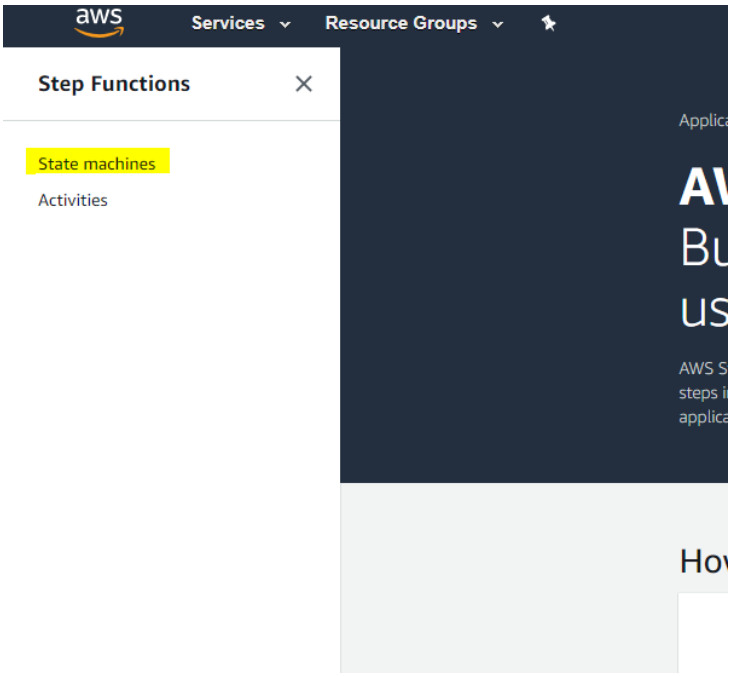
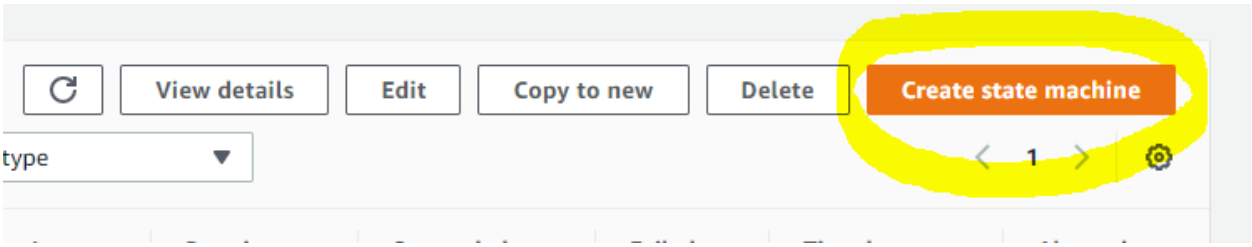
# AWS step function

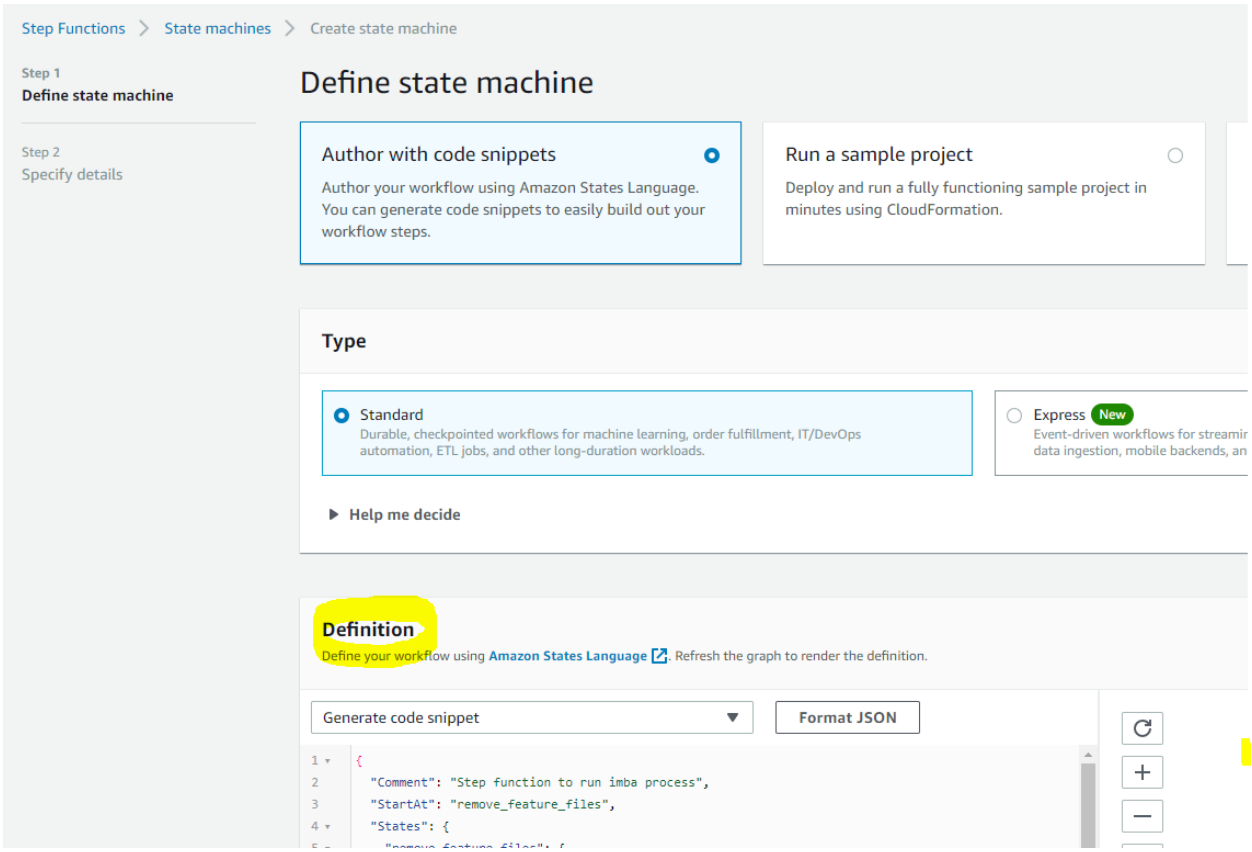
1. Click Step function on AWS console, click State machines on the left side pane, see below:




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:

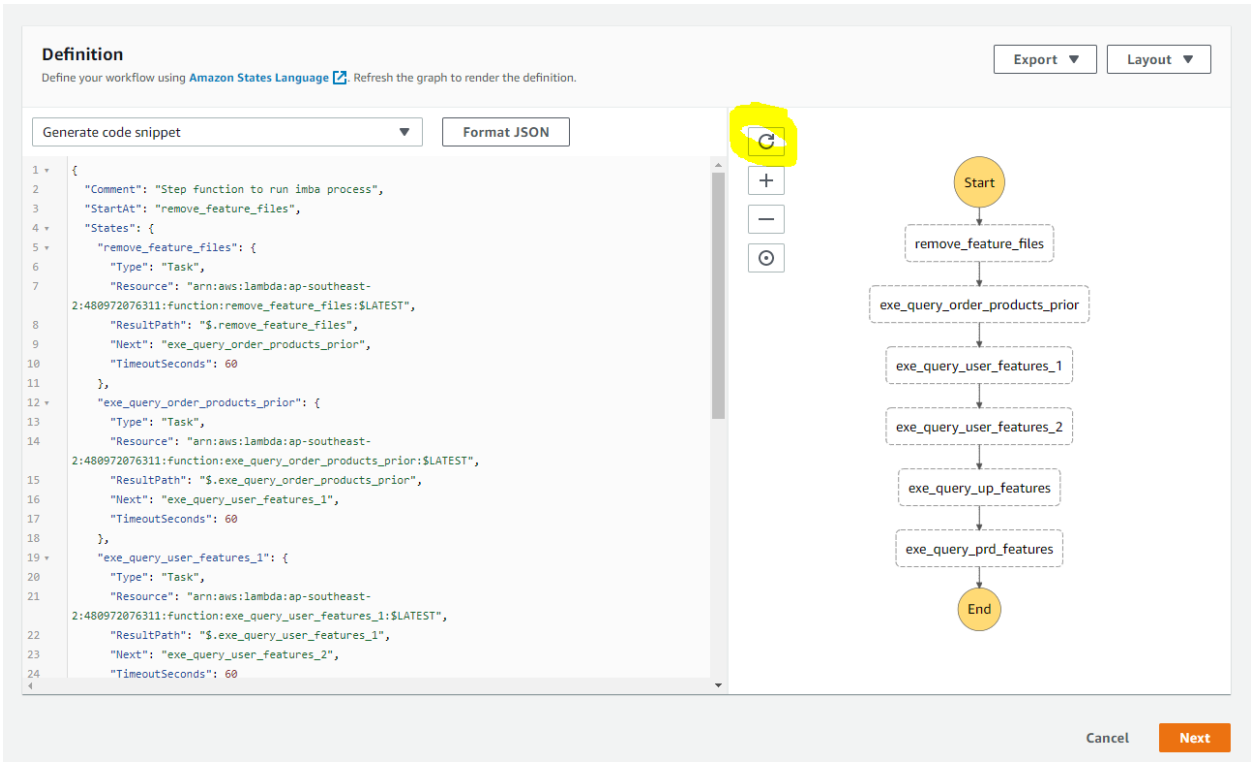




