Amazon Managed Workflows for Apache Airflow（MWAA）ハンズオン

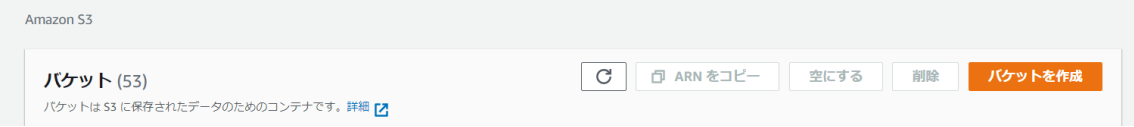
2021/02/19
シニアエバンジェリスト
亀田 治伸

1．アセットのダウンロード

以下のファイルをダウンロードしてください。
https://github.com/harunobukameda/Amazon-Managed-Workflows-for-Apache-Airflow-MWAA-/blob/main/1_cvlog.csv
https://github.com/harunobukameda/Amazon-Managed-Workflows-for-Apache-Airflow-MWAA-/blob/main/handson-athena-job.py

2．作業用 S3 バケットの生成

2-1. バケットの作成を押します。



2-2. 適当な長い名前を入力し、設定はデフォルトのまま[バケットを作成]ボタンを押して、バケットを作成します。

2-3. 作成されたバケットをクリックした画面で[フォルダの作成]を押します。

2-4. ”dags”, ”in0”, “out0”という 3 つのフォルダを作成します。

それぞれ以下の用途で後ほど利用します。

dags：DAG ファイル用

in0：入力 csv データ用

out0：出力 parquet データ用

3. MWAA の環境構築

3-1. MWAA の管理画面、左のペインから[環境]をクリックし、[環境を作成]ボタンを押します。

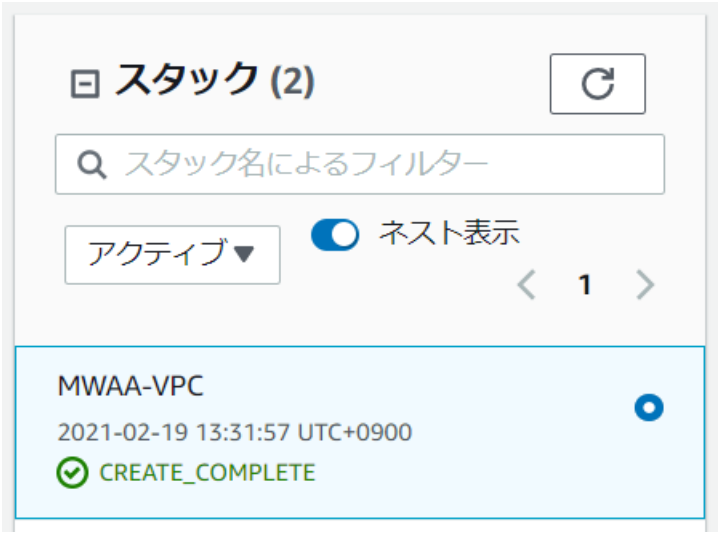
3-2. 名前とバージョンはデフォルトのまま、S3 バケットに先ほど作成したものを設定します

3-3. DAG フォルダに先ほど作成したフォルダ[dags]を含む URI を設定し、[次へ]を押します。

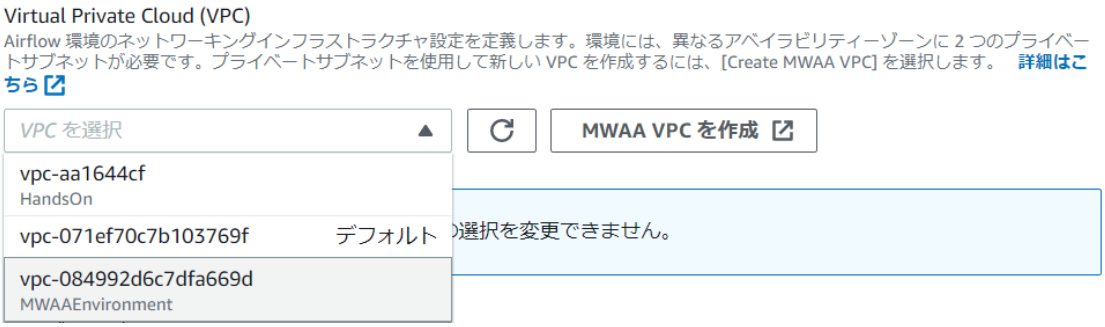
3-4. [MWAA VPC を作成]を押すと別タブで CloudFormation が起動します。デフォルトのまま[スタックの作成]を押してください。

