PUC-Rio
Pós-Graduação em Ciência de Dados e Analytics
Departamento de informática

Engenharia de Dados

Aluno: Vinicius Mattoso Reis da Silva

Setembro
2023

PUC-Rio
Pós-Graduação em Ciência de Dados e Analytics
Departamento de informática

Engenharia de Dados

Relatório Final da Sprint de Engenharia de Dados do Curso de pós graduação em Ciência de Dados e Analytics do departamento de informática da PUC-Rio.

Aluno: Vinicius Mattoso Reis da Silva

Setembro
2023

Conteúdo

1 Descrição dos requerimentos do projeto

Neste trabalho, você deverá ser capaz de construir um pipeline de dados utilizando tecnologias na nuvem. O pipeline irá envolver a busca, coleta, modelagem, carga e análise dos dados.

1.1 Objetivo

Comece pelo objetivo do seu trabalho. Antes de iniciar sua busca pelos dados, pense e descreva claramente qual problema deseja resolver com este MVP. Enumere as perguntas que deseja responder.

É de extrema importância que esta etapa seja feita antes de iniciar qualquer outra etapa.

Uma vez traçado o objetivo e conhecendo bem qual problema se deseja resolver, quais perguntas se deseja responder, é hora de iniciar a busca pelos dados.

Não é necessário atingir todos os objetivos desenhados nesta seção. Assim, não remova perguntas as quais não se conseguiu responder. Deixe a documentação do objetivo intacta e faça uma boa discussão do atingimento deste ao final do trabalho (vide Autoavaliação).

1.2 Detalhamento

1.2.1 Busca pelos dados

Escolha uma base de dados para utilizar em seu MVP de forma que se possa atingir os objetivos traçados na etapa anterior.

1.2.2 Coleta

Uma vez definido o conjunto de dados, devemos coletar e armazená-los na nuvem.

1.2.3 Modelagem

Você deve construir um modelo de dados em Esquema Estrela ou Snowflake, como em um Data Warehouse, ou flat por cada conceito, como em um Data Lake.

1.2.4 Carga

Nesta etapa, será feita a carga dos dados para o Data Warehouse/Data Lake. Na avaliação, nesta etapa, será dado valor pela utilização da ferramenta de ETL (Extração, Transformação e Carga) da plataforma de nuvem utilizada.

1.2.5 Análise

Vamos dividir a etapa de análise em duas: qualidade de dados e solução do problema.

Qualidade de dados

Deve ser feita uma análise da qualidade para cada atributo do conjunto de dados. Existem problemas no conjunto de dados? Caso haja, como esses problemas podem ser resolvidos para que não afetem as respostas das perguntas que quer solucionar?

Solução do problema

Chegou o momento de se solucionar o problema em questão, definido preliminarmente nos objetivos. Deve-se buscar respostas para as perguntas elencadas. Para cada resposta obtida tecnicamente através da análise dos dados deve haver uma discussão do seu resultado, conectando os números obtidos às respostas ao problema a ser solucionado.

2 Objetivo

O Objetivo desse trabalho vai ser avaliar qual é o poço localizado no campo Volvo que possui a maior produção. Para isso, será selecionado 3 poços produtores localizados nesse campo e para fazer as comparações iremos responder as seguintes perguntas:

1. Qual poço possui a maior produção de óleo, água e gás?
2. Qual poço possui a maior razão entre o volume de óleo e água e a maior razão entre o volume de gás e óleo?
3. Qual foi o dia mais produtivo de cada poço considerando os diferentes materiais produzidos?

3 Busca pelos dados

Para a criação deste MVP, foram utilizados os dados de produção do campo de petróleo Volvo. O campo Volvo foi descoberto em 1993 e está localizado na parte central do Mar do Norte. A Figura 1 apresenta um recorte do mapa que mostra a localização do campo, obtido a partir do site "NORWEGIAN PETROLEUM"(https://www.norskpetroleum.no/en/facts/field/volve/).

