AWS Glue Studioワークショップ

2021/03/29

シニアエバンジェリスト

亀田 治伸

1. 以下のURLへアクセスしCloudFormationを起動します。

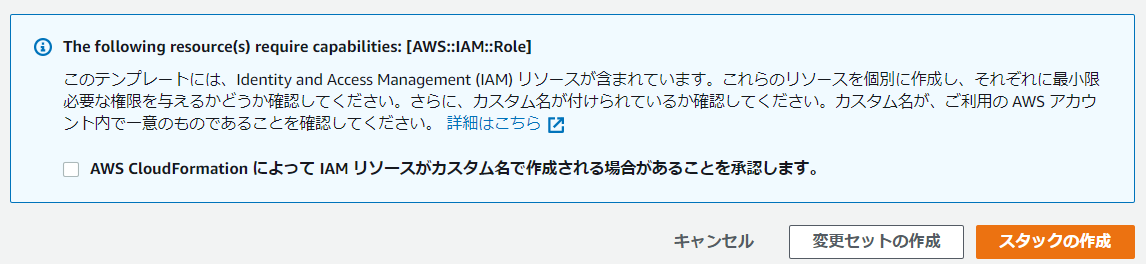
短縮

<https://amzn.to/36n5q1F>

フル

<https://console.aws.amazon.com/cloudformation/home?region=us-east-1#/stacks/create/review?stackName=Glue-Studio-Blog&templateURL=https://aws-bigdata-blog.s3.amazonaws.com/artifacts/gluestudio/cftemplate/CFgluestudio.yaml>

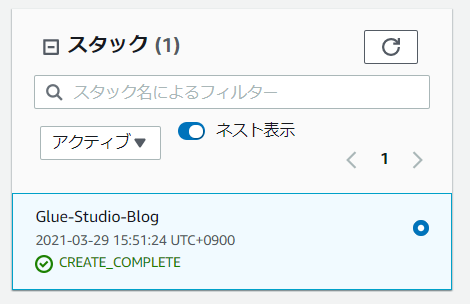
1. 以下にチェックを付けて、「スタックの作成」を押します。



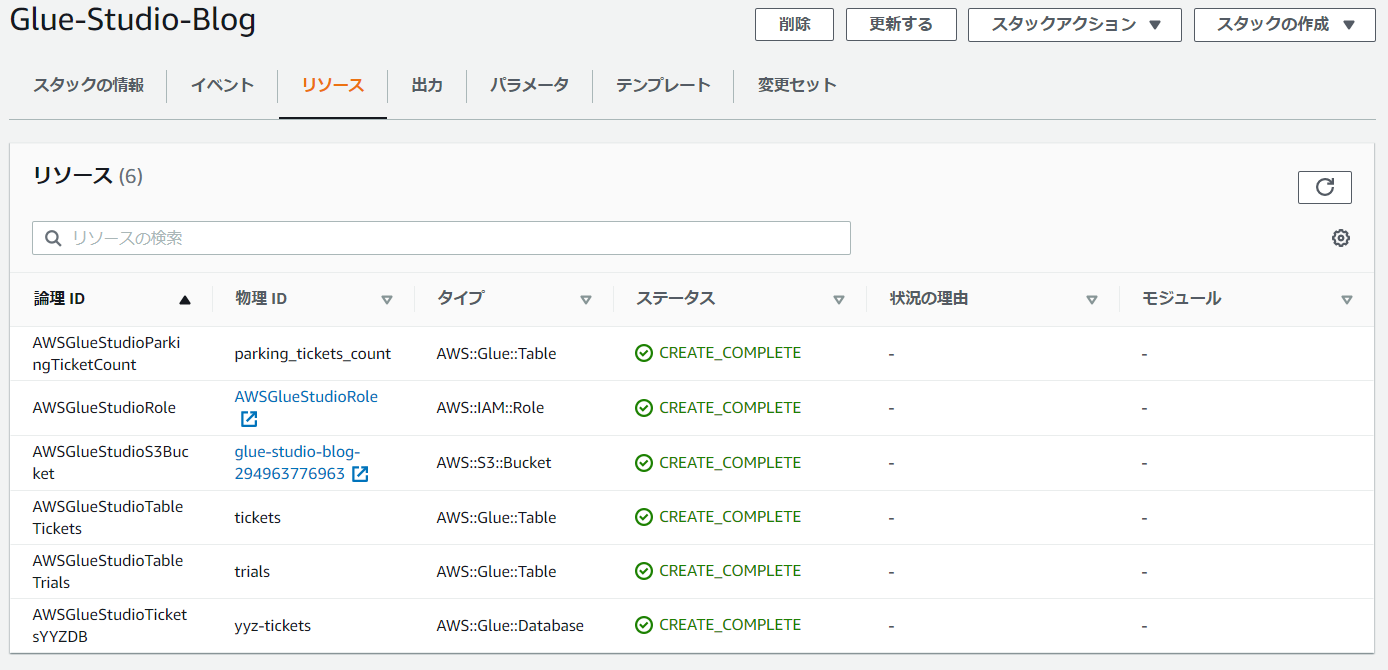
1. 構築作業中となりますのでしばらく待ちます。（たまに更新ボタンを押して画面をリフレッシュしてください）



1. 以下が表示されれば、作成完了です。



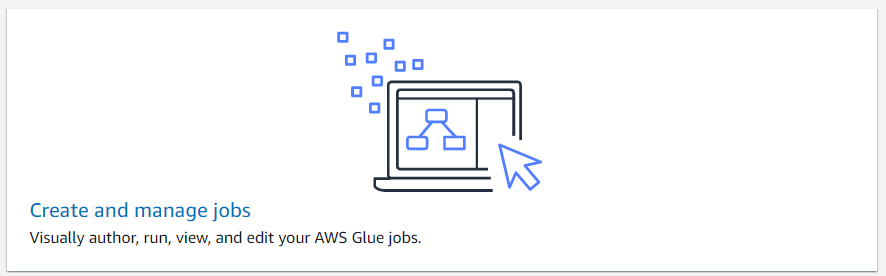
1. リソースタブを見ると、Glueのデータベースが1つ、テーブルが３つ、S3バケット、IAMロールが作成されたことがわかります。



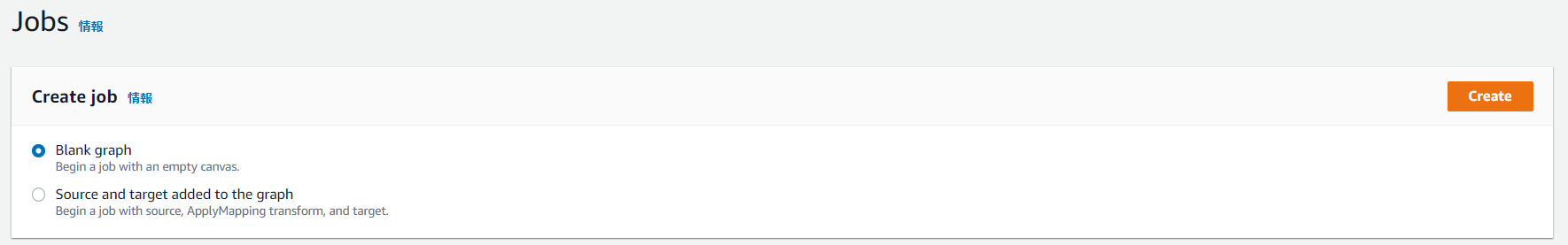
1. AWS Glueのマネージメントコンソールにいきます。（管理者画面上の検索窓にglueと入力してください）
2. 左のペインから「Glue Studio」を選んでください



