Projeto Final Disciplina Infraestrutura Hadoop

Professor: André Ormastroni Victor

Aluno: Rafael Diniz Ramos

Trabalho como Gerente de Compras em uma empresa que presta serviço de Infraestrutura de TI, além da venda de produtos do ramo.

Gerei pelo sistema da empresa (OMIE), relatórios de Compras, Vendas e de Fornecedores, dos últimos dois anos e meio, para analisar os dados e buscar insights para melhores decisões no meu setor.

Os relatórios foram gerados em planilhas.xlsx, transformados em CSV e tratados com Python/Pandas antes de serem carregados no Haddop.

Google Cloud Plataform

- Criei um bucket e fiz o upload dos arquivos csv.
- Crie um cluster usando o serviço Dataproc.

Através de uma conexão SSH, acessei o cluster para manipulação dos arquivos no Hadoop e no Hive.

Em ambiente local:

- Criei um diretório chamado myfiles e em seguida montei nesse diretório, através do GCSFUSE um espelho dos arquivos que estão no bucket no GCP.
  - - mkdir myfiles
  - - gcsfuse arquivos_hadoop myfiles

Em ambiente HDFS:

- Criei um diretório chamado "projeto", no /user do HDFS, para posteriormente copiar os arquivos para a pasta.

  - - hfds dfs -mkdir /user/projeto

- Dentro da pasta myfiles, copiei os arquivos CSV para o destino /user/projeto no HDFS.

  - - hdfs dfs -put compras.csv /user/projeto

  - - hdfs dfs -put vendas.csv /user/projeto

  - - hdfs dfs -put fornecedores.csv /user/projeto

```
rafael_dramos@cluster-9ecc-m:~$ ls
myfiles  sakila-data.sql  sakila-schema.sql  sqoop-1.4.6.bin__hadoop-2.0.4-alpha  sqoop-1.4.6.bin__hadoop-2.0.4-alpha.tar.gz
rafael_dramos@cluster-9ecc-m:~$ cd myfiles
rafael_dramos@cluster-9ecc-m:~/myfiles$ ls
'SnP500 All assets.csv'  compras.csv  drivers.csv  timesheet.csv  vendas.csv  walmart.csv
rafael_dramos@cluster-9ecc-m:~/myfiles$ hdfs dfs -put compras.csv /user/projeto
rafael_dramos@cluster-9ecc-m:~/myfiles$ hdfs dfs -put vendas.csv /user/projeto
rafael_dramos@cluster-9ecc-m:~/myfiles$ hdfs dfs -ls /user/projeto
Found 2 items
-rw-r--r--   1 rafael_dramos hadoop     499858 2023-08-31 01:43 /user/projeto/compras.csv
-rw-r--r--   1 rafael_dramos hadoop     677650 2023-08-31 01:43 /user/projeto/vendas.csv
rafael_dramos@cluster-9ecc-m:~/myfiles$
```

```
rafael_dramos@cluster-9ecc-m:~/myfiles$ hdfs dfs -put fornecedores.csv /user/projeto
rafael_dramos@cluster-9ecc-m:~/myfiles$ hdfs dfs -ls /user/projeto
Found 3 items
-rw-r--r--   1 rafael_dramos hadoop     499858 2023-08-31 01:43 /user/projeto/compras.csv
-rw-r--r--   1 rafael_dramos hadoop     216716 2023-08-31 02:02 /user/projeto/fornecedores.csv
-rw-r--r--   1 rafael_dramos hadoop     677650 2023-08-31 01:43 /user/projeto/vendas.csv
rafael_dramos@cluster-9ecc-m:~/myfiles$
```

HIVE

- Instalação do client JDBC do Hive, para manipulação das tabelas através de SQLLine no HDFS.

  - - beeline -u jdbc:hive2://localhost:10000/default -n rafael_dramos@cluster-9ecc-m -d org.apache.hive.jdbc.HiveDriver

- Criação do banco de dados e conexão
  - - create database adm;
  - - use adm;

- Criação das tabelas e carregamento de dados

CREATE TABLE compras (Descricao STRING, Data STRING, Fornecedor STRING, Nota_Fiscal INT, Quantidade INT, Valor_Mercadoria FLOAT, ipi FLOAT, icms FLOAT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

CREATE TABLE vendas (descricao STRING, data STRING, total INT, imposto_federal STRING, imposto_estadual STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

CREATE TABLE fornecedores (cnpj_cpf STRING, fornecedor STRING, telefone STRING, email STRING, cidade STRING, inscricao_estadual INT, contribuinte_icms STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ';'

STORED AS TEXTFILE;

