Pontifícia Universidade Católica
do Rio de Janeiro

Sprint III –
Engenharia de dados

Aluno : Arthur Ferreira de F. Lopes

Pós-graduação em ciência de dados e Analytics

Professores : Vitor Almeida e Silvo Alonso

Introdução

O número de restaurantes em Nova York está aumentando dia a dia. Muitos estudantes e profissionais ocupados confiam nesses restaurantes devido ao seu estilo de vida agitado. O serviço de entrega de comida online é uma ótima opção para eles. Fornece-lhes boa comida nos seus restaurantes preferidos. Uma empresa agregadora de alimentos, FoodHub, oferece acesso a vários restaurantes por meio de um único aplicativo de smartphone.

O aplicativo permite que os restaurantes recebam um pedido online direto de um cliente. O aplicativo designa um entregador da empresa para retirar o pedido após a confirmação do restaurante. O entregador então usa o mapa para chegar ao restaurante e aguarda o pacote de comida. Assim que a embalagem de comida é entregue ao entregador, ele confirma a retirada no aplicativo e se desloca até o local do cliente para entregar a comida. O entregador confirma a entrega no aplicativo após entregar a embalagem de comida ao cliente. O cliente pode avaliar o pedido no aplicativo. O agregador de alimentos ganha dinheiro arrecadando uma margem fixa do pedido de entrega dos restaurantes.

Coleta

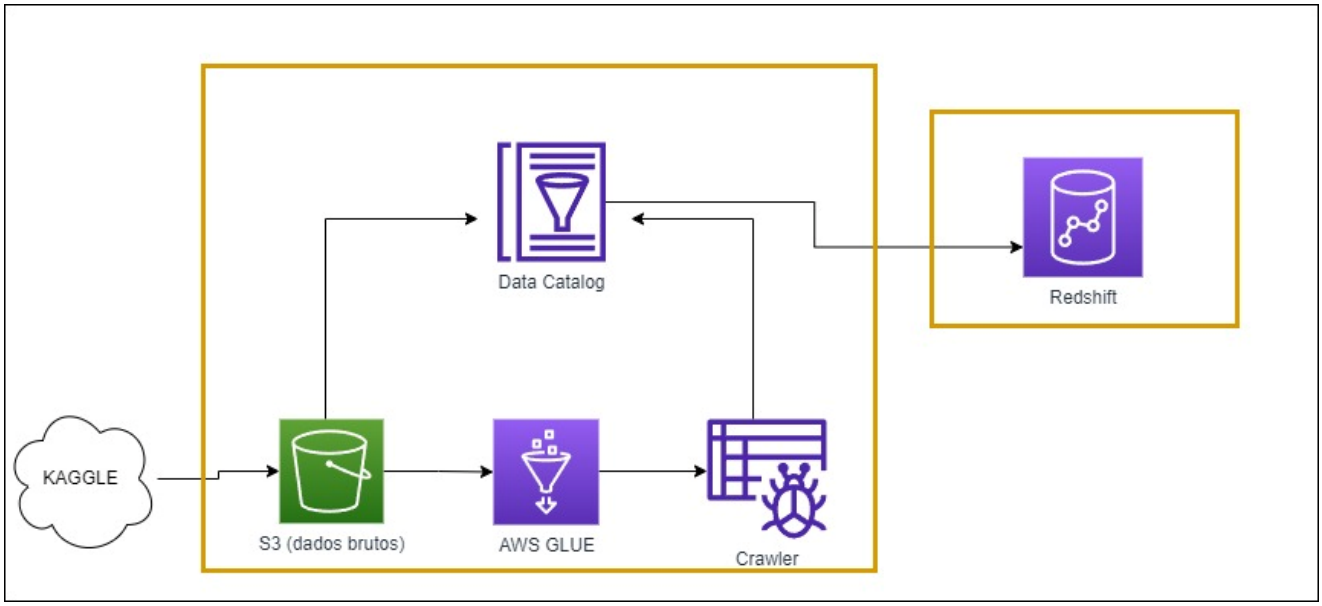
A coleta de dados foi realizada através do repositório de dados Kaggle. Por onde pude ter informações detalhadas do dataset que realizei o downloads.

https://www.kaggle.com/datasets/ahsan81/food-ordering-and-delivery-app-dataset

O dataset foi baixado em apenas um arquivo no formato CSV, onde é um formato que facilita o upload do mesmo na cloud AWS, a plataforma de escolhi para fazer o MVP

Modelagem

Um "pipeline de dados" refere-se a um conjunto de etapas sequenciais ou processos que são usados para coletar, processar, transformar e mover dados de uma fonte para um destino específico. Os pipelines de dados são comumente usados em engenharia de dados, análise de dados e em sistemas de processamento de dados em geral. Eles são projetados para automatizar e facilitar o fluxo de dados através de diferentes estágios. Abaixo, está representado o processo desde a coleta de dados até a utilização dos mesmo para responder alguma pergunta.



Catálogo de dados.

Um catálogo de dados é como uma lista organizada de informações sobre todos os dados que temos. É como um catálogo de uma biblioteca, mas em vez de livros, ele lista todos os nossos dados. Com um catálogo de dados, podemos encontrar rapidamente informações sobre onde os dados estão, como estão estruturados e quem os usa. Isso nos ajuda a gerenciar nossos dados de forma eficaz, economizando tempo e tornando mais fácil usar essas informações importantes para tomar decisões informadas. Em resumo, um catálogo de dados é como um guia útil para nossos dados, tornando mais simples encontrá-los e usá-los. Abaixo irei descrever os dados do dataset escolhido.

O dataset contém os diferente dados relacionados a um pedido de comida em um aplicativo.

As descrição das colunas são :

order_id: ID exclusivo do pedido - Tipo : bigint

customer_id: ID do cliente que pediu a comida – Tipo: bigint

nome_restaurante: Nome do restaurante – Tipo : Varchar

cuisine_type: Culinária pedida pelo cliente – Tipo : Varchar

day_of_the_week: Indica se o pedido é feito em dia de semana ou final de semana (o dia da semana é de segunda a sexta e o final de semana é sábado e domingo) obs: Existe pedidos que não foram classificados pelo cliente – Varchar

rating: Classificação dada pelo cliente em 5 (Existe pedidos que o cliente não informou a classificação) – Tipo: Varchar – Valor minino 3 – Valor Maximo 5

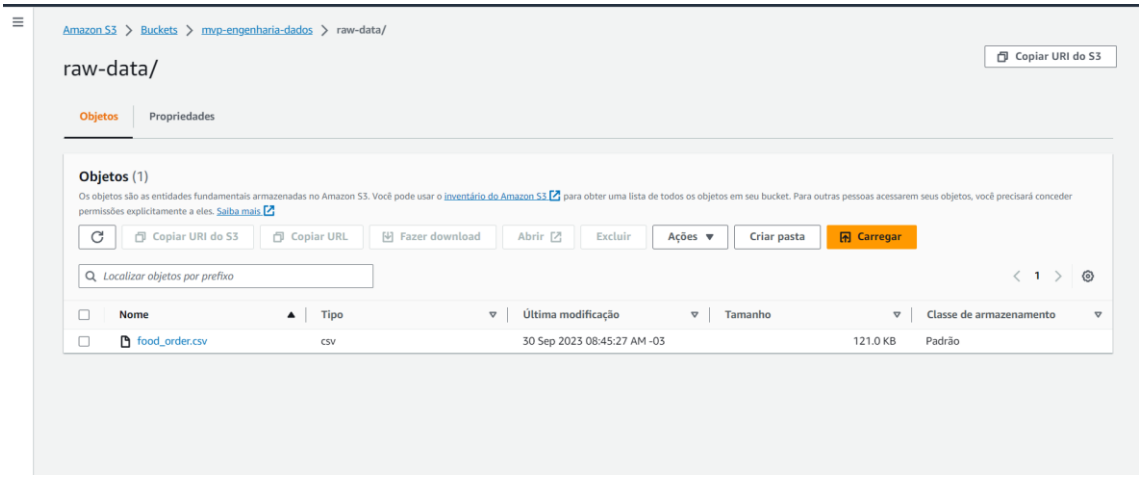
food_preparation_time: Tempo (em minutos) que o restaurante leva para preparar a comida. Isto é calculado tomando a diferença entre os carimbos de data/hora da confirmação do pedido do restaurante e a confirmação de retirada do entregador. – Tipo: bigint - Valor minino 20 – Valor máximo – 35

delivery_time: Tempo (em minutos) que o entregador leva para entregar a embalagem de comida. Isso é calculado tomando a diferença entre os carimbos de data/hora da confirmação de coleta do entregador e as informações de entrega – Tipo : bigint – Valor mínimo 15 – Valor máximo 33

