Pontifícia Universidade Católica do Rio de Janeiro
PUC-RJ

ALEXANDRE DE SOUZA GAMBATI

Elaboração de pipeline de dados utilizando tecnologias na nuvem: Apoio a tomada de decisão em auditorias de regulação de medicamentos.

PÓS GRADUAÇÃO EM CIÊNCIA DE DADOS E ANALYTICS
ENGENHARIA DE DADOS

Rio de Janeiro
2023

RESUMO

GAMBATI, Alexandre de Souza: ***Elaboração de pipeline de dados utilizando tecnologias na nuvem****: Apoio a tomada de decisão em auditorias de regulação de medicamentos.*

O presente trabalho tem por objetivo a elaboração de indicadores de apoio a decisão que poderão ser utilizadas por equipes de auditoria que atuam na área de regulação de preços de medicamentos. Para tanto, foram utilizadas tecnologias em nuvem para a extração, transformação e carga dos dados e modelagem de uma base análitica a partir da qual é possível obter os indicadores necessários.

Palavras-chave: engenharia de dados; ETL; pipeline; medicamentos; regulação

ABSTRACT

GAMBATI, Alexandre de Souza: ***Development of data pipeline using cloud technologies****: Support for decision-making in drug regulation audits.*

This study aims to create decision-support indicators that can be used by audit teams operating in the field of drug price regulation. To achieve this, cloud technologies were employed for data extraction, transformation, and loading, as well as the modeling of an analytical database from which the necessary indicators can be obtained.

Keywords*: data engineering; ETL; pipeline; drugs; regulation.*

Sumário

1. CONTEXTO

1.1 REGULAÇÃO DO MERCADO DE MEDICAMENTOS

Dentre as atribuições da Agência Brasileira de Vigilância Sanitária – ANVISA (**https://www.gov.br/anvisa/pt-br**) se destaca a regulação de preços de medicamentos comercializados no nosso país.

Esse processo é mantido pela Câmara de Regulação de Mercado de Medicamentos – CMED, órgão executivo da ANVISA que edita normatização que deve ser observada pelo mercado de medicamentos do país.

Dentre as regulações normatizadas pela CMED destacam-se para o interesse desse trabalho:

Preços Máximos de Venda ao Consumidor: A CMED edita tabelas de preços máximos de comercialização de medicamentos nos pontos de venda ao consumidor final.

Preços Máximos de Venda ao Governo: A CMED edita tabelas de preços máximos de comercialização de medicamentos quando os órgãos e entidades do governo são os consumidores finais desses produtos.

Ambas as tabelas são atualizadas periódicamente e podem ser acessadas em **https://www.gov.br/anvisa/pt-br/assuntos/medicamentos/cmed/precos**

1.2 AUDITORIAS DE FISCALIZAÇÃO DOS PREÇOS REGULADOS

Tendo em vista as limitações de recursos humanos e materais para a realização de auditorias pela CMED/ANVISA, o controle da regulação até então tem se valido de denúncias e representações encaminhadas em sede de controle social.

Oportunamente, com o advento das Notas Fiscais Eletrônicas (***NFE***), os governos estaduais passaram a dispor de uma fonte de dados que permite conhecer as transações de produtos comercializados, inclusive medicamentos. Através do compartilhamento desses dados entre os órgãos de controle, tem sido possível extrair insights e conhecimento necessário ao apoio a decisão, o que permite que os órgãos de controle dirijam seus esforços de auditoria para onde se demonstrar um maior benefício da atividade de fiscalização.

2. OBJETIVO

Este trabalho tem por objetivo geral elaborar indicadores de apoio a decisão para serem utilizados por equipes de auditoria que atuam na área de regulação de preços de medicamentos.

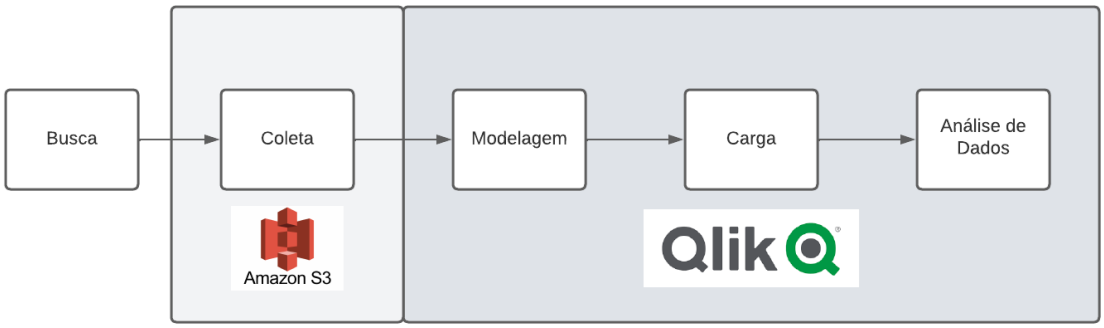
Como questão principal, deve ser avaliado se os dados presentes nas Notas Fiscais Eletrônicas podem ser utilizadas para detectar ocorrências de vendas de medicamentos aos órgãos do Governo por preços superiores aos permitidos pela regulação em vigor.

Dessa forma, caso o objetivo principal se demonstre viável, devem ser buscados indicadores ou gráficos que demonstrem:

1) Quais os órgão do governo que estão realizando maiores volumes de compras de medicamentos a preços superiores aos permitidos pela regulaçao em vigor.

2) Quais empresas estão vendendo os maiores volumes de medicamentos a preços superiores aos permitidos pela regulação em vigor.

Para realizar o processamento necessário para atingir os requisitos do projeto, será elaborado o seguinte pipeline usando tecnologias em nuvem:



3. BUSCA DOS DADOS

Para atender aos requisitos deste projeto, procedeu-se inicialmente a buscas de dados públicos referentes a regulação de medicamentos divulgados no site da Câmara de Regulação de Mercado de Medicamentos – CMED:

https://www.gov.br/anvisa/pt-br/assuntos/medicamentos/cmed/precos

Foram baixados os seguintes arquivos:

Arquivo	Descrição
xls_conformidade_gov_20230901_131509553.xls	Preços máximos de venda ao governo
xls_conformidade_site_20230901_131509553.xls	Preços Máximos de venda ao consumidor

Além dos preços regulados, essas tabelas divulgadas pela CMED apresentam diversas outras informações relativas ao produto regulado, especialmente o dado do EAN (*European Article Number*), composto por 13 digitos que representa o código de barras dos produtos comercializados.

Nessa fase de busca de dados, foi possível observar que os dados das Notas Fiscais Eletrônicas (NFE) armazenadas pelo Governo do Estado do Rio de Janeiro também contém o dado do Código EAN para cada transação de venda, o que gera potencial para obtenção de informações e geração de conhecimento (indicadores). Além disso, as NFEs contém uma diversidade considerável de dados referentes ao produto comercializado, bem como a empresa vendedora, transportadores, clientes, etc.

Os dados das NFEs foram obitdos por meio de acesso controlado a base de dados compartilhada entre o Tribunal de Contas do Estado do Rio de Janeiro e o Governo do Estado do Rio de Janeiro, tendo sido gerados os seguintes arquivos brutos:

Arquivo	Descrição
nfe.csv	Contém informações do cabeçalho da Nota Fiscal
produto_nfe.csv	Contém informações de cada produto comercializado em cada Nota Fiscal.

Há que se registrar que os dados das NFEs são protegidos por sigilo fiscal e possuem informações que não podem ser divulgadas no âmbito desse projeto, devendo esses dados serem devidamente **anonimizados** para uso no âmbito deste trabalho acadêmico.

O conjunto de dados das NFEs baixadas se referem as operações realizadas no Estado do Rio de Janeiro no período de janeiro a julho de 2023.

Por fim, também foram buscadas informações cadastrais de órgãos públicos, mais especificamente o nome do órgão e o respectivo número de CNPJ (Cadastro Nacional de Pessoa Física), para que seja possível correlacionar as operações de venda destinadas a órgãos públicos. Esses dados foram obitdos por meio de acesso controlado a base de dados mantida pelo Tribunal de Contas do Estado do Rio de Janeiro, tendo sido gerados os seguintes arquivos brutos:

Arquivo	Descrição
ug_estado.csv	Contém informações de órgãos do Estado do Rio de Janeiro
ug_municipio.csv	Contém informações de órgãos do Município do Rio de Janeiro

Apesar de se tratarem de dados obtidos por meio de acesso controlado, esses dados são públicos, não requerendo tratamento de anonimização.