O carregamento dos dados na tabela é feito através da sintaxe abaixo:

LOAD DATA INPATH '/user/projeto/compras.csv' overwrite into table compras;

LOAD DATA INPATH '/user/projeto/vendas.csv' overwrite into table vendas;

LOAD DATA INPATH '/user/projeto/fornecedores.csv' overwrite into table fornecedores;

```
0: jdbc:hive2://localhost:10000/default> LOAD DATA INPATH '/user/projeto/compras.csv' overwrite into table compras;
INFO  : Compiling command(queryId=hive_20230831030011_c4fd8b8d-eece-4f51-84d2-891a79d98dc1): LOAD DATA INPATH '/user/projeto/compras.csv' overwrite into table compras
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230831030011_c4fd8b8d-eece-4f51-84d2-891a79d98dc1); Time taken: 0.387 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230831030011_c4fd8b8d-eece-4f51-84d2-891a79d98dc1): LOAD DATA INPATH '/user/projeto/compras.csv' overwrite into table compras
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table adm.compras from hdfs://cluster-9ecc-m/user/projeto/compras.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20230831030011_c4fd8b8d-eece-4f51-84d2-891a79d98dc1); Time taken: 0.397 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.817 seconds)
0: jdbc:hive2://localhost:10000/default>
```

```
0: jdbc:hive2://localhost:10000/default> LOAD DATA INPATH '/user/projeto/vendas.csv' overwrite into table vendas;
INFO  : Compiling command(queryId=hive_20230831030100_e5ece910-d378-4f85-8488-0428ef34420c): LOAD DATA INPATH '/user/projeto/vendas.csv' overwrite into table vendas
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230831030100_e5ece910-d378-4f85-8488-0428ef34420c); Time taken: 0.079 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230831030100_e5ece910-d378-4f85-8488-0428ef34420c): LOAD DATA INPATH '/user/projeto/vendas.csv' overwrite into table vendas
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table adm.vendas from hdfs://cluster-9ecc-m/user/projeto/vendas.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20230831030100_e5ece910-d378-4f85-8488-0428ef34420c); Time taken: 0.327 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.433 seconds)
0: jdbc:hive2://localhost:10000/default>
```
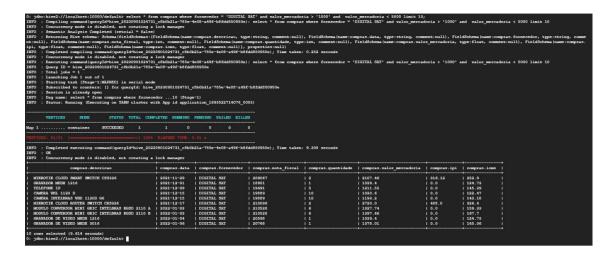
```
0: jdbc:hive2://localhost:10000/default> LOAD DATA INPATH '/user/projeto/fornecedores.csv' overwrite into table fornecedores;
INFO  : Compiling command(queryId=hive_20230831030215_6c10c1d7-b2e0-46a5-8223-fc65209a57b3): LOAD DATA INPATH '/user/projeto/fornecedores.csv' overwrite into table fornecedores
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20230831030215_6c10c1d7-b2e0-46a5-8223-fc65209a57b3); Time taken: 0.095 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230831030215_6c10c1d7-b2e0-46a5-8223-fc65209a57b3): LOAD DATA INPATH '/user/projeto/fornecedores.csv' overwrite into table fornecedores
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table adm.fornecedores from hdfs://cluster-9ecc-m/user/projeto/fornecedores.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20230831030215_6c10c1d7-b2e0-46a5-8223-fc65209a57b3); Time taken: 0.297 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.416 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Vamos agora buscar algumas informações nas tabelas, usando sintaxe de SQL através do HIVE.

Abaixo nós solicitamos somente compras realizadas no fornecedor DIGITAL SAT.

