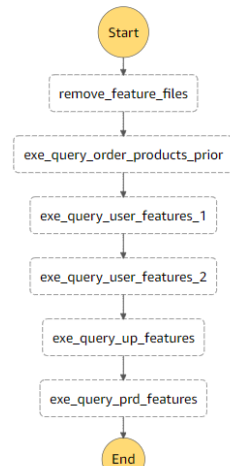
-

⌂

```

1 {
2   "Comment": "Step function to run imba process",
3   "StartAt": "remove_feature_files",
4   "States": {
5     "remove_feature_files": {
6       "Type": "Task",
7       "Resource": "arn:aws:lambda:ap-southeast-
2:480972076311:function:remove_feature_files:$LATEST",
8       "ResultPath": "$.remove_feature_files",
9       "Next": "exe_query_order_products_prior",
10      "TimeoutSeconds": 60
11    },
12    "exe_query_order_products_prior": {
13      "Type": "Task",
14      "Resource": "arn:aws:lambda:ap-southeast-
2:480972076311:function:exe_query_order_products_prior:$LATEST",
15      "ResultPath": "$.exe_query_order_products_prior",
16      "Next": "exe_query_user_features_1",
17      "TimeoutSeconds": 60
18    },
19    "exe_query_user_features_1": {
20      "Type": "Task",
21      "Resource": "arn:aws:lambda:ap-southeast-
2:480972076311:function:exe_query_user_features_1:$LATEST",
22      "ResultPath": "$.exe_query_user_features_1",
23      "Next": "exe_query_user_features_2",
24      "TimeoutSeconds": 60
25    }
26  }
27 }

```



Cancel


Next

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

Definition

Generate code snippet

Format JSON



+

-

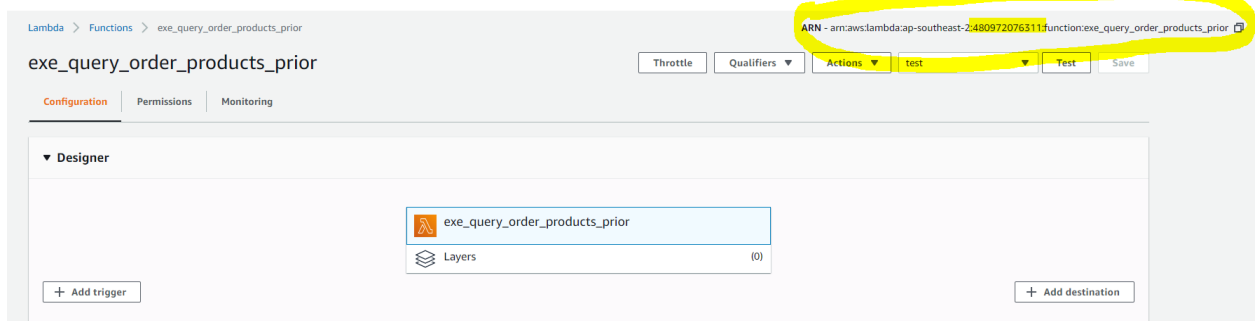
⌂

```

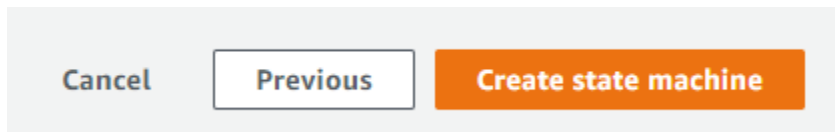
1 {
2   "Comment": "Step function to run imba process",
3   "StartAt": "remove_feature_files",
4   "States": {
5     "remove_feature_files": {
6       "Type": "Task",
7       "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:remove_feature_files:$LATEST",
8       "ResultPath": "$.remove_feature_files",
9       "Next": "exe_query_order_products_prior",
10      "TimeoutSeconds": 60
11    },
12    "exe_query_order_products_prior": {
13      "Type": "Task",
14      "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:exe_query_order_products_prior:$LATEST",
15      "ResultPath": "$.exe_query_order_products_prior",
16      "Next": "exe_query_user_features_1",
17      "TimeoutSeconds": 60
18    },
19    "exe_query_user_features_1": {
20      "Type": "Task",
21      "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:exe_query_user_features_1:$LATEST",
22      "ResultPath": "$.exe_query_user_features_1",
23      "Next": "exe_query_user_features_2",
24      "TimeoutSeconds": 60
25    },
26    "exe_query_user_features_2": {
27      "Type": "Task",