4. COLETA DOS DADOS

Com base na etapa anterior, foram definidos e baixados os seguintes conjuntos de dados:

Arquivo	**Descrição**
xls_conformidade_gov_20230901_131509553.xls	Preços máximos de venda ao governo
xls_conformidade_site_20230901_131509553.xls	Preços Máximos de venda ao consumidor
nfe.csv	Contém informações do cabeçalho da Nota Fiscal
produto_nfe.csv	Contém informações de cada produto comercializado em cada Nota Fiscal.
ug_estado.csv	Contém informações de órgãos do Estado do Rio de Janeiro
ug_municipio.csv	Contém informações de órgãos do Município do Rio de Janeiro

Nessa etapa, os dados devem ser armazenados em serviço que permita a computação em nuvem. Para essa finalidade, optou-se pela utilização do serviço S3 (Simple Storage Service) mantido pela estrutura da AWS (Amazon Web Services).

A seguir, são apresentados os passos necessários para o armazenamento e controle de acesso aos dados coletados:

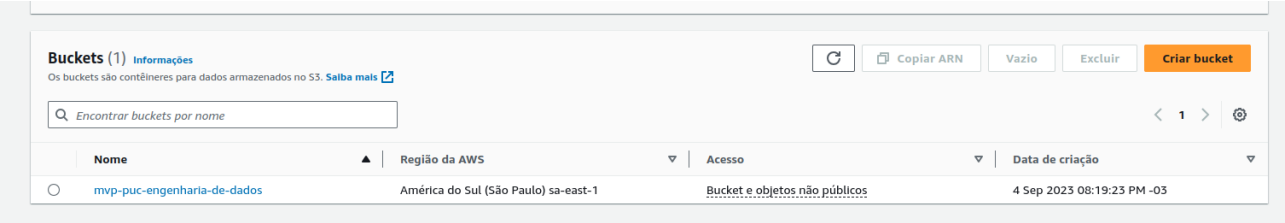
4.1 Armazendo os dados brutos em Nuvem.

(Essa etapa pressupõe que já exista uma conta root já ativa na AWS)

4.1.1 Criar o bucket

Para armazenar os dados em nuvem, utilizaremos o serviço S3 da AWS e para tanto, criaremos um bucket para receber os dados brutos.

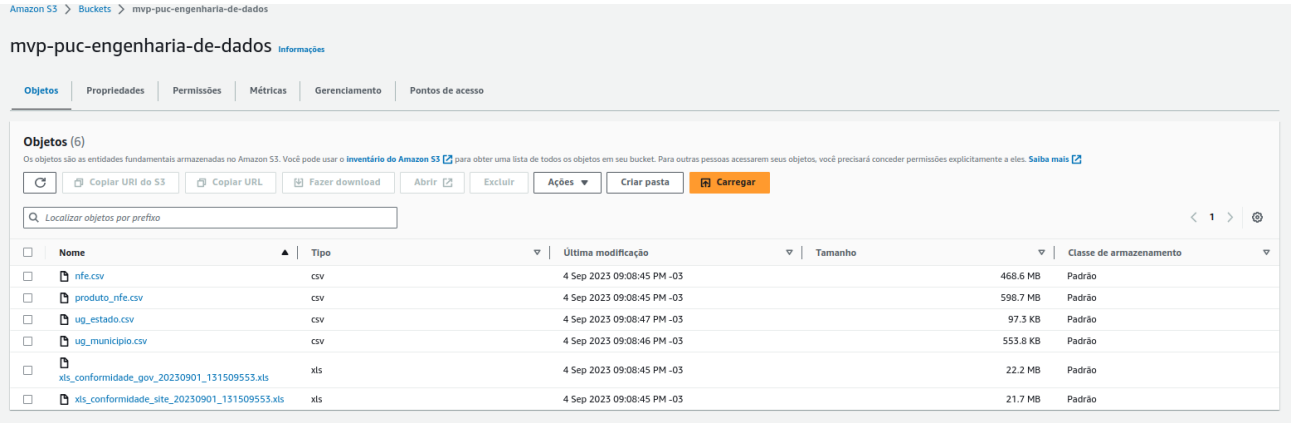
- Digite **S3** no console da AWS
- Clique em **Buckets** em seguida, **Criar Bucket**.
- Nomeie o bucket e dê uma descrição.



A figura anterior apresenta a lista de buckets criados, onde é exibido *mvp-puc-engenharia-de-dados*, que será o bucket utilizado nesta etapa.

4.1.2 Carregar os dados no bucket

- Clique no bucket criado na etapa anterior
- Clique em **carregar**.
- Selecione os arquivos a serem carregados na estrutura do S3
- Confirme clicando em **carregar**.



A figura anterior apresenta os arquivos carregados no bucket.

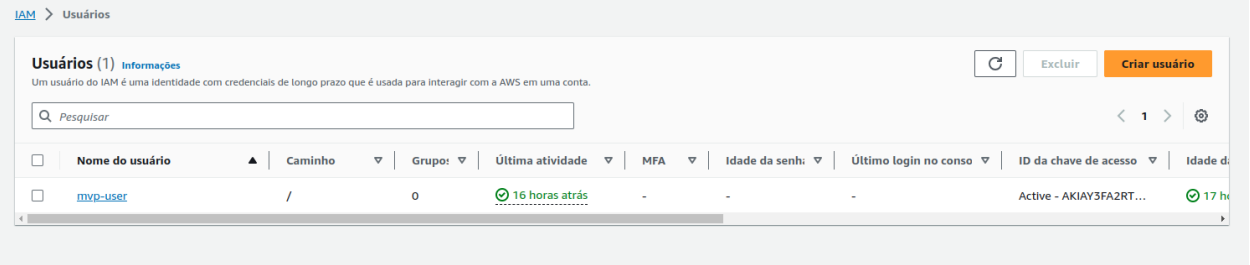
4.2 Controle de Acesso.

4.2.1 Geração de credenciais

Para que os arquivos do bucket sejam acessíveis por aplicações externas à estrutura da AWS, é necessário **criar um usuário** que possua **chaves de acesso**. Para tanto, seguiremos os seguintes passos:

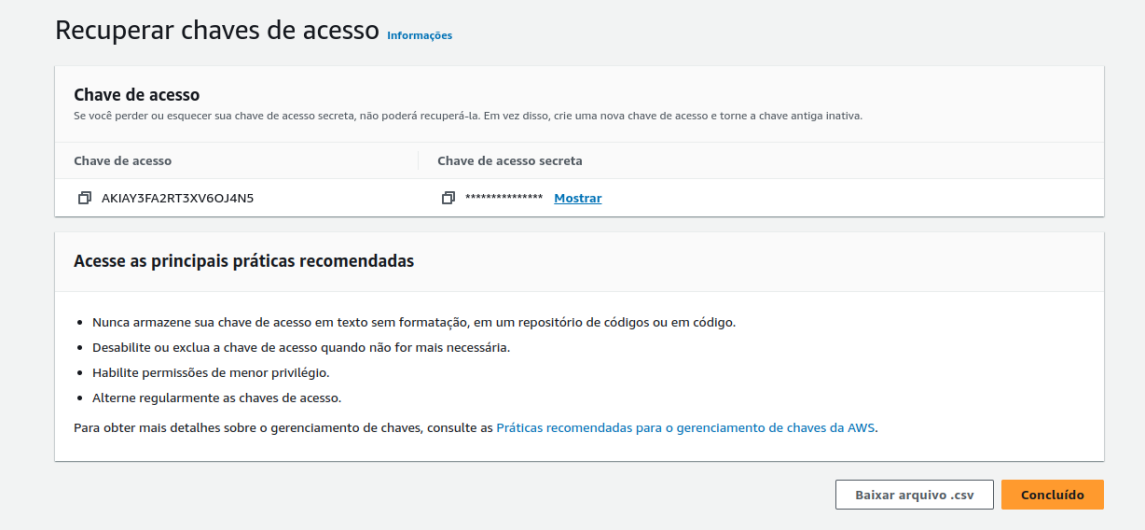
4.2.2 Criação de um usuário.

- Digite **IAM** no console da AWS
- No painel do IAM, vá em **usuários**
- Clique em **criar usuário**
- Em opções de permissões, selecione **anexar políticas diretamente**
- Em políticas e permissões, selecione a política **AmazonS3ReadOnlyAccess**
- Clique em **próximo**, e depois em **criar usuário.**



4.2.3 Gerar chaves de acesso

- Clique no usuário criado na etapa anterior
- Clique em **criar chave de acesso**
- Em caso de uso, selecione **aplicação executada fora da AWS**
- Anote em lugar seguro as **chaves de acesso** e **chave secreta** geradas para esse usuário.



A figura anterior apresenta as chave de acesso e chave secreta que posteriormente serão usadas para criar uma conexão com a aplicação externa.

5. IMPORTAÇÃO DOS DADOS NO QLIK CLOUD

Uma vez coletados e armazenados os dados em serviço de computação em nuvem, procederemos a importação desses dados para a plataforma Qlik Cloud, que é uma solução serverless para modelagem, ETL e análise de dados.