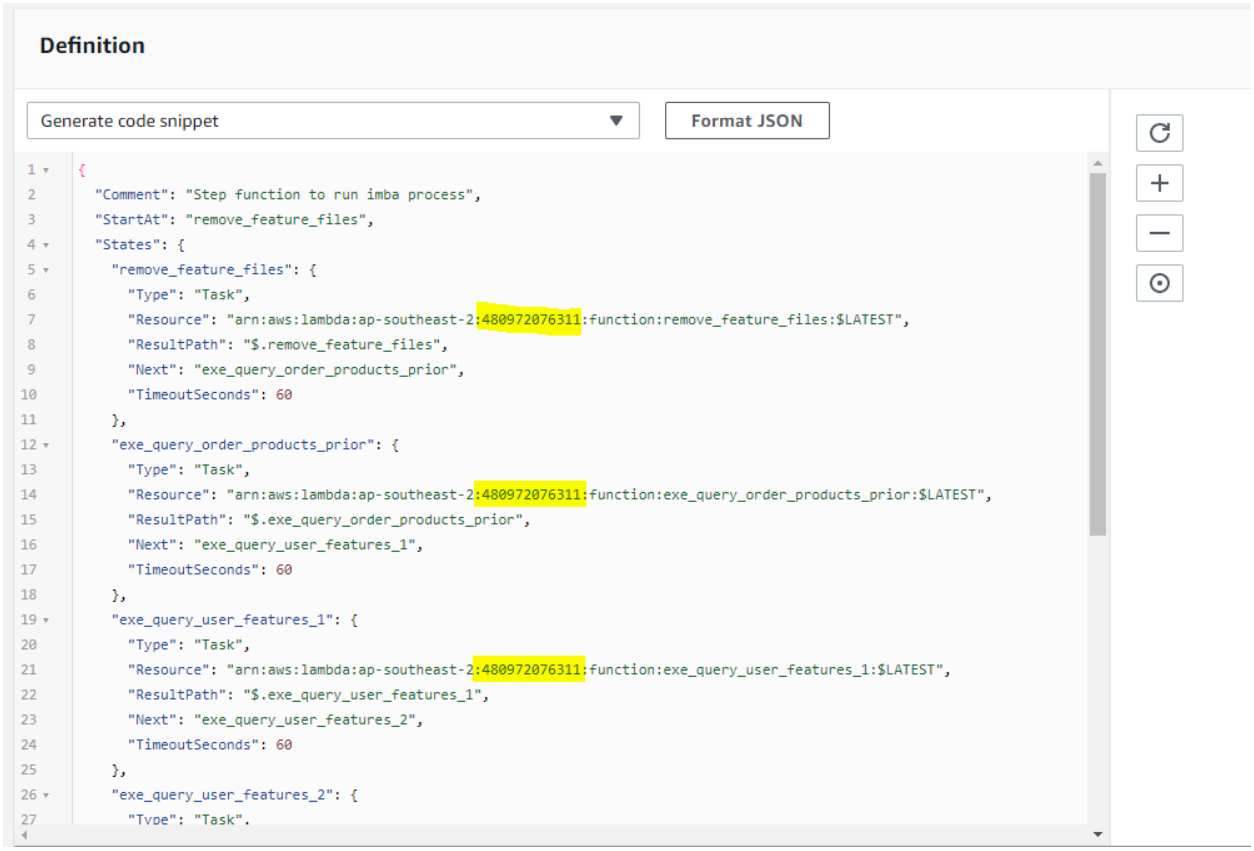
```
{
 "Comment": "Step function to run imba process",
 "StartAt": "remove_feature_files",
 "States": {
  "remove_feature_files": {
   "Type": "Task",
   "Resource": "arn:aws:lambda:ap-southeast-2:480972076311:function:remove_feature_files:$LATEST",