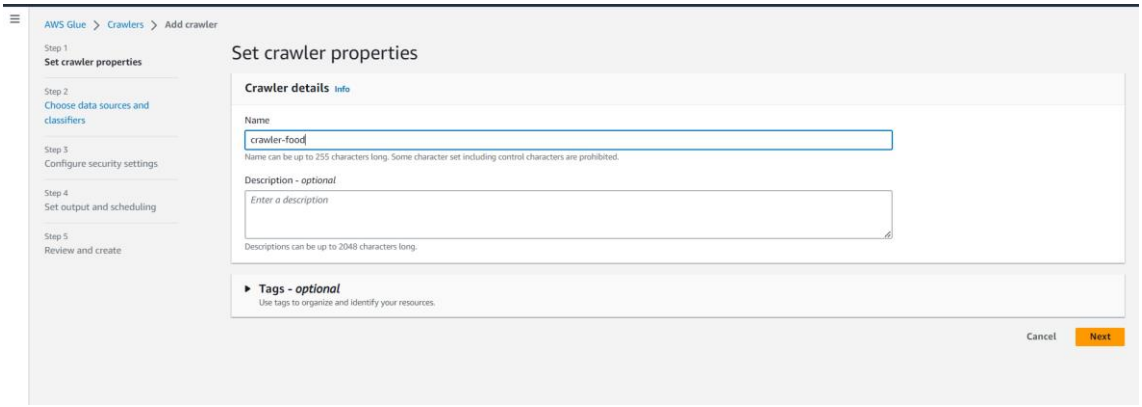
Carga dos dados

Como visto acima, os dados foram obtidos através de um repositório na internet e que somente depois do downloads pude iniciar a carga dos dados , abaixo, irei explicar cada passo do pipeline :

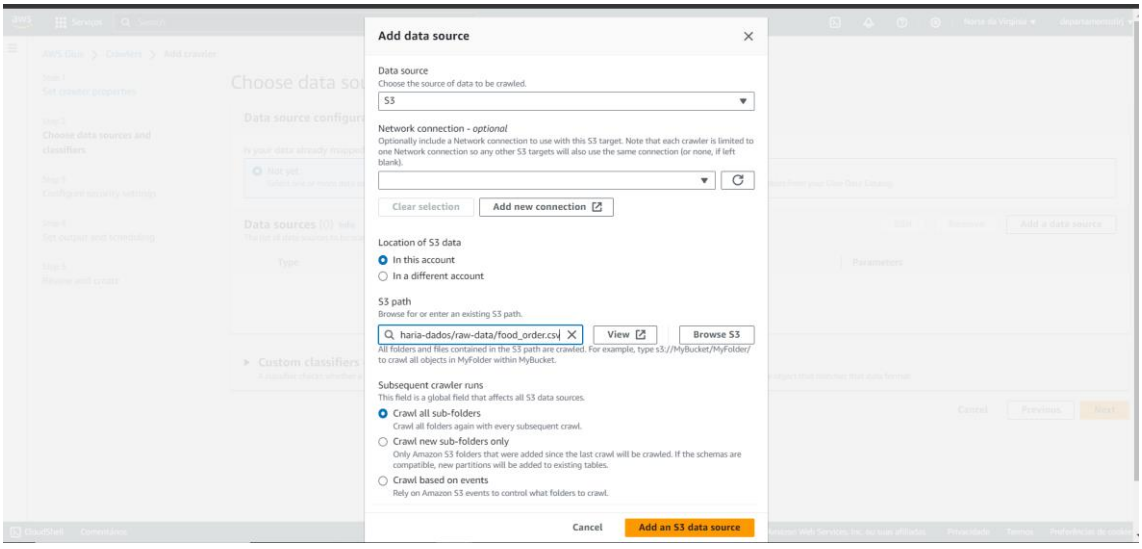
1- A Ingestão de dados no S3 (bucket) foi feita através do upload no arquivo na plataforma. Realizei a criação da pasta (mvp-engdados) e subir o arquivo CSV na casa raw-data, dessa forma consigo organizar os arquivos brutos, aqueles que ainda não foram processados.



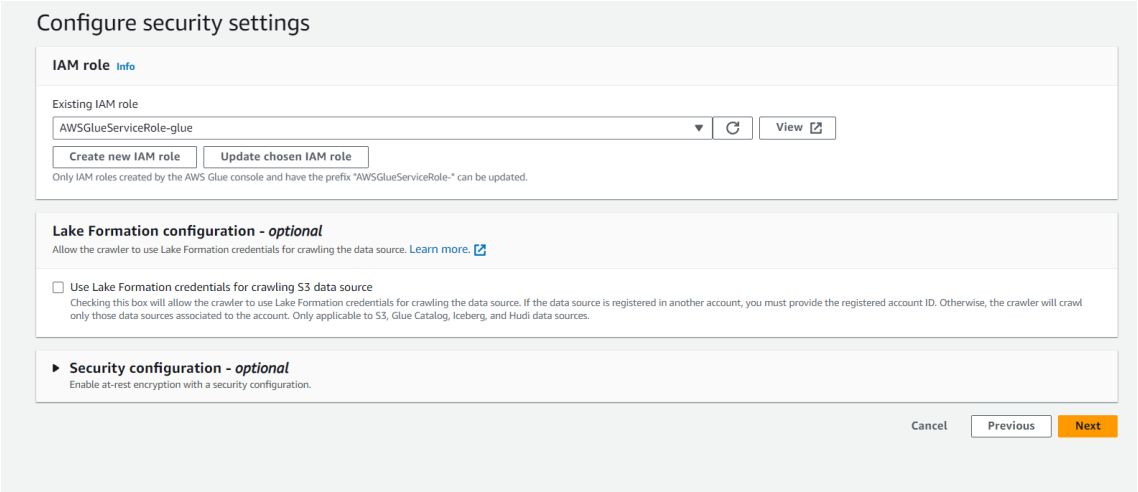
2- Para a próxima etapa, foi necessário criar um crawler , que nada mais é do que uma ferramenta do AWS GLUE onde é possível processar e catalogar os dados e os metadados.