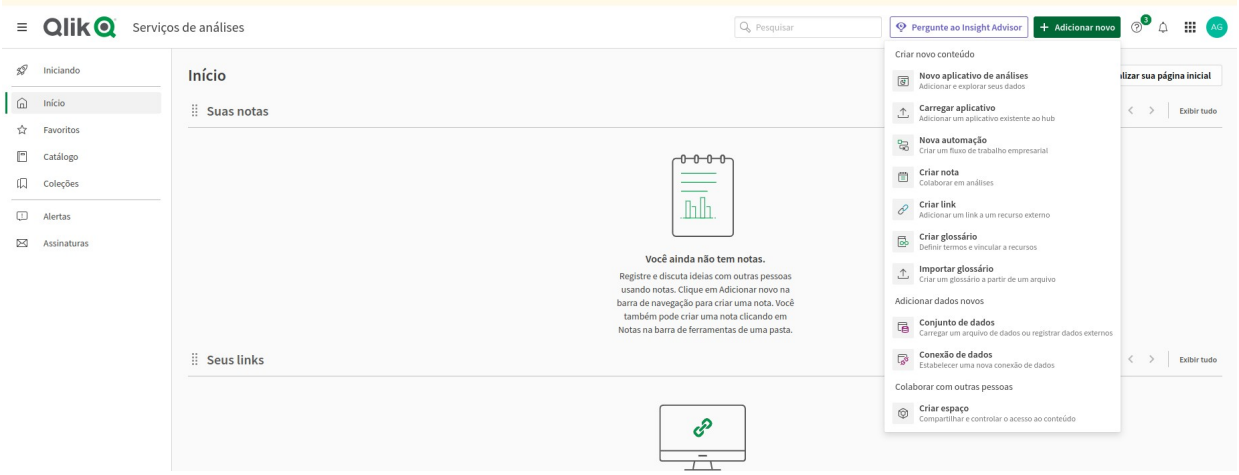
Criaremos uma conexão com o serviço S3 da AWS e importaremos os dados para o Qlik Cloud para realização das etapas finais do presente trabalho.

5.1 Criação da Aplicação Qlik Cloud

(Essa etapa pressupõe que já exista uma conta já ativa no Qlik Cloud)

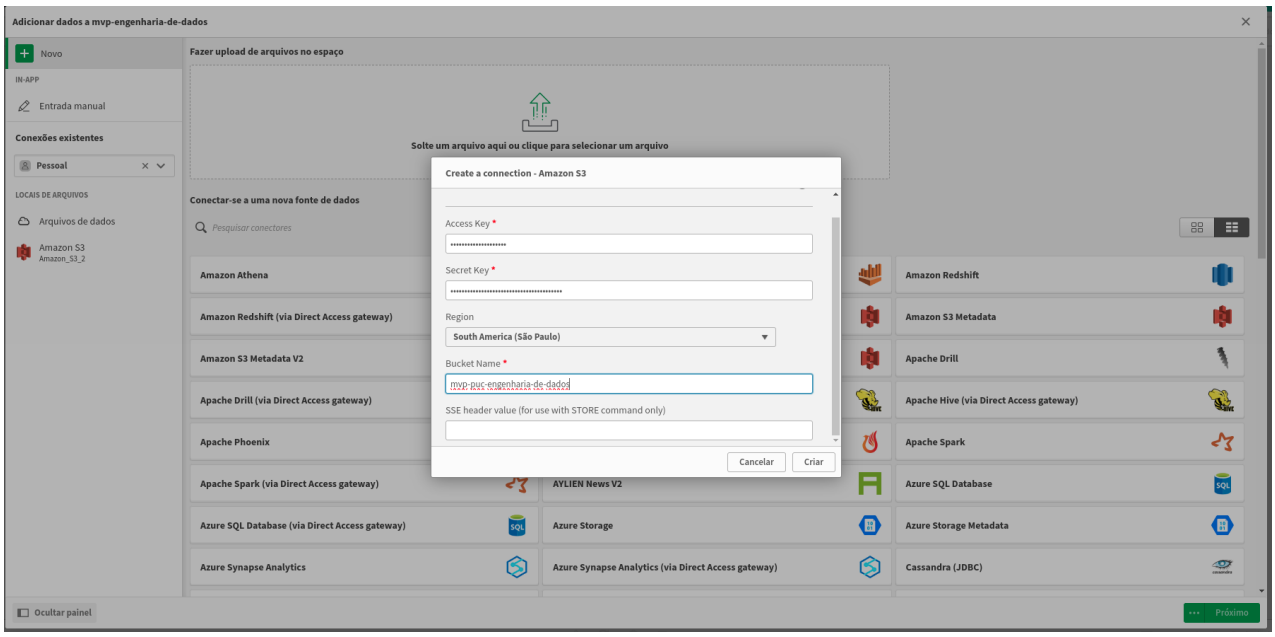
Para iniciar uma aplicação no Qlik Cloud, siga os seguintes passos:

- Clique em **+Adicionar novo**
- Clique em **novo aplicativo de análises**
- Nomeie a aplicação



5.1.1 Criação da conexão com o serviço AWS S3

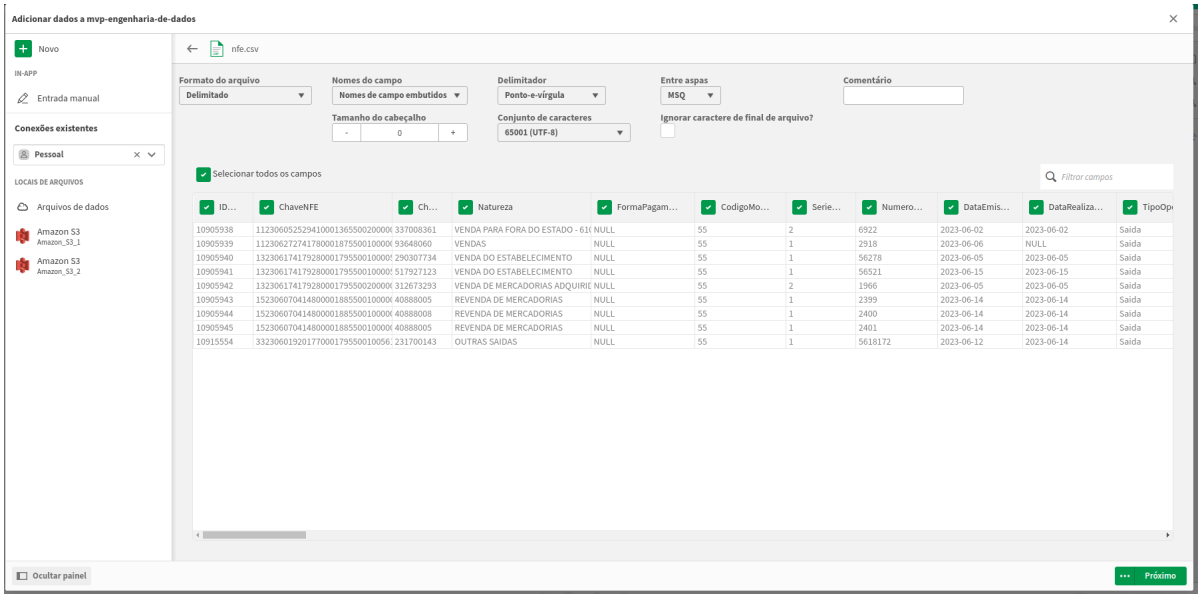
- Abra a aplicação recém-criada.
- Selecione **adicionar dados de outras fontes**
- Em conectar-se a uma fonte de dados, selecione Amazon S3
- Forneça a chave de acesso, chave secreta e região e nome do bucket na S3



Nesse ponto, será criada uma conexão com o bucket do projeto na estrutura da S3. Os arquivos do bucket serão listados e então poderá ser feita a importação de cada um deles para o Qlik.

5.1.2 Importação dos arquivos

- Selecione a conexão criada no passo anterior.
- Selecione o primeiro arquivo do bucket.
- Será apresentado um preview da importação. Verifique a qualidade da importação.
- Clique em próximo.



Repita esse processo para cada um dos arquivos do bucket.

5.1.2.1 Importação de planilhas excel

Atentar que as planilhas excel utilizadas nesse projeto possuem muitas linhas informativas antes do cabeçalho dos dados. Portanto, ao carregar esses arquivos deve ser informado o número da linha que inicia o cabeçalho dos dados para permitir uma importação correta dos dados.