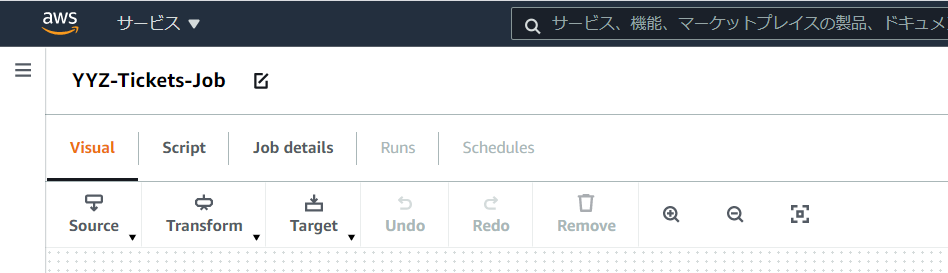
1. 「Create and Manage jobs」を選びます



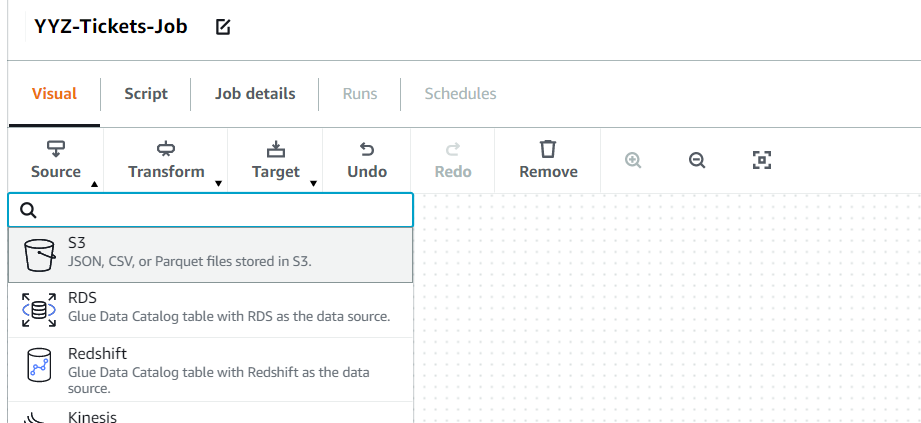
1. 「Blank Graph」を選び、「Create」をおします



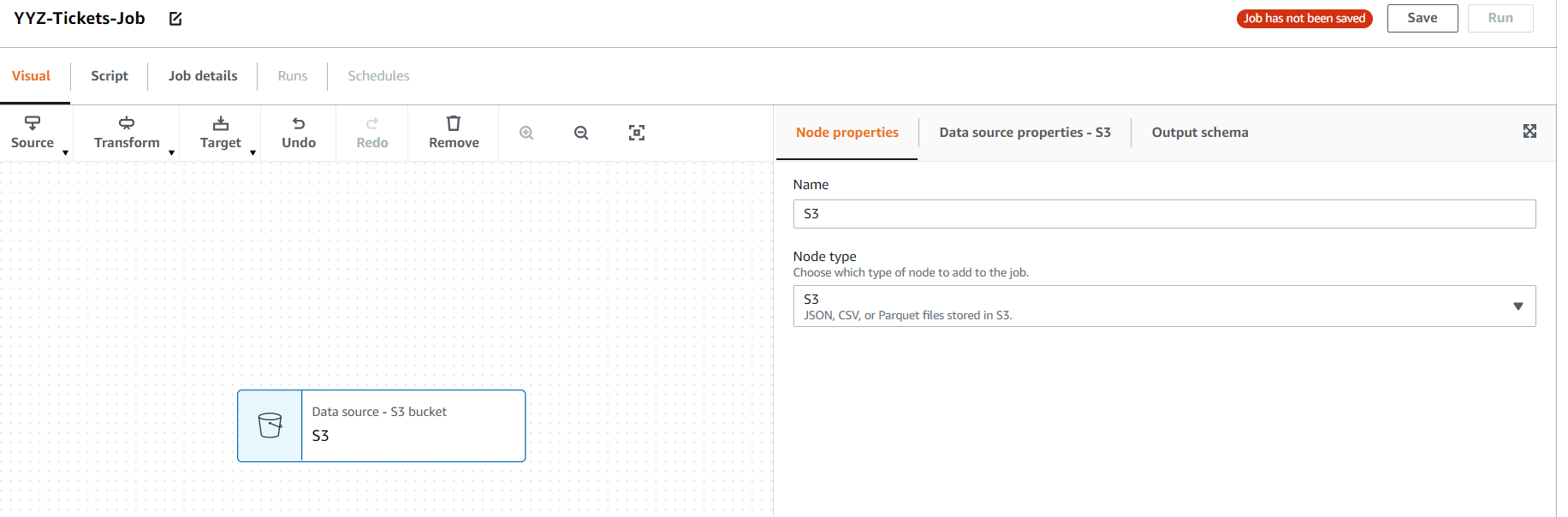
1. 左上Job名を「YYZ-Tickets-Job」に変更します。（名前は任意でなんでもOKです）



