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

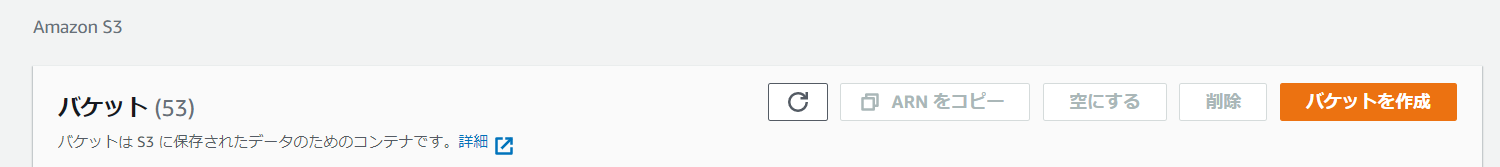
以下のファイルをダウンロードしてください。

<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>

1. 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。

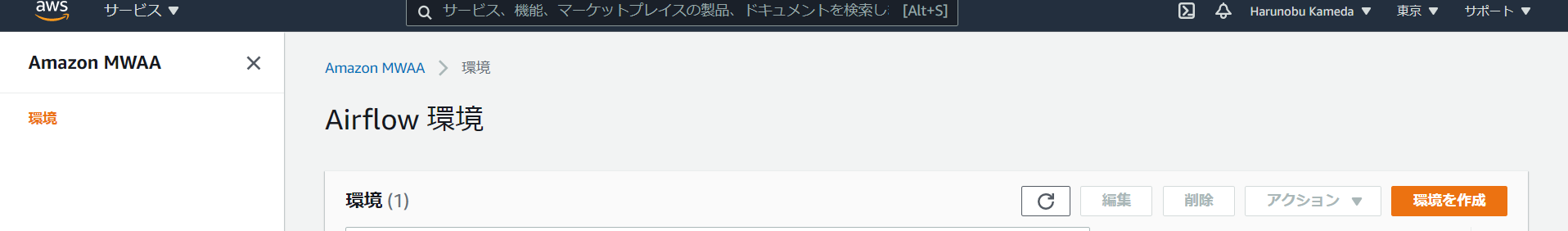
dags：DAGファイル用

in0：入力csvデータ用

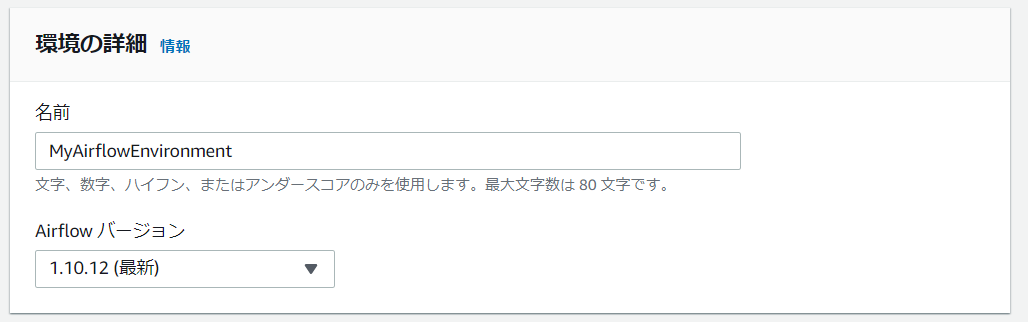
out0：出力parquetデータ用

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。

