PUC RIO

PÓS GRADUAÇÃO EM CIÊNCIA DE DADOS E ANALYTICS

RELATÓRIO SPRINT 3 – MVP – ENGENHARIA DE DADOS

Aluno: Rubens L. Cirino

Sumário

Índice de Figuras

Índice de Tabelas

Objetivo

O objetivo do presente trabalho é o de pesquisar uma base de dados de um supermercado que contêm filiais físicas em algumas regiões e, também, dois canais de venda: pessoa jurídica (hotéis, cafés e restaurantes) e pessoa física (varejo). Vamos fazer algumas comparações entre produtos de forma que as respostas possam vir a melhorar as tomadas de decisões estratégicas sobre o negócio. O Board do supermercado está interessado em saber o volume de vendas de cada produto por canal e por região. Também buscam saber se há alguma associação entre compras de produtos congelados e/ou frescos e os canais de vendas usados.

Detalhamento

Busca de dados.

Seguindo as orientações que veem sendo passadas pela coordenação, fizemos uma busca nos links que oferecem bases dados abertas.

O dataset usado neste projeto será o Wholesale customers, proveniente do *UCI Machine Learning Repository*. O presente dataset é um subset de uma base de dados maior, referida em Abreu, N. (2011). "Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional" – Mestrado em Marketing , ISCTE-IUL, Lisboa. Seu objetivo é mostrar o volume anual de vendas de um comércio atacadista, expresso em unidades monetárias, de diversas categorias de produtos. Para mais detalhes sobre o dataset, consulte: https://archive.ics.uci.edu/ml/datasets/Wholesale+customers#

Informações sobre os atributos:

Channel – qual o tipo de canal de venda foi usado para a compra:

a. Horeca (Hotel/Restaurante/Café) – pessoa jurídica

b. Varejo – pessoa física;

Region:

a. Lisboa

b. Porto

c. Outra Região;

Fresh – são as vendas anuais em unidades de dinheiro com produtos frescos;

Milk – são as vendas anuais em unidades de dinheiro de laticínios;

Grocery – são as vendas anuais em unidades de dinheiro de produtos de mercado;

Frozen – são as vendas anuais em unidades de dinheiro de congelados;

Detergents_Paper – são as vendas anuais em unidades de dinheiro em produtos de higiene e limpeza;

Delicassen – são as vendas anuais em unidades de dinheiro de produtos classificados como delicatessen;

Coleta

Neste projeto, usamos a plataforma da Google.

Começamos com a instanciação do Data Fusion

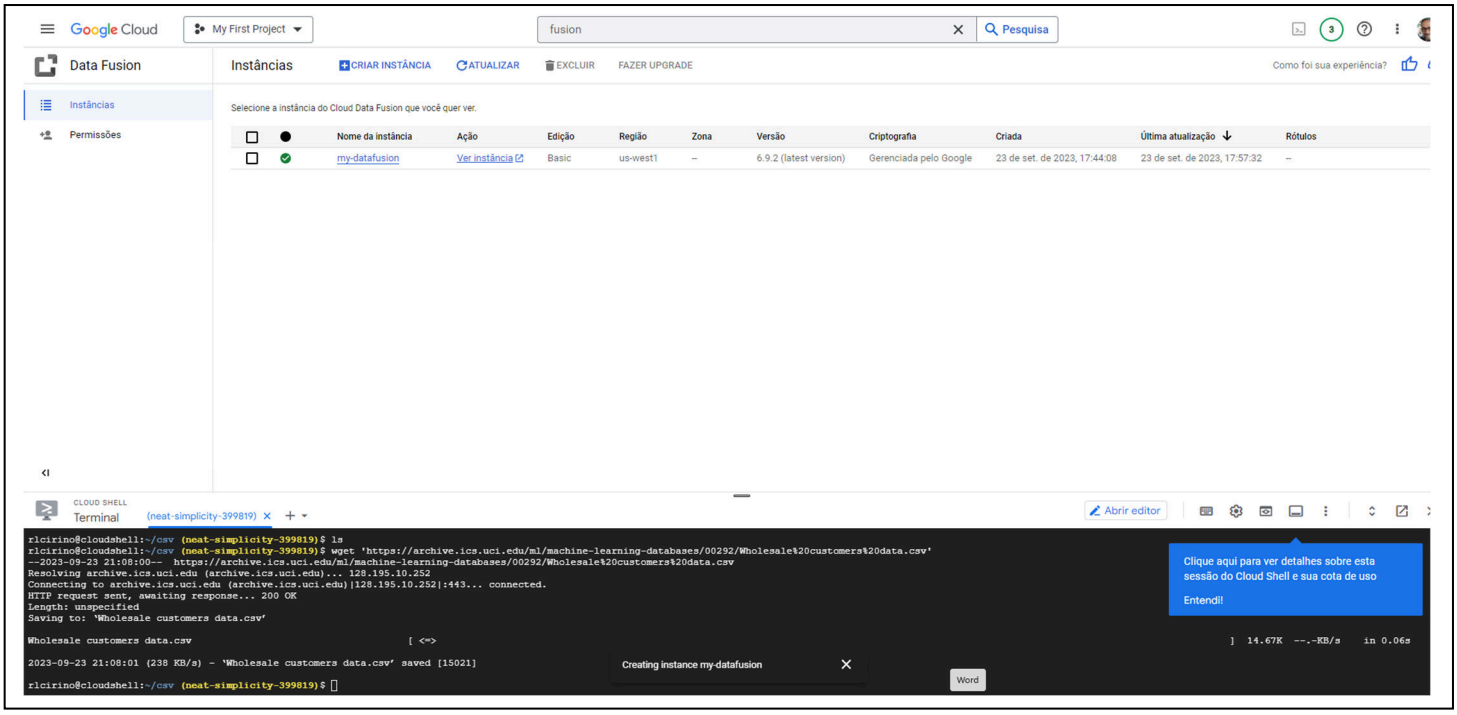

Figura 1: Instanciação do Data Fusion

Depois fizemos a criação do bucket e carregamos o arquivo de dados

Em seguida, criamos o bucket temporário.