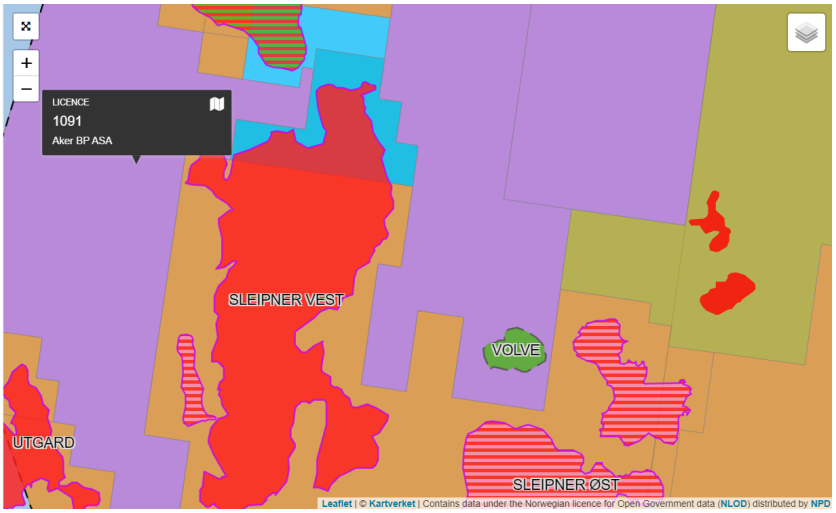

Figura 1: Recorte do mapa com a localização do campo Volvo retirado do site:https://www.norskpetroleum.no/en/facts/field/volve/

O reservatório presente neste campo é composto principalmente por arenito e está situado a uma profundidade que varia entre 2.700 e 3.100 metros. Este campo foi equipado para iniciar a produção em 2016 e encerrou sua fase de produção em 2018. Entre os dados disponíveis referentes a este campo, foram selecionados os dados de produção de três poços específicos: 15/9-F-12 H, 15/9-F-1 C e 15/9-F-15 D.

4 Coleta

Conforme mencionado na seção anterior, a base de dados selecionada para este trabalho consiste nos dados de produção do campo Volvo. Para adquirir esses dados, foram seguidos os procedimentos recomendados pela empresa Equinor, proprietária dos dados. Os dados estão hospedados na plataforma de nuvem Azure da Microsoft e estão disponíveis para download no formato .csv.

5 Modelagem

Para o processo de modelagem de dados, foram desenvolvidos vários esquemas, começando com um esquema mais genérico que descreve a sequência de etapas necessárias para a criação do Data Warehouse.

