

MVP – Engenharia de dados – Matheus Motta Dias Ferreira

Objetivo

O meu objetivo é explorar o conjunto de dados para traçar o perfil dos participantes da campanha de marketing realizada e responder as seguintes perguntas usando apenas queries em SQL:

Qual cliente mais velho e mais novo contatado que aceitaram a proposta de investimento?

Qual a média de idade das pessoas contatadas?

Quantas pessoas foram contatadas, quantas fizeram o investimento, quantas não fizeram e a taxa de sucesso?

Qual a duração média de chamada que a campanha teve sucesso e qual tempo médio que não teve sucesso?

Qual saldo médio da conta das pessoas que fizeram o investimento e saldo médio da conta das pessoas que não fizeram o investimento?

Levando em conta emprego, nível de educação e idade, qual demográfico a campanha teve mais sucesso?

Detalhamento

Optei por usar o mesmo conjunto de dados do meu MVP da sprint II, visto que já tenho familiaridade com os dados, pois eles já foram pré-processados por mim, por isso, posso atestar sua qualidade e a integridade.

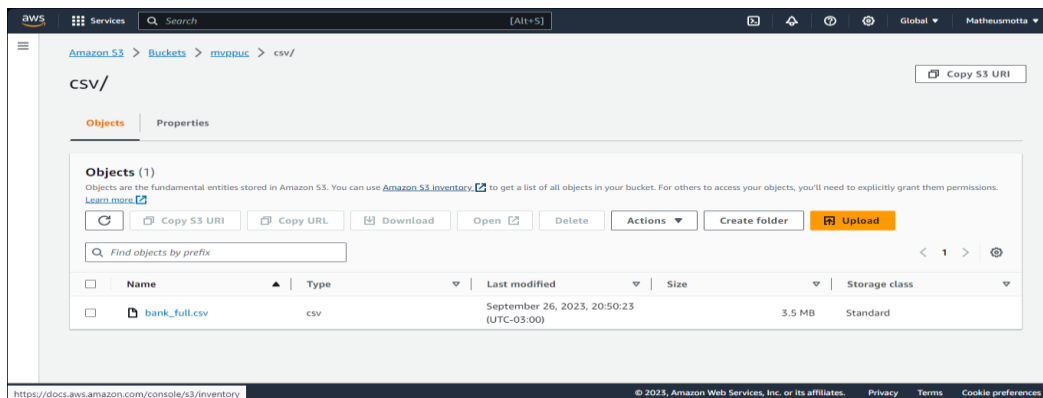
Os dados são referentes a uma campanha de marketing executada através de ligações telefônicas feitas por representantes um banco português. O objetivo de campanha é atrair cliente para realização de investimentos na modalidade de prazo fixo (term deposit).

Nesse MVP vou utilizar exclusivamente as ferramentas da AWS para guardar os dados em nuvem (bucket S3), realizar o job de ETL (glue) e fazer consultas em SQL (redshift)

1. Processo de ETL utilizando Glue

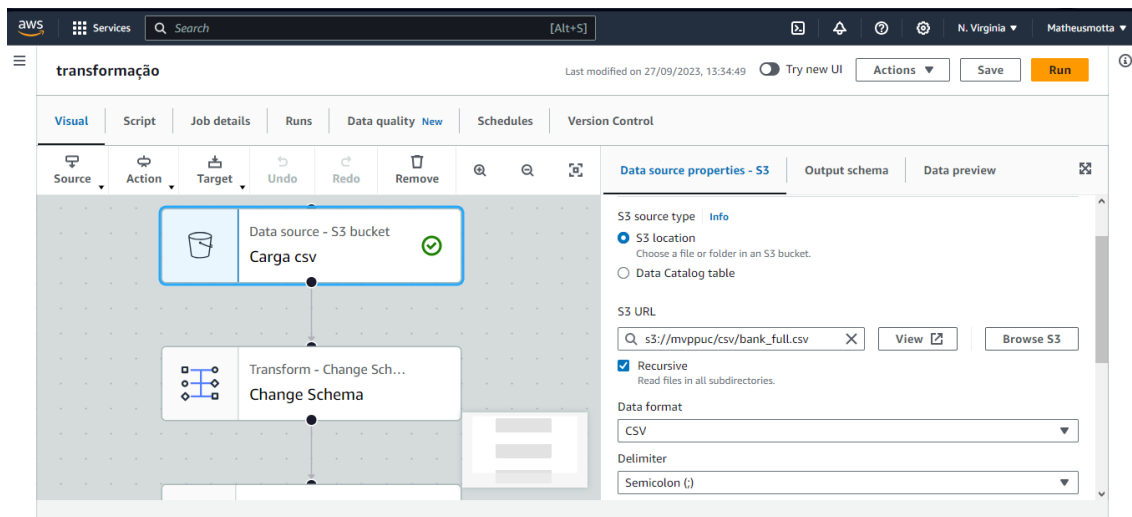
Utilizarei a interface visual do glue para realizar o job de ETL, mas antes é necessário subir o conjunto de dados a ser trabalhado para o ambiente cloud da Amazon.