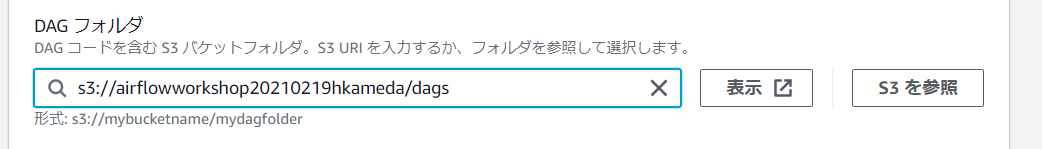


3-2. 名前とバージョンはデフォルトのまま、S3バケットに先ほど作成したものを設定します





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。



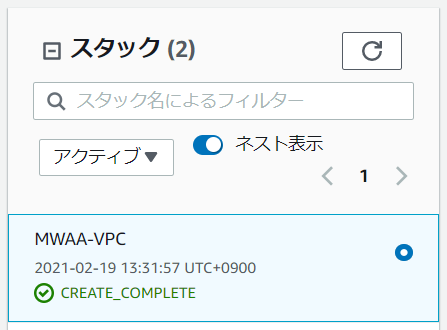
3-4．[MWAA VPCを作成]を押すと別タブでCloudFormationが起動します。デフォルトのまま[スタックの作成]を押してください。



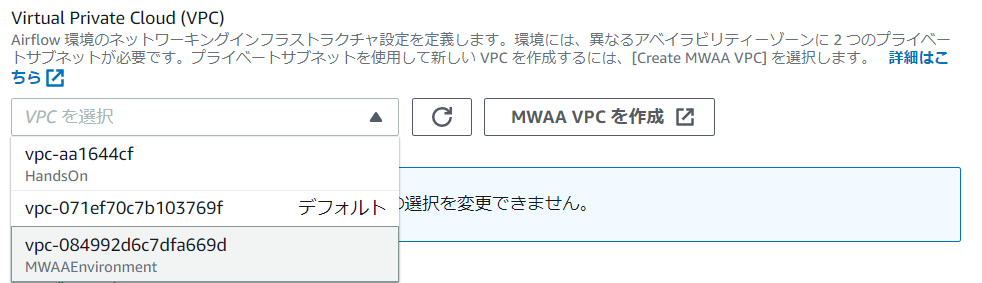
3-5. 作成が完了するまでしばらく待ちます。



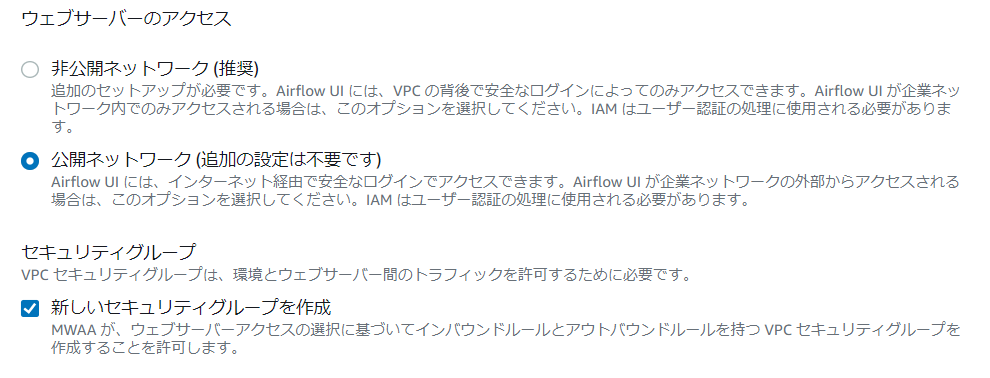
3-6. 以下のようにCreate\_CompleteとなったらMWAAの画面に戻ります。