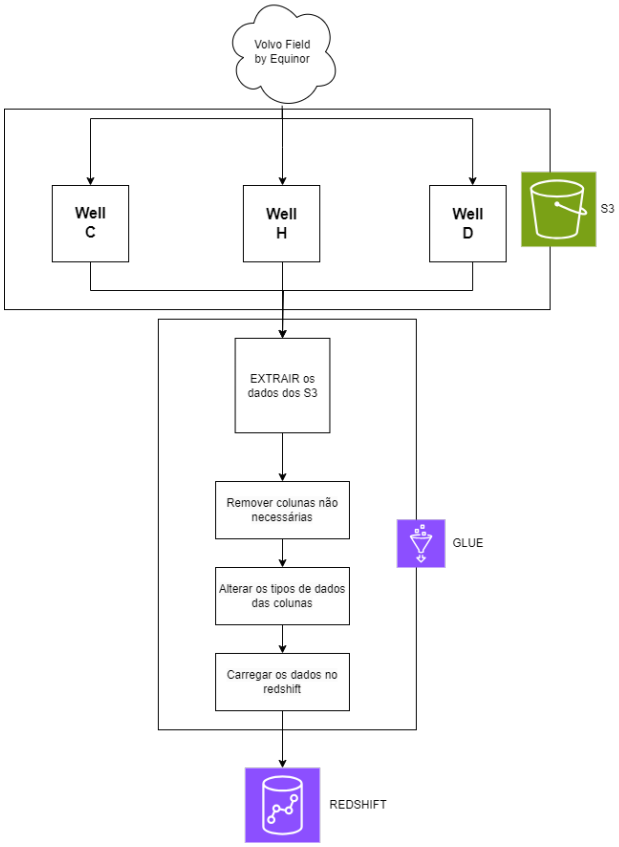

Figura 2: Esquema genérico de todas as etapas para carregar os dados na nuvem até a criação do redshift.

Após a conclusão do processo de criação do Redshift, podemos considerar a implementação de um modelo "estrela"usando os dados disponíveis. Mesmo que não tenhamos uma coluna de dimensão específica para a localização, podemos presumir essa dimensão, uma vez que todos os dados estão relacionados ao campo Volvo.

No diagrama apresentado na Figura 3, os dados medidos no poço estão organizados por dia de produção. Associado a esse fato, temos as dimensões de localização, que indica a localização do poço, a dimensão do poço, que identifica qual poço esteve em produção em um determinado dia, e, por fim,

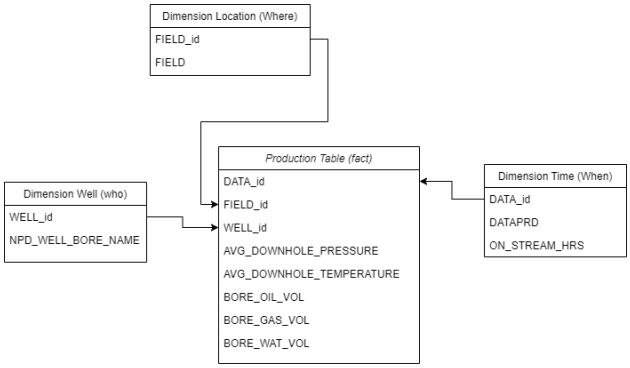

Figura 3: Captura de tela do Glue feito para fazer a conexão do s3 ao redshift.

a dimensão de tempo, que contém a data da produção e também a quantidade de horas durante as quais um determinado poço esteve em produção nesse dia.

6 Carga

A plataforma de nuvem selecionada para a carga de dados foi a AWS da Amazon. Nesta seção, as etapas realizadas para a carga de dados serão apresentadas por meio de capturas de tela.

1. Foi criado um "bucket s3"com o nome "mvp-oil-production-data-engineer"e dentro desse bucket foi criada uma pasta "oil-production".

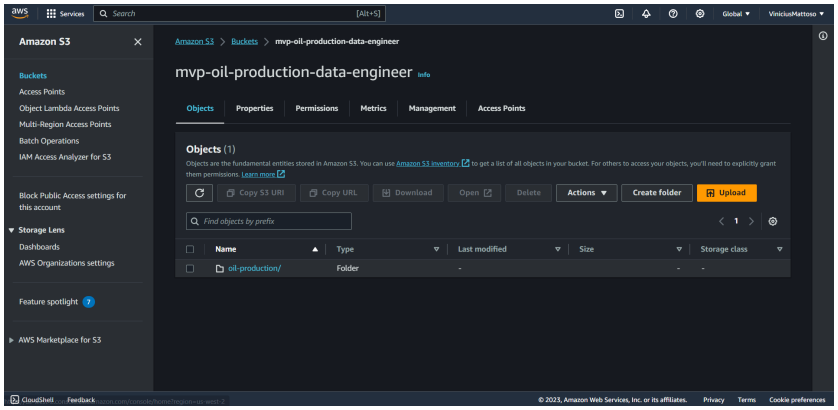

Figura 4: Captura de tela do bucket s3.

2. Dentro da pasta "oil-production"foi carregado 3 arquivos .csv, onde cada arquivo era referente a um poço produtor do campo em análise.