O conjunto de dados em CSV já se encontrava em um diretório local e foi inserido em um bucket do S3



1.1 Processo de extração

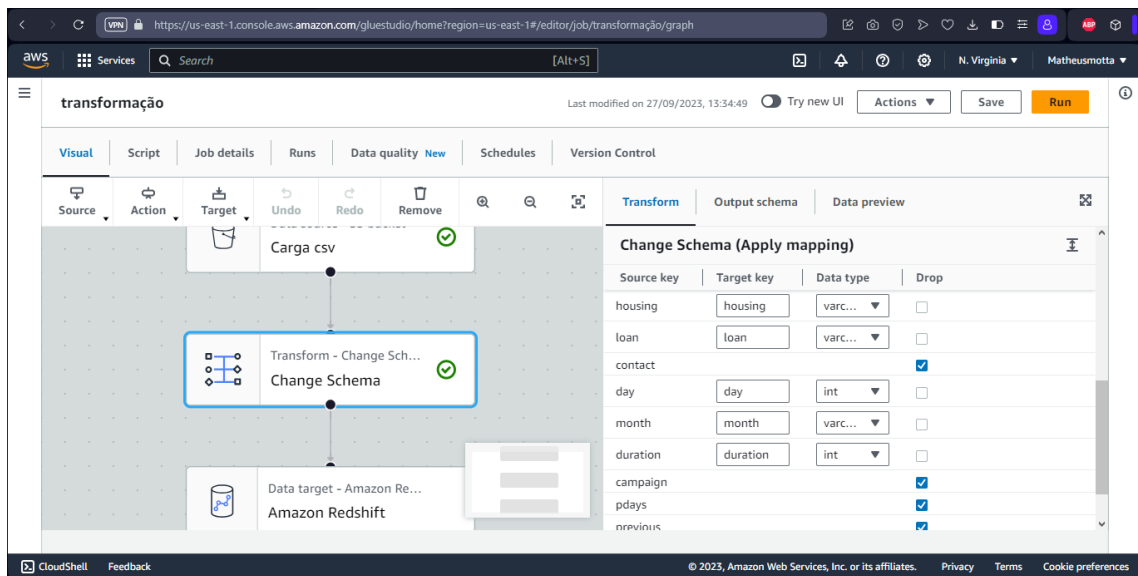
Nessa etapa foram “setados” os parâmetros da fonte de dados como local da fonte de dados, formato de dados, delimitadores e se a primeira linha funciona como cabeçalho.