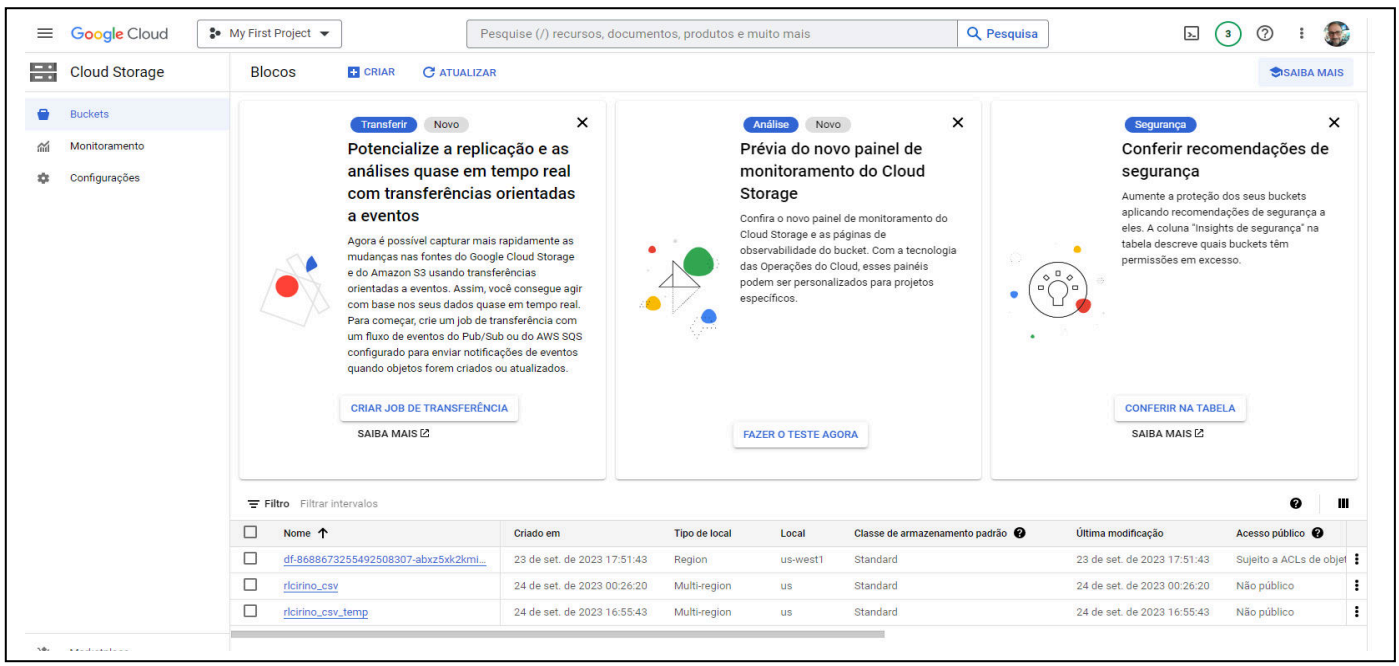

Figura 2: Criação do bucket temporário

Após, criamos o dataset no Bigquery

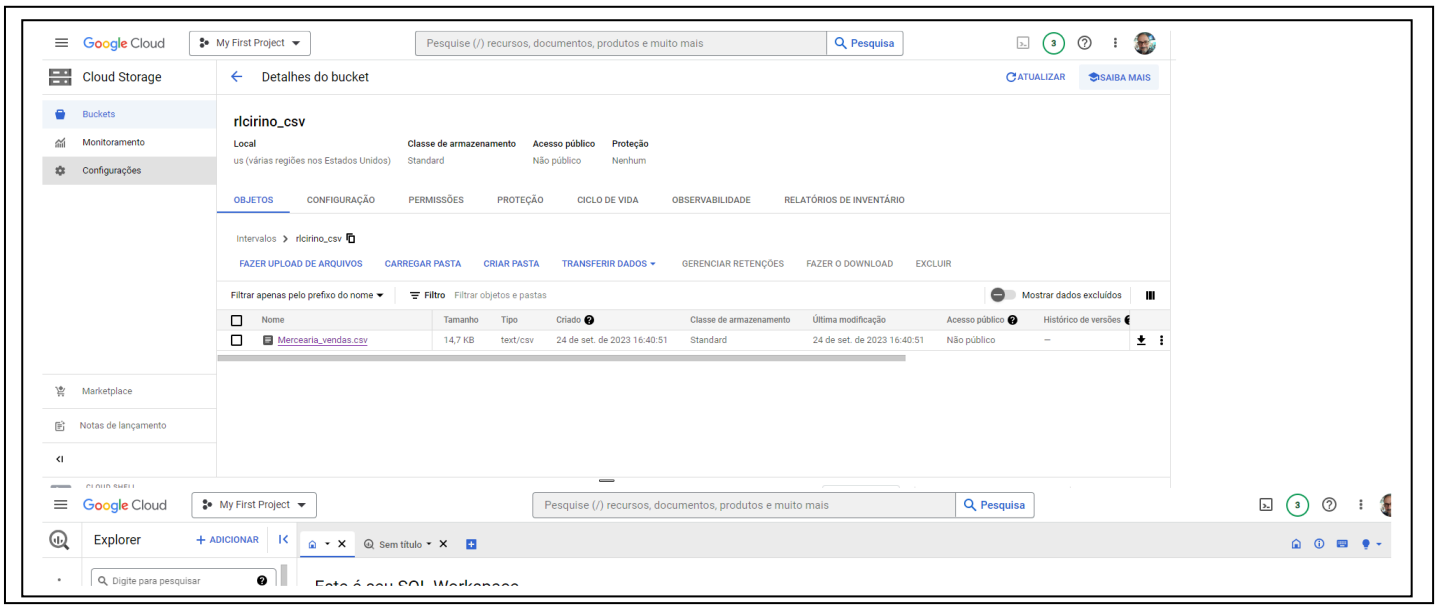

Figura 3: Criação do dataset no Bigquery

Modelagem

Para o presente projeto, usamos uma única tabela que contém todos registros.

Usamos a ferramenta Dataplex para montar o catálogo de dados da base utilizada.