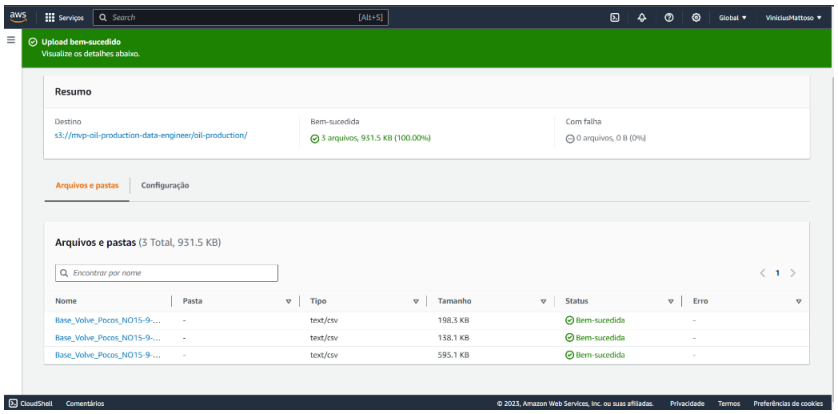

Figura 5: Captura de tela dos arquivos .csv carregados no bucket.

3. Foi feito o esquema dentro do AWS Glue para fazer a transferência do s3 para um redshift.

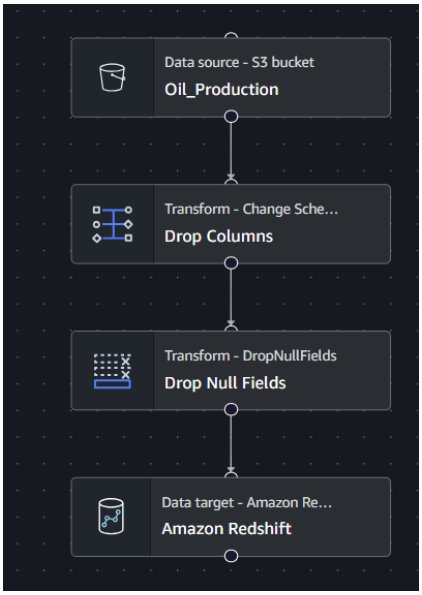

Figura 6: Captura de tela do Glue feito para fazer a conexão do s3 ao redshift.

4. Criar a tabela dentro do public para ser preenchida quando o ETL JOB for exectuado.

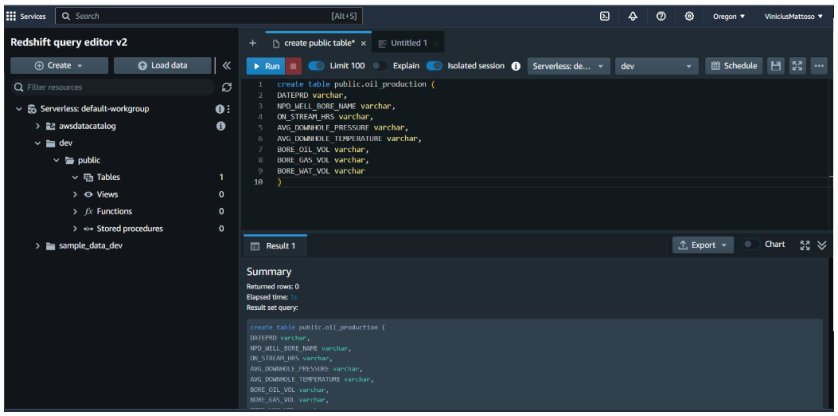

Figura 7: Captura de tela do comando SQL utilizado para ciar a tabela oil production.

7 Análise

Para criar o catálogo de dados, foi criado um Crawler para coletar informações do S3 em análise.