1.2 Transformação

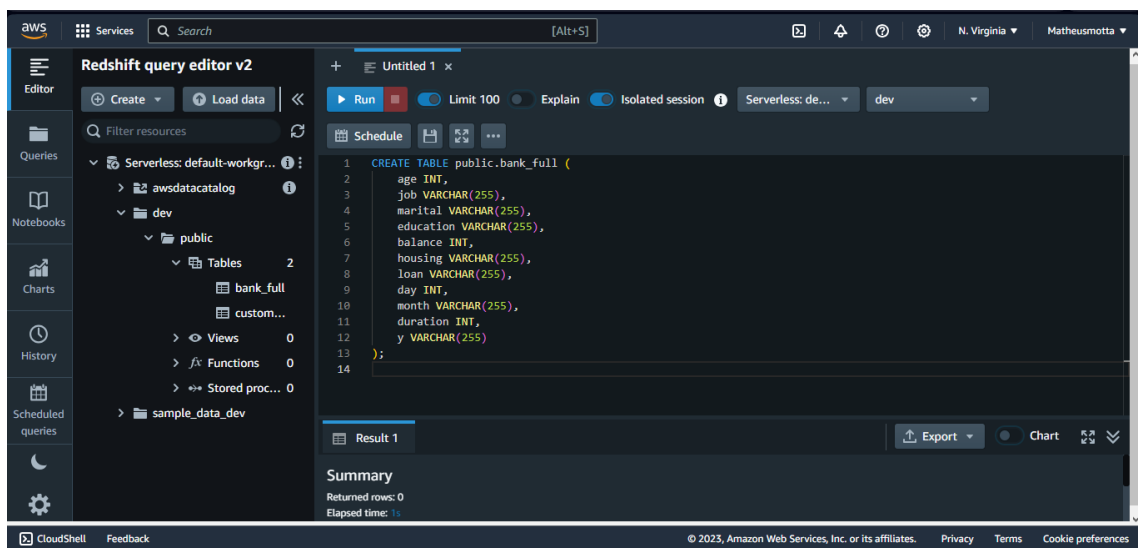
Nessa segunda etapa é realizado um trabalho de transformação (modelagem) dos dados contidos no CSV. Nessa etapa são definidos os tipos de dados contidos em cada coluna da tabela e quais colunas serão transportadas para o alvo, que nesse caso é o Redshift.

Defini as colunas com valores numéricos como “int” para realizar cálculos utilizando consultas em SQL e defini colunas que contém apenas texto como “varchar”. Optei por não carregar algumas colunas, pois elas não são tão relevantes para a análise a ser realizada.



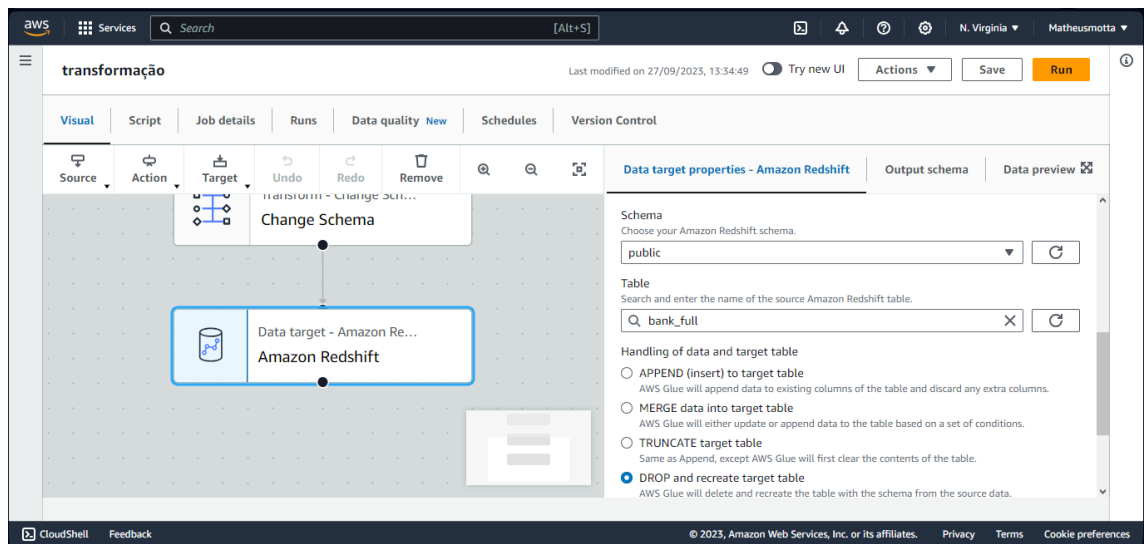
1.3 Carga de dados e criação da tabela no Redshift Query Editor

Nessa terceira etapa, é necessário definir o “target” dessa fonte de dados, por isso criei uma tabela no Redshift utilizando a seguinte query:



Agora podemos “setar” essa tabela como destino no Glue, assim como definir outros parâmetros como juntar dados novos a tabela, limpar a tabela e acrescentar os dados previamente “setados” ou apagar a tabela e recria-la utilizando o “schema” da fonte de dados.