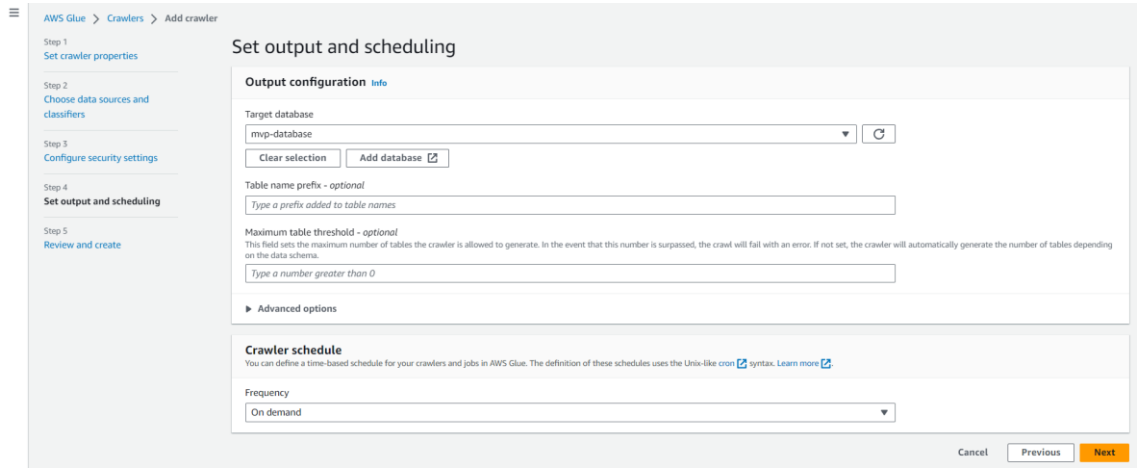
2.1- Nessa etapa será necessário localizar o arquivo ou pasta que contém os dados e informar a localização



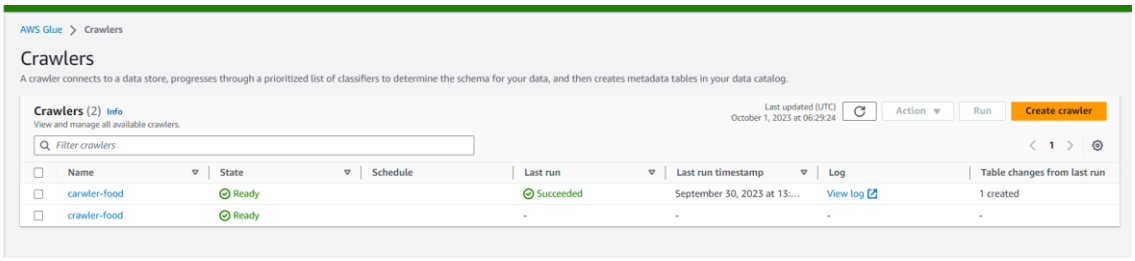
2.2 - – Será necessário configuração algumas permissões de segurança e liberar no IAM para que prosseguíssemos no processo.



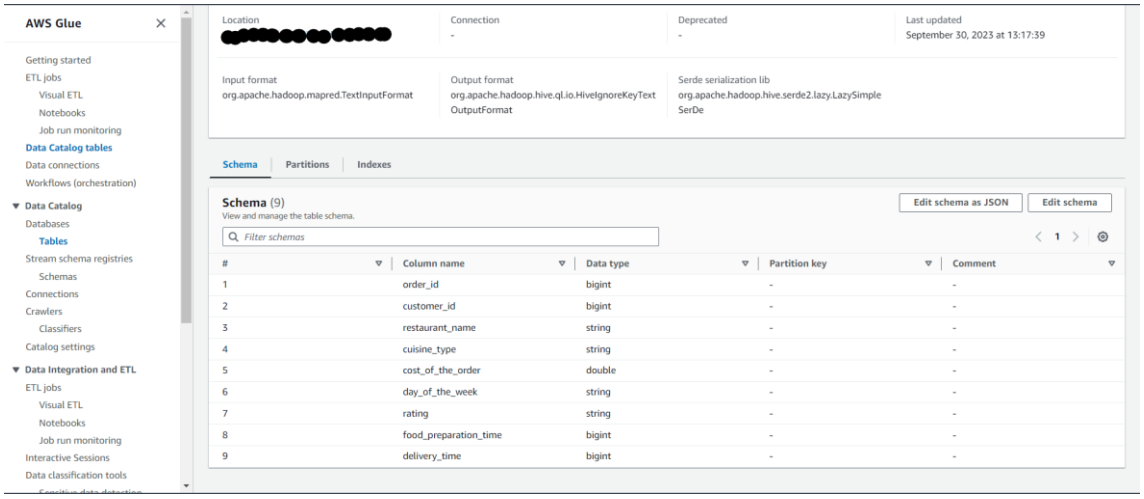
2.2- Teremos também que criar ou apenas informar o local de saída dos dados processados. No meu caso criei uma pasta com o nome “MVP-database”



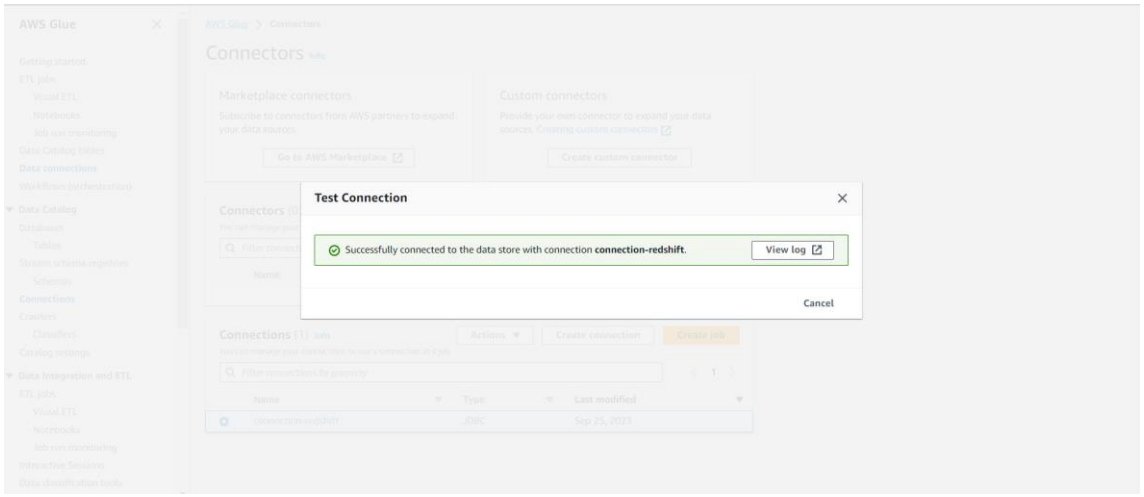
2.3- Após o cralwer criado será necessário executa-lo para que ele possa rodar e fazer o catalogo dos dados.



3 – Após os dados serem catalogados, ele ficam armazenados no “data catalog “ , onde é possível obter informações da tabela processada, como mostrado a imagem abaixo. É possível ter informações do metadados.

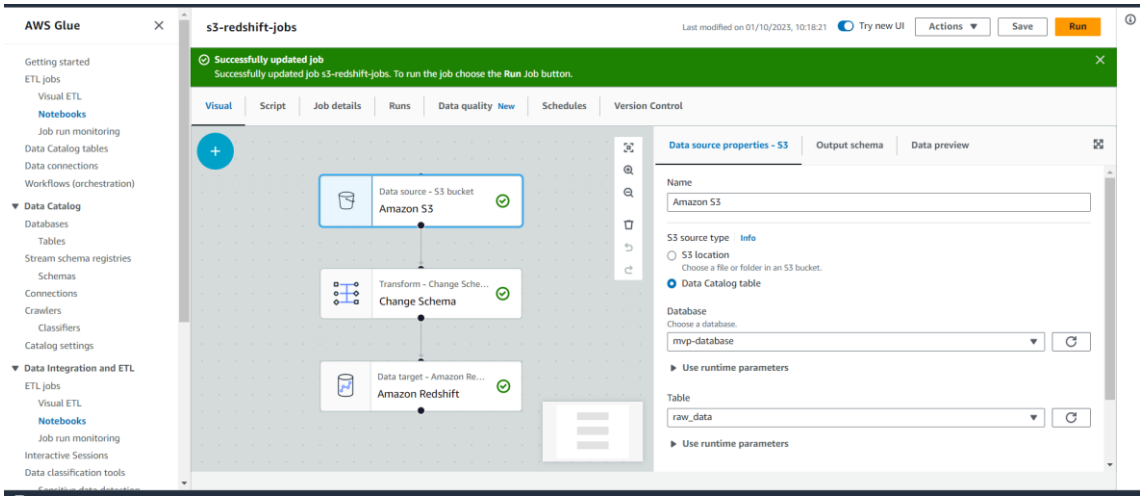


4 – Agora com os dados já processados será necessário criar uma conexão para podermos conectar no redshift através do Job e assim injetar nossos dados dentro do banco. Obs: É sempre uma boa prática testar o conector criado antes de executar o job que irá ser utilizado.