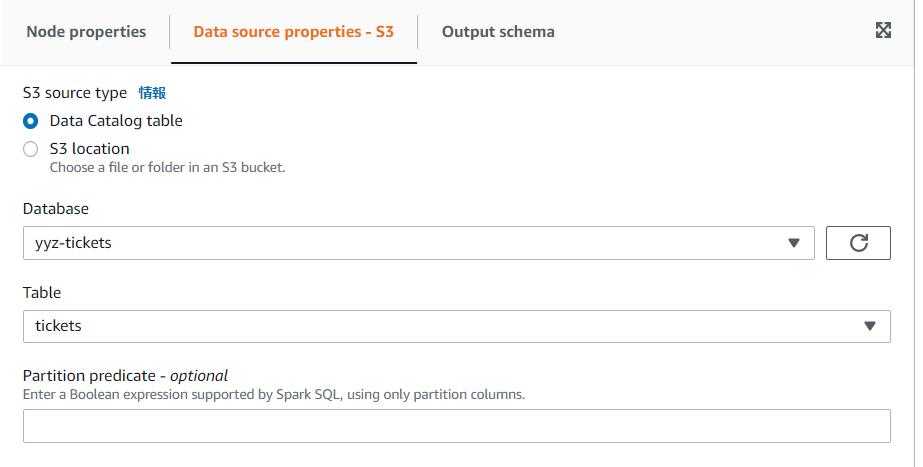
1. 「Visual」→「Source」→「S3」を選びます



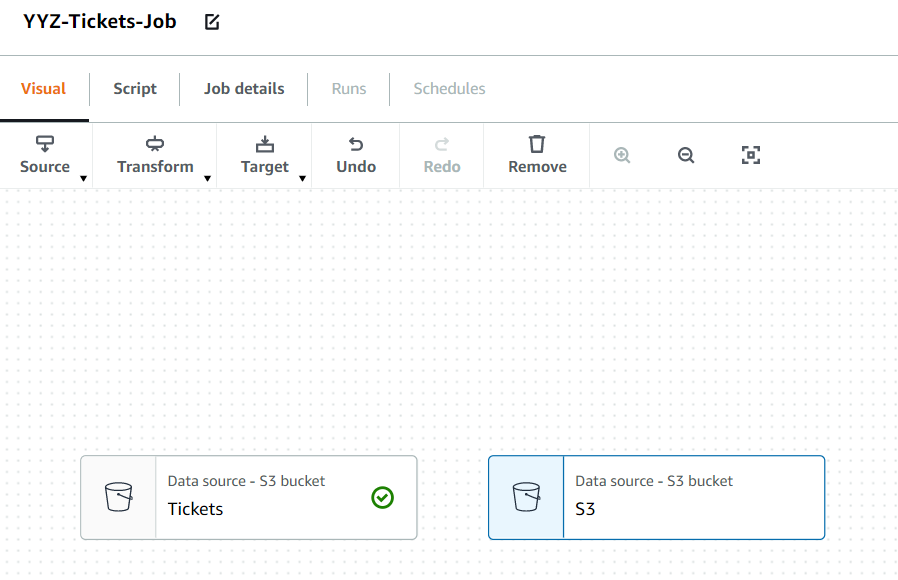
1. グラフ領域のS3を選んで、画面右の「Note Properties」タブを選び、「Name」にTicketsと入力します



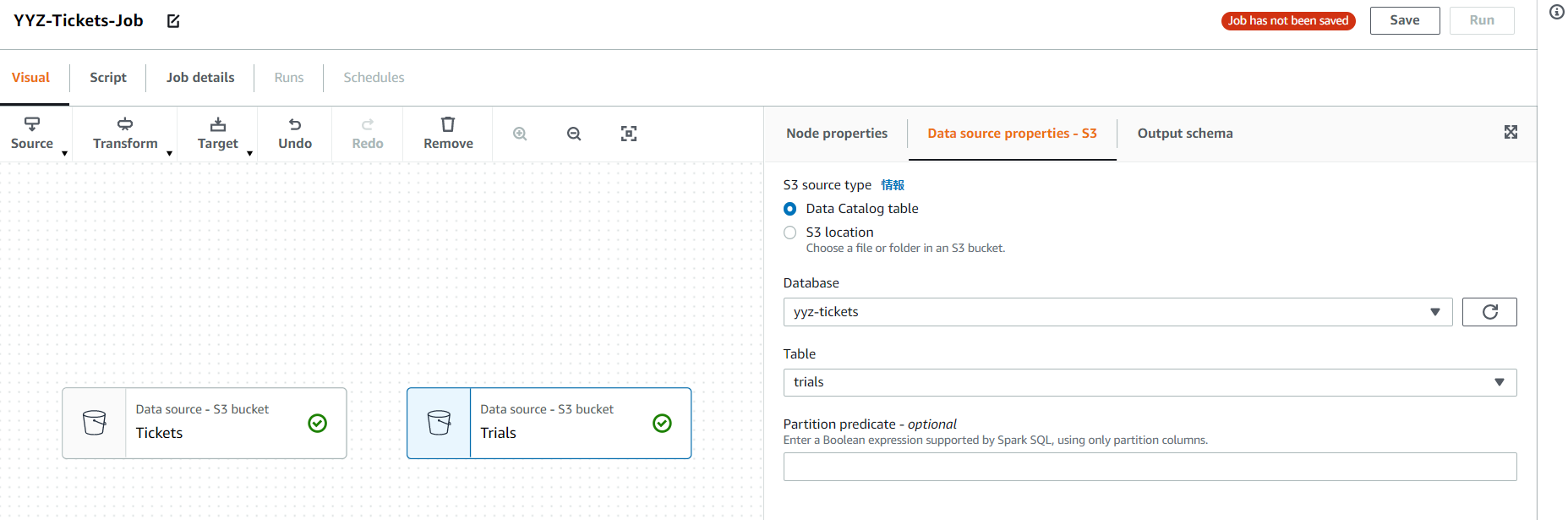
1. 「Data source properties S3」タブを開き、「Database」にyyz-tickets、「Table」にticketsを選びます。