Nesse caso, como a tabela estava vazia, optei por usar a função DROP para evitar possíveis conflitos de “schema” ao carregar os dados na tabela.



O print a seguir demonstra a carga bem sucedida, além de outros testes feitos anteriormente.

transformação

Last modified on 27/09/2023, 13:34:49 Try new UI Actions Save Run

Visual Script Job details Runs Data quality New Schedules Version Control

Job runs (1/6) Info

Last updated (UTC) September 28, 2023 at 01:37:06 View details Stop job run

Table View Card View

Filter job runs by property

Run status	Retries	Start time	End time	Duration	Capacity (DPUs)	Worker type	Glue version
✓ Succeeded	0	09/27/2023 13:34:56	09/27/2023 13:37:15	2 m 3 s	10 DPUs	G.1X	4.0
✓ Succeeded	0	09/27/2023 13:06:30	09/27/2023 13:08:39	1 m 46 s	10 DPUs	G.1X	4.0
✓ Succeeded	0	09/27/2023 12:58:22	09/27/2023 13:01:40	3 m 1 s	10 DPUs	G.1X	4.0
✓ Succeeded	0	09/26/2023 22:49:03	09/26/2023 22:51:42	2 m 21 s	10 DPUs	G.1X	4.0
✓ Succeeded	0	09/26/2023 22:40:45	09/26/2023 22:43:10	2 m 9 s	10 DPUs	G.1X	4.0

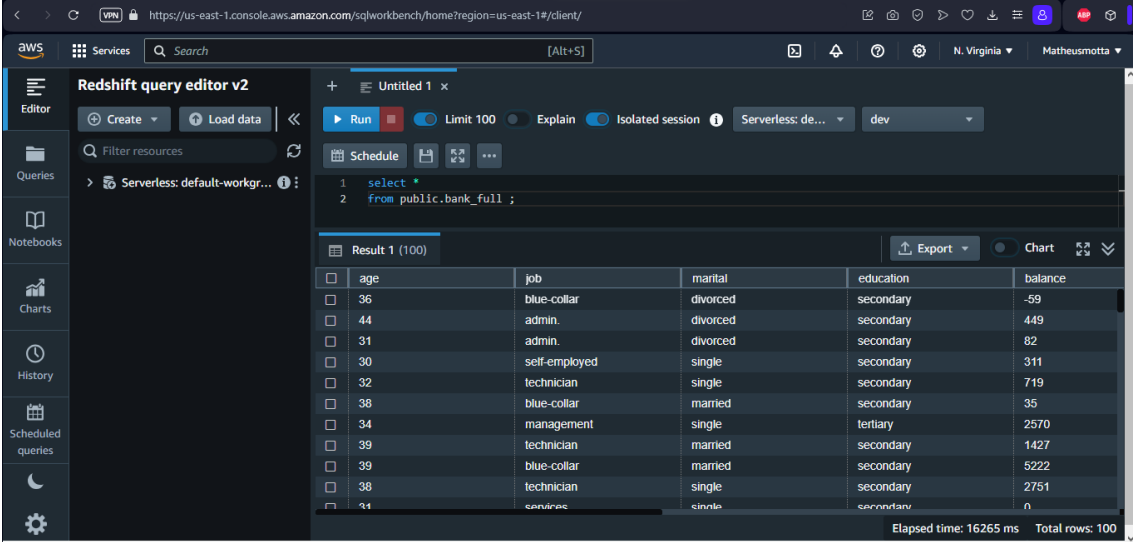
09/27/2023 13:34:56

Job name Id Run status Glue version

transformação jr_9105ff478c4f7a1147042e464fcb6b237e2f 68809b6ba515fcd40feea27fcd79 ✓ Succeeded 4.0

2. Analise

A query a seguir mostra que a tabela foi criada com sucesso e apresenta os seguintes campos:



	age	job	marital	education	balance
36		blue-collar	divorced	secondary	-59
44		admin.	divorced	secondary	449
31		admin.	divorced	secondary	82
30		self-employed	single	secondary	311
32		technician	single	secondary	719
38		blue-collar	married	secondary	35
34		management	single	tertiary	2570
39		technician	married	secondary	1427
39		blue-collar	married	secondary	5222
38		technician	single	secondary	2751
31		retired	single	secondary	0