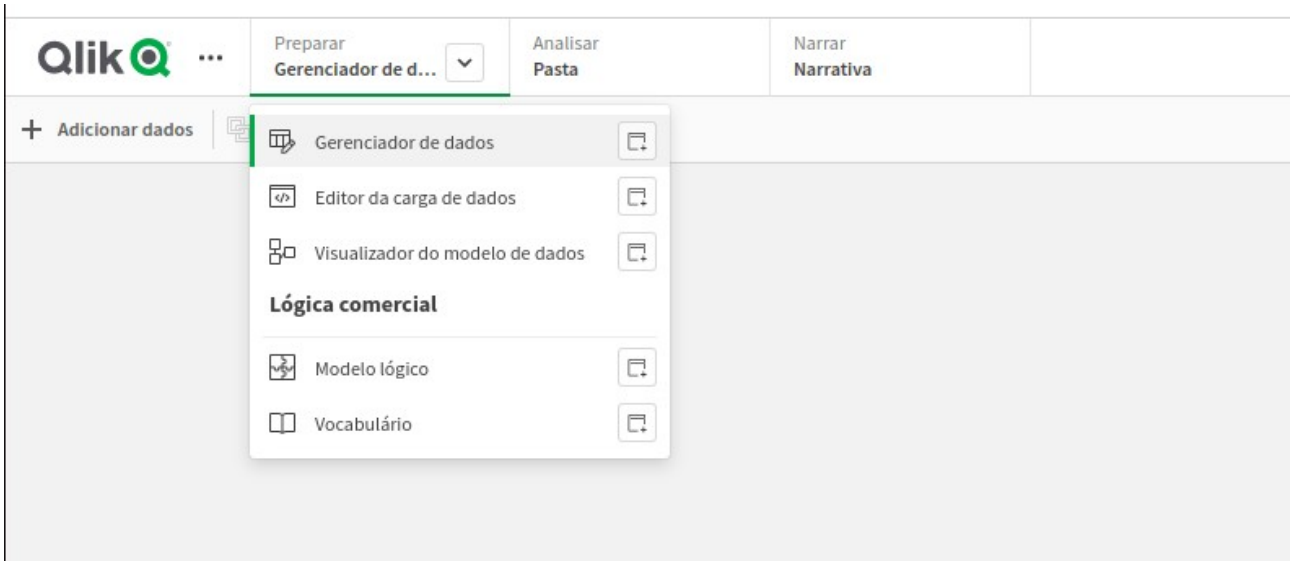
6. MODELAGEM E CARGA DOS DADOS

Uma vez criada a conexão com o serviço de armazenamento de dados em nuvem e feita a importaçao dos arquivos necessários, o próximo passo é realizar o processo de extração, transformação e carga dos dados (ETL). O processo de ETL deve ser realizado de modo a gerar um esquema estrela (star schema) ou floco de neve (snow flake) composto de tabelas de dimensões e de fatos que serão utilizadas pelas ferramentas analíticas para apresentar os indicadores que atendam aos requisitos do projeto.

Como estratégia geral, será buscado criar tabelas de dimensões referentes a medicamentos, produtos, pessoas jurídicas (emitentes, destinatários, transportadores, unidades gestoras) e tabelas de fatos relativos as vendas e regulaçao de medicamentos.

Para essa finalidade o Qlik Cloud dispõe de ferramentas baseadas em scripts de importação que permitem criar tabelas de destino no modelo que são construídas a partir da carga dos dados de interesse nas tabelas de origem. Também é possível realizar transformações nos dados de origem e fazer operaçoes de junção entre tabelas para gerar a modelagem desejada. A partir das tabelas criadas, a própria aplicação do Qlik Cloud irá criar vínculos com base em campos que possuem o mesmo nome em tabelas distintas.

Para realizar essa operação deve ser aberta a aplicação do Qlik Cloud e solicitar a opção **Editor de Carga de Dados** que se encontra na guia preparar.



No caso deste projeto, optou-se pela elaboração de scripts manualmente, e portanto foi excluida a seção de código de importação gerada automaticamente pela aplicação.

6.1 Scripts de ETL

São documentados os scripts do Qlik Cloud utilizados para a fase de ETL, onde serão instruídas as operações de extração, transformação e carga de dados a serem realizadas pela aplicação. As operaçoes de ETL podem ser divididas em seções no Qlik Cloud, apenas para fins de melhor organização e compreensão do modelo. Serão criadas três sessões: “Medicamentos Regulados”, “Notas Fiscais Eletronicas” e “Unidades Gestoras”.

6.1.1 Scripts da Seção “Medicamentos Regulados”

Inicialmente, optou-se por modelar os dados que originalmente se encontram nos arquivos de regulação de medicamentos. Os dados originais são compostos de duas tabelas de regulação de preços: preços de venda ao governo e preços de venda ao consumidor.

Será criada uma única tabela de dimensão, referente ao conjunto de medicamentos regulados, por meio de uma operação de FULL OUTER JOIN entre as tabelas de origem.

Além disso, é realizada uma operação de anonimização dos dados do CNPJ da empresa fabricante, usando-se para tanto da função HASH128() que gera um código Hash que representa o valor original do CNPJ. Vale registrar que a função HASH128() é determinística e gera o mesmo valor codificado para uma dada entrada. Dessa forma, será possível gerar campos CNPJ anonimizados que poderão ser funcionar como chaves estrangeiras normalmente, como seriam se não fossem anonimizados. A função da anonimização servirá apenas para garantir o sigilo das empresas cujos dados são tratados neste trabalho.

Feitas essas considerações, assim foi implementado o script de modelagem da tabela dimensão de medicamentos:

```
// Tabela de Dimensão Medicamento Regulado Governo
Dim_Medicamento:
LOAD
    "PRINCÍPIO ATIVO",
    hash128(CNPJ) as CNPJ_Anonimizado,
    LABORATÓRIO,
    "EAN 1" as EAN,
    PRODUTO,
    APRESENTAÇÃO
FROM [lib://Amazon_S3_2/xls_conformidade_gov_20230901_131509553.xls]
(biff, embedded labels, header is 43 lines, table is Planilha1$);

// FULL OUTER JOIN com a tabela de preços ao consumidor.
OUTER JOIN (Dim_Medicamento)
LOAD
    "PRINCÍPIO ATIVO",
    hash128(CNPJ) as CNPJ_Anonimizado,
    LABORATÓRIO,
    "EAN 1" as EAN,
    PRODUTO,
    APRESENTAÇÃO
FROM [lib://Amazon_S3_2/xls_conformidade_site_20230901_131509553.xls]
(biff, embedded labels, header is 29 lines, table is Planilha1$);
```

Em seguida, é criada a tabela de fato de regulacao de medicamento, também através de operação de FULL OUTER JOIN entre as tabelas de origem, selecionando-se os preços de venda ao governo e ao consumidor final das tabelas de origem.

Dessa forma, assim foi implementado o script de modelagem da tabela fato de regulacao de medicamento:

```
// Tabela de Fato regulacao medicamento.
//Optou-se pela utilização de tabela de fato,
//a fim de permitir uso futuro de preços para cada data de vigencia da regulação
Fact_regulacao_medicamento:
LOAD
    "EAN 1" as EAN,
    "PMVG 20%" as "Valor Máximo Governo"
FROM [lib://Amazon_S3_2/xls_conformidade_gov_20230901_131509553.xls]
(biff, embedded labels, header is 43 lines, table is Planilha1$);

// FULL OUTER JOIN com a tabela Fact_regulacao_medicamento
OUTER JOIN (Fact_regulacao_medicamento)
LOAD
    "EAN 1" as EAN,
    "PMC 20%" as "Valor Máximo Consumidor"
FROM [lib://Amazon_S3_2/xls_conformidade_site_20230901_131509553.xls]
(biff, embedded labels, header is 29 lines, table is Planilha1$);
```

6.1.2 Scripts da Seção “Notas Fiscais Eletronicas”

As notas fiscais eletronicas possuem uma diversidade de informações. O arquivo original nfe.csv contém informações gerais sobre a venda realizada, como pessoas físicas e jurídicas (vendedores, clientes, transportadores, etc). Por sua vez, o arquivo produto_nfe.csv contém

informações de cada produto entregue em cada operação de venda, que é relacionado ao arquivo de notas fiscais através do ChaveNFE (chave única). A cardinalidade entre essas tabelas é 1:n (uma nota pode ter vários produtos).

Nessa seção também são anonimizados os valores de CNPJs através do uso da função HASH128() e omitidas informaçoes que possam identificar as pessoas físicas e jurídicas relacionadas, como nomes das empresas, endereços, telefones, etc.

A tabela de **Fact_NFE** contem as datas de emissão e realização da operaçao de venda e possui campos para estabelecer relações com as tabelas de dimensões de vendedores (**Dim_Emitentes**) e clientes (**Dim_Destinatarios**).

Além disso, através do campo ChaveNFE estabelece-se relacionamento com a tabela **Fact_Produto_NFE**, que contém o código do produto, e o preço de venda unitário e total do produto vendido em cada operação de venda registrado na tabela **Fact_NFE**.