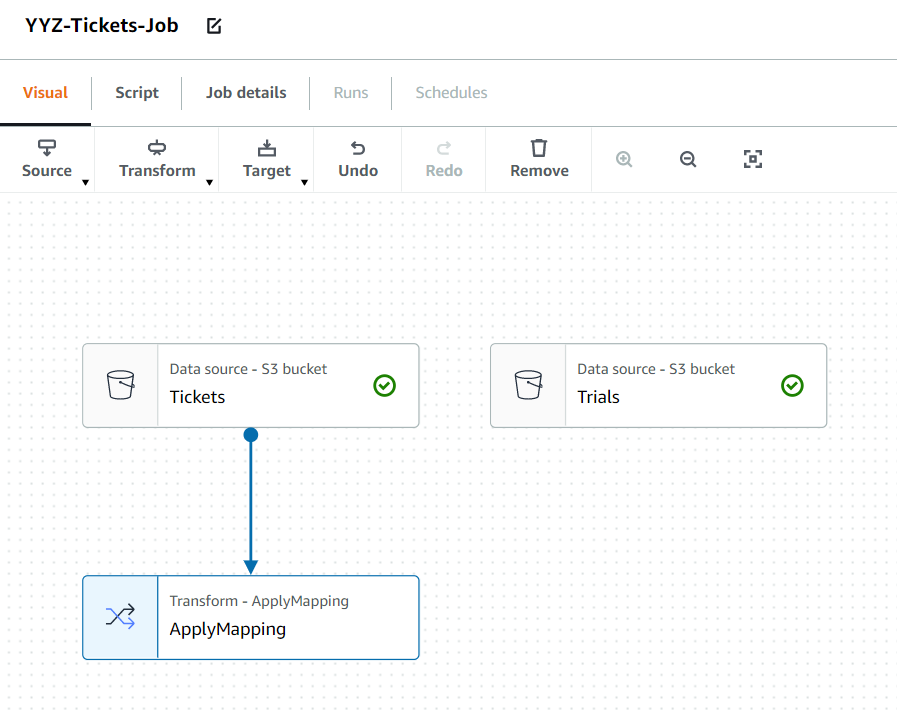
1. 2個目の「Source」→「S3」を追加します。



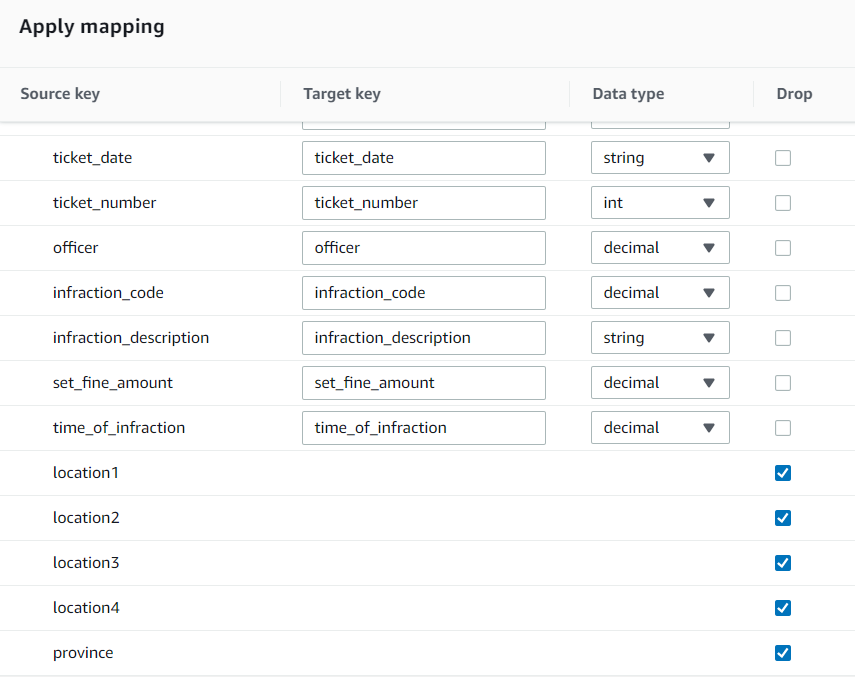
1. 「Name」にTrialsと設定し、「Database」には先ほどと同じyyz-tickets、「Table」にtrialsを選びます



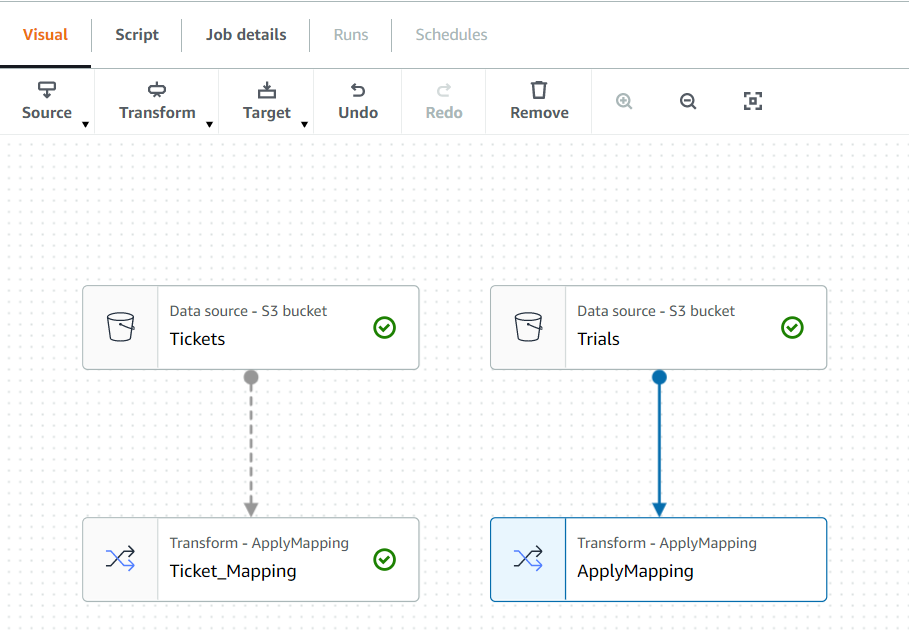
1. グラフのTicketsを選び（薄い水色になります）、「Transform」からApplyMappingを選びます。



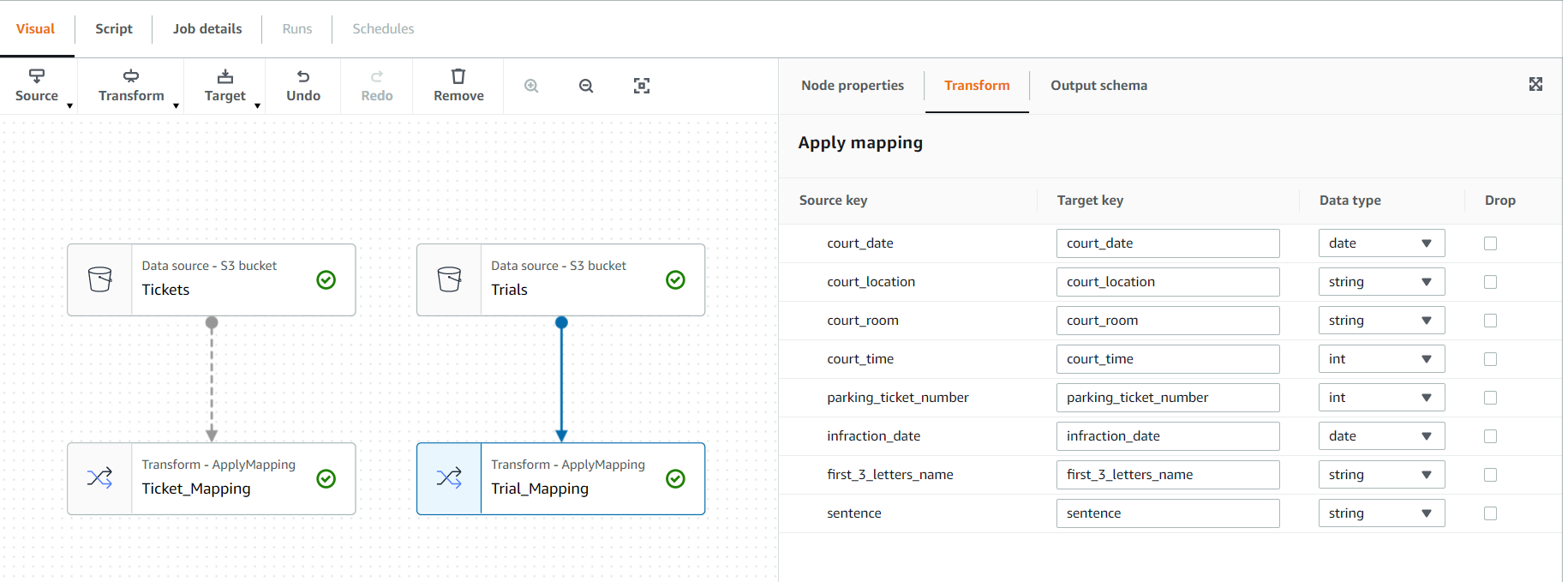
1. 「Name」をTicket\_Mappingとし、「Transform」タブでticket\_numberをdecimalからintに変更します。さらに、Location1、Location2、Location3、Location4、Provinceのカラムをdropします