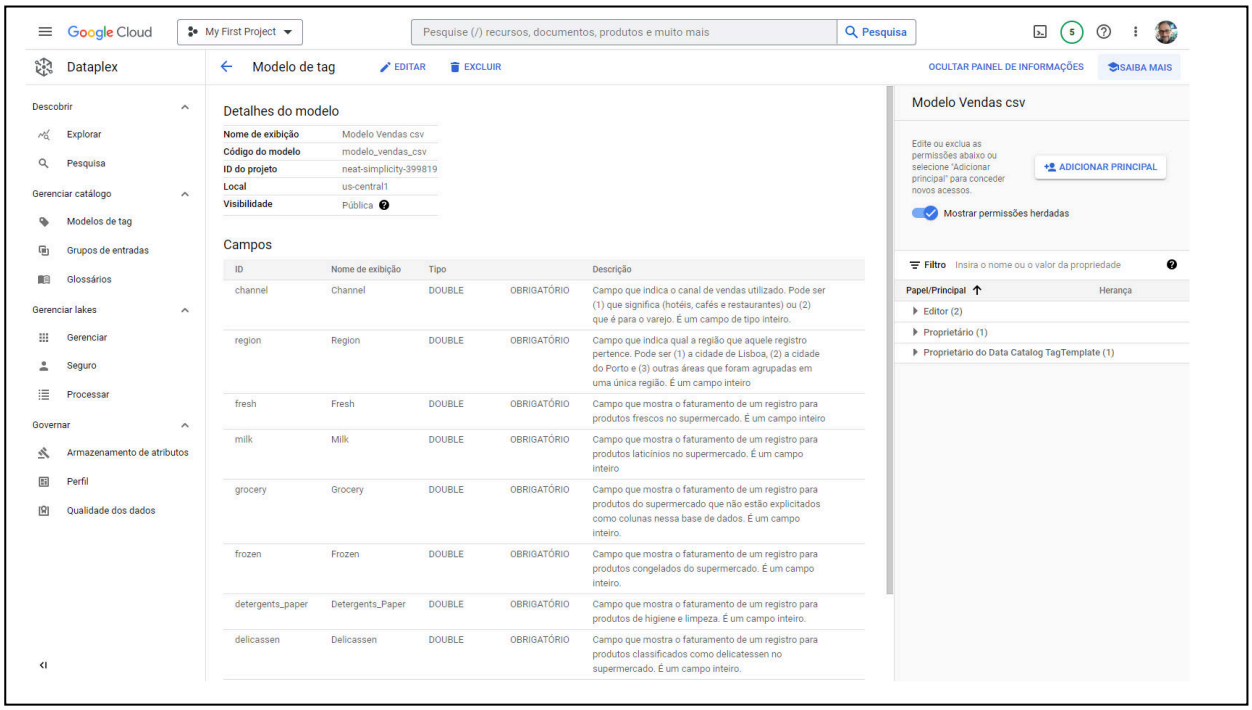

Figura 4: Criação do catálogo de dados no Dataplex

Após, fizemos uma edição do esquema da tabela.

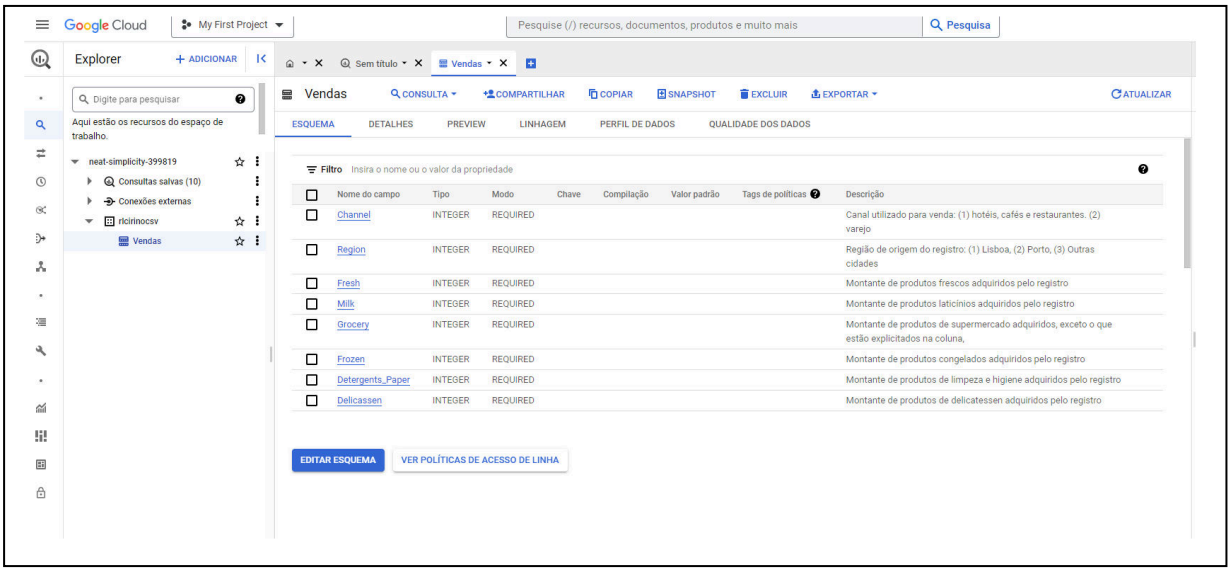

Figura 5: Edição do esquema da tabela

Carga

Nesta etapa fizemos a carga dos dados para o Data Warehouse. Utilizamos a ferramenta de ETL – Data Fusion. A escolha desta ferramenta se deu à partir da aula de dúvidas conduzida pelo professor. Dentre as ferramentas vistas, o Data Fusion, nos pareceu mais adequado.

O diagrama montado na ferramenta ETL ficou como apresentado abaixo:

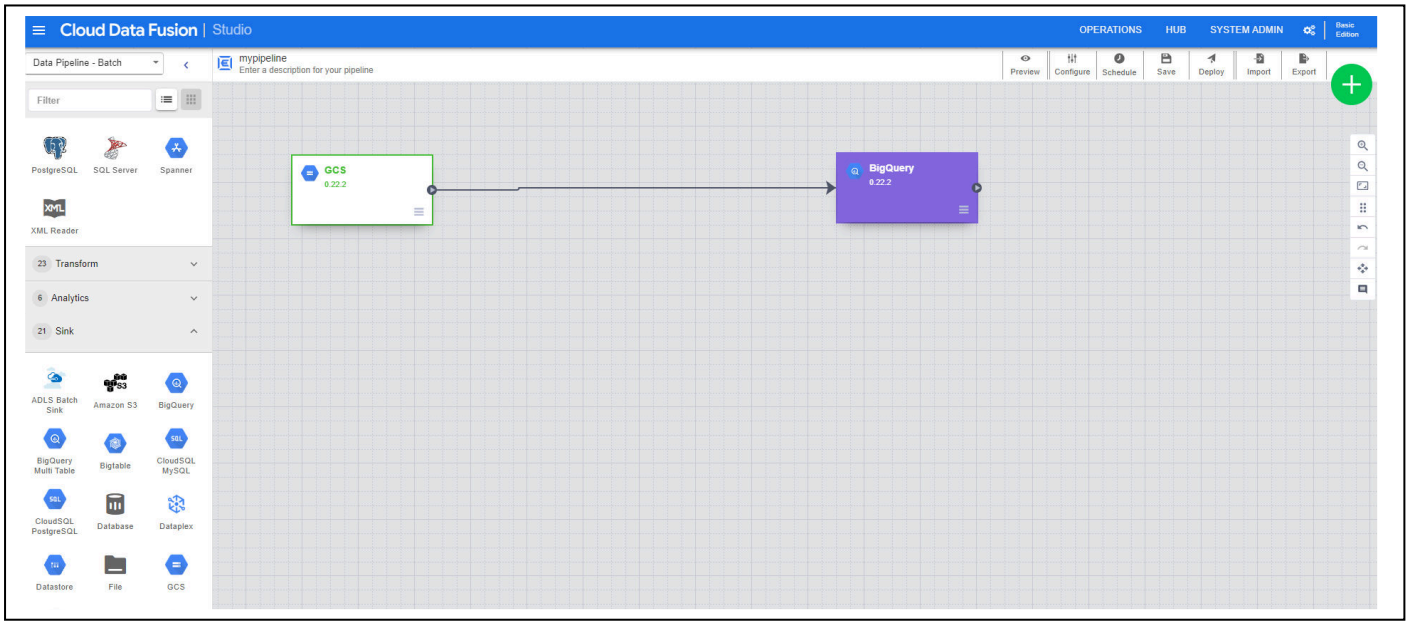

Figura 6: Diagrama de fluxo montado no Data Fusion

Na parte de extração usamos o GCS que é ofertado pela ferramenta. Os dados são extraídos da área da Google Cloud. A descrição do GCS ficou como mostrado à seguir:

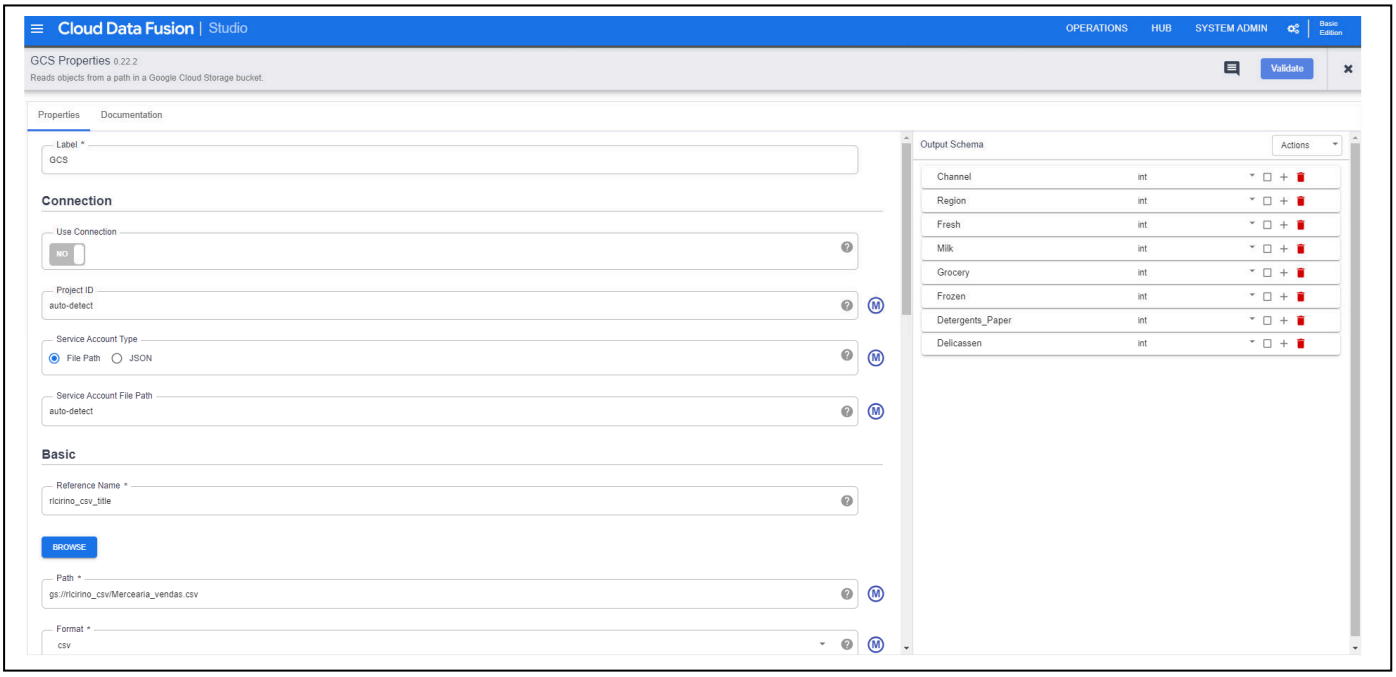

Figura 7: Extração dos dados utilizando-se o GCS

A base utilizada para o projeto já se encontrava, no site original, totalmente acertada. Não houve a necessidade de se fazer uma transformação na ferramenta ETL. Caso fosse necessário, iríamos utilizar o Wrangler para efetuarmos uma retirada de coluna ou substituição de algum conteúdo presente na tabela.

Após a instanciação do GCS, fizemos o carregamento do Bigquery (próximas duas imagens).