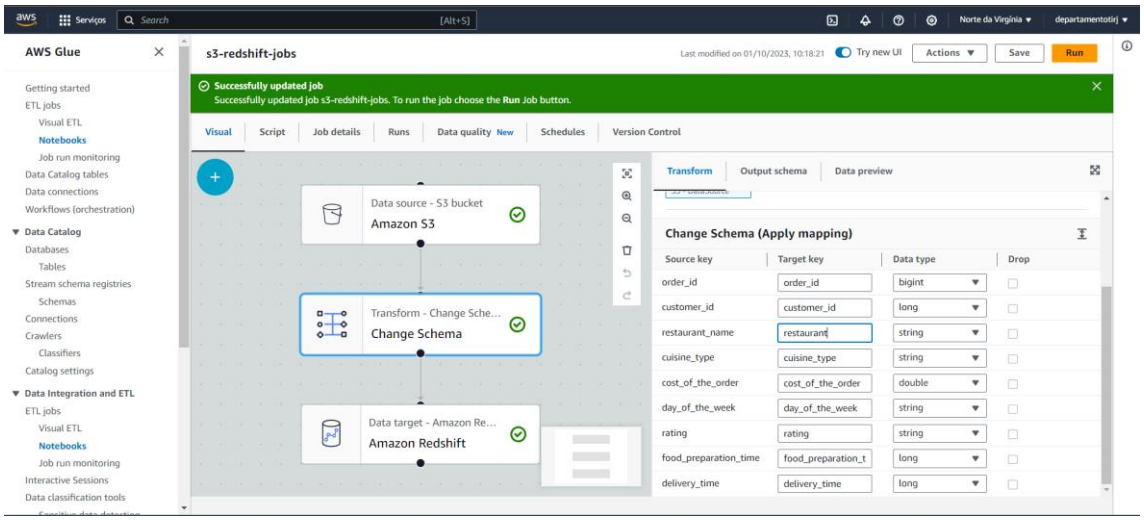


5 – Agora que já subimos o arquivo no S3, já criamos o crawler para catalogar os dados, conseguimos visualizar os dados catalogados e a conexão com o conector foi validada, agora é o momento de criar o Job e subir os dados do S3 para o redshift.

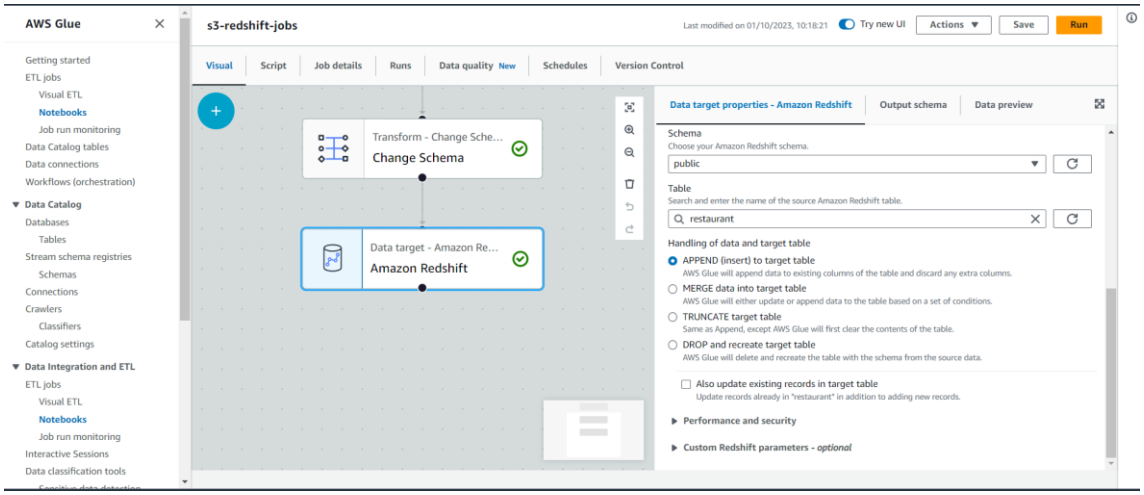
5.1- Para iniciar a criação do Job, adicionei o S3(bucket) onde estão salvo os dados processados e optei por escolher a opção de catalogo de dados, onde está salva a table processada pelo crawler



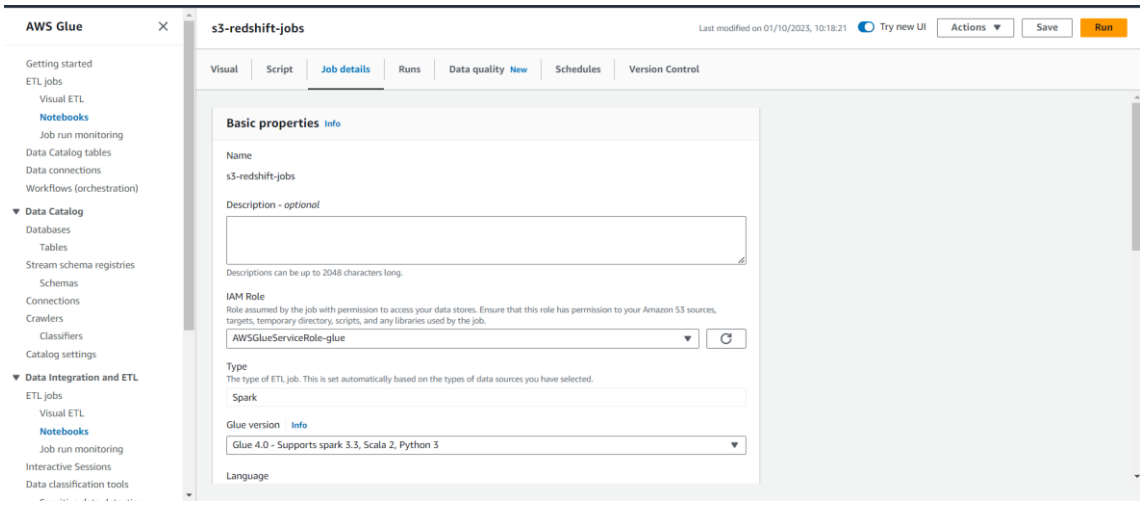
5.2 – Nessa etapa irei aplicar algumas transformações no nome de cada e coluna e também no tipo de dados de cada uma delas.