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

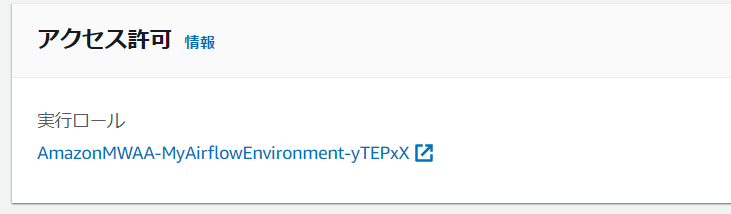


3-9．その他設定は全てデフォルトで[環境を作成]を押します

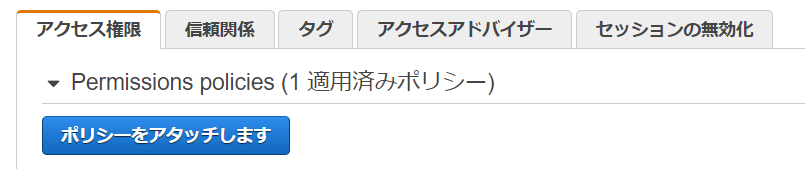
3-10. 20-30分待つと以下のように環境作成が完了します



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



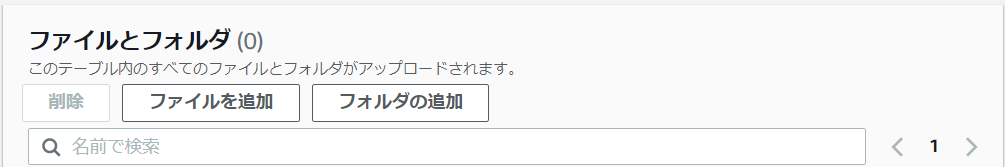
3-12. [ポリシーをアタッチ]をクリックします

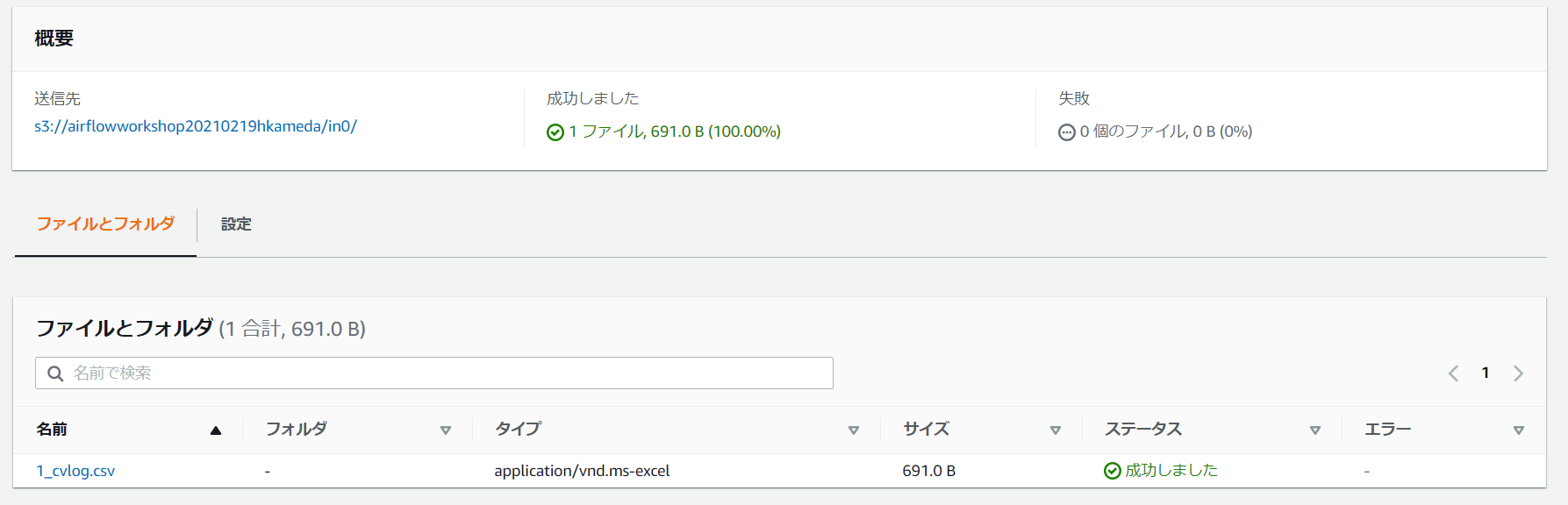


3-13. 以下2つのポリシーをアタッチします

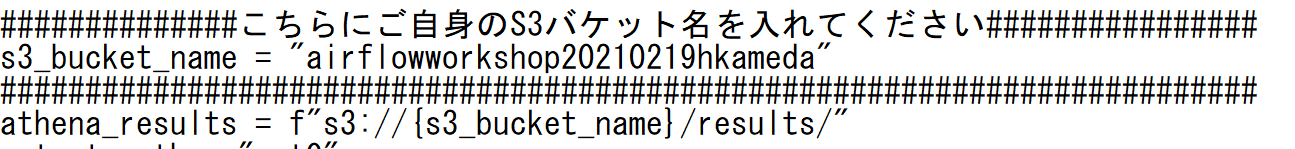