Por sua vez, é criada a tabela de dimensão **Dim_Produto** contém o código do produto, relacionando-se com a tabela **Fact_Produto_NFE.** Uma parcela considerável dos registros da tabela de **Dim_Produto** informaçao ao código EAN do produto. Considerando que o conjunto de dados de produtos em Notas Fiscais Eletronicas se referem a produtos em geral, apenas em alguns casos haverá identificação com códigos EAN presentes na tabela **Fact_regulacao_medicamento.** Não obstante, esse subconjunto é o que representa as relações de interesse deste projeto.

Feitas essas considerações, assim foi implementado o script de modelagem das tabelas **Fact_NFE, Dim_Emitentes, Dim_Destinatarios, Dim_Produto e Fact_Produto_NFE:**

```
// Tabela de Fato (NFE)
Fact_NFE:
LOAD
    [ChaveNFE],
    Date([DataEmissao]) as [DataEmissao],
    Date([DataRealizacao]) AS [DataRealizacao],
    hash128([CNPJEmitente]) as [CNPJEmitente_Anonimizado],
    hash128([CNPJDestinatario]) as [CNPJDestinatario_Anonimizado]
FROM [lib://Amazon_S3_2/nfe.csv]
(txt, utf8, embedded labels, delimiter is ';', msq);

// Tabela de Dimensão Emitentes (apenas registros distintos)
Dim_Emitentes:
LOAD DISTINCT
    hash128([CNPJEmitente]) as [CNPJEmitente_Anonimizado],
    hash128([RazaoSocialEmitente]) as [RazaoSocialEmitente_Anonimizado],
    [BairroEmitente],
    [codigoMunicipioEmitente],
    [MunicipioEmitente],
    [UFEmitente],
    [CEPEmitente],
    [CodigoPaisEmitente],
    [PaisEmitente]
FROM [lib://Amazon_S3_2/nfe.csv]
(txt, utf8, embedded labels, delimiter is ';', msq);

// Tabela de Dimensão Destinatários (apenas registros distintos)
Dim_Destinatarios:
LOAD DISTINCT
    hash128([CNPJDestinatario]) as [CNPJDestinatario_Anonimizado],
    hash128([RazaoSocialDestinatario]) as [RazaoSocialDestinatario_Anonimizado],
    [BairroDestinatario],
    [CodigoMunicipioDestinatario],
    [MunicipioDestinatario],
    [UFDestinatario],
    [CEPDestinatario],
    [CodigoPaisDestinatario],
    [PaisDestinatario]
FROM [lib://Amazon_S3_2/nfe.csv]
(txt, utf8, embedded labels, delimiter is ';', msq);
```

```
Dim_Produto:
LOAD DISTINCT
    CodigoProduto,
    DescricaoProduto,
    CleanTributavel as EAN
FROM [lib://Amazon_S3_2/produto_nfe.csv]
(txt, utf8, embedded labels, delimiter is ';', msq);

Fact_Produto_NFE:
LOAD
    ChaveNFE,
    CodigoProduto,
    ValorUnitario,
    ValorTotalBruto
FROM [lib://Amazon_S3_2/produto_nfe.csv]
(txt, utf8, embedded labels, delimiter is ';', msq);
```

6.1.3 Scripts da Seção “Unidades Gestoras”

Nessa seção são importadas informações relativas as unidades gestoras dos governos estaduais e municipais do Estado do Rio de Janeiro.

Apesar de não haver necessidade de sigilo desses dados, as informações de CNPJs (Cgc) devem também ser anonimizadas através do uso da função HASH128(), a fim de permitir o relacionamento com os valores presentes nas tabelas de destinatários (**Dim_Destinatarios**) que já foram anonimizados na seção anterior.

Assim, é criada a tabela **Dim_unidade_gestora** através de operação de FULL OUTER JOIN entre as tabelas de origem, selecionando-se as informaçoes que identificam as unidades gestoras. O script de modelagem é apresentado a seguir:

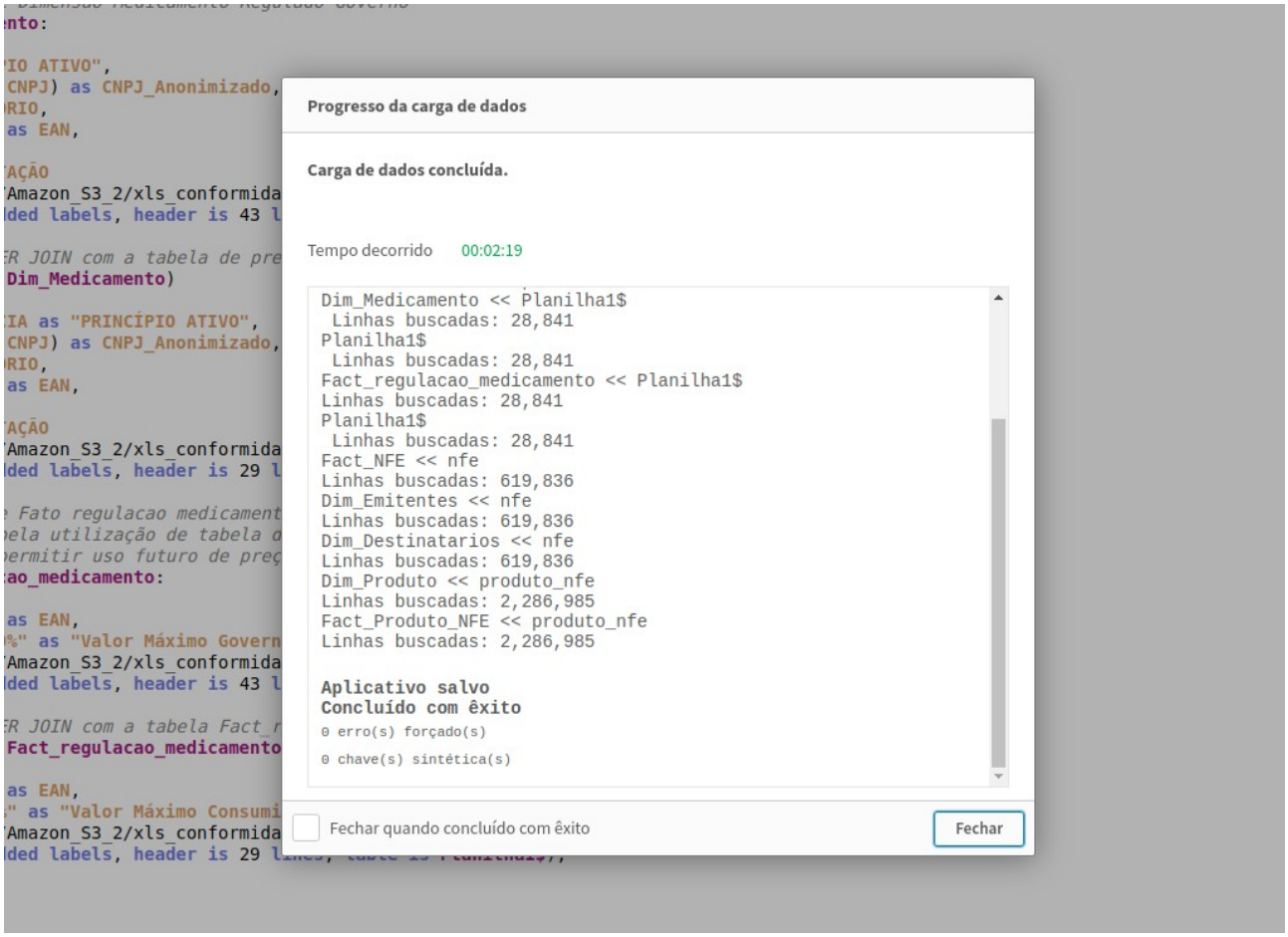
```
Dim_unidade_gestora:
LOAD DISTINCT
    NomeUnidade,
    SiglaUnidade,
    hash128([Cgc]) as [CNPJDestinatario_Anonimizado]
FROM [lib://Amazon_S3_2/ug_estado.csv]
(txt, utf8, embedded labels, delimiter is ';', msq);

OUTER JOIN (Dim_unidade_gestora)
LOAD DISTINCT
    NomeUnidade,
    SiglaUnidade,
    hash128([Cgc]) as [CNPJDestinatario_Anonimizado]
FROM [lib://Amazon_S3_2/ug_municipio.csv]
(txt, utf8, embedded labels, delimiter is ';', msq);
```

6.2 Carga dos Dados

Após a implementação dos scripts de ETL demonstrados no passo anterior, deve-se proceder a carga dos dados. Na aplicação do Qlik Cloud, vá em **Carregar Dados** (canto direito da tela).