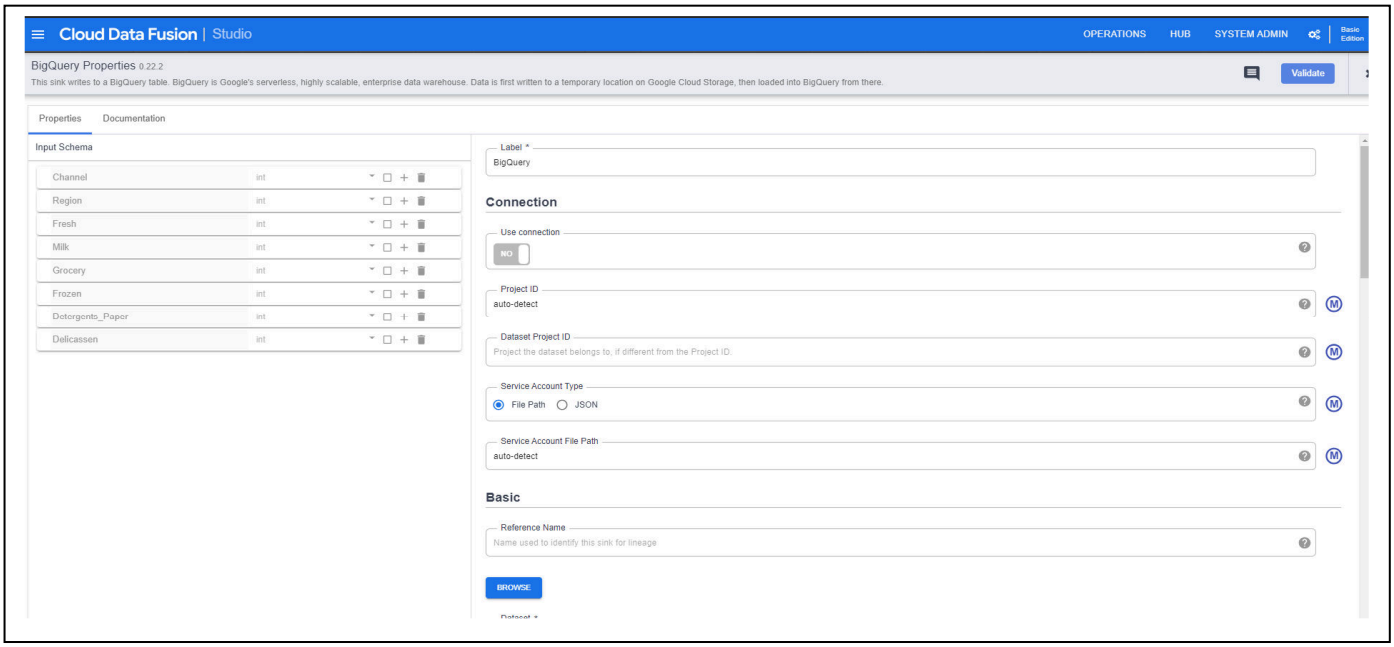

Figura 8: Carregamento no Bigquery (parte 1)

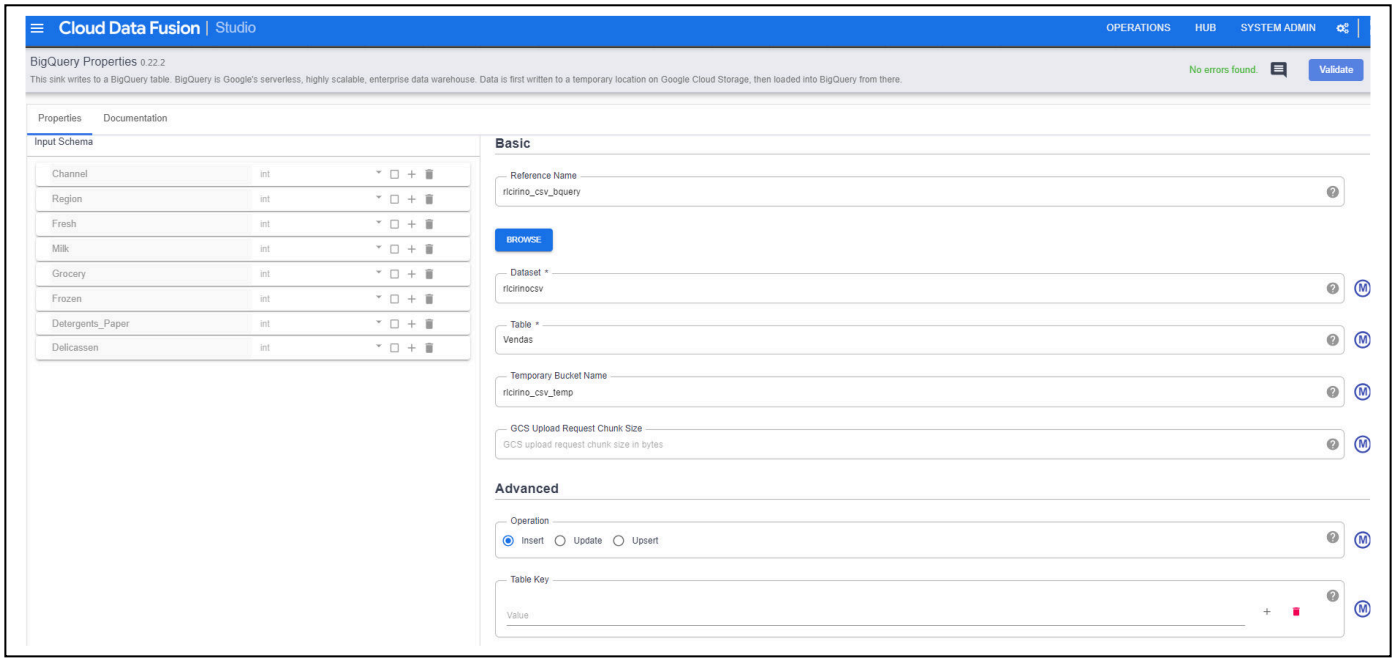

Figura 9: Carregamento no Bigquery (parte 2)

Em seguida, fizemos o deploy do fluxo, como mostrado na figura à seguir:

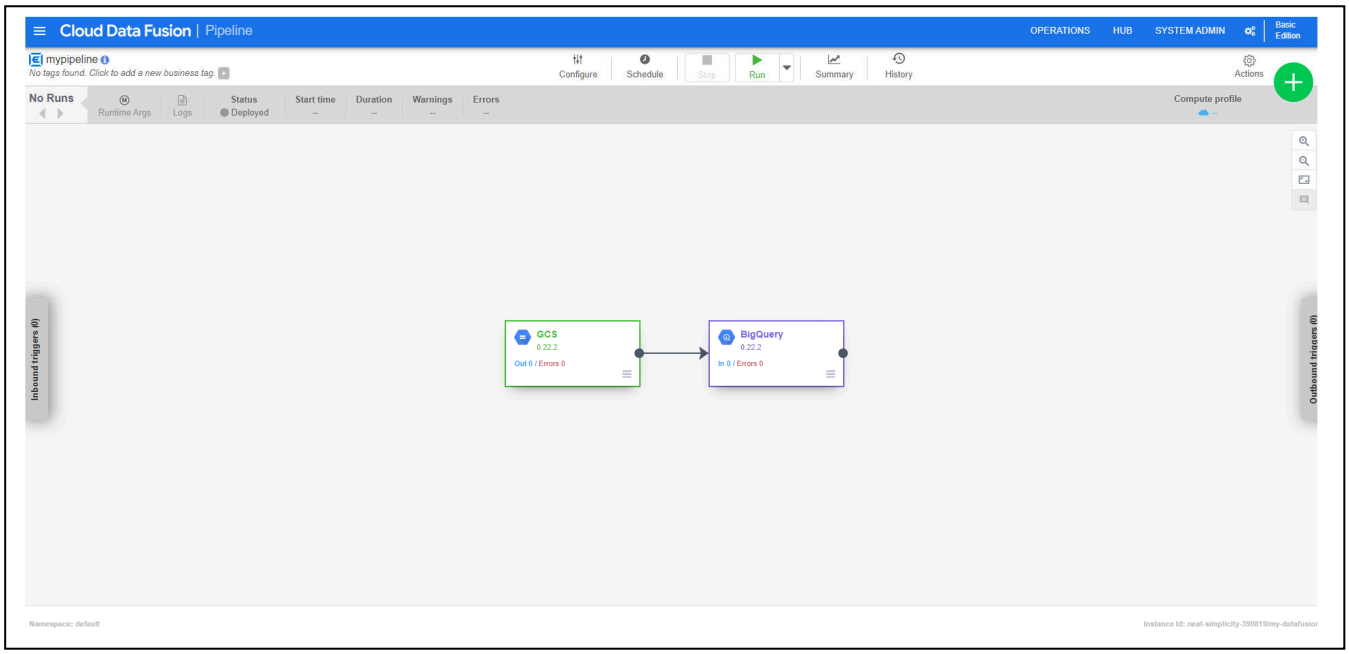

Figura 10: Deploy do fluxo

Com o fluxo com o deploy feito, o Bigquery pode ser carregado como mostra a figura à seguir:

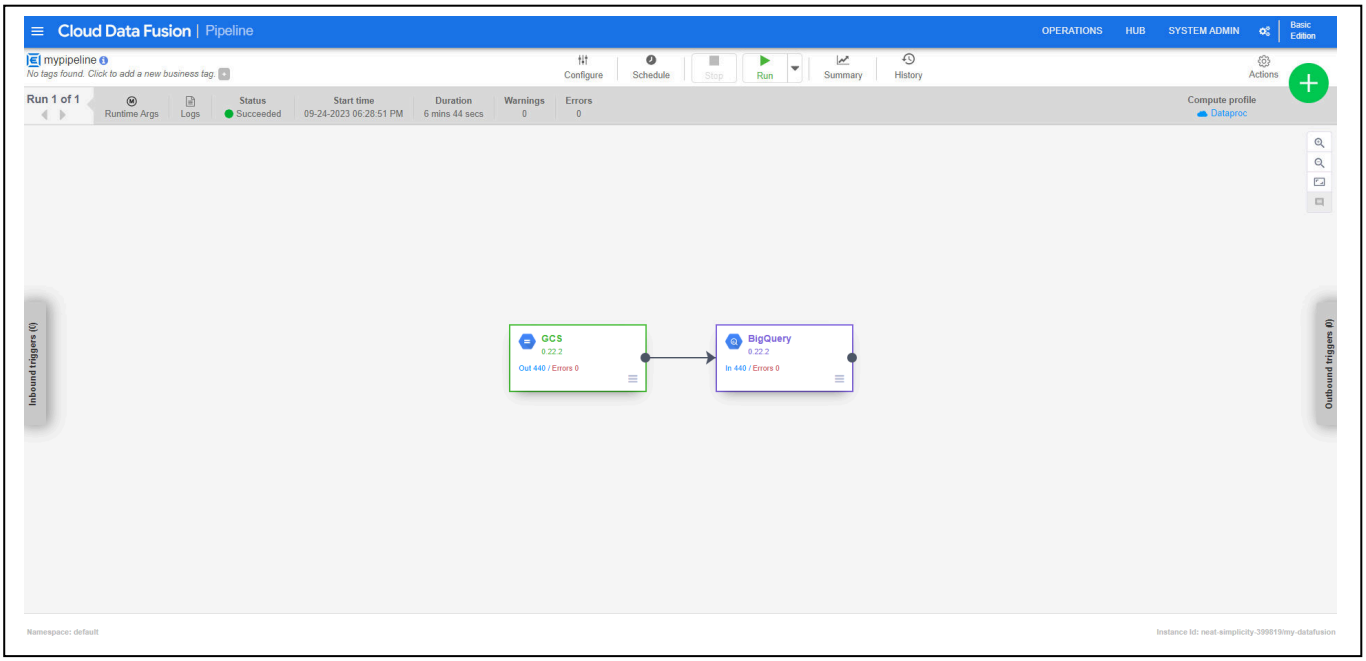

Figura 11: Fluxo com o Deploy feito

Análise

a) Qualidade de dados

O conjunto de dados que foi trabalhado neste projeto não contém registros em branco. Trata-se de uma base onde o conteúdo refere-se a valores monetários de vendas de um supermercado. A base que foi disponibilizada no site para ser usada, é um subconjunto de uma base maior. A verificação de que não há registros em branco foi feita utilizando-se o ambiente Google Colab através de um código Python, como mostra a figura abaixo:

```
# Mostra as informações do dataset (tipos de dados)
print(mercearia.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 440 entries, 0 to 439
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Channel           440 non-null    int64
 1   Region            440 non-null    int64
 2   Fresh             440 non-null    int64
 3   Milk              440 non-null    int64
 4   Grocery           440 non-null    int64
 5   Frozen            440 non-null    int64
 6   Detergents_Paper  440 non-null    int64
 7   Delicassen        440 non-null    int64
dtypes: int64(8)
memory usage: 27.6 KB
None
```

Figura 12: Informações do dataset utilizado na análise

b) Solução do Problema

Seguindo o checklist disponibilizado para a entrega do presente MVP, vamos buscar as respostas para as perguntas elencadas no item “Objetivo”. Para a primeira pergunta: O Board do supermercado está interessado em saber o volume de vendas de cada produto por canal e por região.
Para isso fizemos as seguintes consultas SQL para levantarmos os dados e gerarmos a tabela que se segue.

