Glue のテストについて

Glue とは

- データの分類、クリーニング、加工を優れたコスト効果で容易に行い、さまざまなデータストア間およびデータストリーム間でデータを確実に移動するための、完全マネージド型 ETL (Extract/Transform/Load、抽出/変換/ロード) サービス

Glue Job

- Glue Job には Spark と Python Shell の 2 つのジョブタイプがある
- Spark タイプは、Apache Spark を使用したデータの分散処理が可能なため、大規模データの ETL 処理に向いている
- Python Shell はその名の通り Python 環境を使用したスクリプトを実行できるので Spark を使うほどではないが Glue Job として実行させたい場合などに使用する
- Glue の ETL ジョブはサーバーレスで実行される Spark であるため、ローカル環境で気軽に動作確認をしたりすることができない
- そのための開発環境として Glue には開発エンドポイント

Glue Job の料金

- **DPU 時間あたり 0.44USD** が 1 秒単位で課金され、Apache Spark タイプの ETL ジョブごとに最小 1 分 (Glue version 2.0) または最小 10 分 (Glue version 0.9/1.0)
- Lambda は 512MB で 0.0000000083USD/ミリ秒→ 0.0000083 USD/秒 → **0.02988 USD/時**
- 基本的に大規模データを並列処理させるような用途のため、それなりのお値段
- Lambda と違いローカルで実行しにくいため、開発時などに Glue 上で実際に動かして動作確認して・・・としていると気がついたら結構な金額になってしまうことも

Glue 開発エンドポイント

- 開発エンドポイントは、AWS Glue スクリプトの作成およびテストに使用できる環境
- ただし実際に使用しようとすると**お高くなりがち**
    - 開発エンドポイントのプロビジョニング時には、5 個の DPU が割り当てられ

ます。開発エンドポイントを 24 分 (5 分の 2 時間) 実行すると、DPU 時間あたり 0.44 USD で 5 DPU * 2/5 時間、つまり 0.88 USD が請求されます。

- しかもこのエンドポイントは EC2 みたいに一時停止とかできないので、使う間はずっと起動させっぱなしにしないといけない。止めたい場合は削除する

ということでローカル実行環境を作る

[AWSGlue ライブラリを使用した ETL スクリプトのローカルでの開発とテスト](https://docs.aws.amazon.com/ja_jp/glue/latest/dg/aws-glue-programming-etl-libraries.html)

上記にいくつか方法があるが、ローカルに環境構築するのはめんどくさいので Docker イメージを使ったコンテナ環境を作る