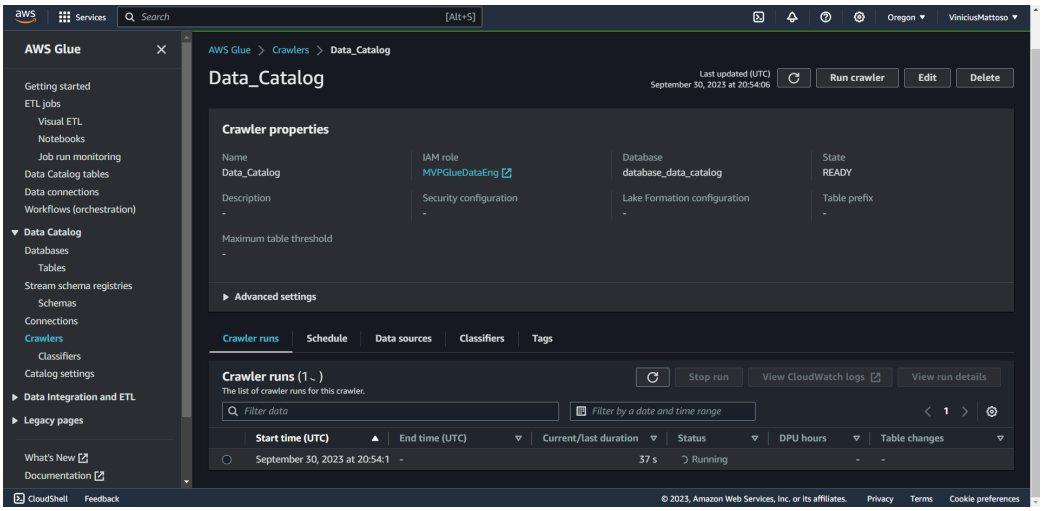

Figura 8: Captura de tela da criação do Crawler para gerar o catalogo de dados.

Após a criação e execução do Crawler obtemos os seguintes schemas:

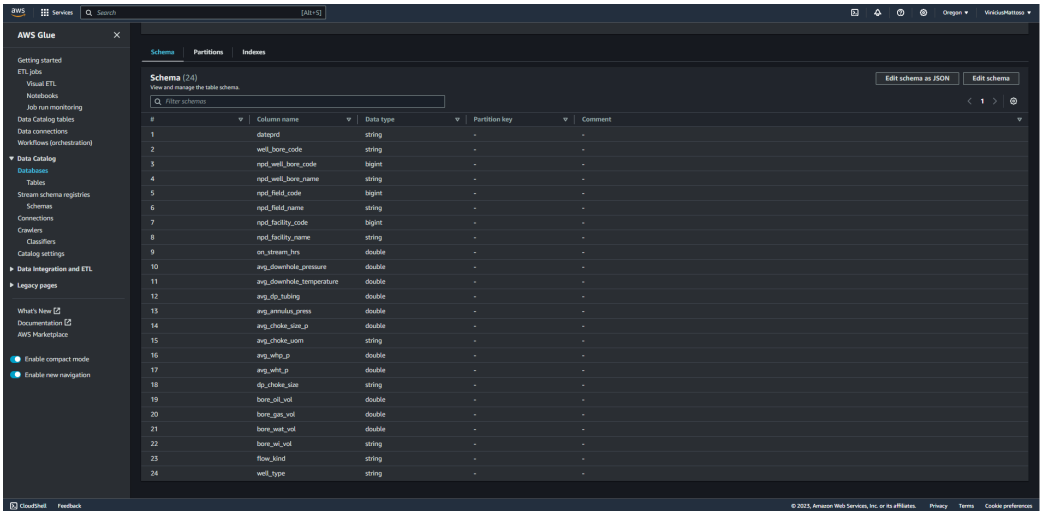

Figura 9: Captura de tela com os schemas identificados no catalogo de dados.

Antes de começar a responder as perguntas iniciais desse trabalho, podemos realizar um query para exibir todos os dados da base.