```
0: jdbc:hive2://localhost:10000/default> select * from compras where fornecedor = 'DIGITAL SAT' limit 10;
INFO  : Compiling command(queryId=hive_20230901021513_466d6506-7988-4e50-a396-a7f1c56222ee): select * from compras where fornecedor = 'DIGITAL SAT' limit 10
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:compras.descricao, type:string, comment:null), FieldSchema(name:compras.data, type:string, comment:null), FieldSchema(name:compras.fornecedor, type:string, comment:null), FieldSchema(name:compras.nota_fiscal, type:int, comment:null), FieldSchema(name:compras.quantidade, type:int, comment:null), FieldSchema(name:compras.valor_mercadoria, type:float, comment:null), FieldSchema(name:compras.ipi, type:float, comment:null), FieldSchema(name:compras.icms, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230901021513_466d6506-7988-4e50-a396-a7f1c56222ee); Time taken: 0.196 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230901021513_466d6506-7988-4e50-a396-a7f1c56222ee): select * from compras where fornecedor = 'DIGITAL SAT' limit 10
INFO  : Query ID = hive_20230901021513_466d6506-7988-4e50-a396-a7f1c56222ee
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20230901021513_466d6506-7988-4e50-a396-a7f1c56222ee
INFO  : Session is already open
INFO  : Dag name: select * from compras where fornecedor ...10 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1692522714076_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.37 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230901021513_466d6506-7988-4e50-a396-a7f1c56222ee); Time taken: 8.531 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------------------------+---------------+--------------------+--------------------+-------------------+-------------------------+-------------+--------------+
|           compras.descricao          | compras.data  | compras.fornecedor | compras.nota_fiscal | compras.quantidade | compras.valor_mercadoria | compras.ipi | compras.icms |
+--------------------------------------+---------------+--------------------+--------------------+-------------------+-------------------------+-------------+--------------+
| MIKROTIK CLOUD SMART SWITCH CSS326    | 2021-11-29    | DIGITAL SAT        | 209067             | 2                 | 2107.46                 | 316.12      | 252.9        |
| CAIXA HERMETICA 10X10                 | 2021-11-29    | DIGITAL SAT        | 18758              | 20                | 207.6                   | 0.0         | 24.91        |
| CONECTOR CONEX 1000 RJ45 CAT5E        | 2021-11-30    | DIGITAL SAT        | 18837              | 2                 | 51.44                   | 0.0         | 6.17         |
| CONECTOR CONEX 1000 RJ45 CAT6         | 2021-11-30    | DIGITAL SAT        | 18837              | 4                 | 115.64                  | 0.0         | 13.88        |
| MODULO INTELIGENTE PORTARIA MIP 1000  | 2021-11-30    | DIGITAL SAT        | 18837              | 1                 | 489.01                  | 0.0         | 58.68        |
| MINI RACK DESMONTAVEL 12U 570MM       | 2021-12-01    | DIGITAL SAT        | 18951              | 1                 | 622.81                  | 0.0         | 74.74        |
| GRAVADOR MMDX 1216                    | 2021-12-01    | DIGITAL SAT        | 18951              | 1                 | 1039.4                  | 0.0         | 124.73       |
| BATERIA MINI 12V A23 ELGIN            | 2021-12-01    | DIGITAL SAT        | 18983              | 15                | 37.05                   | 0.0         | 4.45         |
| VOICE PANEL 50 PORTAS CAT3 1U PRETO NEXANS | 2021-12-01 | DIGITAL SAT       | 18983              | 1                 | 504.9                   | 0.0         | 60.59        |
| CONJ SENSOR MAGN REED BASC 2000       | 2021-12-01    | DIGITAL SAT        | 18984              | 1                 | 33.24                   | 0.0         | 3.99         |
+--------------------------------------+---------------+--------------------+--------------------+-------------------+-------------------------+-------------+--------------+
10 rows selected (8.802 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Abaixo nós buscamos somente compras realizadas no fornecedor DIGITAL SAT, entre os valores 1000 e 5000 reais.
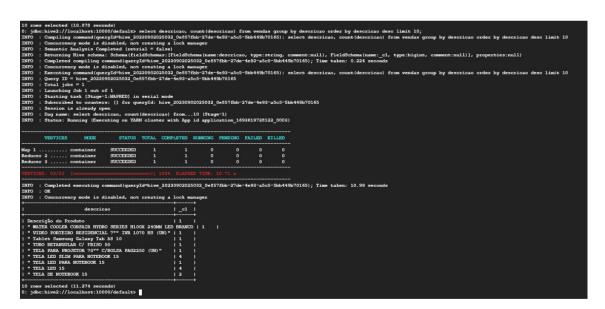

Abaixo nós buscamos pelas colunas nota fiscal e valor de mercadoria, compras realizadas acima de 5.000 reais e abaixo de 10.000 reais.

Abaixo nós selecionamos por fornecedor e valor de mercadoria, compras realizadas na Pauta Distribuidora, ordenados do maior ao menor valor.

Abaixo nós solicitamos uma contagem pela descrição, para saber quantas vendas foram realizadas por produto.

Abaixo nós podemos verificar o valor médio de venda por produto.

Abaixo nós fizemos uma junção de dados entre as tabelas Compras e Fornecedores. Pegamos o nome do fornecedor da tabela Compras e unimos com as informações de telefone, e-mail e cidade, da tabela Fornecedor.