INSTITUTO INFNET
ENGENHARIA DE DADOS

FELIPE MELO BUARQUE DE GUSMÃO

INFRAESTRUTURA HADOOP

RIO DE JANEIRO
2025

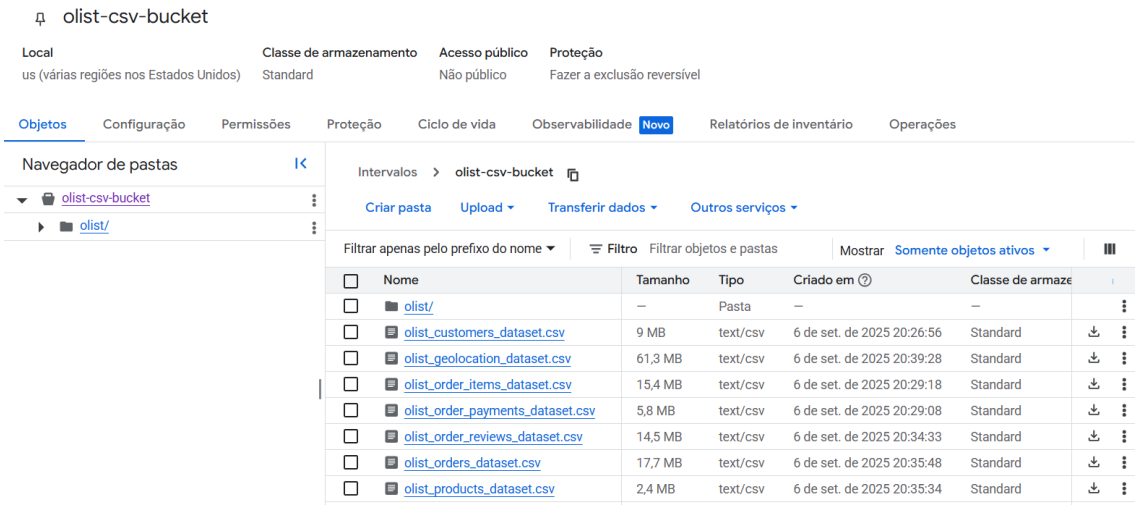
Link repositório: [felipemelobginfnet/-pos_hadoop]

Foi criada uma VM utilizando dataproc do GCP.

Script de conexão, login via SSH utilizando a UI do GCP no navegador.

Ingestão de dados no bucket via UI do GCP.

beeline -u jdbc:hive2://127.0.0.1:10000 -n felipe_gusmao

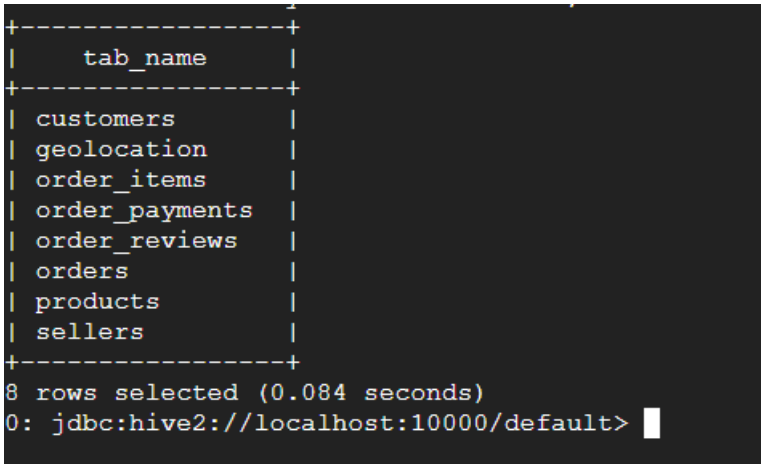


Bucket no GCS: gs://olist-csv-bucket/.