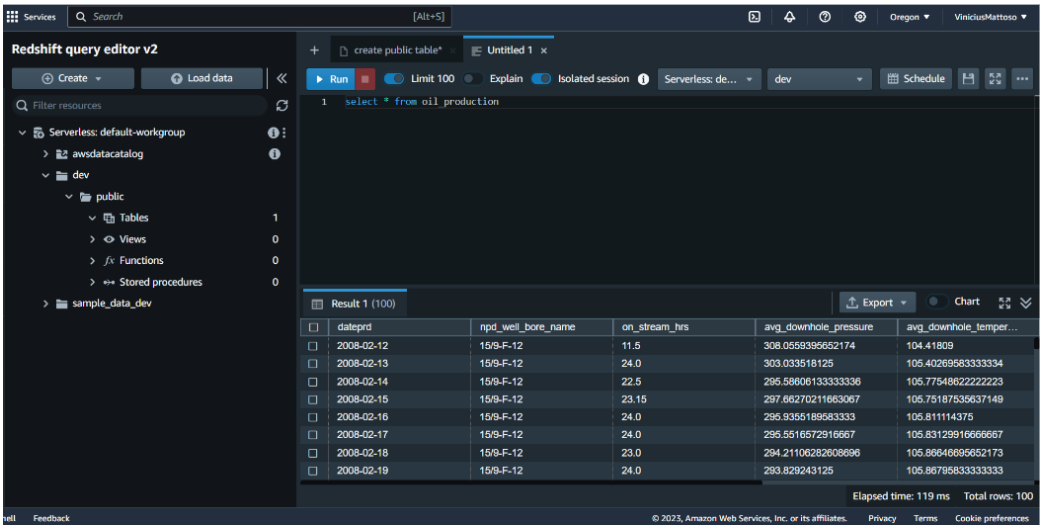

Figura 10: Captura de tela do SELECT * de toda a base de dados.

7.1 Qual poco possui a maior produção de óleo, água e gás?

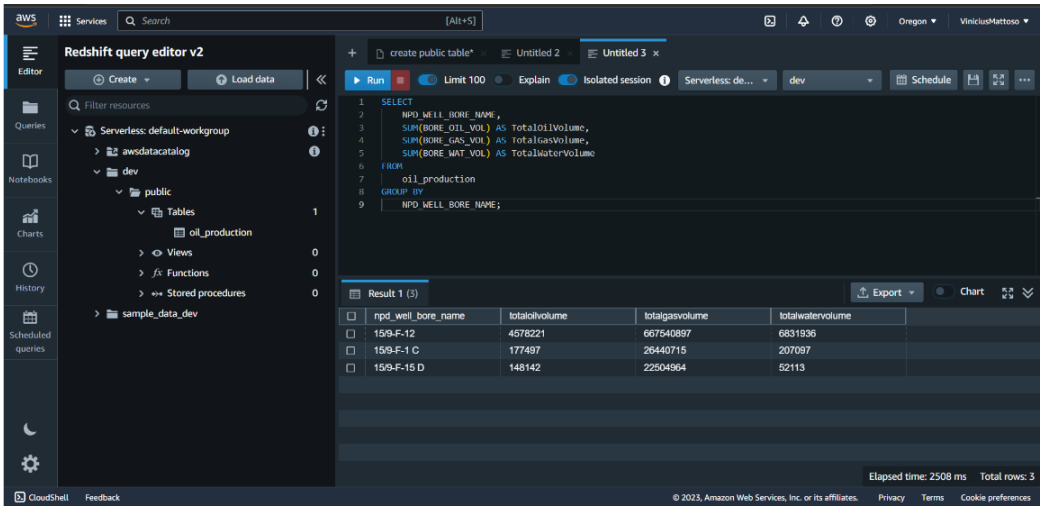

Figura 11: Captura de tela da query feita para saber qual é o poço mais produtivo.

Com base na query feita podemos concluir que o poço que mais produziu oléo, água e gás foi o poço 15/9-F-12 H.