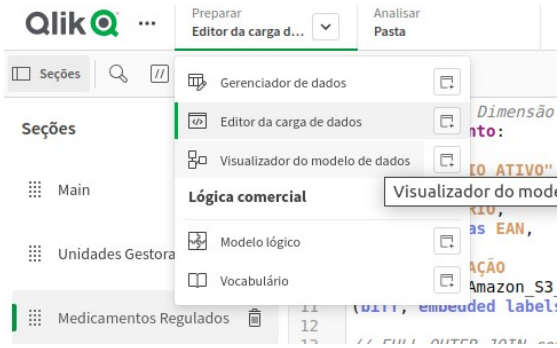
Aguarde a conclusão do processo de carga dos dados. A seguir é apresentado o resultado do processamento.



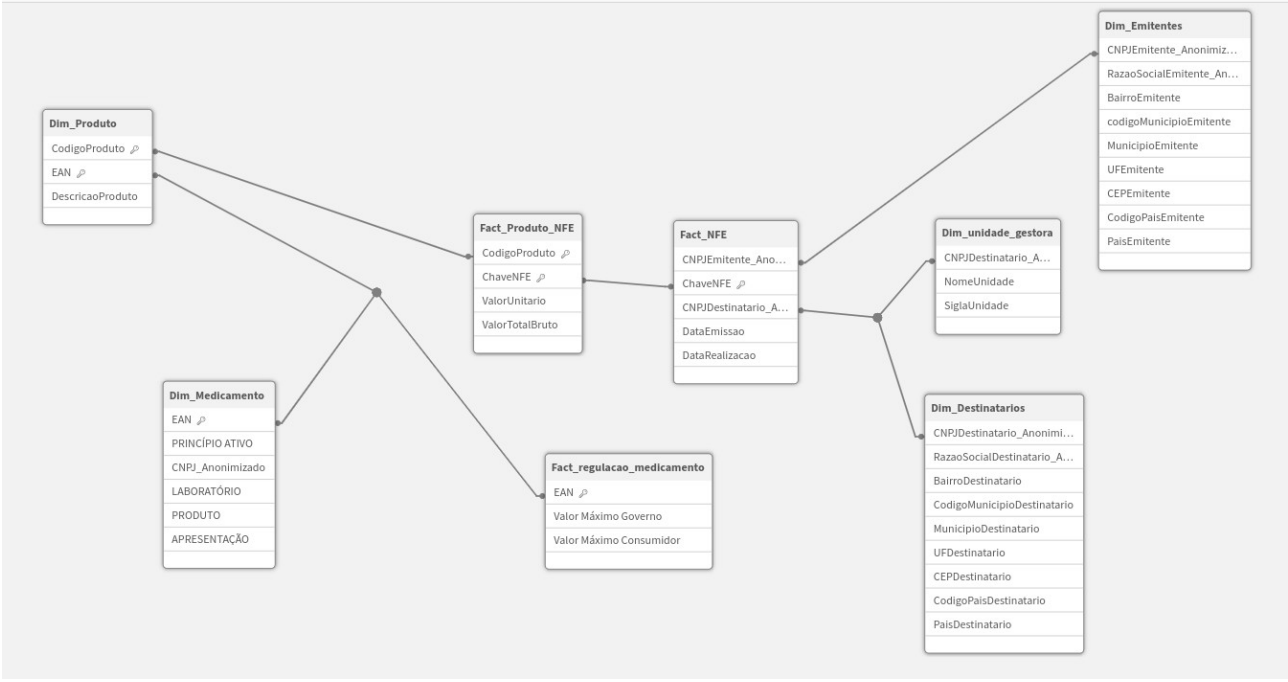
Nesta fase é importante observar se foram criadas chaves sintéticas (chaves compostas), o que não é desejável e pode gerar complexidade no modelo criado. Oportunamente, o presente modelo foi gerado apenas com chaves simples entre as tabelas criadas.

6.3 Modelo de Dados

Para visualizar o modelo criado, vá na seção Preparar e clique em **visualizador do modelo de dados**.



A seguir é apresentado o modelo de dados criado através dos scripts de ETL:



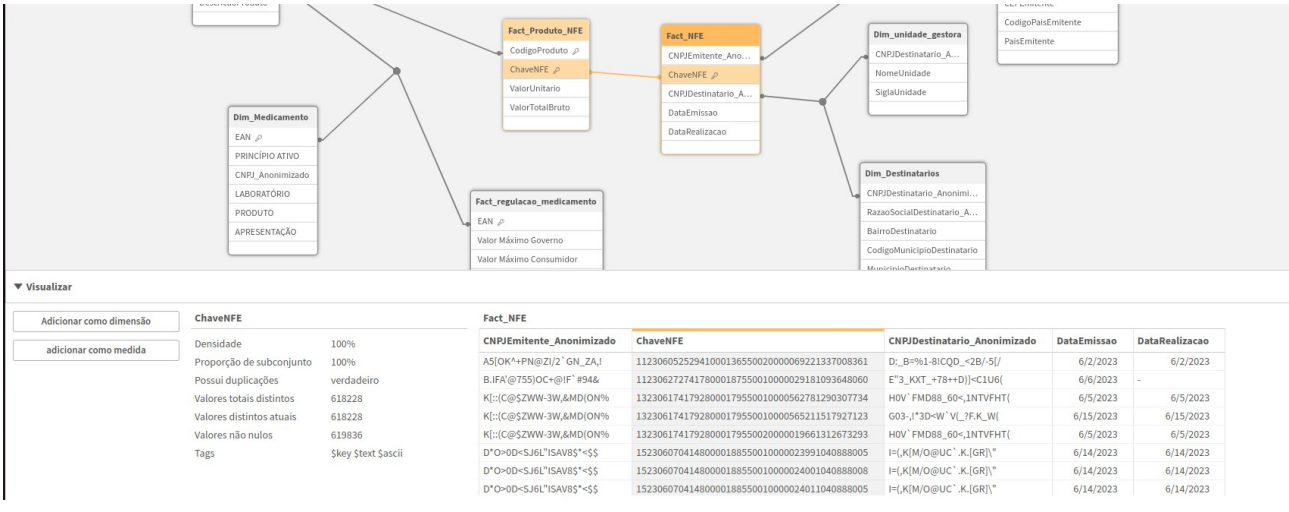
Modelo de dados usando o esquema de floco de neve (snow flake).

7. ANÁLISE DE DADOS

7.1 Análise de qualidade dos dados

Nesta etapa, iniciaremos pela análise da qualidade do conjunto de dados gerados pelos processos de ETL descritos nas etapas anteriores.

O Qlik Cloud possui ferramentas que facilitam a visualização da qualidade dos dados. Ao clicar no elemento **visualizar** situado no canto inferior esquerdo, será aberta uma tela que exibirá as informaçoes de qualidade da tabela ou campo selecionado na visualizaçao.



É possível verificar várias verificações, tais como:

- se há valores duplicados em um campo que represente uma chave única;
- proporção de um subconjunto de registros entre tabelas relacionadas por chaves (em alguns casos é esperado que nem todos os campos de chave correspondam entre as tabelas);
- valores distintos em um campo;
- valores nulos;
- formato do valor de um campo;

Feitas as checagens de qualidade e não se detectando maiores questões, passa-se a à fase de elaboraçao de painel de indicadores.

7.2 Elaboração de painel de indicadores

Nesta etapa, serão elaborados gráficos e tabelas que apresentem as informações requeridas para atender aos requisitos do projeto.

Além da seleção de dimensões e medidas para gerar as tabelas e gráficos necessários, serão utilizadas funções set analysis do Qlik Cloud, que permitem realizar operações em subconjuntos dos dados ou mesmo condicionais.

Por exemplo, para determinar o total de compras de um órgão público, poderiamos usar a expressão simples **Sum(ValorUnitario)**. De modo mais avançado, para determinar o total de compras em desrespeito a regulação, ou seja, o total de compras em que o valor unitário transacionado foi superior ao valor fixado na regulação, seria possível usar a expressão **Sum({<ValorUnitario = {"> [Valor Maximo Governo]"}>} ValorUnitario)**, que realiza a operação de soma apenas para os registros que satisfazem a condição.

Contudo, no projeto em análise o conjunto de dados é consideravelmente grande de modo que a ferramenta apresentou um **erro de limite de memória**. Então como forma de contornar essa limitação, preliminarmente foi gerada uma tabela contendo informações de produtos e preços, para persistir uma coluna em que se verifica se houve desrespeito a regulação.