   "ResultPath": "$.remove_feature_files",
   "Next":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

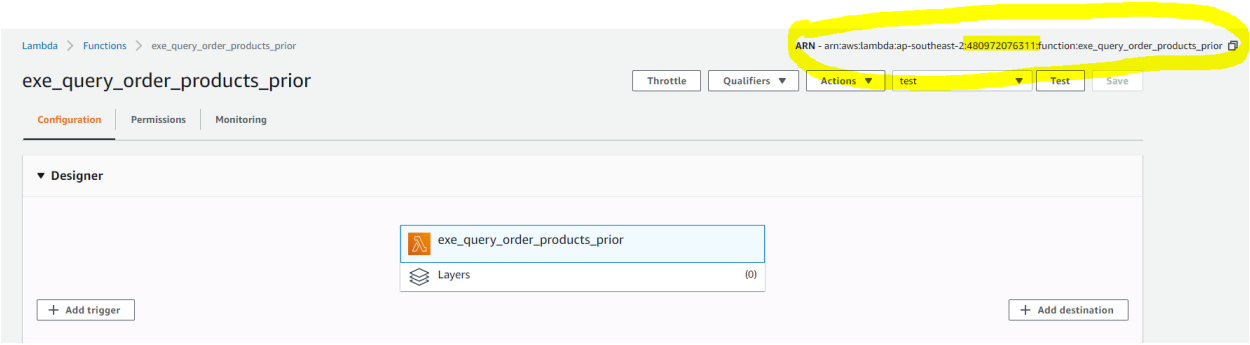
4. Click the refresh button(marked as yellow below), you should see the execution graph:



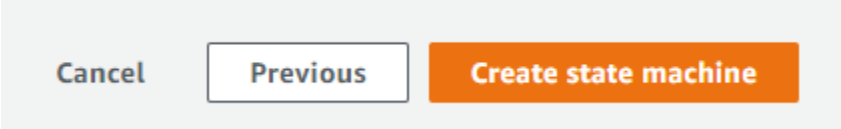

5. In the json object, you need to replace all the account_id in