3-14．先ほどダウンロードした[1\_cvlog.csv]をS3バケットのin0にアップします。





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



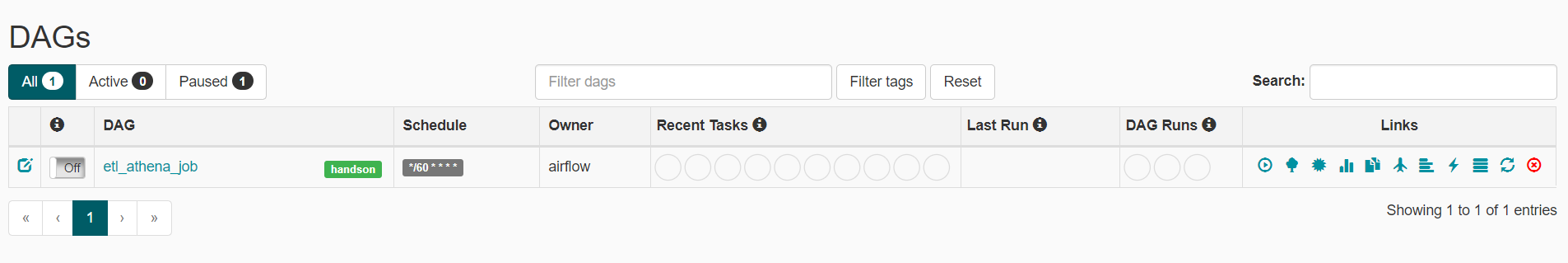
3-16. このファイルを今度はdagsのフォルダにアップロードします



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

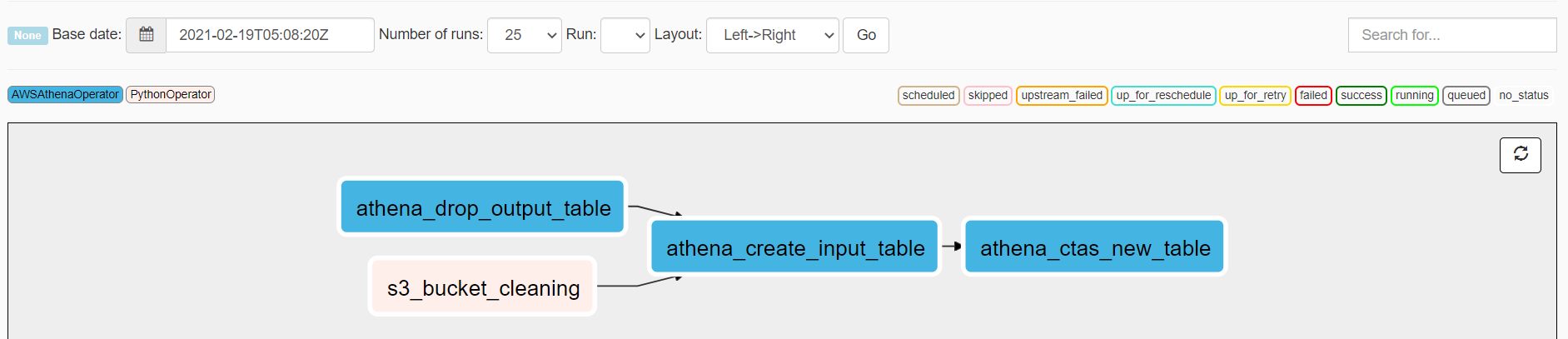


3-18. 正しく設定できていればすでに一つDAGが作成されています。（S3バケットのdagsファイルをもとに生成されている）



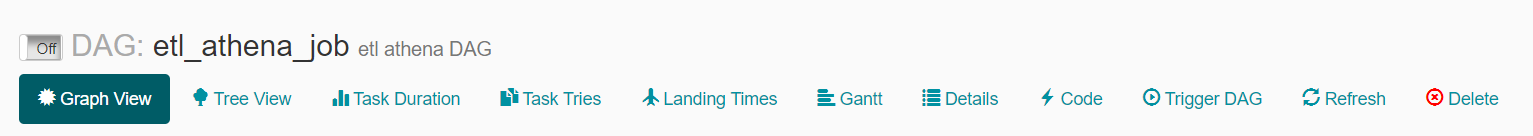
3-19. DAGをクリックして開きます。Tree Viewが表示されていますが、Graph Viewに切り替えます。



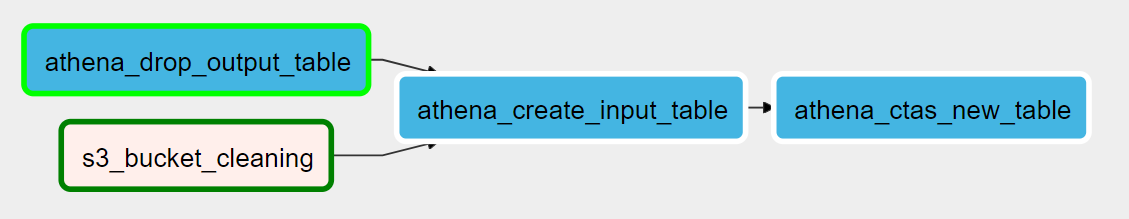