1 - age (int)

2 - job (varchar: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')

3 – marital (varchar: 'divorced', 'married', 'single', 'unknown'; obs: 'divorced' pode significar divorciado ou viúvo)

4 - Education (varchar: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')

5 – balance (int)

6 – housing, tem imóvel financiado? (varchar: 'no', 'yes', 'unknown')

7 – loan, tem emprestimo pessoal? (varchar: 'no', 'yes', 'unknown')

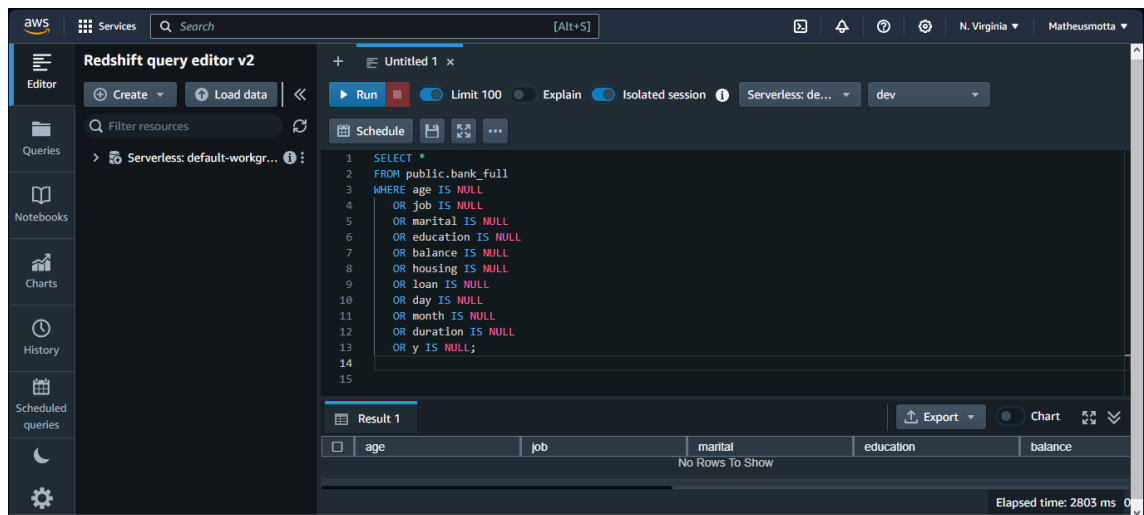
8 – day (int)

9 – Month (varchar)

10 – Duration, duração da chamada em segundos (int)

11 – y, essa coluna aponta se o cliente fez o investimento (1) ou não (0) (int)

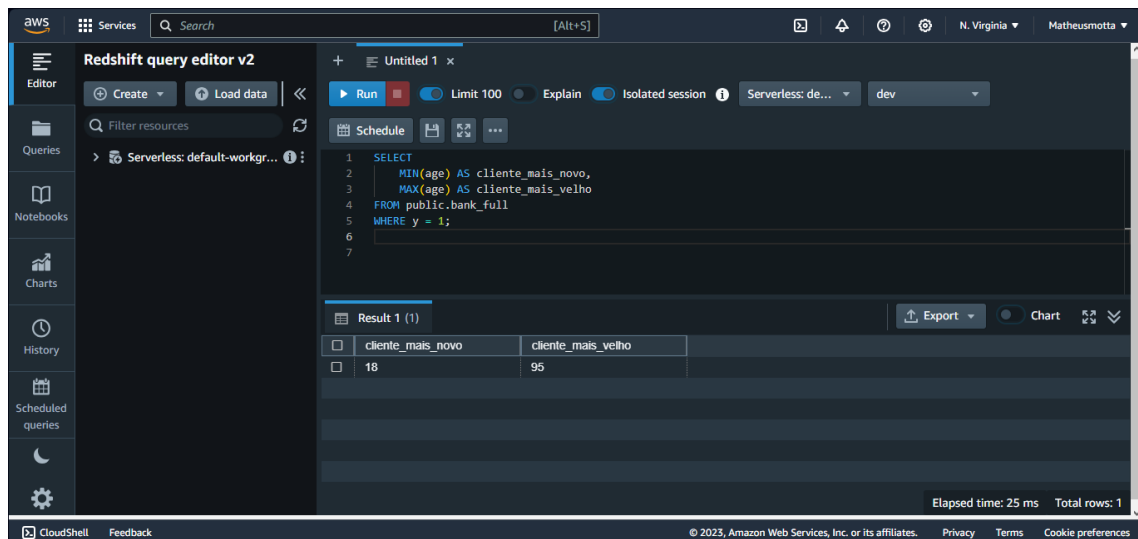
Por esse ser um conjunto de dados pré-processado por mim, posso atestar a qualidade e integridade dos dados, posso comprovar que não existem valores em branco ou “nulos” em toda a tabela através da seguinte query:



Essa query é responsável por buscar valores em branco em cada uma das colunas, e caso haja algum valor em branco ou “nulo”, a query retornará a linha com o valor em branco ou “nulo”. Note que a query foi executada com sucesso e que o resultado é “No Rows To Show” atestando a qualidade dos dados inseridos na tabela.

3. Respondendo perguntas elencadas no objetivo

Qual cliente mais velho e mais novo contatado que aceitaram a proposta de investimento?



Qual a média de idade das pessoas contatadas?

The screenshot shows the AWS Redshift Query Editor v2 interface. The query editor is open with a single query in 'Untitled 1'.

```
1 SELECT AVG(age) AS media_idade
2 FROM public.bank_full
3
4
```

The query has been executed, and the results are displayed in a table with one row:

media_idade
40

The interface also shows a sidebar with navigation options like Editor, Queries, Notebooks, Charts, History, and Scheduled queries. The bottom status bar indicates 'Elapsed time: 168 ms' and 'Total rows: 1'.

Quantas pessoas foram contatadas, quantas fizeram o investimento, quantas não fizeram e a taxa de sucesso?

The screenshot shows the AWS Redshift Query Editor v2 interface with a more complex query in 'Untitled 2'.

```
1 SELECT
2   COUNT(*) AS total_pessoas,
3   SUM(CASE WHEN y = 1 THEN 1 ELSE 0 END) AS sucesso,
4   SUM(CASE WHEN y = 1 THEN 0 ELSE 1 END) AS falha,
5   (SUM(CASE WHEN y = 1 THEN 1 ELSE 0 END) * 1.0 / COUNT(*)) AS taxa_sucesso
6 FROM public.bank_full;
7
```

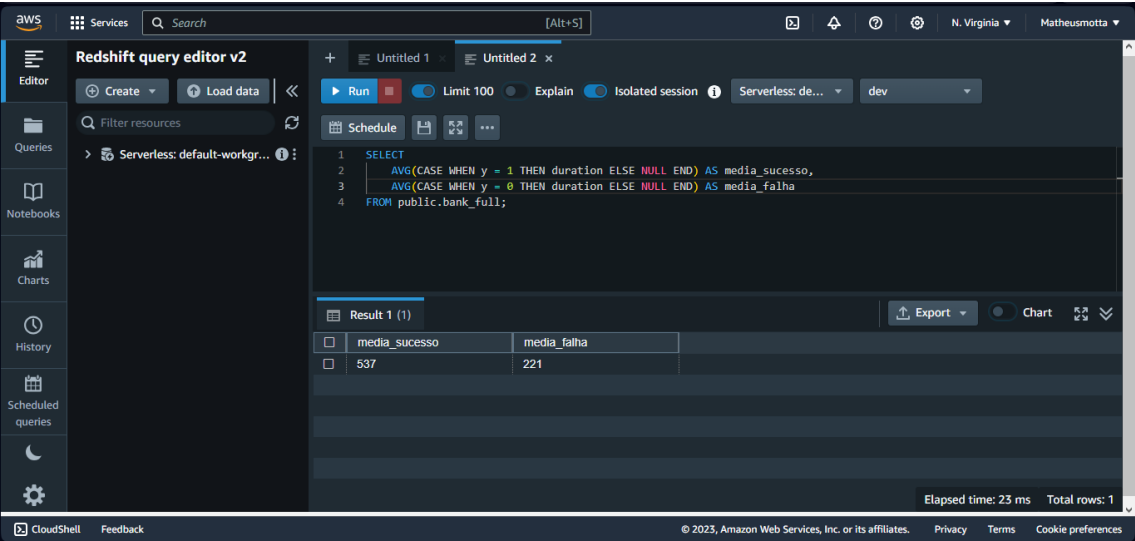
The query has been executed, and the results are displayed in a table with one row:

total_pessoas	sucesso	falha	taxa_sucesso
45211	5289	39922	0.11698480458295547

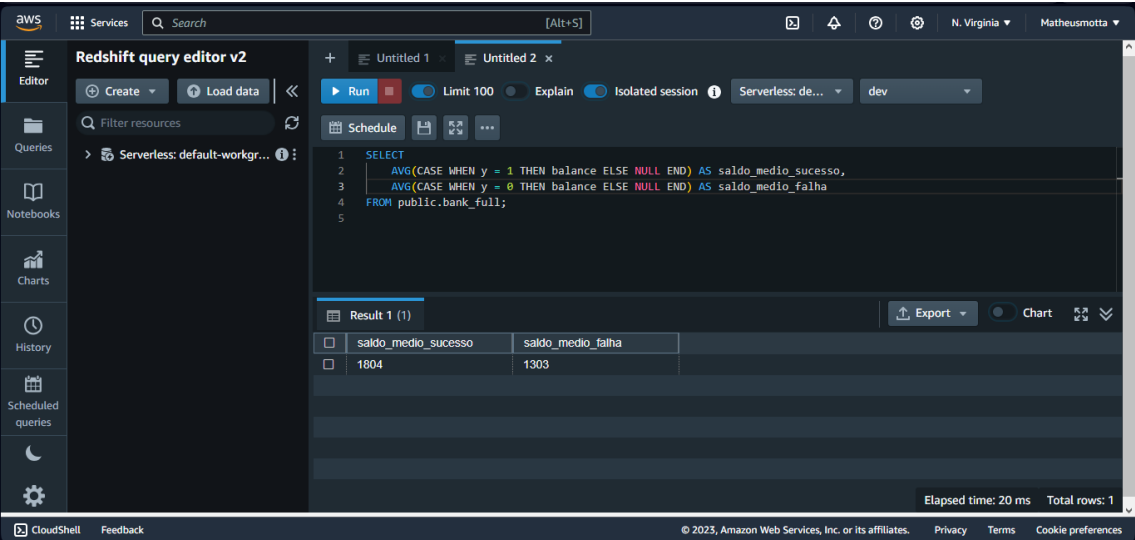
The interface also shows a sidebar with navigation options like Editor, Queries, Notebooks, Charts, History, and Scheduled queries. The bottom status bar indicates 'Elapsed time: 135 ms' and 'Total rows: 1'.

A query anterior apresenta um nível de complexidade ligeiramente maior que as outras, uma vez que são realizadas expressões algébricas. Na quinta linha da query é utilizada a multiplicação pelo fator 1.0 para garantir que a divisão seja feita como “FLOAT”, para que seja obtido um resultado decimal, já que os números estão “setados” como “int”. Logo, obtivemos resultado de 11,6% de taxa de sucesso.