the five lambda ARN to your account id, see below:



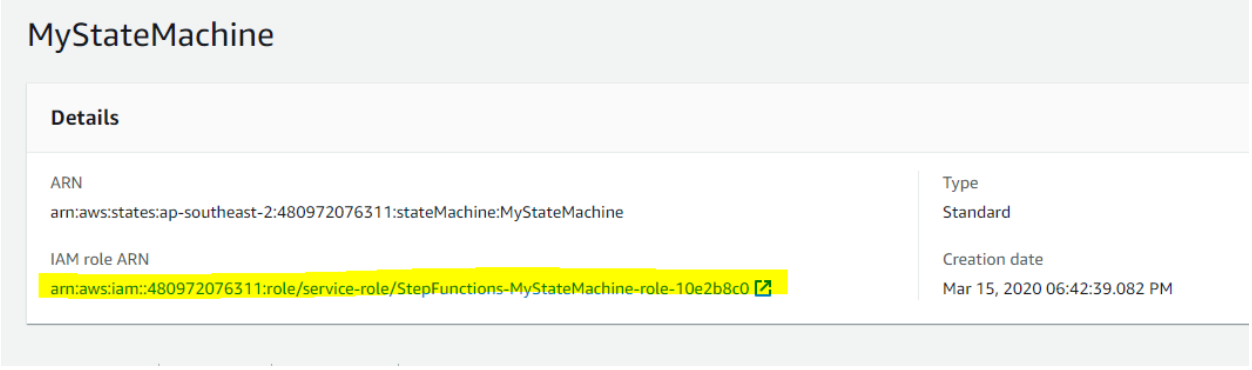
The account id or the lambda arn is located in your lambda function details, see below for one example:



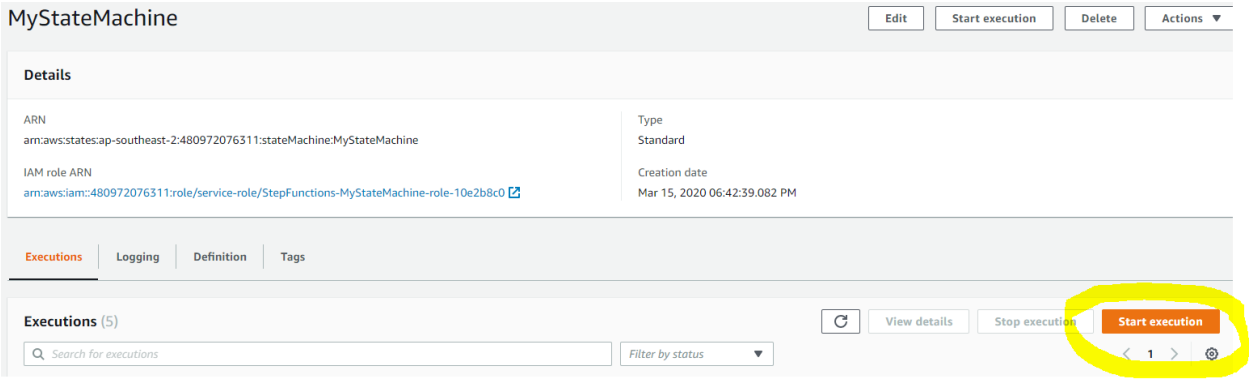

6. Click next, give your state machine a name and then click Create state machine in the bottom:




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




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

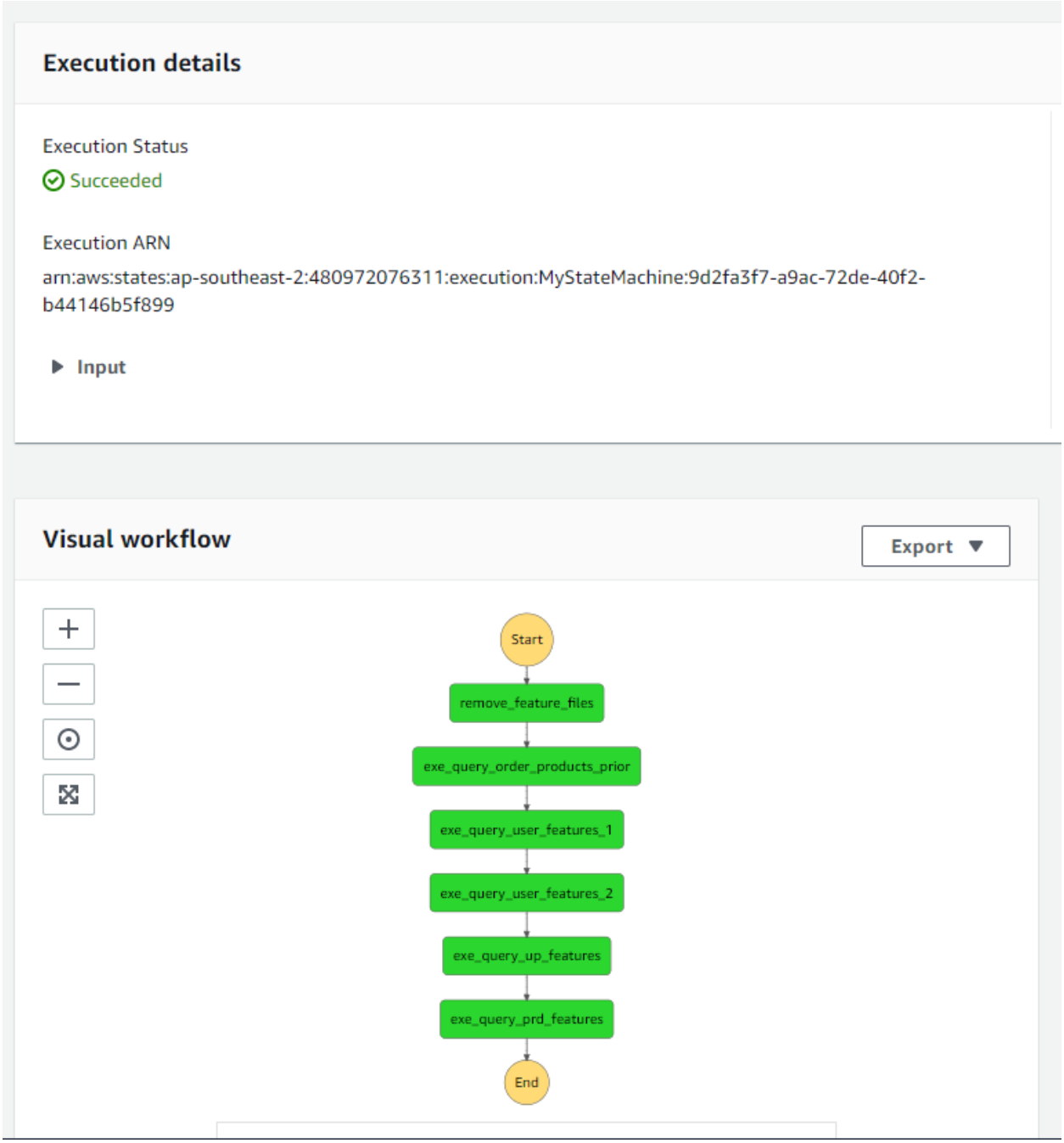



Put below json as input and click Start execution:

```
{
 "bucket": "<your s3 bucket>",
 "prefix": "features/",
```

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