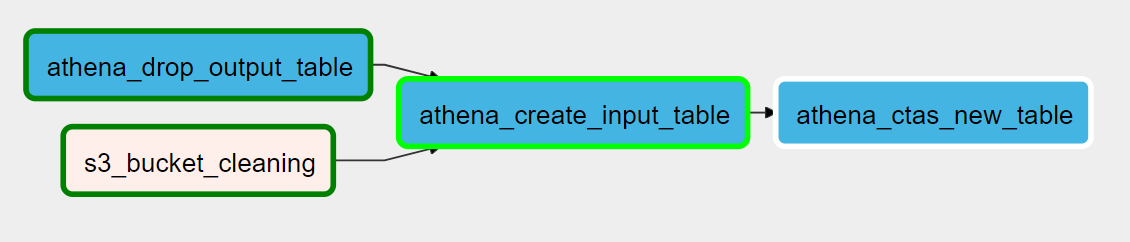
これが作業フローとなります。まず最初に、Athenaのテーブルの削除（繰り返し実行用）とS3バケットのアウトプットファイルの削除（繰り返し実行用）を行い、その後S3バケットのインプットファイルをもとにAthenaでテーブルを作成します。

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。



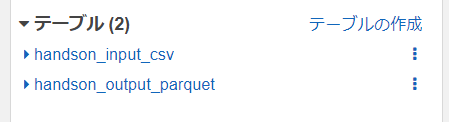
作業中が明るいみどり、作業完了が濃い緑で表示されます。





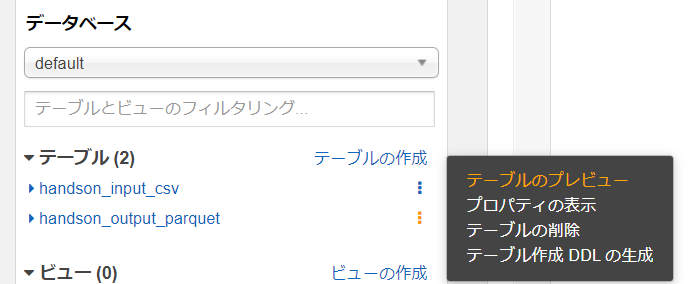
4. Athenaでのデータ確認

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（途中でエラーで止まりますので、VPCを手で削除した後再実行してください）