7.2 Qual poço possui a maior razão entre o volume de óleo e água e a maior razão entre o volume de gás e óleo?

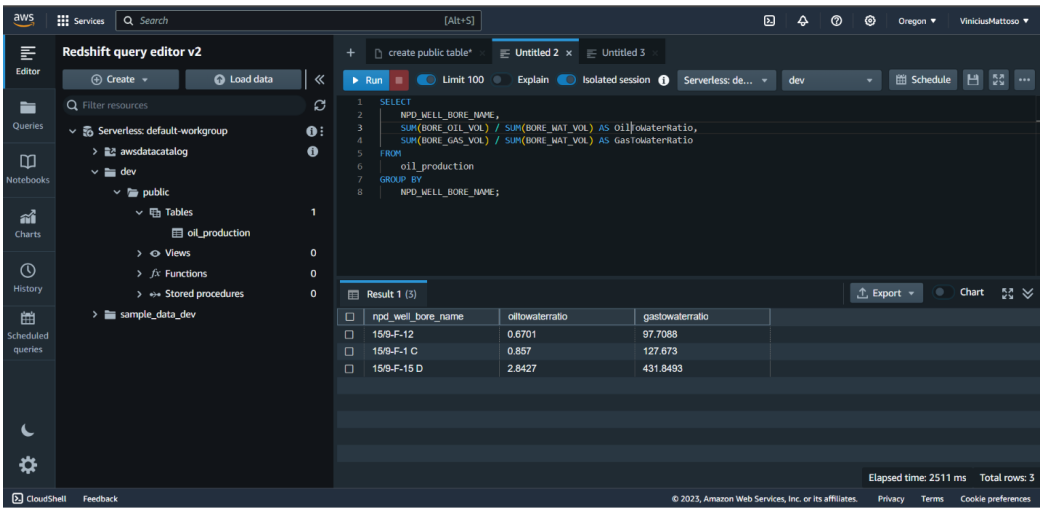

Figura 12: Captura de tela da query feita para saber a razão entre óleo e água e a razão entre gás e água.

Com base na consulta realizada, podemos concluir que, embora o poço 15/9-F-12 H tenha sido o poço que mais produziu óleo, a análise da relação entre óleo e água indica que esse poço produziu mais água do que óleo. Apenas o poço 15/9-F-15 D produziu mais óleo do que água. No que diz respeito à relação entre gás e água, todos os poços produziram um volume de gás muito maior do que o volume de água. Novamente, o poço 15/9-F-15 D se destacou, produzindo mais gás do que água. Essa análise é de grande relevância para a indústria, uma vez que os produtos comercializados são óleo e gás, e quanto maior a presença de água na mistura, mais complexo se torna o processo de separação das fases produzidas.