3-5. 作成が完了するまでしばらく待ちます。

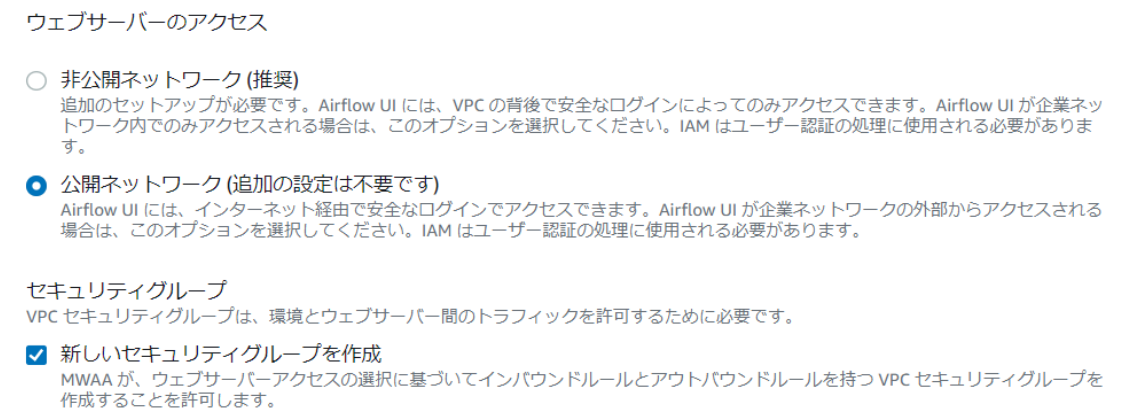
3-6. 以下のように Create_Complete となったら MWAA の画面に戻ります。