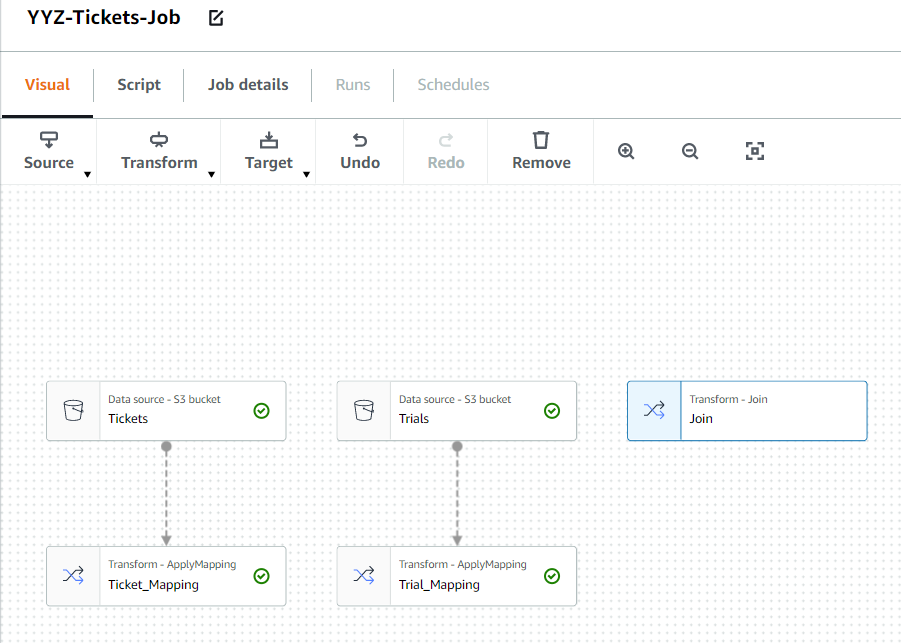
1. 同様に、TrialsのS3の下にも、ApplyMappingをぶら下げます。



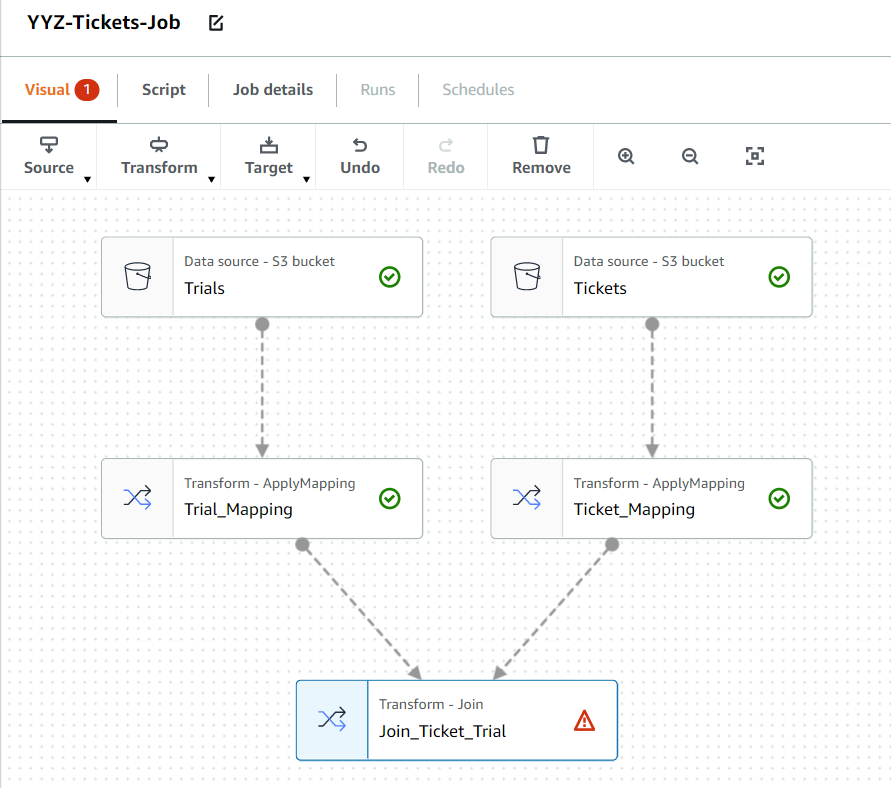
1. 「Name」をTrial\_Mappingとし、parking\_ticket\_number の型を long から int に変更します



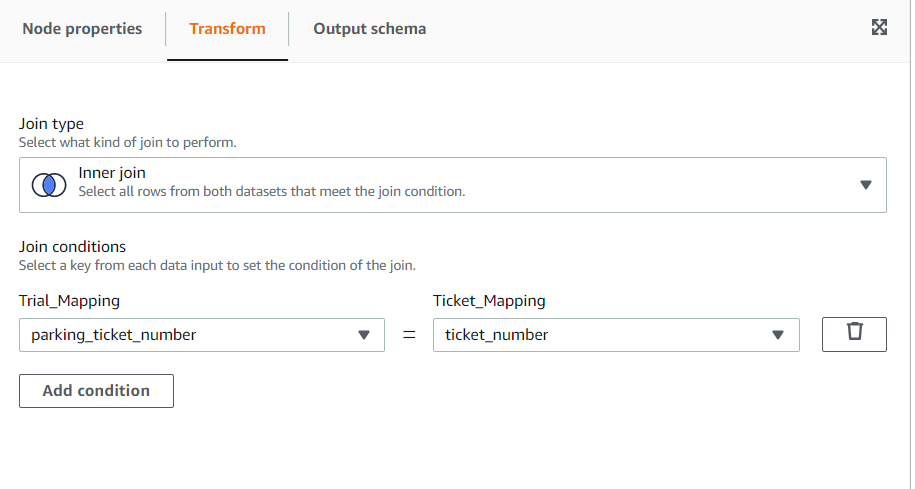
1. 以上でテーブル連結に必要な型の統一や、不要カラムの削除設定が完了しました。「Transform」→「Join」を選択します



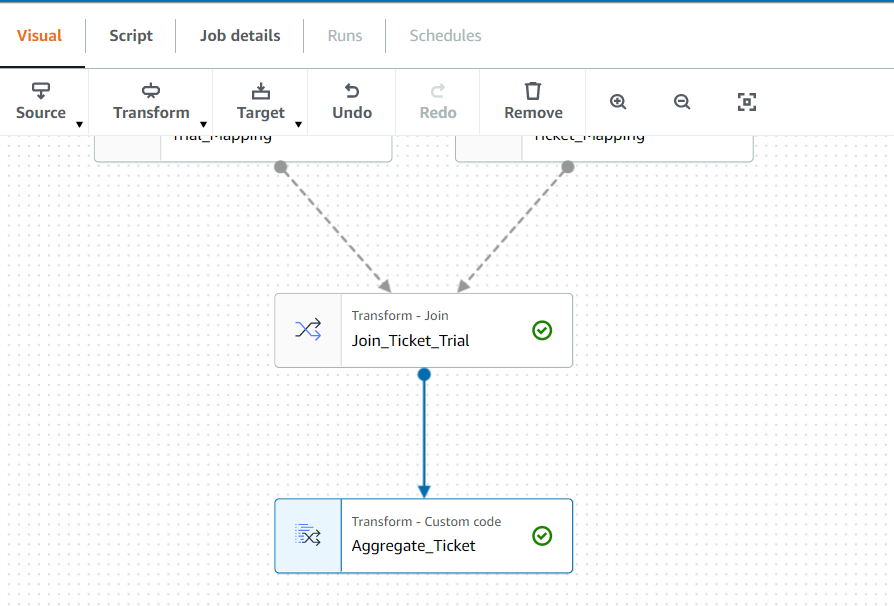
1. 「Name」をJoin\_Ticket\_Trialと設定し、「Note parents」に先ほど作成した2つのTransformを選びます。