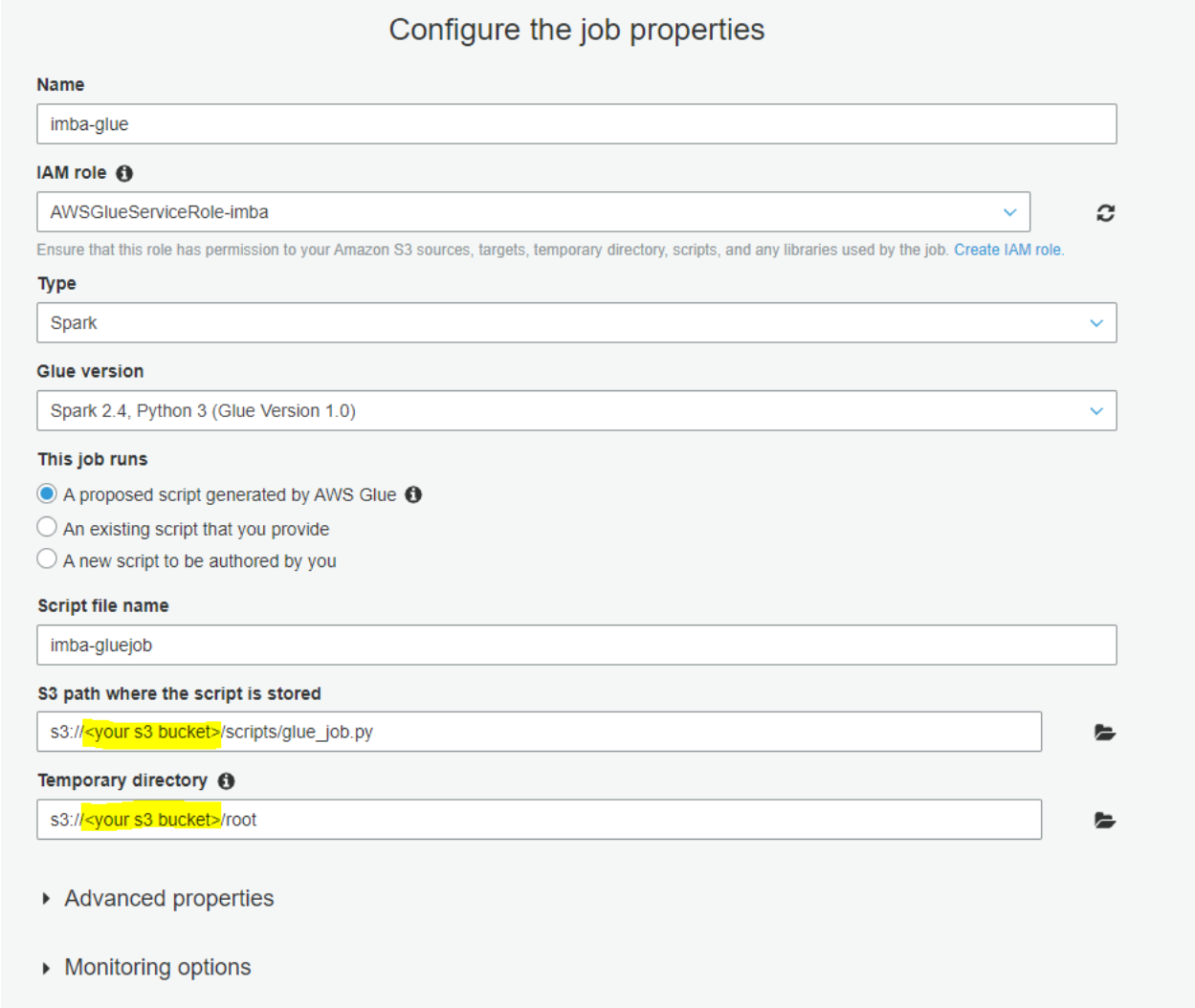
10. You should see result similar to below if everything is configured correctly:



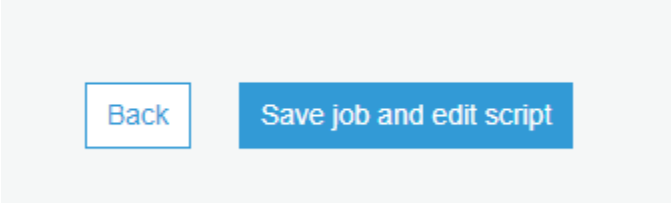
## AWS Glue

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:

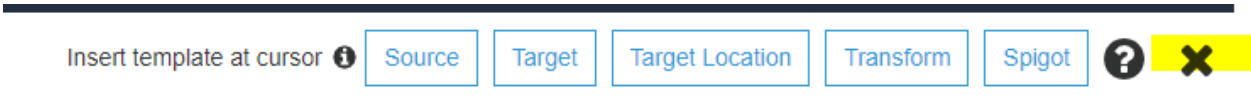

3. Click Add job and you should fill in the details similar to below, name the job to “imba-glue”, create a new IAM role or re-use an existing one (you just need to make sure AmazonS3FullAccess and AWSGlueServiceRole is attached). Make sure you select “An existing script that you provide” for “this job 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.



