

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

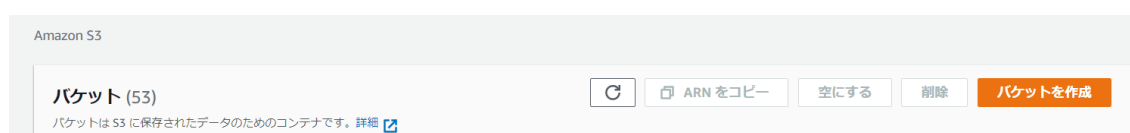
以下のファイルをダウンロードしてください。

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。

A screenshot of the 'Create Bucket' form in the Amazon S3 console. The form is titled '一般的な設定' (General Settings). It has two main sections. The first section is 'バケット名' (Bucket Name), which contains a text input field with the value 'airflowworkshop20210219hkameda'. Below the input field is a note: 'バケット名は一意である必要があり、スペース、または大文字を含めることはできません。バケットの命名規則をご参照ください' (Bucket names must be unique and cannot contain spaces or uppercase letters. See the bucket naming convention). The second section is 'リージョン' (Region), which has a dropdown menu showing 'アジアパシフィック (東京) ap-northeast-1'. At the bottom of the form, there's a section titled '既存のバケットから設定をコピー - オプション' (Copy settings from existing buckets - optional), with a note: '次の設定のバケット設定のみがコピーされます。' (Only the bucket settings for the following settings will be copied). Below this is a button labeled 'バケットを選択する' (Select bucket).

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。

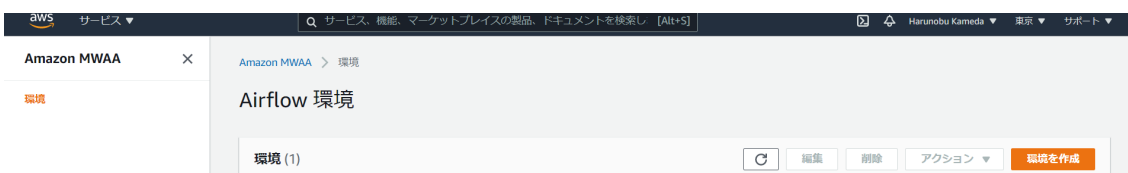
dags : DAG ファイル用

in0 : 入力 csv データ用

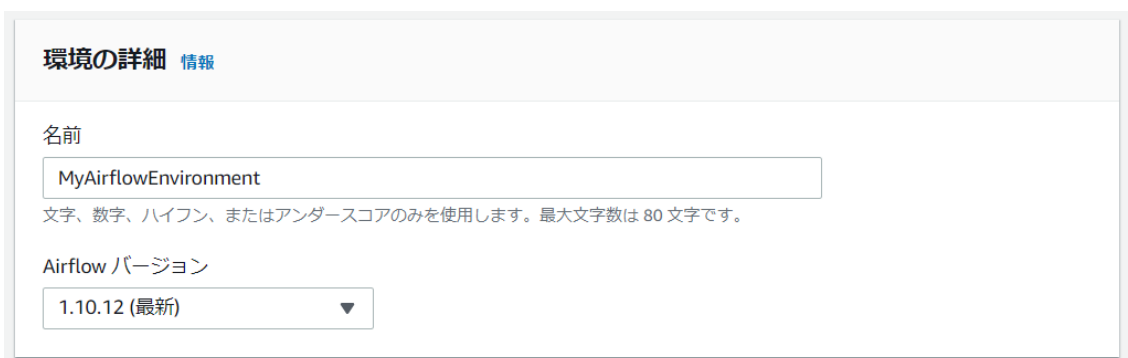
out0 : 出力 parquet データ用

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。

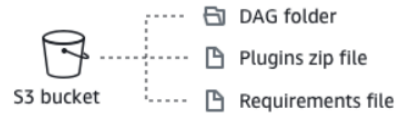


3-2. 名前とバージョンはデフォルトのまま、S3 バケットに先ほど作成したものを設定します



Amazon S3 の DAG コード 情報

Amazon MWAA は、お客様の Amazon S3 バケットを使用して DAG とサポートファイルをロードします。S3 バケットと DAG フォルダのパス、plugins.zip、requirements.txt を指定します。



i DAG コードを保存する S3 バケットを作成または指定します。バケット名のバージョンングを有効にし、「airflow-」で始まる必要があります。で新しいバケットを作成できます [Amazon S3 コンソール](#)

S3 バケット

ソースコードが保存されている S3 バケット。S3 URI を入力するか、バケットを参照して選択してください。

×

表示

S3 を参照

形式: s3://mybucketname

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。

DAG フォルダ

DAG コードを含む S3 バケットフォルダ。S3 URI を入力するか、フォルダを参照して選択します。

×

表示

S3 を参照

形式: s3://mybucketname/mydagfolder

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。

詳細設定を構成

ネットワーク 情報

Virtual Private Cloud (VPC)