7.3 Qual foi o dia mais produtivo de cada poço considerando os diferentes materiais produzidos?

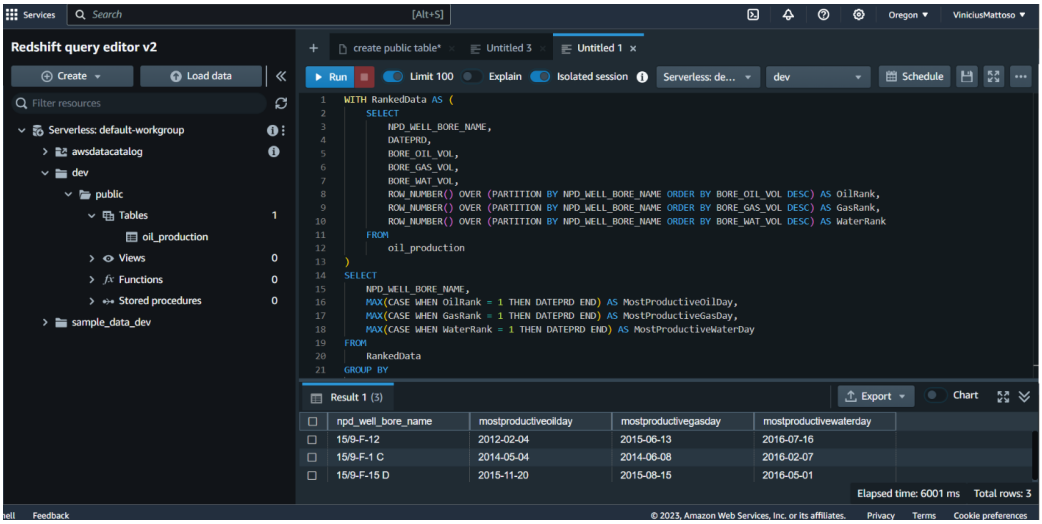

Figura 13: Captura de tela da query feita para saber qual foi o dia de maior produção de cada fase em cada poço.

8 Dificuldades encontradas

Nessa seção será apresentada quais foram as dificuldades encontradas ao longo desse projeto, assim como os passos feitos para superar essas dificuldades.

8.1 Remoção de dados faltantes

Na base de dados encontrada, foram identificadas algumas colunas com valores nulos. Uma vez que os dados se referem à produção de óleo, não é apropriado assumir valores constantes ou médias para preencher os dados ausentes. Portanto, para corrigir esses valores nulos, optou-se por remover todos os dados do dia em questão caso alguma informação estivesse faltando. Para realizar essa remoção, foi implementada uma transformação denominada "Drop Null Fields,"conforme ilustrado na Figura 6.

8.2 Carregar a tabela criada no public

Outro problema que surgiu durante o processo foi relacionado à carga da tabela "oil production"criada. Ao ajustar os tipos de dados para serem mais compatíveis com a base de dados, a carga resultou na criação de novas colunas e atribuiu valores NULL às colunas originalmente criadas.

Para resolver esse outro problema encontrado, ao invés de colocar os tipos

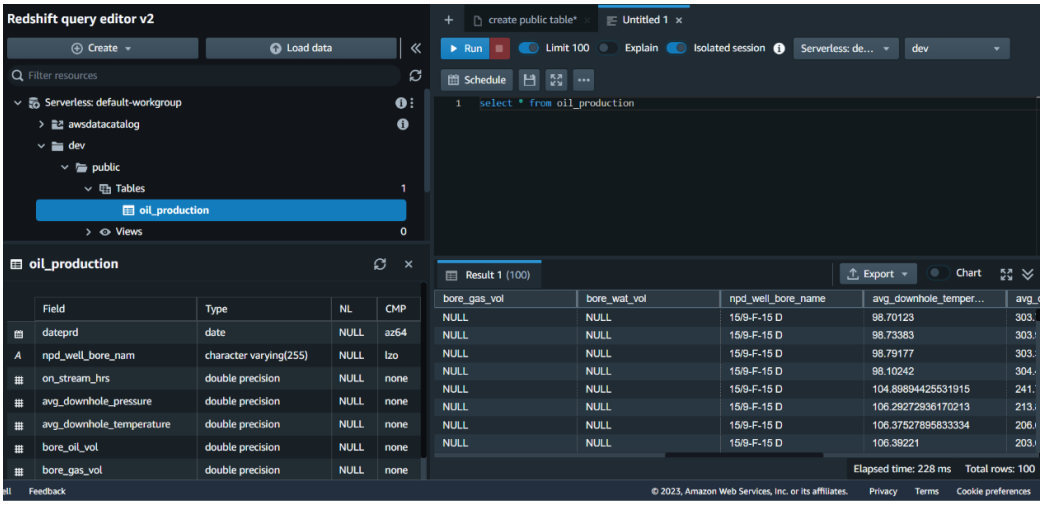

Figura 14: Captura de tela do problema encontrado na hora de fazer a carga da tabela criada no public.

de dados coerentes para cada uma das colunas da tabela, foi colocado como se todas as colunas fossem do tipo varchar.