```

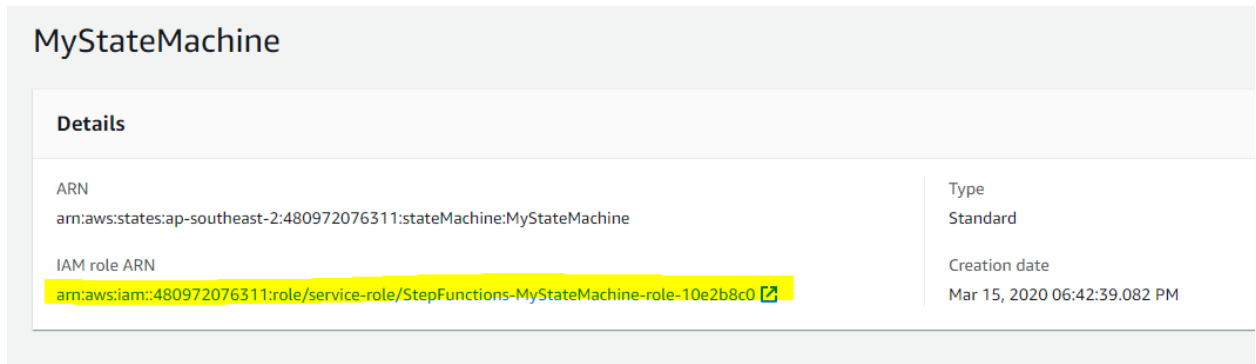
The account id or the lambda arn is located in your lambda function details, see below for one example:



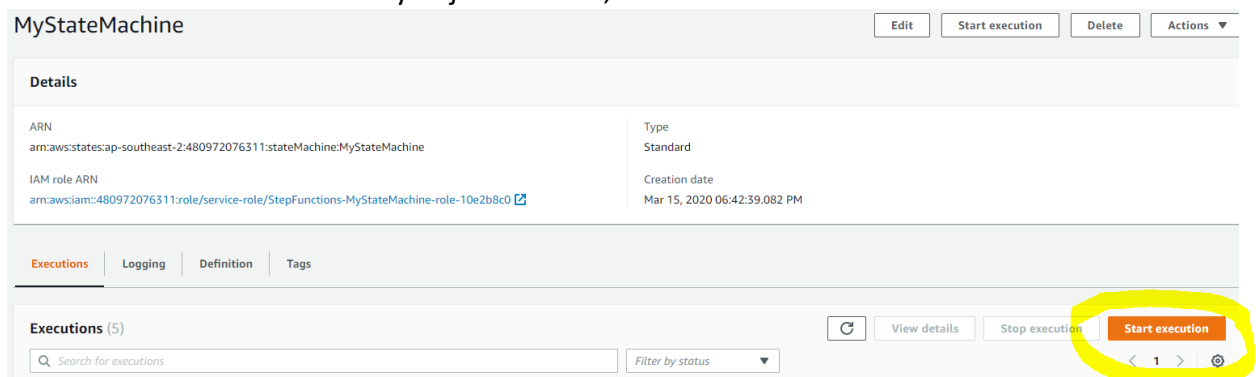
- Click next, give your state machine a name and then click Create state machine in the bottom:



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



- Attach AWSLambdaFullAccess permission to this role.
- Go back to the state machine you just created, click Start execution:

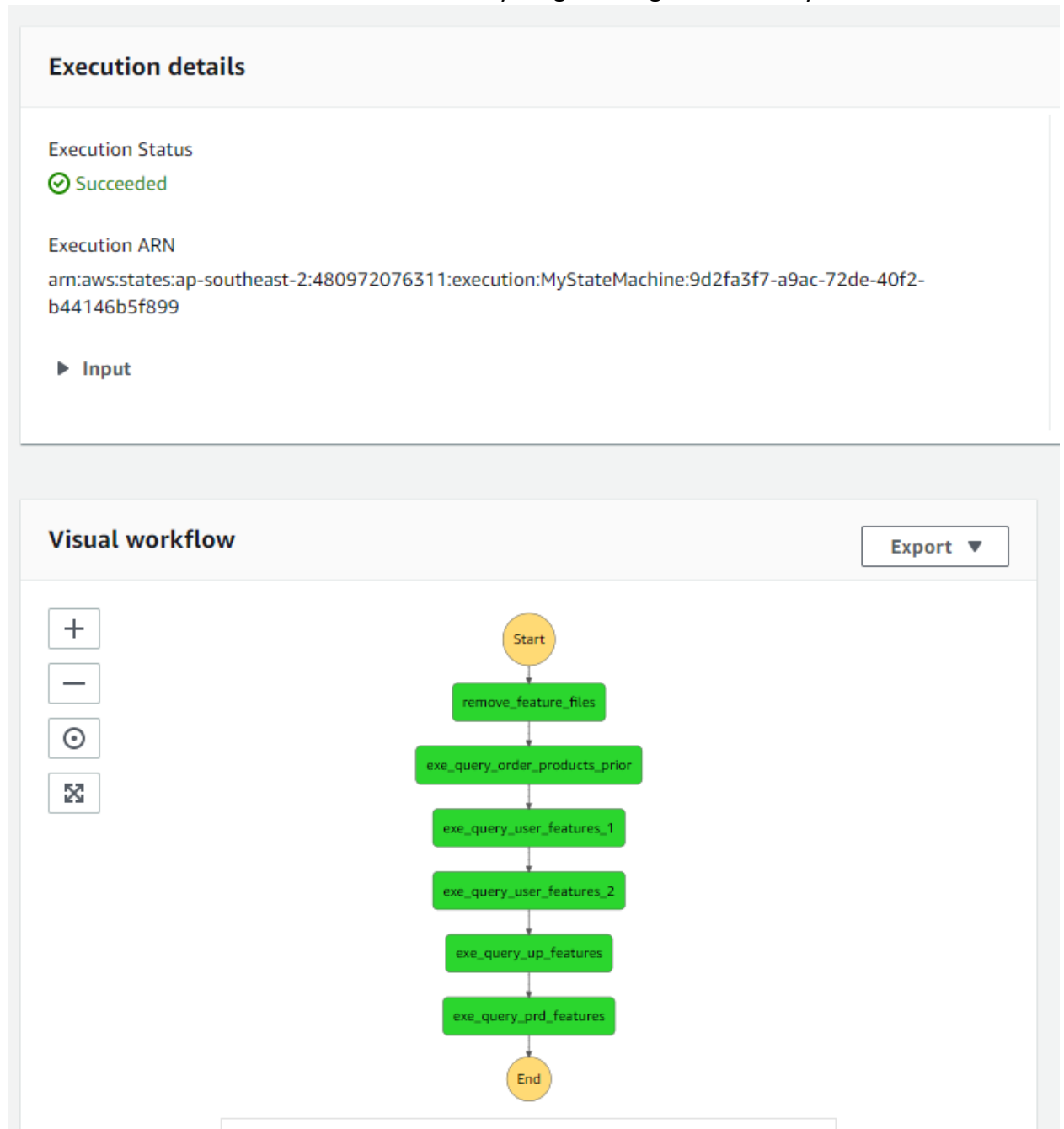


Put below json as input and click Start execution:

```
{
  "bucket": "<your red 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:



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

IAM role ⓘ

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


Glue version

This job runs


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

Script file name

S3 path where the script is stored

Temporary directory ⓘ

▶ Advanced properties
 ▶ Monitoring options

- Click Save job and edit script:

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

Insert template at cursor ⓘ

- 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 seconds.

Reduced start times for AWS Glue Spark jobs (preview): Glue Spark jobs will start in under a minute. [Learn more](#)

Add job

Action ▾

Filter by tags and attributes

<input checked="" type="checkbox"/>	Name	Run job
<input checked="" type="checkbox"/>	imba-g	Stop job run
		Choose job triggers
		Delete
		Edit job
		Edit script
		Reset job bookmark
		Create development endpoint

History

Details

Script

Metrics

View run metrics

Rewind job bookmark

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-xxxxxxxxxxxx.csv
- 8. Download this file to your local desktop and rename it as “data.csv”