Airflow 環境のネットワーキングインフラストラクチャ設定を定義します。環境には、異なるアベイラビリティゾーンに 2 つのプライベートサブネットが必要です。プライベートサブネットを使用して新しい VPC を作成するには、[Create MWAA VPC] を選択します。 [詳細はこちら](#)

VPC を選択 ▼

🔄

MWAA VPC を作成

3-5. 作成が完了するまでしばらく待ちます。

MWWA-VPC

削除

更新する

スタックアクション ▼

スタックの作成 ▼

スタックの情報

イベント

リソース

出力

パラメータ

テンプレート

変更セット

イベント (1)

検索イベント

タイムスタンプ	論理 ID	ステータス	状況の理由
2021-02-19 13:31:57 UTC+0900	MWAA-VPC	CREATE_IN_PROGRESS	User Initiated

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

スタック (2)

検索

スタック名によるフィルター

アクティブ ▼

ネスト表示

<

1

>

MWAA-VPC

2021-02-19 13:31:57 UTC+0900

✓

CREATE_COMPLETE

3-7. 作成された VPC を選びます。

Virtual Private Cloud (VPC)

Airflow 環境のネットワーキングインフラストラクチャ設定を定義します。環境には、異なるアベイラビリティゾーンに 2 つのプライベートサブネットが必要です。プライベートサブネットを使用して新しい VPC を作成するには、[Create MWAA VPC] を選択します。 [詳細はこちら](#)

VPC を選択 ▲

vpc-aa1644cf

HandsOn

vpc-071ef70c7b103769f

デフォルト

vpc-084992d6c7dfa669d

MWAAEnvironment

MWAA VPC を作成

選択を変更できません。

3-8. ネットワーク設定を公開に変更します

- 非公開ネットワーク (推奨)

- 公開ネットワーク (追加の設定は不要です)

VPC セキュリティグループは、環境とウェブサーバー間のトラフィックを許可するために必要です。

- ✓ 新しいセキュリティグループを作成

3-10. 20-30 分待つと以下のように環境作成が完了します

3-12. [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。

ファイルとフォルダ (0)
このテーブル内のすべてのファイルとフォルダがアップロードされます。

削除 ファイルを追加 フォルダの追加

🔍 名前で検索

概要

送信先 s3://airflowworkshop20210219hkameda/in0/	成功しました 🟢 1 ファイル, 691.0 B (100.00%)	失敗 🔴 0 個のファイル, 0 B (0%)
---	---------------------------------------	----------------------------

ファイルとフォルダ 設定

ファイルとフォルダ (1 合計, 691.0 B)

🔍 名前で検索

名前	フォルダ	タイプ	サイズ	ステータス	エラー
1_cvlog.csv	-	application/vnd.ms-excel	691.0 B	🟢 成功しました	-

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>の部分を実際の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

概要

送信先 s3://airflowworkshop20210219hkameda/dags/	成功しました 🟢 1 ファイル, 2.6 KB (100.00%)	失敗 🔴 0 個のファイル, 0 B (0%)
--	--------------------------------------	----------------------------

ファイルとフォルダ 設定

ファイルとフォルダ (1 合計, 2.6 KB)

🔍 名前で検索

名前	フォルダ	タイプ	サイズ	ステータス	エラー
handson-athena-job.py	-	-	2.6 KB	🟢 成功しました	-

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

環境 (1)

🔍 環境を検索

名前	状態	作成日	Airflow バージョン	Airflow UI
MyAirflowEnvironment	🟢 利用可能	Feb 19, 2021 13:39:01 (UTC...	1.10.12	Airflow UI を開く

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)

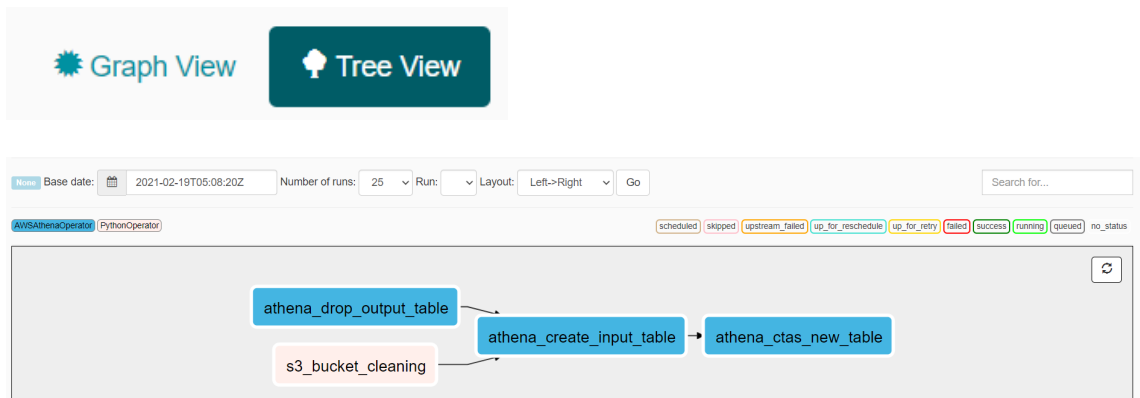
DAGs

☒ All ☐ Active ☐ Paused

	DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
<input checked="" type="checkbox"/>	etl_athena_job	hourly	airflow	○ ○ ○ ○ ○ ○ ○ ○ ○ ○		○ ○ ○	<input type="button" value="🔍"/> <input type="button" value="📊"/> <input type="button" value="📅"/> <input type="button" value="🔗"/> <input type="button" value="🔄"/> <input type="button" value="🗑️"/>