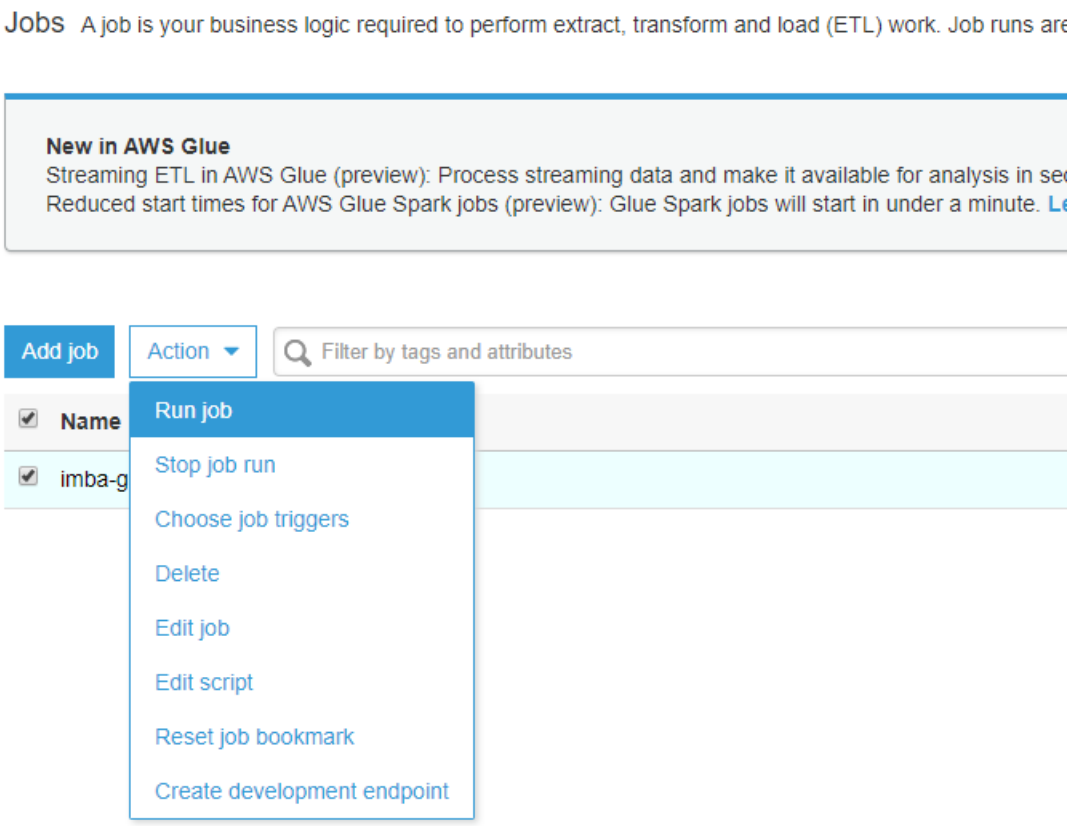

4. Click Save job and edit script:

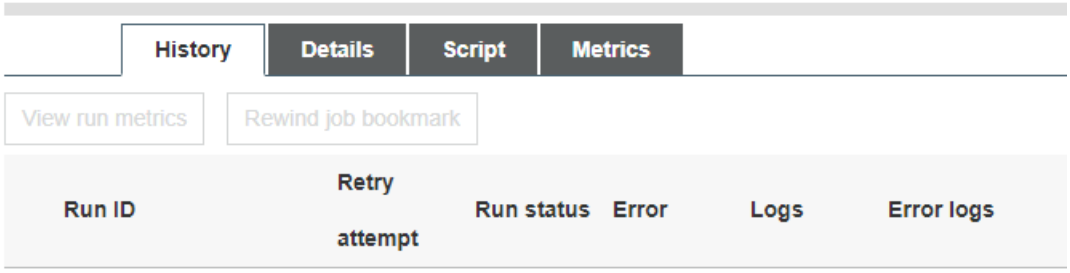



5. Have a look at the script and close it by clicking the top right X button:




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




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”