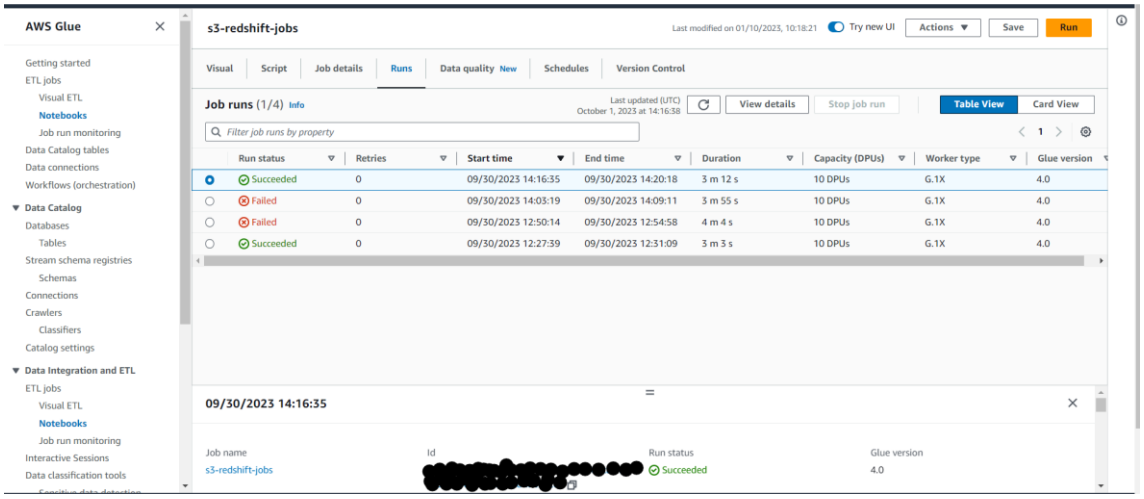
5.3 – Na etapa para salva a saída dos dados, foi necessário selecionar o conector entre o s3 e o Redshift, que foi criado anteriormente, e também selecionar a tabela dentro do redshift que esse novos dados irão pertencer. No meu caso , criei uma tabela chamada “restaurant” dentro do schema “Public” e optei pela opção do Job dar um insert dos dados na tabela especificada.



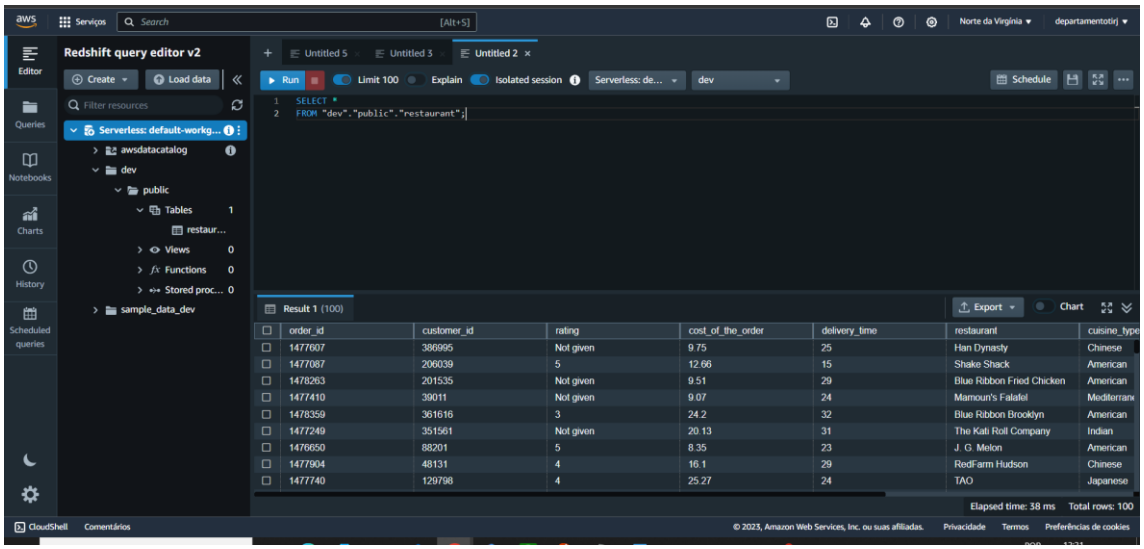
5.4 - Para finalizar o Job e coloca-lo em produção, Precisei acessar a aba “Job Details” e selecionar a IAM ROLE que permite obter esses acessos.



5.5 – Agora basta rodar o Job e aguardar que ele faça a ingestão de dados no redshift.



6 – Após a execução do Jobs, é possível visualizar e realizar consulta dentro do redshift.



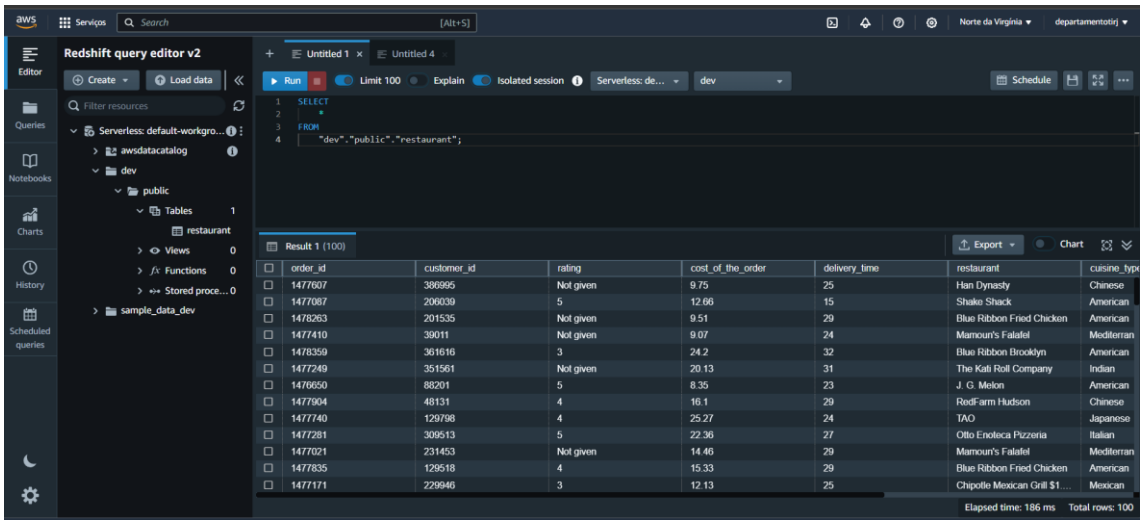
Análise de dados

A análise de dados é o processo de examinar informações para encontrar respostas, identificar tendências e tomar decisões mais inteligentes. Ela envolve olhar para números, fatos e figuras para descobrir o que eles nos dizem. A análise de dados é uma ferramenta valiosa em muitos campos, ajudando a resolver problemas, melhorar processos e impulsionar o sucesso. É como usar um superpoder para entender e agir com base nos dados disponíveis. Com isso, desenvolvi algumas query que é possível extrair informações para novas tomadas de decisões como mostrarei abaixo.

1 – Listando todas informações do dataset e entender como ele está sendo distribuído.

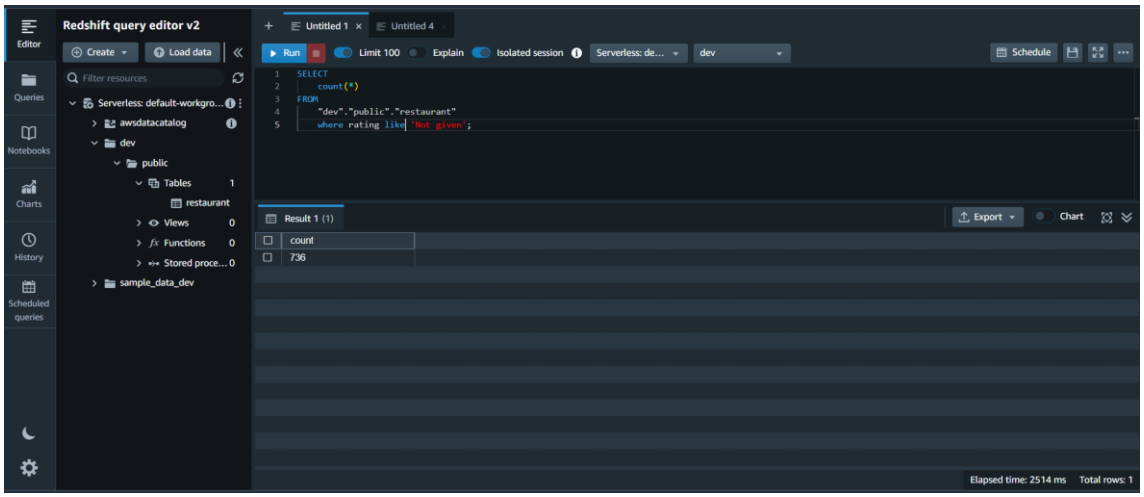
O dataset possui 1898 pedidos.