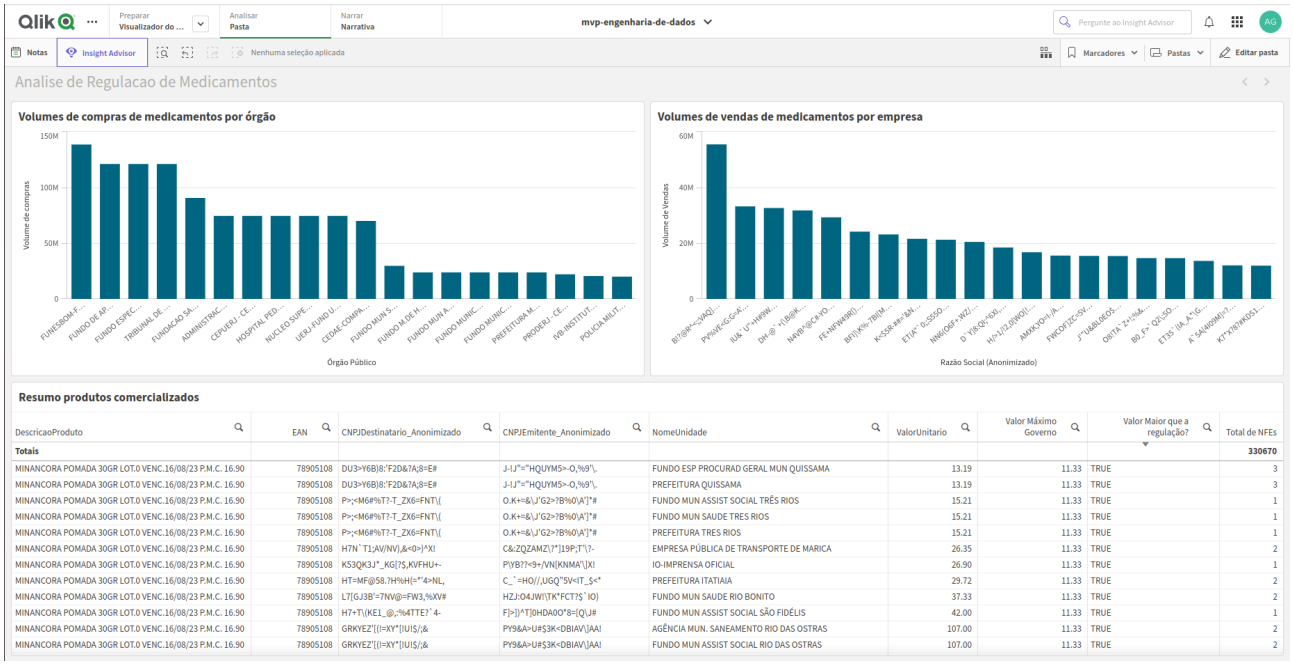
Foi então gerada a tabela Resumo Produtos Comercializados, que apresenta os dados referentes a **DescricaoProduto**, **CNPJDestinarioAnonimizado**, **CNPJEmitenteAnonimizado**, **EAN**, **NomeUnidade**, **ValorUnitario**, [**Valor Máximo Governo]**. Foi incluida ainda como dimensão a coluna [**Valor Maior que a regulação?]** que apresenta o resultado da expressão **=IF(ValorUnitario > [Valor Máximo Governo], 'TRUE', 'FALSE')**. Foi incluída, ainda, a coluna [**Total de NFEs]**, que consiste da agregação **Count(DISTINCT ChaveNFE)**.

Em seguida, foram gerados os seguintes gráficos:

Volumes de compras de medicamentos por órgão, que relaciona as colunas NomeUnidade como dimensão e [**Volume de Compras**] como medida, definda pela expressão **SUM(ValorUnitario)**.

Volumes de Vendas de medicamentos por empresa, que relaciona as colunas **RazaoSocialEmitente_Anonimizado** como dimensão e [**Volume de Vendas**] como medida, definda pela expressão **SUM(ValorUnitario)**.

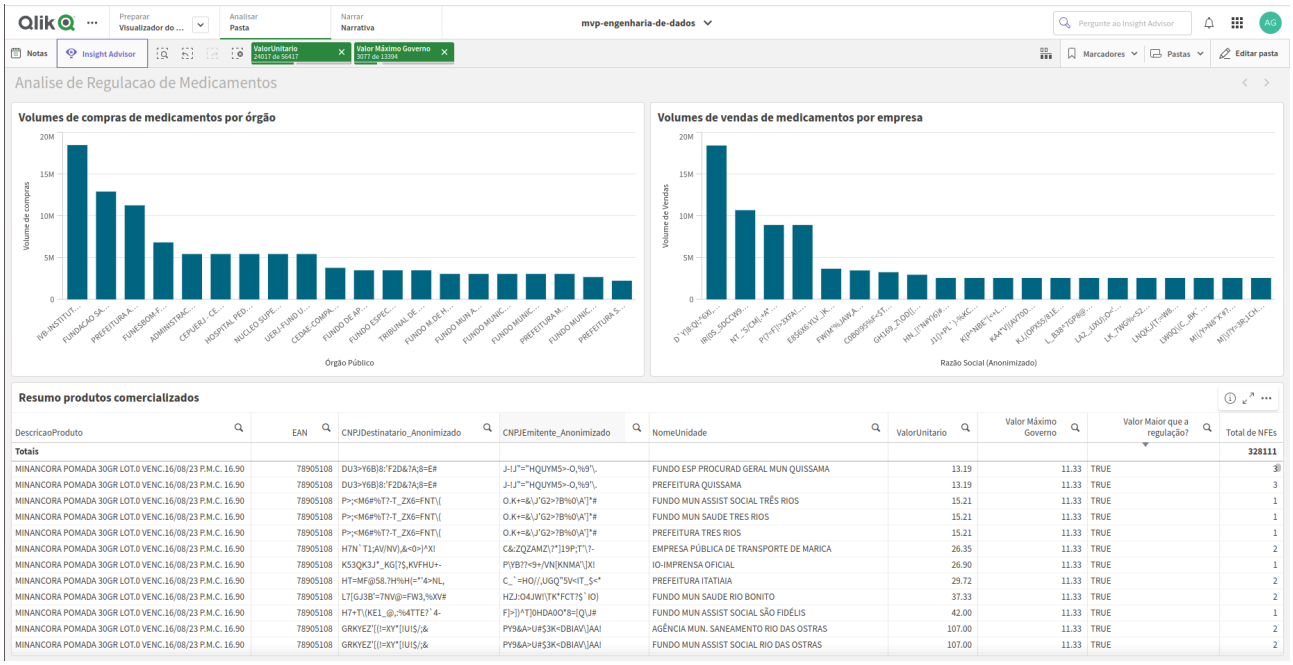
Como resultado, foi elaborado o seguinte painel.



DescricaoProduto	EAN	CNPJDestinatario_Anonimizado	CNPJEmitente_Anonimizado	NomeUnidade	ValorUnitario	Valor Máximo Governo	Valor Maior que a regulação?	Total de NFEs
Totais								**330670**
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	DU3>Y6B)8:'F2D&7A;8=E#	J-IJ"='HQUYM5>-O,%9'\,	FUNDO ESP PROCURAD GERAL MUN QUISSAMA	13.19	11.33	TRUE	3
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	DU3>Y6B)8:'F2D&7A;8=E#	J-IJ"='HQUYM5>-O,%9'\,	PREFEITURA QUISSAMA	13.19	11.33	TRUE	3
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	P>;<M6F%T?-T_ZX6=FNT\(	O.K+=&\J'G2>?B%0\A"]*#	FUNDO MUN ASSIST SOCIAL TRÊS RIOS	15.21	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	P>;<M6F%T?-T_ZX6=FNT\(	O.K+=&\J'G2>?B%0\A"]*#	FUNDO MUN SAUDE TRES RIOS	15.21	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	P>;<M6F%T?-T_ZX6=FNT\(	O.K+=&\J'G2>?B%0\A"]*#	PREFEITURA TRES RIOS	15.21	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	H7N` T1;AV/NV),&<0>)^XI	C&;ZQZAMZ\]*]19P;T\?-	EMPRESA PÚBLICA DE TRANSPORTE DE MARICA	26.35	11.33	TRUE	2
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	K53QK3J*_KG[?S,KVFHU+-	PJYB7?<9+/VN[KNMA\\]XI	IO-IMPRENSA OFICIAL	26.90	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	HT=MF@58.7H%H(="'4>NL,	C_` =HO//,UGQ"5V<IT_$<*	PREFEITURA ITATIAIA	29.72	11.33	TRUE	2
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	L7[GJ3B'=7NV@=FW3,%XV#	HZJ:O4JW\(TK*FCT?$` IO)	FUNDO MUN SAUDE RIO BONITO	37.33	11.33	TRUE	2
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	H7+T\(KE1_@,:%4TTE?`4-	F]>])^T]0HDA0O*8=(Q\J#	FUNDO MUN ASSIST SOCIAL SÃO FIDÉLIS	42.00	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	GRKYEZ"[(I=XY*[!U1$/;&	PY9&A>U#$3K<DBIAV\]AA!	AGÊNCIA MUN. SANEAMENTO RIO DAS OSTRAS	107.00	11.33	TRUE	2
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	GRKYEZ"[(I=XY*[!U1$/;&	PY9&A>U#$3K<DBIAV\]AA!	FUNDO MUN ASSIST SOCIAL RIO DAS OSTRAS	107.00	11.33	TRUE	2

Inicialmente não houve a aplicação de filtros, apenas tendo sido marcadas as opções para não apresentar registros com valores nulos.