1. 「Transform」タブの「Join Conditions」で「Add condition」ボタンをおし、parking\_ticket\_number とticket\_numberを結びつけます



1. Transform –Joinを選択した状態で、「Transform」→「Custom Transform」を選び、「Name」をAggregate\_Ticketとします。



1. 「Transform」タブの「MyTransform」部分をAggregate\_Ticketsに置換します。そして2行目以降に以下をペーストします。

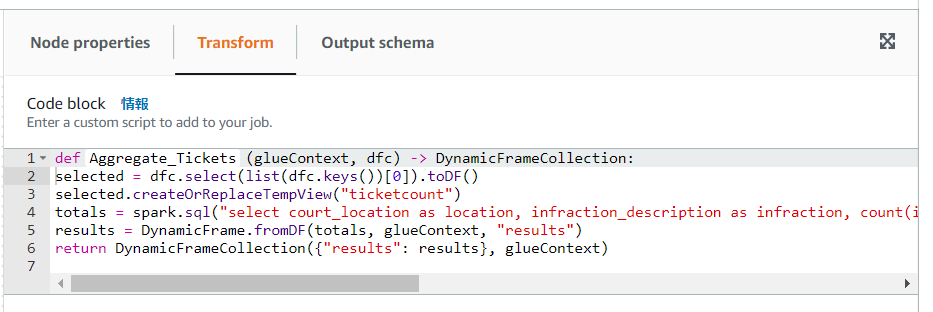
selected = dfc.select(list(dfc.keys())[0]).toDF()

selected.createOrReplaceTempView("ticketcount")

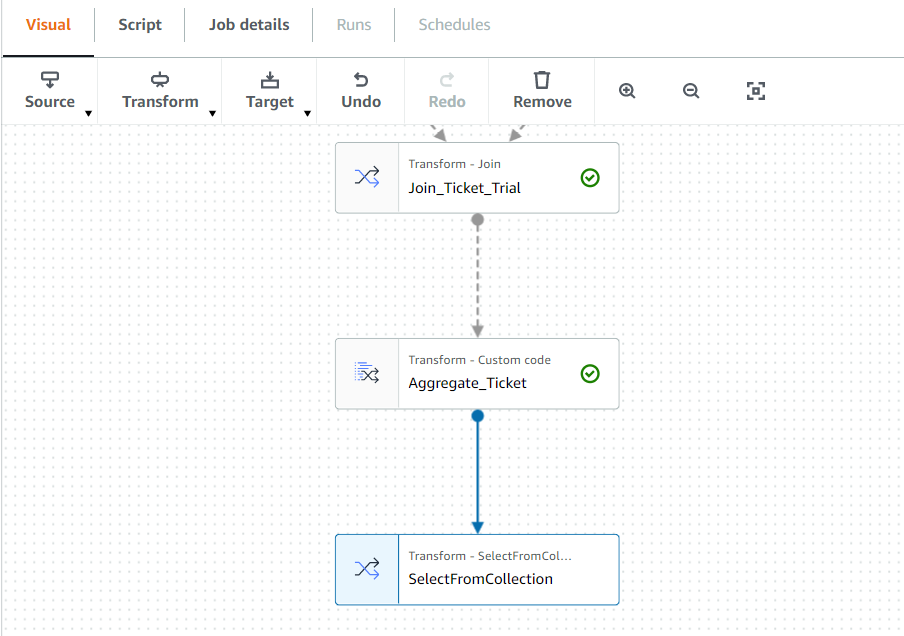
totals = spark.sql("select court\_location as location, infraction\_description as infraction, count(infraction\_code) as total FROM ticketcount group by infraction\_description, infraction\_code, court\_location order by court\_location asc")

results = DynamicFrame.fromDF(totals, glueContext, "results")

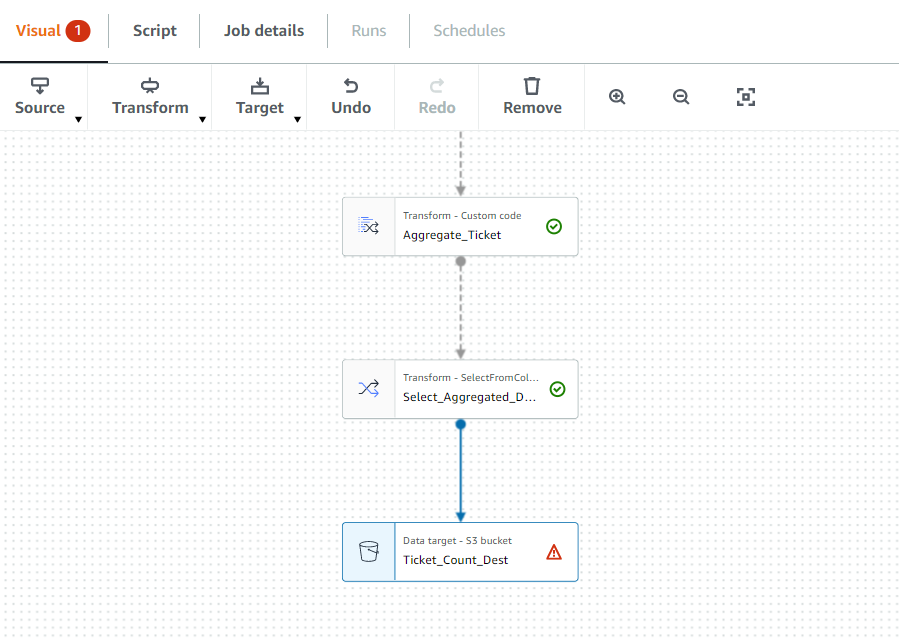
return DynamicFrameCollection({"results": results}, glueContext)



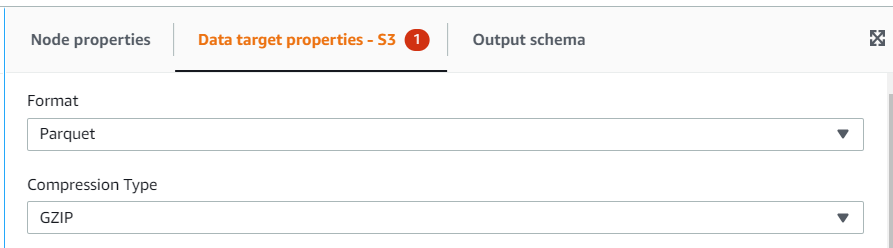
1. グラフ領域で「Aggregate\_Ticket」を選んで「Transform」→「SelectFormCollection」を選びます。