3-7. 作成された VPC を選びます。



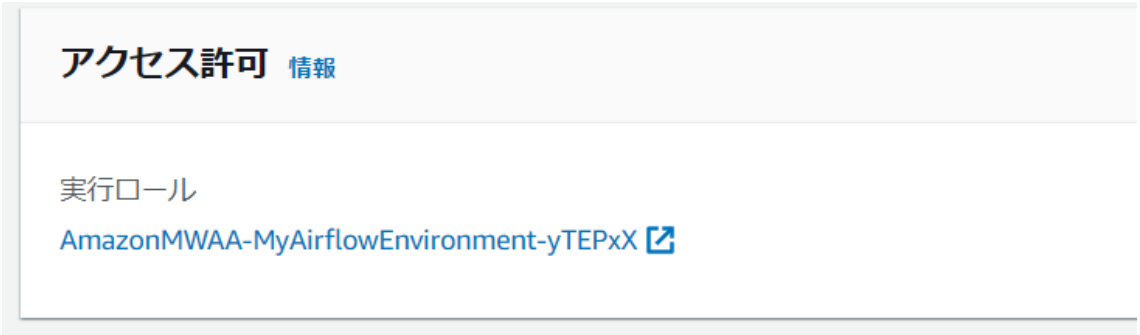
3-8. ネットワーク設定を公開に変更します



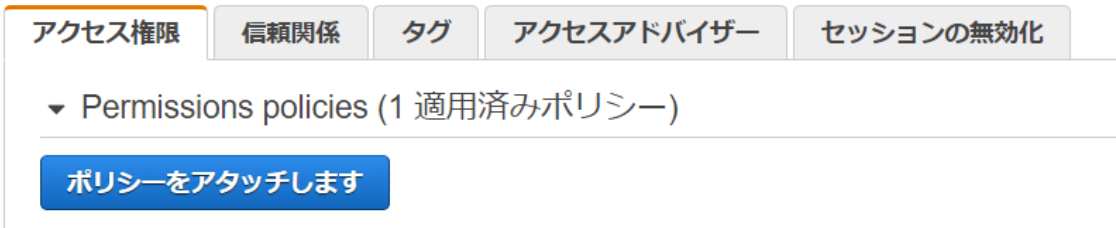
3-9. その他設定は全てデフォルトで[環境を作成]を押します

3-10. 20-30 分待つと以下のように環境作成が完了します

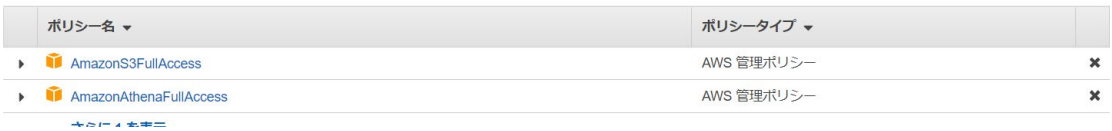
3-11. 出来上がった環境をクリックし、画面下部のアクセス許可で表示されている実行ロールをクリックします



3-12. [ポリシーをアタッチ]をクリックします



3-13. 以下 2 つのポリシーをアタッチします



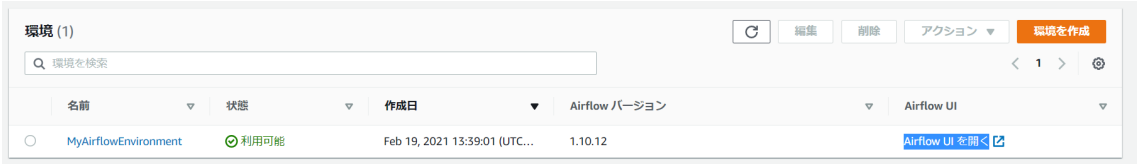
3-14. 先ほどダウンロードした[1_cvlog.csv]を S3 バケットの in0 にアップします。

3-15. 先ほどダウンロードした[handson-athena-job.py]を適当なエディタで開き、<your bucket name>の部分をバケット名で置換します。この時バケット名は”で囲みます

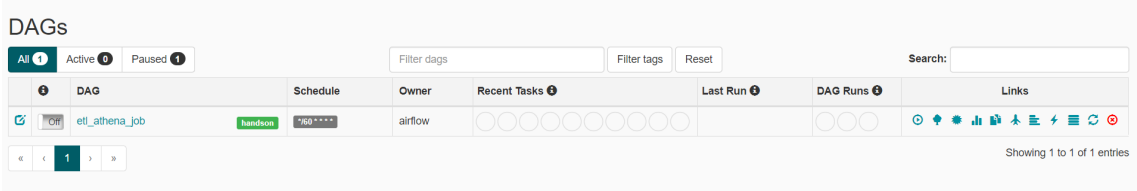
```
###############こちらにご自身のS3バケット名を入れてください################
s3_bucket_name = "airflowworkshop20210219hkameda"
################################################################################
athena_results = f"s3://{s3_bucket_name}/results/"
```

3-16. このファイルを今度は dags のフォルダにアップロードします

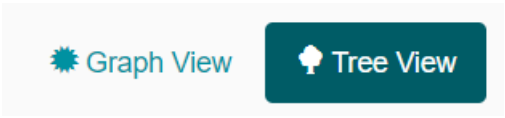
3-17. Airflow の画面から、[Airflow UI を開く]をクリックします