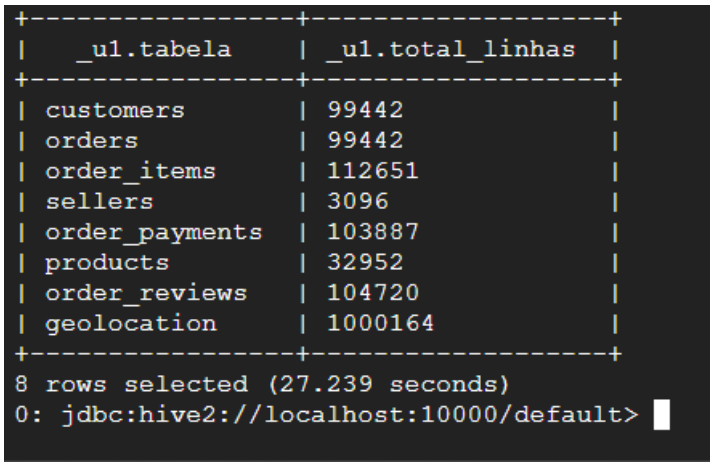
Script de criação das tabelas:

/Script_criacao_tabelas.sql

Tabelas contidas no banco:



Tabelas estão populadas



```
+-----------------+--------------------+
|   _u1.tabela    | _u1.total_linhas   |
+-----------------+--------------------+
| customers       | 99442              |
| orders          | 99442              |
| order_items     | 112651             |
| sellers         | 3096               |
| order_payments  | 103887             |
| products        | 32952              |
| order_reviews   | 104720             |
| geolocation     | 1000164            |
+-----------------+--------------------+
8 rows selected (27.239 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Perguntas Analíticas:

*por questões de estouro de memória da instância ser a mínima possível, não foi possível fazer consultas complexas, sendo uma proposta para próximos passos, criação de tabelas pensadas para inteligência de negócio, já agregando informações importantes para a área de negócio.

Pedidos por estado: Qual estado possui mais pedidos?

1. Pedidos por estado

```
SELECT c.customer_state, COUNT(*) AS total_pedidos
FROM orders o
JOIN customers c ON o.customer_id = c.customer_id
GROUP BY c.customer_state
ORDER BY total_pedidos DESC;
```

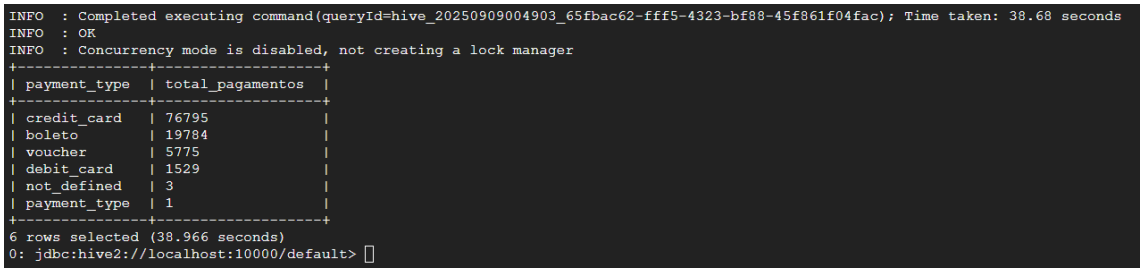
```
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------+-----------------+
| c.customer_state   | total_pedidos   |
+--------------------+-----------------+
| SP                 | 41746           |
| RJ                 | 12852           |
| MG                 | 11635           |
| RS                 | 5466            |
| PR                 | 5045            |
| SC                 | 3637            |
| BA                 | 3380            |
| DF                 | 2140            |
| ES                 | 2033            |
| GO                 | 2020            |
| PE                 | 1652            |
| CE                 | 1336            |
| PA                 | 975             |
| MT                 | 907             |
| MA                 | 747             |
| MS                 | 715             |
| PB                 | 536             |
| PI                 | 495             |
| RN                 | 485             |
| AL                 | 413             |
| SE                 | 350             |
| TO                 | 280             |
| RO                 | 253             |
| AM                 | 148             |
| AC                 | 81              |
| AP                 | 68              |
| RR                 | 46              |
| customer_state     | 1               |
+--------------------+-----------------+
28 rows selected (73.337 seconds)
0: jdbc:hive2://localhost:10000/default>
Display all 966 possibilities? (y or n)
```

Com isso vemos que o estado de São Paulo é o que contém mais pedidos de forma isolada, possuindo mais do triplo do segundo estado com mais pedidos, que é o do Rio de Janeiro, seguidos por Minas Gerais. Assim demonstrando a concentração de pedidos na região sudeste do Brasil.

Métodos de pagamento mais utilizados: Quais formas de pagamento são mais comuns?

2. Métodos de pagamento mais utilizados

```
SELECT payment_type, COUNT(*) AS total_pagamentos
FROM order_payments
GROUP BY payment_type
ORDER BY total_pagamentos DESC;
```