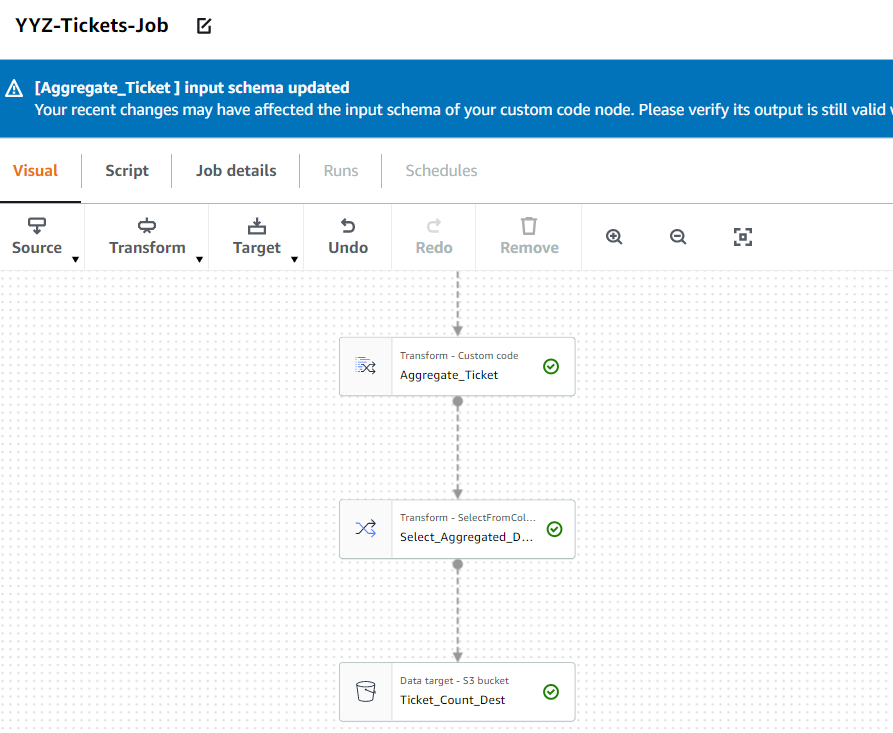
1. 「Name」をSelect\_Aggregated\_Dataとします。
2. 「Target」→「S3」を選び、「Name」をTicket\_Count\_Destとします。



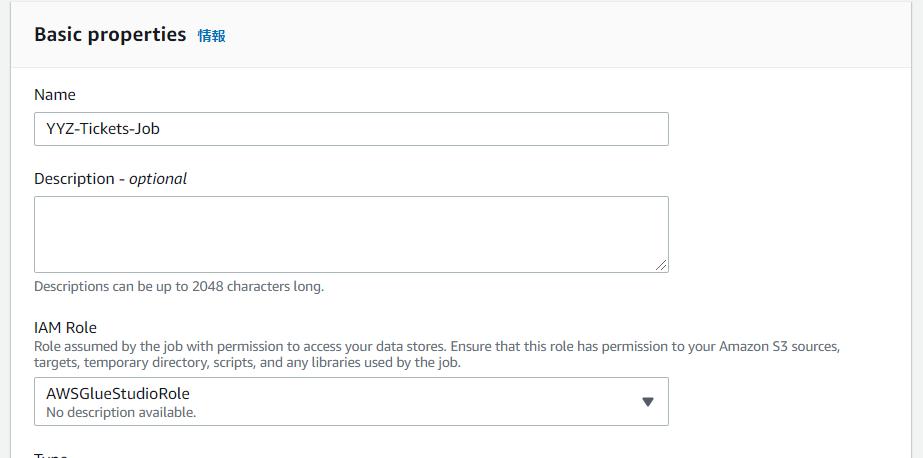
1. 「Data target properties –S3」タブで、「Format」をParquetにし、Compression TypeをGZIPにします。



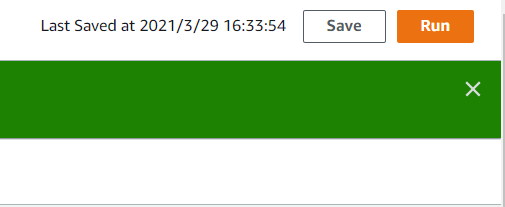
1. 今日作成されたS3バケットを選び、後ろに「parking\_tickets\_count/」を付け加えます。（文字列は / で終わることに注意してください）
2. グラフ領域の上にあるタブで「Job details」を選びます。



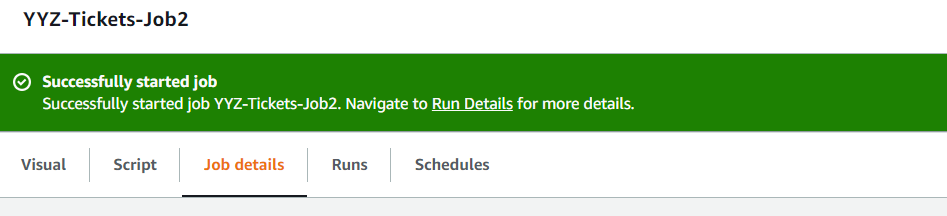
1. 今日作成されたIAMRoleを選びます



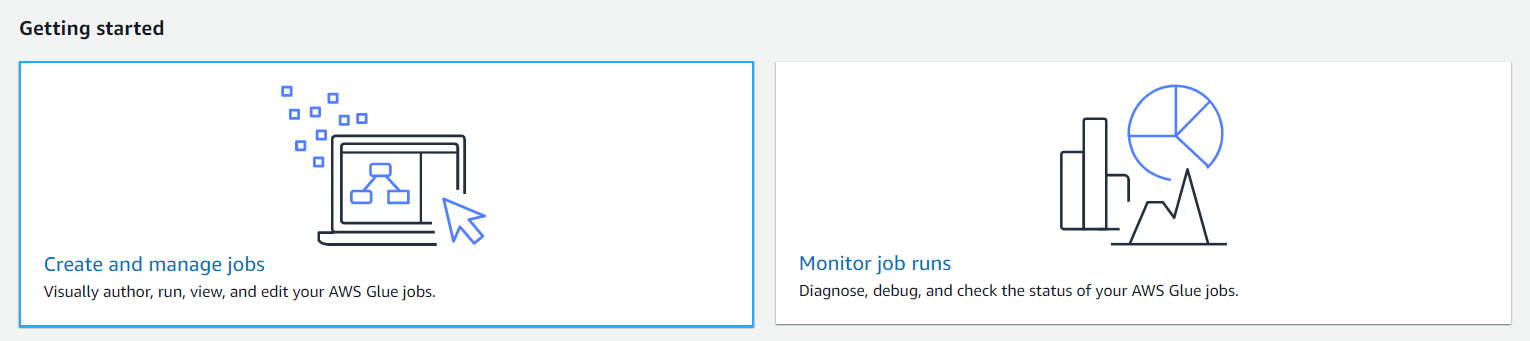
1. 「Job bookmark」をDisableに設定し、画面右上（ブラウザのスクロールバーで上に戻ります）で「Save」ボタンをおします。
2. 「Save」が終わったら「Run」ボタンをおします