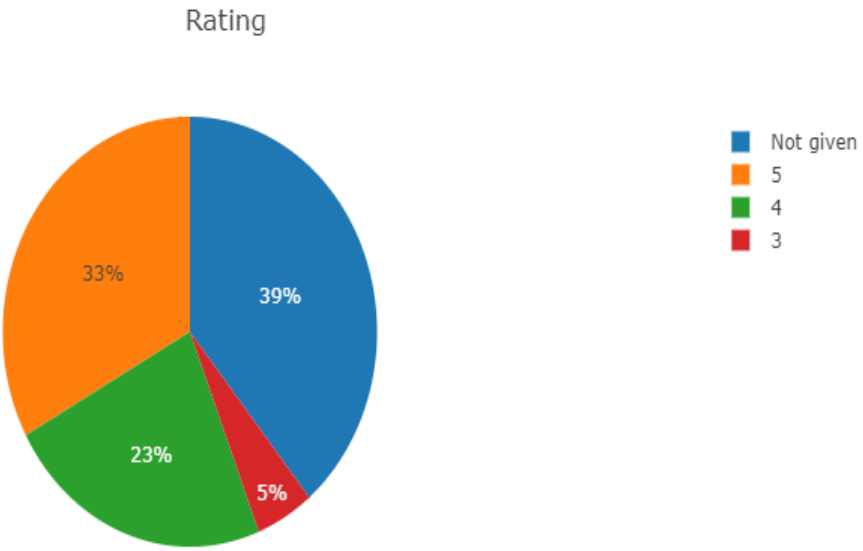
```
SELECT
    *
FROM
    "dev"."public"."restaurant";
```



2 – Todos os pedidos possui alguma tipo de classificação entre 3 e 5 e se o usuário não quiser dar alguma classificação, a classificação fica registrado no banco de dados como “Not given”.

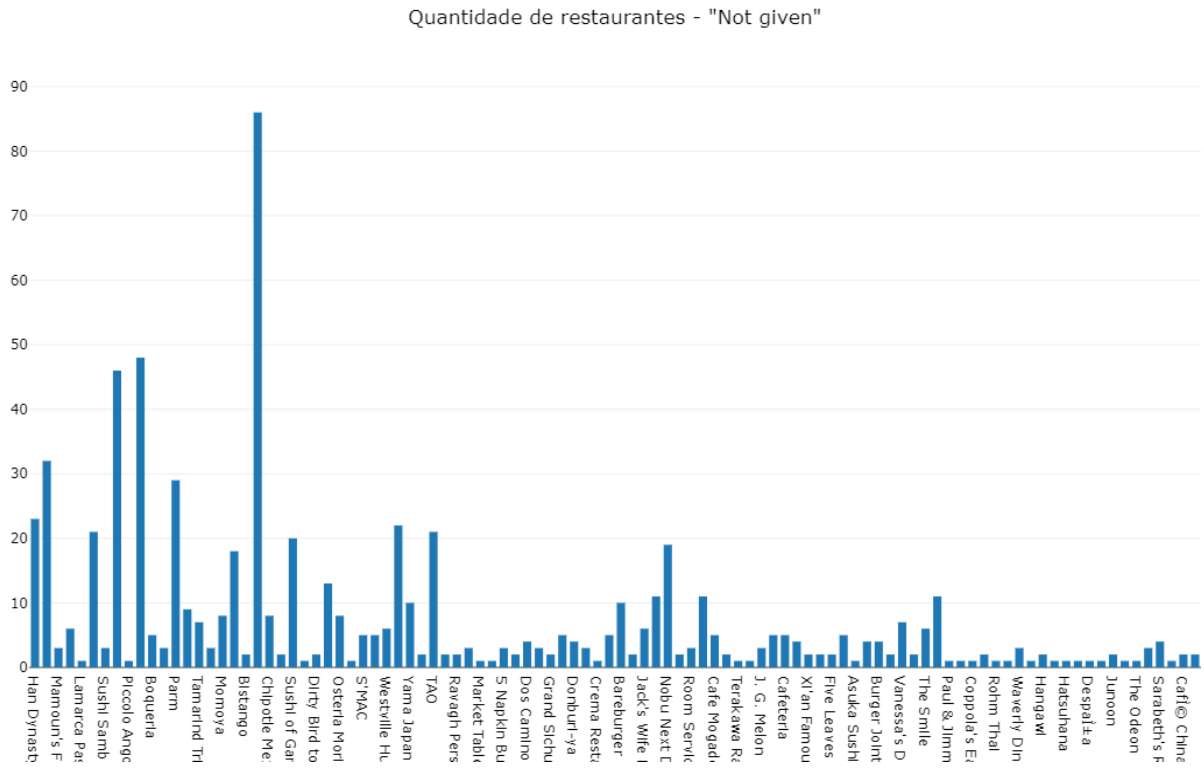
Nessa analise, percebi que a proporção de pedidos não classificados **“ Not given”** chegou a ter uma porcentagem expressiva de 39% que corresponde num total de 736 pedidos.



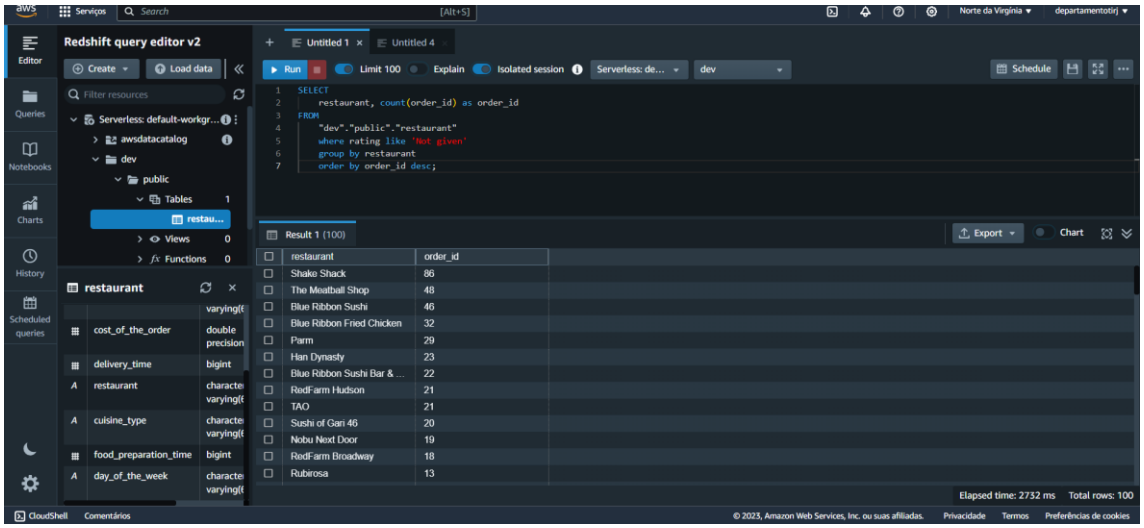


3 - Para entender melhor o motivo dessa alta taxa, listei todo os restaurantes e verifiquei a quantidade de pedidos com a classificaçã o “ Not given” de cada um dele e com isso.