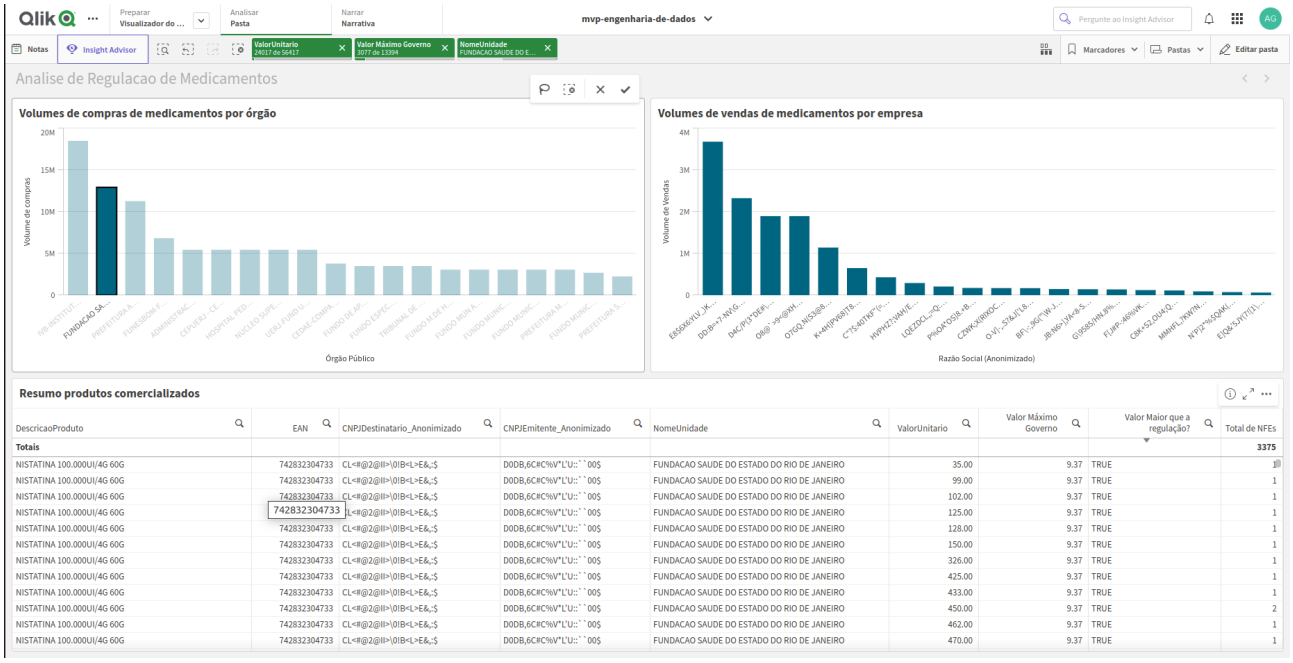
Para avaliar as situações de desrespeito a regulação, basta selecionar a coluna [**Valor Maior que a regulação?]** do gráfico e usar o filtro (lupa) para escolher os valores iguais a TRUE. A figura a seguir apresenta o resultado deste filtro.



DescricaoProduto	EAN	CNPJDestinatario_Anonimizado	CNPJEmitente_Anonimizado	NomeUnidade	ValorUnitario	Valor Máximo Governo	Valor Maior que a regulação?	Total de NFEs
Totais								**328111**
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	DU3>Y6B)8:'F2D&?A;8=E#	J-I.J"="HQUYM5>-O,%9'\,	FUNDO ESP PROCURAD GERAL MUN QUISSAMA	13.19	11.33	TRUE	3
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	DU3>Y6B)8:'F2D&?A;8=E#	J-I.J"="HQUYM5>-O,%9'\,	PREFEITURA QUISSAMA	13.19	11.33	TRUE	3
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	P>;<M6F%T?-T_ZX6=FNT\(	O.K+=&\.J'G2>?B%0\A"]*#	FUNDO MUN ASSIST SOCIAL TRÊS RIOS	15.21	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	P>;<M6F%T?-T_ZX6=FNT\(	O.K+=&\.J'G2>?B%0\A"]*#	FUNDO MUN SAUDE TRES RIOS	15.21	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	P>;<M6F%T?-T_ZX6=FNT\(	O.K+=&\.J'G2>?B%0\A"]*#	PREFEITURA TRES RIOS	15.21	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	H7N` T1;AV/NV),&<0>)^XI	C&;ZQZAMZ\?*]19P;T\?-	EMPRESA PÚBLICA DE TRANSPORTE DE MARICA	26.35	11.33	TRUE	2
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	K53QK3J*_KG[?S,KVFHU+-	PJYB?7<9+/VN[KNMA\']XI	IO-IMPRENSA OFICIAL	26.90	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	HT=MF@58.7H%H(=""4>NL,	C_`-=HO//,UGQ"5V<IT_$<*	PREFEITURA ITATIAIA	29.72	11.33	TRUE	2
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	L7[GJ3B'=7NV@=FW3,%XV#	HZJ:O4JW\'(TK*FCT?$` IO)	FUNDO MUN SAUDE RIO BONITO	37.33	11.33	TRUE	2
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	H7+T\(KE1_@,:%4TTE?`4-	F]>]]\^T]0HDA0O*8=[Q]J#	FUNDO MUN ASSIST SOCIAL SÃO FIDÉLIS	42.00	11.33	TRUE	1
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	GRKYEZ"[(I=XY*[!U!$/;&	PY9&A>U#$3K<DBIAV\]AAI	AGÊNCIA MUN. SANEAMENTO RIO DAS OSTRAS	107.00	11.33	TRUE	2
MINANCORA POMADA 30GR LOT.0 VENC.16/08/23 P.M.C. 16.90	78905108	GRKYEZ"[(I=XY*[!U!$/;&	PY9&A>U#$3K<DBIAV\]AAI	FUNDO MUN ASSIST SOCIAL RIO DAS OSTRAS	107.00	11.33	TRUE	2

Agora já é possível visualizar os órgãos públicos e empresas relacionados a operações de vendas de medicamentos por preços superiores a regulação.

É possível utilizar outros filtros. Agora vamos selecionar o segundo órgão público que operou o maior volume de compras de medicamentos em desrespeito a regulação e observar os empresas que participaram dessas vendas.



DescricaoProduto	EAN	CNPJDestinatario_Anonimizado	CNPJEmitente_Anonimizado	NomeUnidade	ValorUnitario	Valor Máximo Governo	Valor Maior que a regulação?	Total de NFEs
Totais								**3375**
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	35.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	99.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	102.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	125.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	128.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	150.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	326.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	425.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	433.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	450.00	9.37	TRUE	2
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	462.00	9.37	TRUE	1
NISTATINA 100.000UI/4G 60G	742832304733	CL<#@2@II>\0IB<L>E&,:$	D0DB,6C#C%V*L'U::`` 00$	FUNDACAO SAUDE DO ESTADO DO RIO DE JANEIRO	470.00	9.37	TRUE	1

A partir de combinações de filtros entre os elementos do painel é possível a geração de maior conhecimento sobre as relações de interesse para os objetivos das equipes de auditoria.

8. CONCLUSÕES

Através do presente trabalho foi possível elaborar um pipeline de coleta, modelagem, carga e análise de dados, utilizando-se de tecnologias de computação em nuvem (S3 da AWS e Qlick Cloud).

Foi verificado que, em conjunto com outros dados públicos, os dados presentes nas Notas Fiscais Eletrônicas podem ser utilizados para detectar ocorrências de vendas de medicamentos aos órgãos do Governo por preços superiores aos permitidos pela regulação em vigor.

Através de gráficos e tabelas gerados pela aplicação criada no Qlik Sense foi possível identificar quais os órgão do governo e empresas (anonimizadas neste trabalho) que estão relacionados a operações de vendas de medicamentos em desrespeito a regulação.

Esses gráficos e tabelas que podem ser utilizados pelas equipes de auditoria como ferramentas de apoio a decisão, permitindo a geração de conhecimento e consequentemente a focalização de esforços nas situações que evidenciam maior impacto para os objetivos das fiscalizações a serem realizadas.