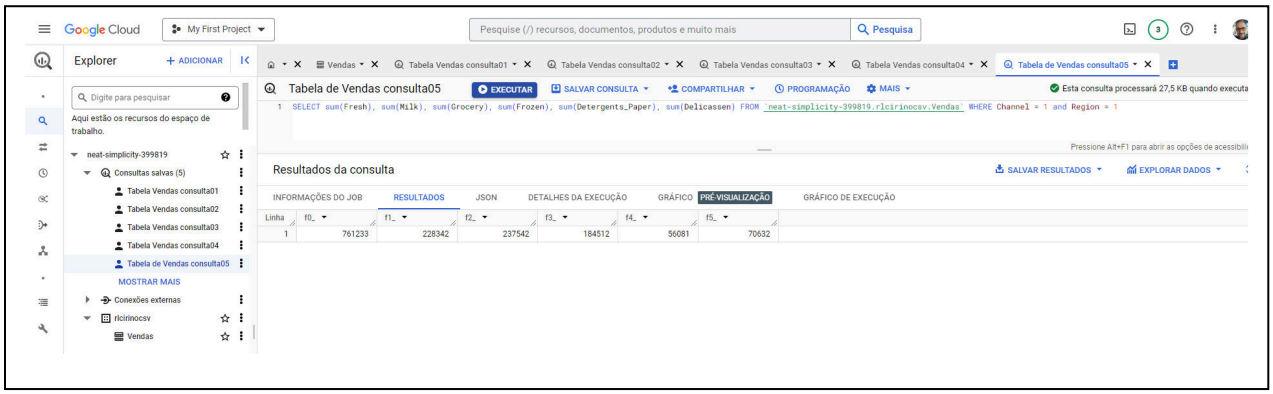

Figura 13: Valores das vendas somadas por produto pelo canal 1 para a região 1

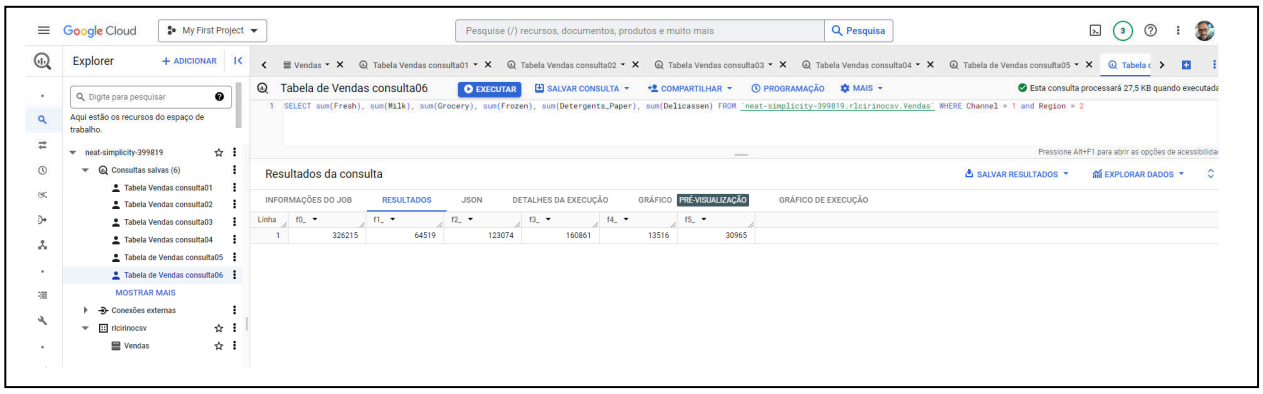

Figura 14: Valores das vendas somadas por produto pelo canal 1 para a região 2

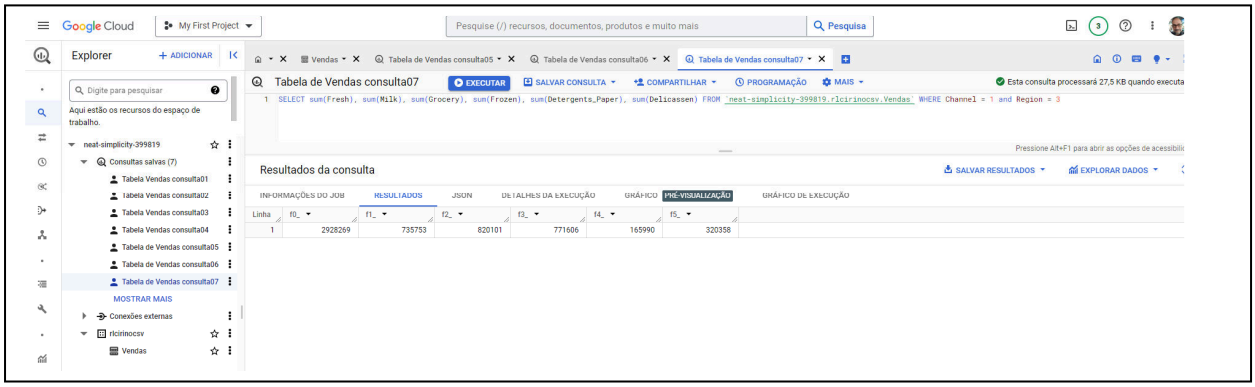

Figura 15: Valores das vendas somadas por produto pelo canal 1 para a região 3

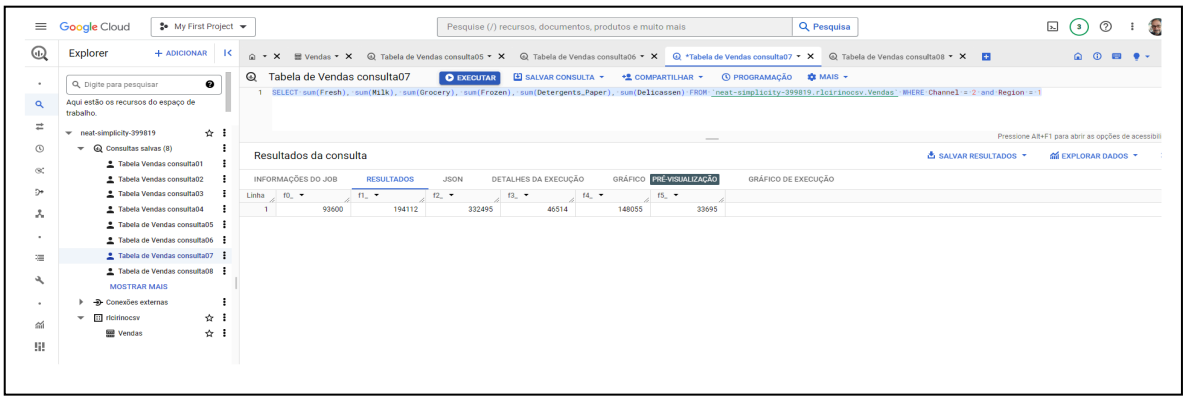

Figura 16: Valores das vendas somadas por produto pelo canal 2 para a região 1

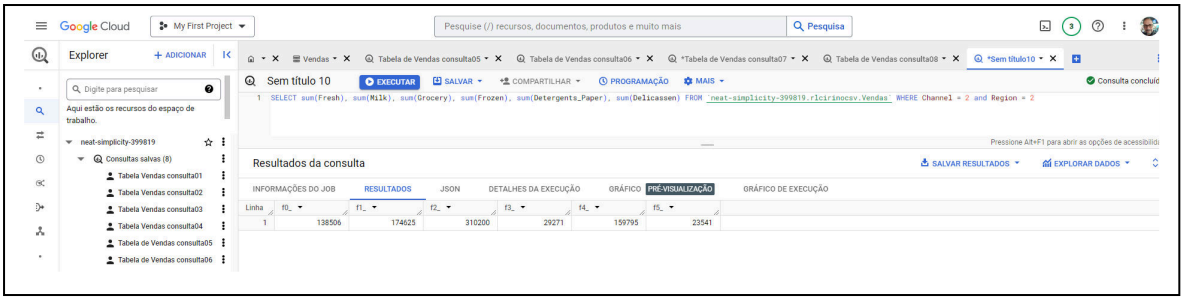

Figura 17: Valores das vendas somadas por produto pelo canal 2 para a região 2

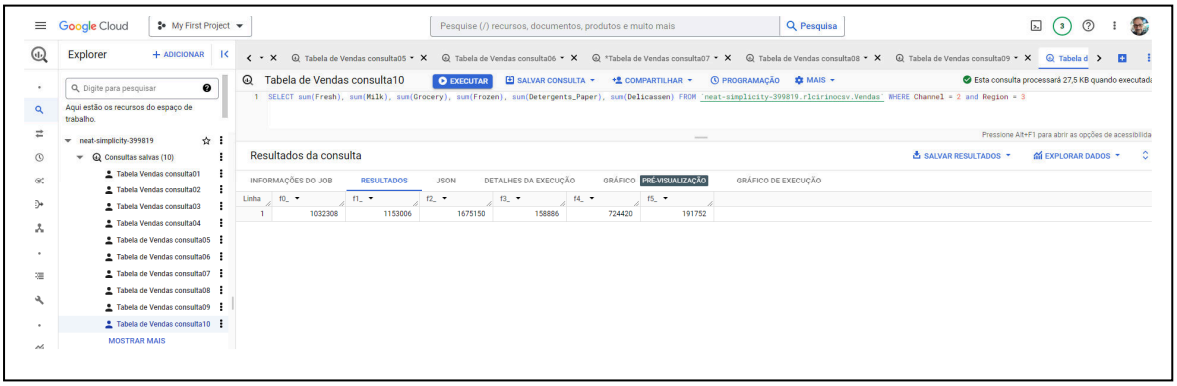

Figura 18: Valores das vendas somadas por produto pelo canal 2 para a região 3

A seguir vamos gerar uma tabela com o consolidado dos resultados de todas as consultas ao banco de dados:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
C1R1	761233	228342	237542	184512	56081	70632
C1R2	326215	64519	123074	160861	13516	30965
C1R3	2928269	735753	820101	771606	165990	320358

C2R1	93600	194112	332495	46514	148055	33695
C2R2	138506	174625	310200	29271	159795	23541
C2R3	1032308	1153006	1675150	158886	724420	191752

Tabela 1: consolidado dos resultados de todas as consultas ao banco de dados

Legenda:

C1R1 – canal 1 região 1	C2R1 – canal 2 região 1
C1R2 – canal 1 região 2	C2R2 – canal 2 região 2
C1R3 – canal 1 região 3	C2R3 – canal 2 região 3

Fazendo uma análise para buscar responder à primeira pergunta feita pelo Board, verifica-se que há diferenças relevantes, mas já esperadas, na distribuição das vendas por canal e por região. Quando se compara os resultados dos canais dentro de uma mesma região nota-se que:

O canal 1 (que atende hotéis, restaurantes e cafés) teve um desempenho melhor, em quase todos os itens, que o canal 2 (que atende o varejo) na região 1 (cidade de Lisboa). A linha “Variação %” (tabela abaixo) mostra o percentual do valor obtido pelo canal 2 em relação ao canal 1. Observa-se que os itens Grocery e Detergents_Paper foram os únicos que apresentaram um desempenho inferior no valor das compras. Os resultados obtidos podem, talvez, mostrar uma tendência de perfil de compras dos clientes. Para itens comuns de mercado e higiene, o canal mais usado é o do varejo. Já para os demais itens, que envolvem alimentos, o canal que atende estabelecimentos comerciais, foi superior em todos eles. Já para a região 2 (cidade do Porto), o resultado diverge pois os percentuais que mostram a diferença do volume entre os canais 1 e 2 ficam mais próximos. Vale ressaltar a diferença no item Detergents_Paper onde o canal de varejo conseguiu um volume muito superior ao canal 1. Entendemos que neste caso, o perfil do consumidor da cidade do Porto é diferente do de Lisboa. A cidade do Porto é bem menor que Lisboa. Talvez esse seja um fator determinante para uma preferência de escolha pelo canal do varejo (canal 2).