O primeiro restaurante da lista é o **Shake Shack** e possui 86 pedidos sem classificação e o restaurante que está na **10º Sushi of Gari 46** possui 20 pedidos nessa classificação. Existe exatamente 134 restaurantes nesta lista.

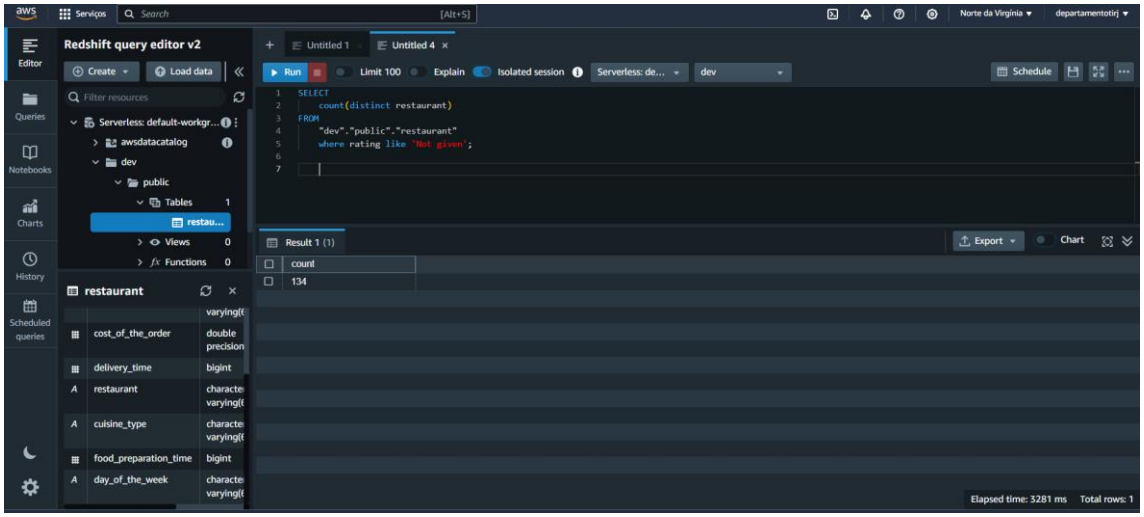



```
SELECT
    restaurant, count(order_id) as order_id
FROM
    "dev"."public"."restaurant"
    where rating like 'Not given'
    group by restaurant
    order by order_id desc;
```

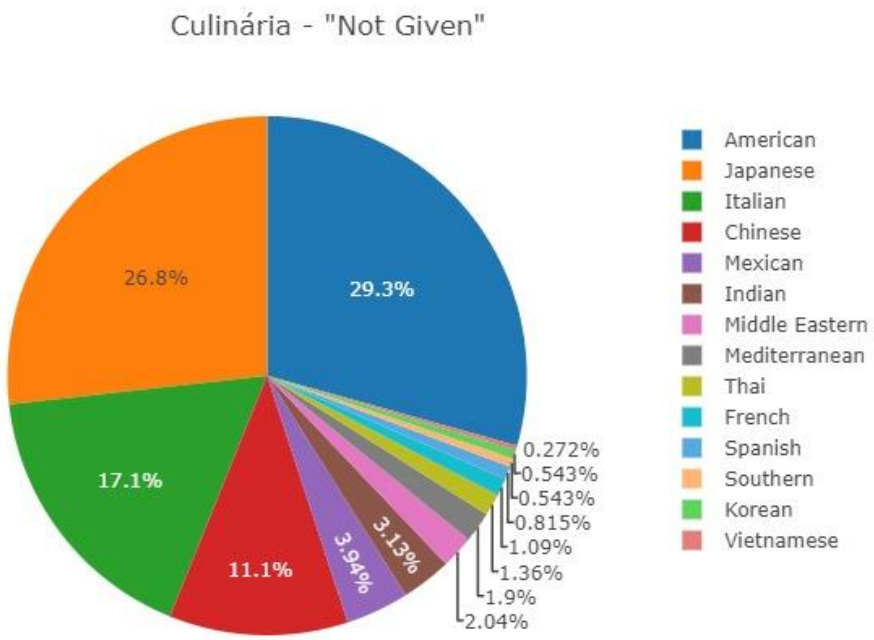


Listando a quantidade de restaurante com possui a classificação “ Not given”

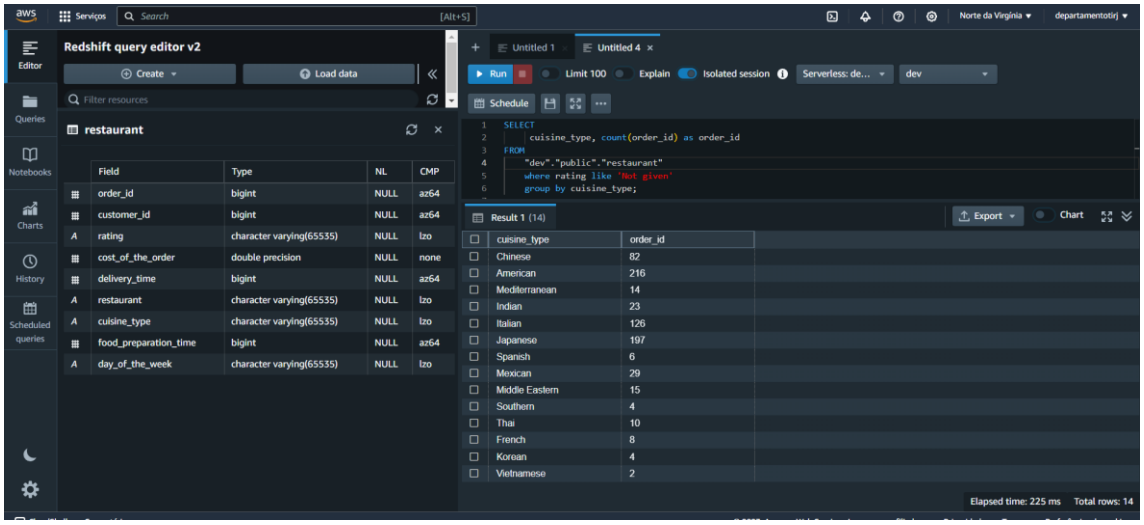
```
SELECT
    count(distinct restaurant)
FROM
    "dev"."public"."restaurant"
    where rating like 'Not given';
```