```
INFO  : Completed executing command(queryId=hive_20250909004903_65fbac62-fff5-4323-bf88-45f861f04fac); Time taken: 38.68 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------+-------------------+
| payment_type  | total_pagamentos  |
+---------------+-------------------+
| credit_card   | 76795             |
| boleto        | 19784             |
| voucher       | 5775              |
| debit_card    | 1529              |
| not_defined   | 3                 |
| payment_type  | 1                 |
+---------------+-------------------+
6 rows selected (38.966 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Método de pagamento mais utilizado é o cartão de crédito e o time de precificação precisa estar ciente disto, pois há taxas a serem cobradas pelo uso, que geram despesas financeiras que podem corroer a margem de lucro. Além do time financeiro que deve estar de olho no ciclo financeiro, pois é comum que compras de cartão de crédito sejam parceladas e isso afeta o fluxo de caixa e a saúde financeira da empresa.

Status dos pedidos: Qual é a distribuição de status dos pedidos (entregue, cancelado, etc.)?

3. Status dos pedidos

SELECT order_status, COUNT(*) AS total_pedidos

FROM orders

GROUP BY order_status

ORDER BY total_pedidos DESC;

```
INFO  : Completed executing command(queryId=hive_20250909005127_694a2719-c1fe-492c-9e59-5593860f4bdd); Time taken: 13.04 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------+----------------+
| order_status  | total_pedidos  |
+---------------+----------------+
| delivered     | 96478          |
| shipped       | 1107           |
| canceled      | 625            |
| unavailable   | 609            |
| invoiced      | 314            |
| processing    | 301            |
| created       | 5              |
| approved      | 2              |
| order_status  | 1              |
+---------------+----------------+
9 rows selected (13.327 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Com isso vemos que a operação em geral está muito boa, pois as taxas de cancelamento são muito baixas comparadas a quantidade de pedidos gerais.

Vendedores por estado: Quantos vendedores estão em cada estado?

4. Vendedores por estado

```
SELECT seller_state, COUNT(*) AS total_vendedores

FROM sellers

GROUP BY seller_state

ORDER BY total_vendedores DESC;
```

```
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------+--------------------+
| seller_state  | total_vendedores   |
+---------------+--------------------+
| SP            | 1849               |
| PR            | 349                |
| MG            | 244                |
| SC            | 190                |
| RJ            | 171                |
| RS            | 129                |
| GO            | 40                 |
| DF            | 30                 |
| ES            | 23                 |
| BA            | 19                 |
| CE            | 13                 |
| PE            | 9                  |
| PB            | 6                  |
| RN            | 5                  |
| MS            | 5                  |
| MT            | 4                  |
| SE            | 2                  |
| RO            | 2                  |
| AC            | 1                  |
| AM            | 1                  |
| MA            | 1                  |
| PA            | 1                  |
| PI            | 1                  |
| seller_state  | 1                  |
+---------------+--------------------+
24 rows selected (20.546 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Como esperado, São Paulo possui a maior quantidade de vendedores, porém RJ que é o segundo maior em vendas, está apenas em quinto lugar em quantidade de vendedores, o que pode apontar uma oportunidade e deverá ser comunicado ao time de hunters de sellers, pois captar mais vendedores locais deve ser muito bom em questões logísticas para venda dentro do RJ.

Produtos mais populares por categoria: Quais categorias de produtos têm mais registros e vendas?

5. Top 10 categorias com mais produtos

SELECT product_category_name, COUNT(*) AS total_produtos

FROM products

GROUP BY product_category_name

ORDER BY total_produtos DESC

LIMIT 10;

```
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------------+-----------------+
|  product_category_name   | total_produtos  |
+--------------------------+-----------------+
| cama_mesa_banho          | 3029            |
| esporte_lazer            | 2867            |
| moveis_decoracao         | 2657            |
| beleza_saude             | 2444            |
| utilidades_domesticas    | 2335            |
| automotivo               | 1900            |
| informatica_acessorios   | 1639            |
| brinquedos               | 1411            |
| relogios_presentes       | 1329            |
| telefonia                | 1134            |
+--------------------------+-----------------+
10 rows selected (7.303 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Vemos que possui um grande mix de cameba, utilidade domésticas e beleza e saúde que em geral são comprados por mulheres e pode-se criar kits de produtos que visem aumentar vendas dessas categorias que são correlacionadas.

Proposta de Evolução do Trabalho

Criar tabelas agregadas para consultas rápidas de pedidos, produtos e vendedores.