Quando levamos a análise para a região 3, vemos que os números são bem diferentes. A região 3 engloba várias outras cidades. Seus números foram colocados juntos a fim de diminuir a complexidade da tabela. Contudo, entendemos que talvez ficasse mais equilibrado se o autor da tabela tivesse deixado, talvez, mais 3 regiões específicas. Os números da região 3 são condizentes com a decisão de juntar outras cidades menores onde o supermercado atua. Há um certo equilíbrio entre as compras feitas pelos canais disponibilizados.

Um ponto que nos chamou a atenção foi o fato de que em qualquer região, as compras de Grocery e Detergents_Papers sempre foram maiores no canal do varejo (canal 2). Isso pode nos levar a uma conclusão sobre o perfil de preferência do cliente do supermercado.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
C1R1	761233	228342	237542	184512	56081	70632
C2R1	93600	194112	332495	46514	148055	33695
Variação %	12,30	85,01	139,97	25,21	264,00	47,71
C1R2	326215	64519	123074	160861	13516	30965
C2R2	138506	174625	310200	29271	159795	23541
Variação %	42,46	270,66	252,04	18,20	1182,27	76,02
C1R3	2928269	735753	820101	771606	165990	320358
C2R3	1032308	1153006	1675150	158886	724420	191752
Variação %	35,25	156,71	204,26	20,59	436,42	59,86

Tabela 2: Comparação de resultados entre canais e regiões

Com relação à segunda pergunta colocada: se há alguma associação entre compras de produtos congelados e/ou frescos e os canais de vendas usados. As respostas obtidas das consultas não demonstram, claramente, uma preferência de compra entre produtos frescos ou/e congelados e o canal de vendas. Em todas as regiões o canal preferido sempre foi o canal 1. O canal do varejo foi sempre apresentou um resultado menor para os itens analisados independente da região observada.

Avaliando, de forma geral, as perguntas feitas sobre a base de dados, entendemos que um volume maior de registros não alteraria os percentuais encontrados. Contudo, acreditamos que um desmembramento da Região 3 em outras regiões, explicitando mais algumas cidades, talvez trouxesse uma visão mais clara do perfil de cliente que o supermercado tem na região que atua.

Autoavaliação

Esta sprint foi um desafio. Embora tenha feito um estudo do material distribuído, entendo que a falta de experiência nas ferramentas utilizadas, principalmente, para o MVP, me trouxe um fator complicador a mais. Procurei realizar um trabalho, sem muitos requintes, de maneira que pudesse mitigar o risco de algo dar errado e eu não conseguir resolver dentro do tempo disponível para a conclusão. Os requisitos do checklist do trabalho foram atendidos. As transmissões via Zoom foram críticas para mostrar as possibilidades. A utilização das ferramentas, não somente desta sprint, normalmente,

têm vários detalhes que se não fossem apresentados nas transmissões teriam gerado um impacto maior na minha alocação de tempo para resolver as ações do MVP.