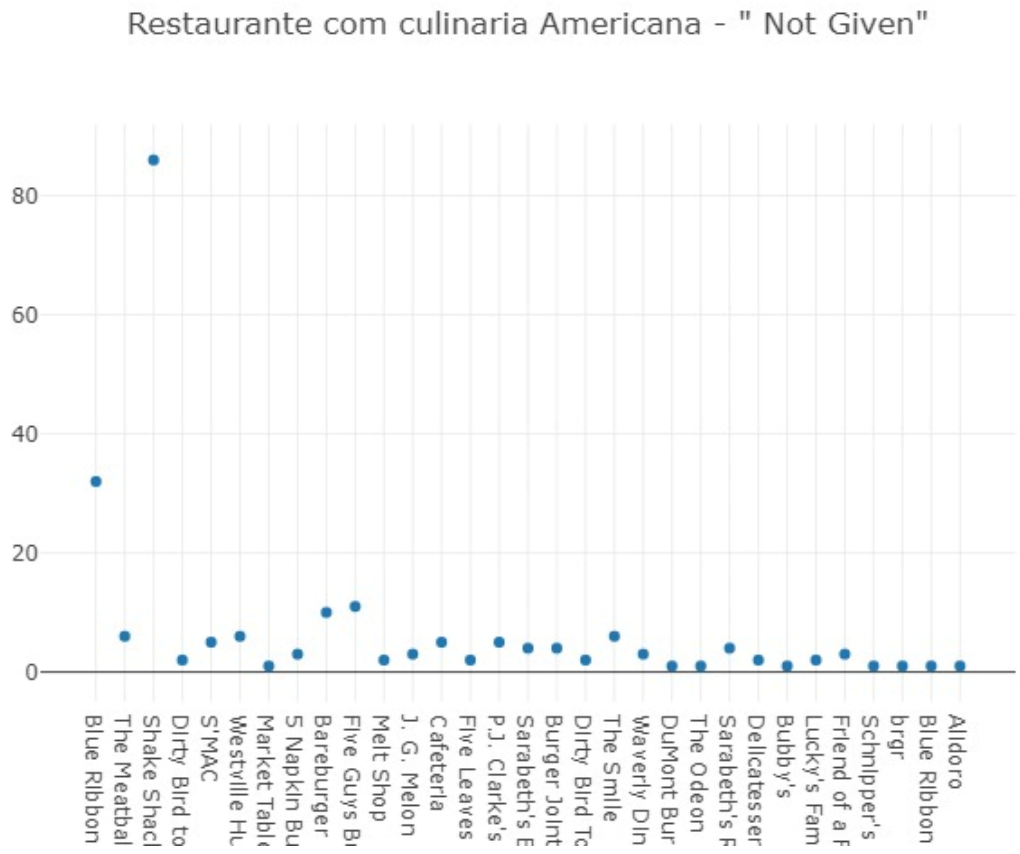
4 – Após analisar os restaurantes, agora diminuir o filtro e pesquisei quais são os tipos de culinária que estão com índice ele levado de “Not given”



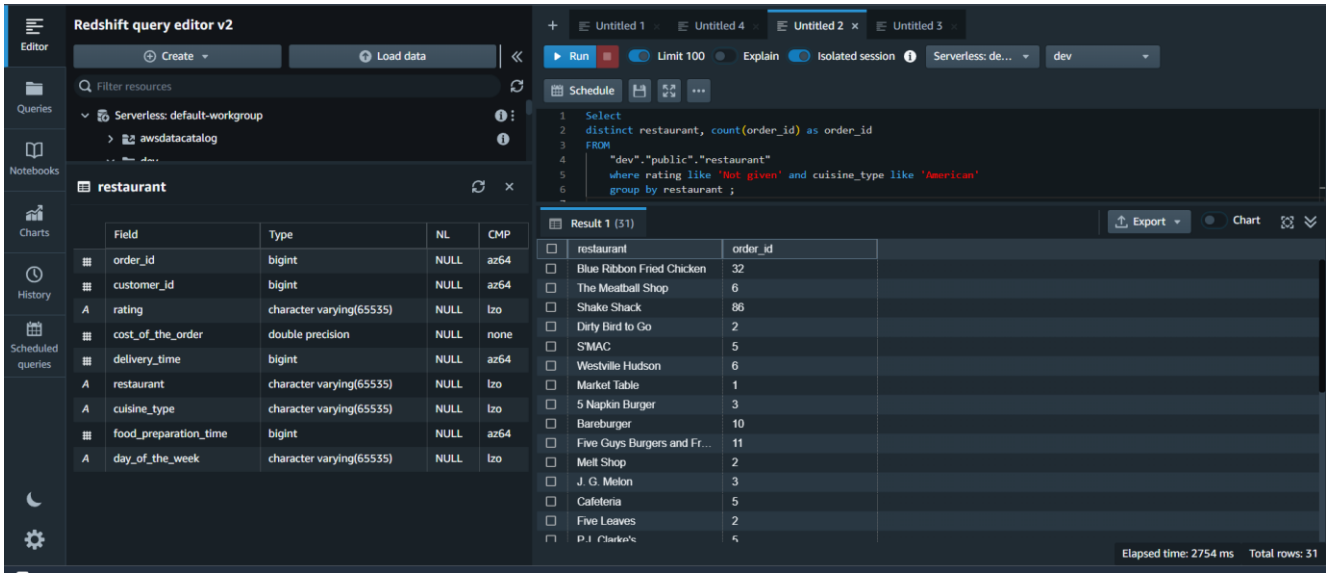
```
SELECT
    cuisine_type, count(order_id) as order_id
FROM
   "dev"."public"."restaurant"
   where rating like 'Not given'
   group by cuisine_type;
```



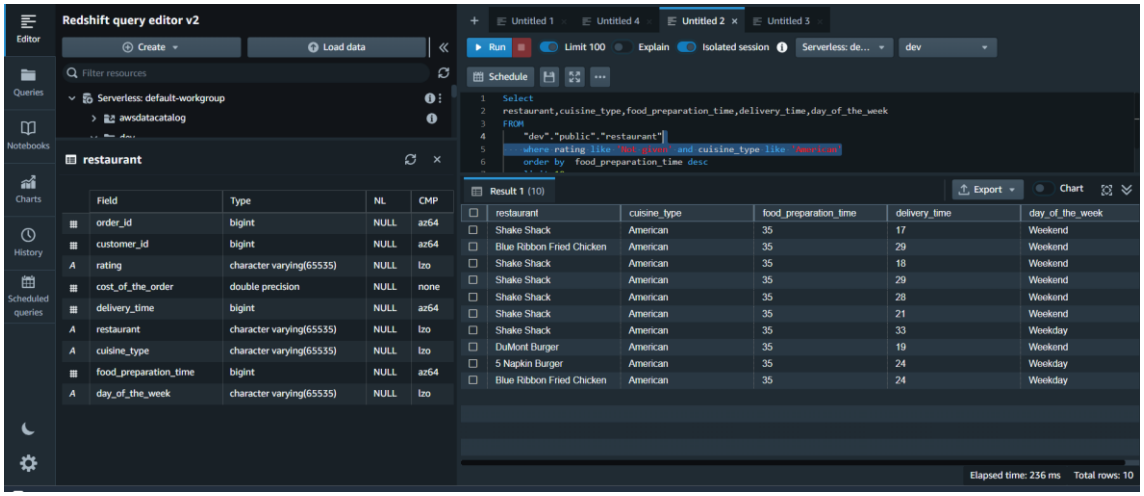
5 – Continuando com a analise mais a fundo, verifiquei pode existem 31 restaurantes que vendem comida America, em um total de 134 restaurantes e que o restaurante **Shake Shack** possui o maior índice de pedido de culinária américa sem classificação.



```
Select
distinct restaurant, count(order_id) as order_id
FROM
    "dev"."public"."restaurant"
    where rating like 'Not given' and cuisine_type like 'American'
    group by restaurant ;
```



6 – A última análise feita em cima dos dados, me mostrou que o motivo do alto índice de “Not Given “ em cima da Culinária Americana está vindo do Restaurante “ Shake Shack” que em outra analise aparecia de forma significativa. Analise também mostrou que entre 10 pedidos com maior tempo de preparo, 6 foram feito pela Shake Shack



```
Select
restaurant,cuisine_type,food_preparation_time,delivery_time,day_of_the_week
FROM
   "dev"."public"."restaurant"
   where rating like 'Not given' and cuisine_type like 'American'
   order by  food_preparation_time desc
   limit 10;
```

Conclusão

Com a análise feita em cima do dataset, objetivo para entender o motivo da alta classificação dos pedido com “ Not given” se deu pela Culinaria américa vim do Restaurante Shake Shack, onde nele o tempo de preparação dos alimentos alcançou o máximo de 35 min, tempo que se reflete na maioria dos pedidos do restaurante.