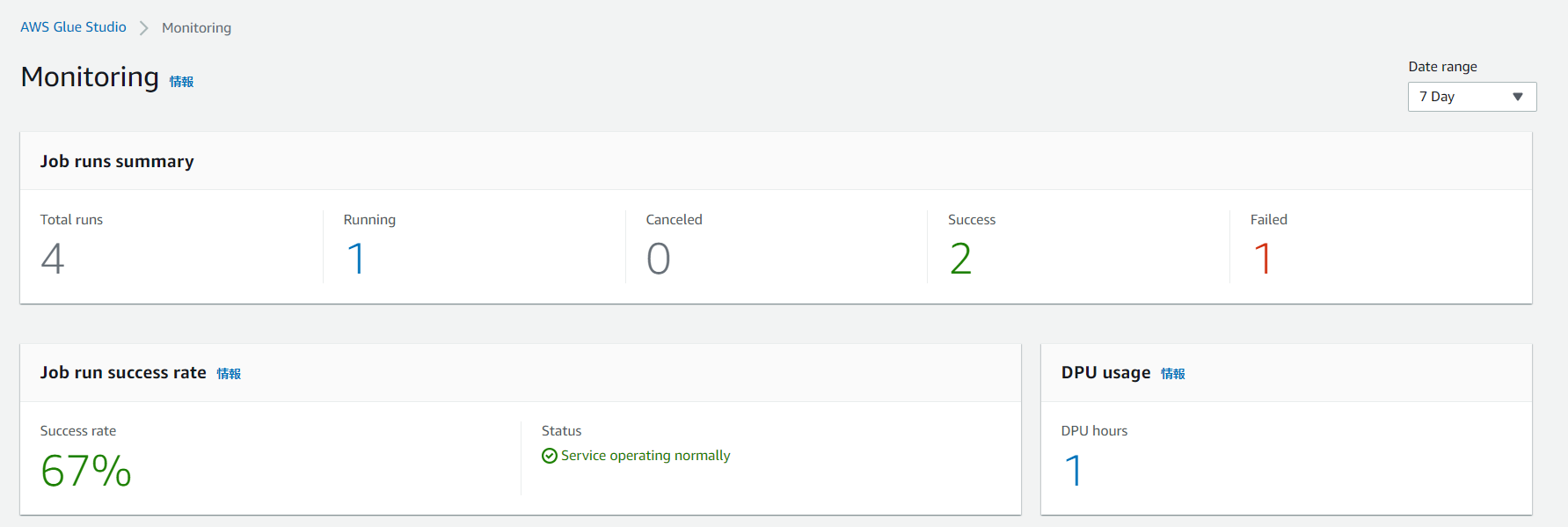
1. 以下の通りJobが開始されます。



1. ブラウザの別タブで再度Glue Studioを開き、「Monitor job runs」を押して下さい



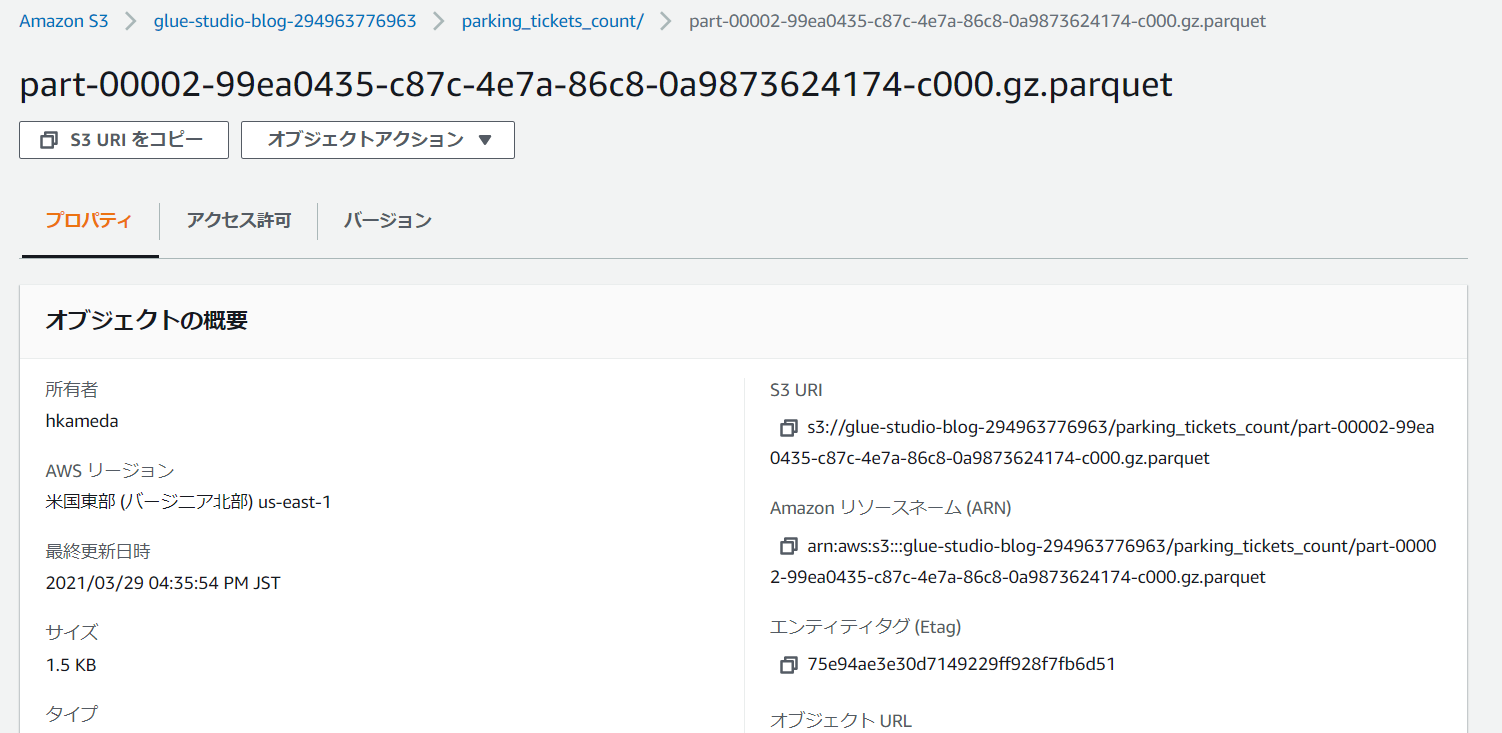
1. 以下のようにJobの実行状況が可視化されます。何度かブラウザをリロードしながらjobが完了するまで待ってください



1. Jobが完了したら今日さ宇製されたS3バケットへ移動し、フォルダができていることを確認します。



1. それでもいいのでフォルダの中に作成されているgzip圧縮されたparquet形式のオブジェクトをクリックします



1. オブジェクトアクションで「S3 Selectを使用したクエリ」を選んで実行してください。入力形式はApache Parquet、出力形式は任意です。



1. 以下のようにデータ変換やJoinが完了し正しくデータが生成されていることがわかります。



おつかれさまでした。以下を削除してさい。

* Job（Glue Studioではなく、Glueの画面からジョブを削除）
* CloudFormationスタック