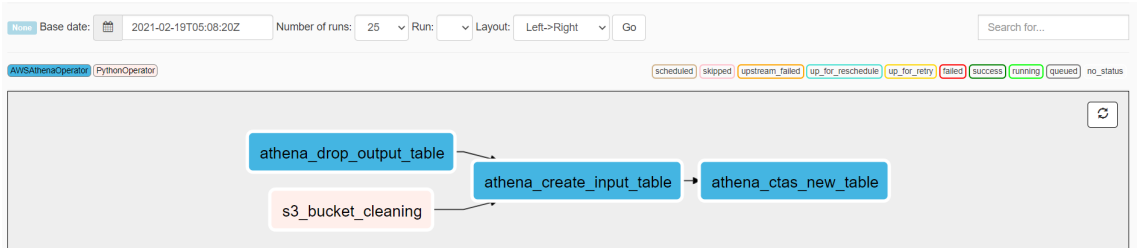


3-18. 正しく設定できていればすでに一つ DAG が作成されています。(S3 バケットの dags ファイルをもとに生成されている)



3-19. DAG をクリックして開きます。Tree View が表示されていますが、Graph View に切り替えます。

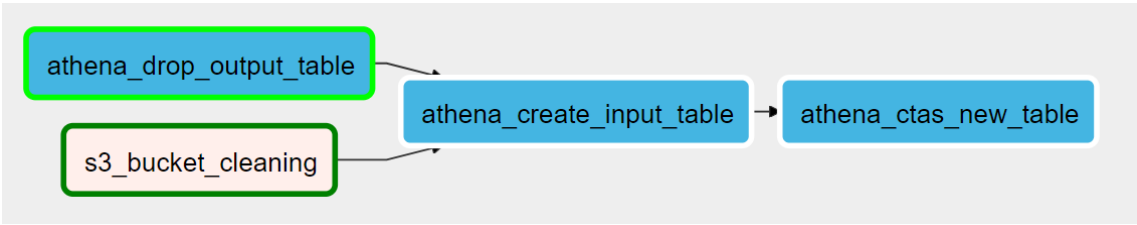


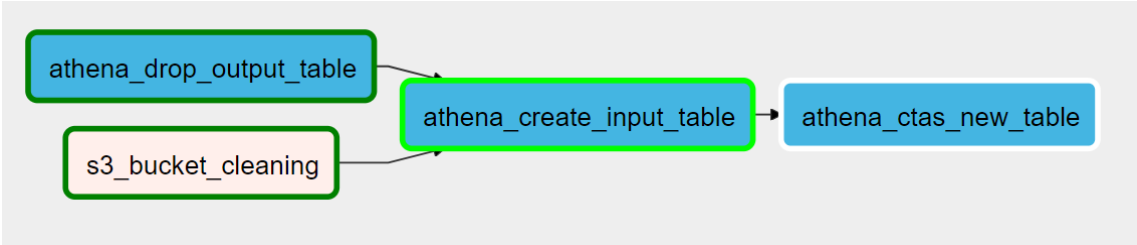


これが作業フローとなります。まず最初に、Athena のテーブルの削除（繰り返し実行用）と S3 バケットのアウトプットファイルの削除（繰り返し実行用）を行い、その後 S3 バケットのインプットファイルをもとに Athena でテーブルを作成します。

3-20. 左上の off スイッチをオンにすると実行が開始されます。

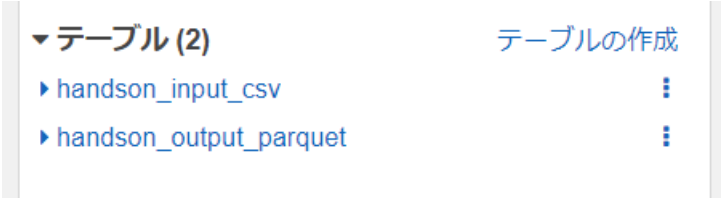
作業中が明るいみどり、作業完了が濃い緑で表示されます。





4. Athena でのデータ確認

4-1. 作業が成功すると Athena の default データベースで 2 つのテーブルが確認できます。



前者は、MWAA が S3 バケットから取り込んだ csv を Athena 経由でテーブル登録したもの、後者が DAG ワークフローで Parquet に変換されたものです。

4-2. それぞれテーブルのプレビューでデータが正しく格納されているか確認します。

4-3. S3 バケットの out0 フォルダにファイルが生成されていることを確認します。

5.お疲れ様でした！

以下の順番で片づけを行います。

1．S3 バケット
2．MWAA の環境
3．CloudFormaiton スタック