Showing 1 to 1 of 1 entries

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。

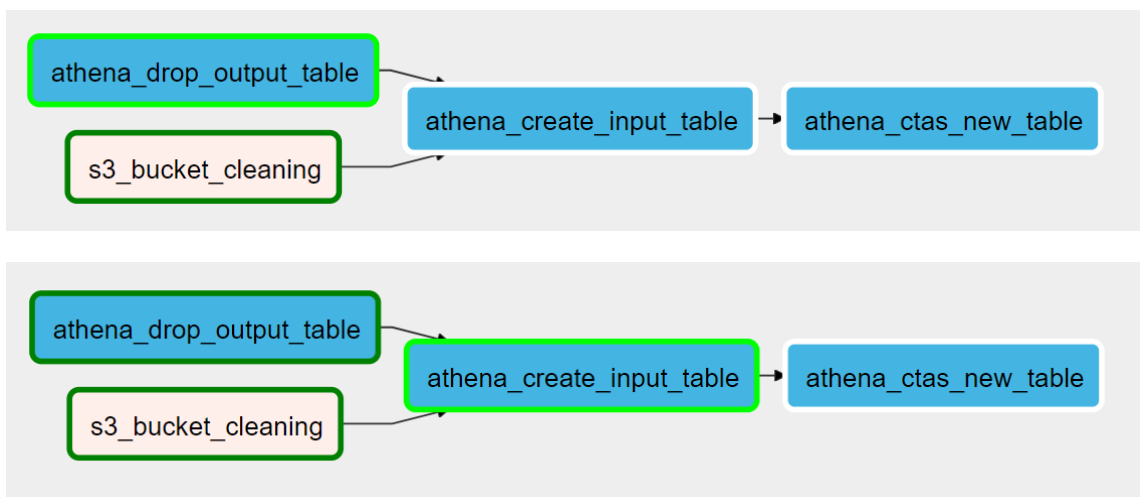


これが作業フローとなります。まず最初に、Athena のテーブルの削除（繰り返し実行用）と S3 バケットのアウトプットファイルの削除（繰り返し実行用）を行い、その後 S3 バケットのインプットファイルをもとに Athena でテーブルを作成します。

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。

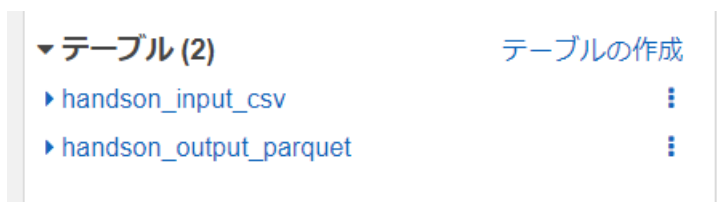


作業中が明るいみどり、作業完了が濃い緑で表示されます。



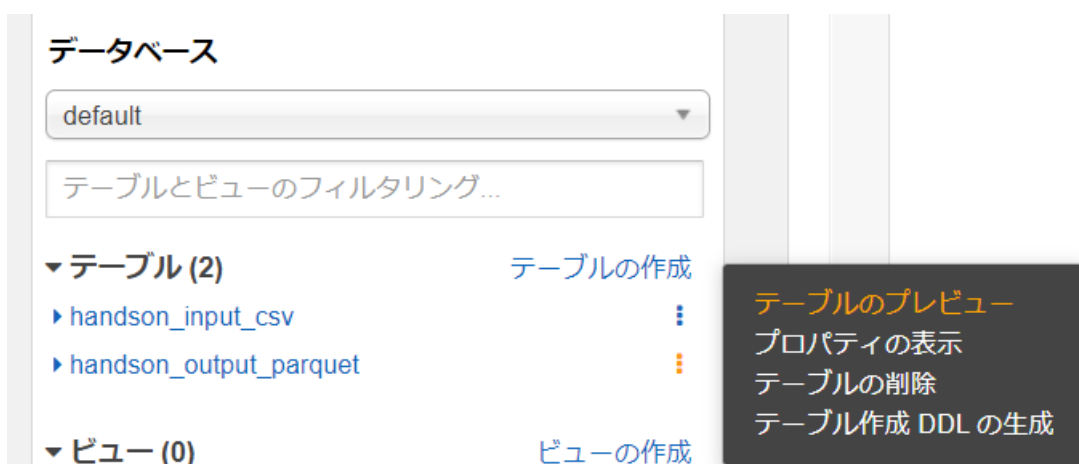
4. Athena でのデータ確認

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. CloudFormation スタック