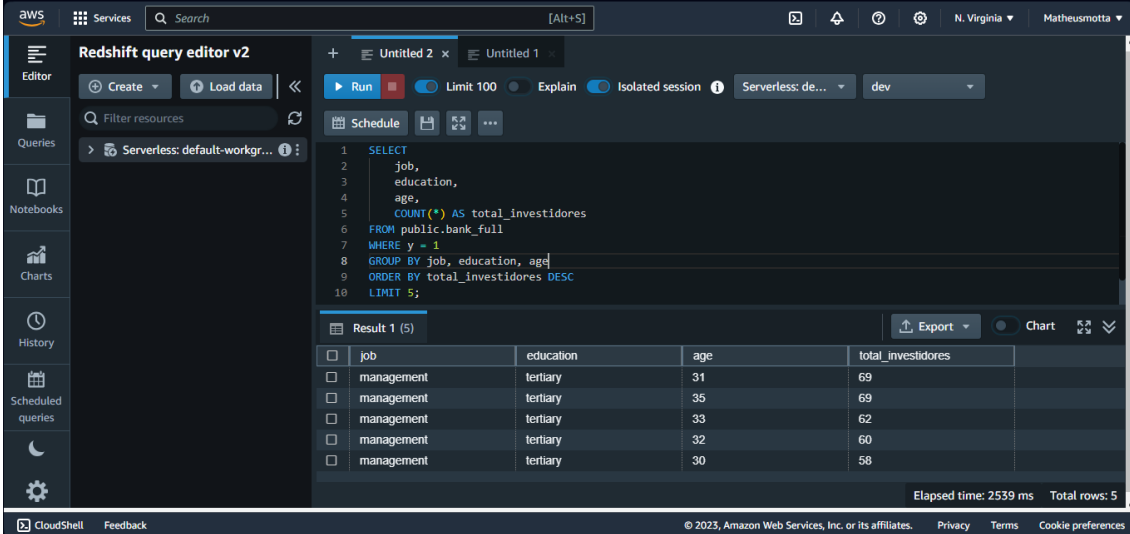
Qual a duração média de chamada que campanha teve sucesso e qual tempo médio que não teve sucesso?



Qual saldo médio da conta das pessoas que fizeram o investimento e saldo médio da conta das pessoas que não fizeram o investimento?



Levando em conta emprego, nível de educação e idade, qual demográfico a campanha teve mais sucesso?



The screenshot shows the AWS Redshift Query Editor v2 interface. The SQL query is as follows:

```
1 SELECT
2   job,
3   education,
4   age,
5   COUNT(*) AS total_investidores
6 FROM public.bank_full
7 WHERE y = 1
8 GROUP BY job, education, age
9 ORDER BY total_investidores DESC
10 LIMIT 5;
```

The results are displayed in a table with 5 rows and 5 columns:

job	education	age	total_investidores	
management	tertiary	31	69	
management	tertiary	35	69	
management	tertiary	33	62	
management	tertiary	32	60	
management	tertiary	30	58	

Elapsed time: 2539 ms Total rows: 5

Conclusão

A partir dos resultados das consultas em SQL podemos inferir que o quanto maior a duração da chamada, maiores são as chances de o cliente aceitar a proposta de investimento, a ligações de sucesso tem duração média aproximada de 10 minutos, enquanto as sem sucesso levam menos da metade desse tempo. Apesar de fazer contato com clientes entre 18 e 95 anos, e ambas as idades aceitarem as propostas de investimento, a idade média da taxa de sucesso é de 40 anos. Quando levamos em conta o saldo da pessoa em banco, podemos inferir que quanto mais dinheiro em conta, maiores são as chances da pessoa realizar o investimento. Além disso, ao rodar a query agrupando a taxa de sucesso da campanha levando em conta a profissão, nível de educação e idade do cliente, podemos concluir que quanto mais alto o nível de educação, mais sucesso profissional, maiores são as chances de sucesso da ligação.

Apesar de simples, essas queries nos retornam informações valiosíssimas a respeito do perfil dos clientes e taxa de sucesso. Ademais, são capazes de fornecer insights para numa nova campanha de escopo parecido, como por exemplo, dar uma atenção especial para esses clientes, dessa maneira, os custos para a campanha podem ser reduzidos, esforços podem ser poupados e a taxa de sucesso seria possivelmente maior.

Autoavaliação

Todas as perguntas feitas no início do trabalho foram respondidas com precisão, enfrentei algumas barreiras durante o processo de criação de “roles” e conexões entre o bucket S3 e o Glue, mas nada que me impedisse de continuar o trabalho. Utilizar uma base de dados conhecida e trabalhada anteriormente para realizar esse trabalho pode me trazer grandes benefícios no que tange a portfólio, pois assim consigo atacar e explorar esse conjunto de dados de várias maneiras diferentes e montar um portfólio completo, desde soluções de armazenamento em nuvem, passando por algoritmos de Machine Learning para solucionar problemas de classificação até um processo de análise de dados mais profunda, que será tema da próxima sprint.