DISCIPLINA

Infraestrutura Hadoop

Pós-Graduação em Engenharia de Dados: Big Data

Aluno: Leonardo Bastos Moraes

Professor (a): André Ormastroni

Tema: Relação entre dados demográficos dos Estados e o Pagamento de Royalties do Petróleo no Brasil.

1. Dicionário de Dados:

Tabela.1: Pagamento Royalties		
Nome	**Descrição**	**Tipo de Dado**
UF	Unidades Federativas	string
Total	Total de Royalties do Petróleo (R$)	float

```
0: jdbc:hive2://localhost:10000/default> select * from pagamento_royalties;
INFO  : Compiling command(queryId=hive_20230904132259_8b670f94-1293-4289-88d2-5be54210b82b): select * from pagamento_royalties
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:pagamento_royalties.uf, type:string, comment:null), FieldSchema(name:pagamento_royalties.total, type
:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230904132259_8b670f94-1293-4289-88d2-5be54210b82b); Time taken: 0.139 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904132259_8b670f94-1293-4289-88d2-5be54210b82b): select * from pagamento_royalties
INFO  : Completed executing command(queryId=hive_20230904132259_8b670f94-1293-4289-88d2-5be54210b82b); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------------+------------------------+
| pagamento_royalties.uf | pagamento_royalties.total |
+----------------------+------------------------+
| UF                   | NULL                   |
| DF                   | 3.20117043E10          |
| RJ                   | 3.18074757E10          |
| ES                   | 7.740799E9             |
| SP                   | 6.1524009E9            |
| RN                   | 2.36448717E9           |
| BA                   | 2.35812659E9           |
| SE                   | 1.58791616E9           |
| AM                   | 1.45727987E9           |
| AL                   | 5.9955789E8            |
| RS                   | 4.6976112E8            |
| CE                   | 3.37681376E8           |
| MA                   | 3.29627648E8           |
| PE                   | 2.99824032E8           |
| SC                   | 2.81323584E8           |
| PB                   | 1.66319904E8           |
| MG                   | 8.7406392E7            |
| PR                   | 4.6649296E7            |
| PA                   | 8291489.0              |
| AP                   | 1463204.0              |
+----------------------+------------------------+
20 rows selected (0.171 seconds)
0: jdbc:hive2://localhost:10000/default> describe pagamento_royalties;
```

Tabela.2: Dados Demográficos		
Nome	**Descrição**	**Tipo de Dado**
estados_do_brasil	Unidades Federativas	string
area_territorial	Área territorial dos Estados	float
populacao_residente	População Residente no Estados	float
densidade_demografica	Densidade Demográfica dos Estados	float
IDH	Índica de Desenvolvimento Humano dos Estados	float
rendimento_mensal_domicili ar_per_capita	Rendimento domiciliar per capita por Estado	float
total_de_veiculos	Total de veículos por Estado	float

```
INFO  : Executing command(queryId=hive_20230904154219_ef8d170a-dd7a-4a75-abf1-e05744b32ebe): SELECT estados_do_brasil, populacao_residente, rendimento_mensal_domiciliar_per_capita, total_de_veiculos FROM dados_demograficos
INFO  : Completed executing command(queryId=hive_20230904154219_ef8d170a-dd7a-4a75-abf1-e05744b32ebe); Time taken: 0.0 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------+---------------------+-----------------------------------------+-------------------+
| estados_do_brasil | populacao_residente | rendimento_mensal_domiciliar_per_capita | total_de_veiculos |
+-------------------+---------------------+-----------------------------------------+-------------------+
| Estados do Brasil | NULL                | NULL                                    | NULL              |
| AL                | 3127511.0           | 935.0                                   | 1034187.0         |
| AP                | 733508.0            | 1177.0                                  | 232691.0          |
| AM                | 3941175.0           | 965.0                                   | 1069794.0         |
| BA                | 1.4136417E7         | 1010.0                                  | 4887673.0         |
| CE                | 8791688.0           | 1050.0                                  | 3625994.0         |
| DF                | 2817068.0           | 2913.0                                  | 2021627.0         |
| ES                | 3833486.0           | 1723.0                                  | 2248960.0         |
| MA                | 6775152.0           | 814.0                                   | 2031236.0         |
| MG                | 2.0538718E7         | 1529.0                                  | 1.3028883E7       |
| PA                | 8116132.0           | 1061.0                                  | 2478988.0         |
| PB                | 3974495.0           | 1096.0                                  | 1523167.0         |
| PR                | 1.1443208E7         | 1846.0                                  | 8575905.0         |
| PE                | 9058155.0           | 1010.0                                  | 3439164.0         |
| RJ                | 1.6054524E7         | 1971.0                                  | 7475503.0         |
| RN                | 3302406.0           | 1267.0                                  | 1495094.0         |
| RS                | 1.0880506E7         | 2087.0                                  | 7869630.0         |
| SC                | 7609601.0           | 2018.0                                  | 5974106.0         |
| SP                | 4.442046E7          | 2148.0                                  | 3.2293192E7       |
| SE                | 2209558.0           | 1187.0                                  | 907388.0          |
+-------------------+---------------------+-----------------------------------------+-------------------+
20 rows selected (0.154 seconds)
0: jdbc:hive2://localhost:10000/default> describe dados_demograficos;
INFO  : Compiling command(queryId=hive_20230904154303_09abc63b-48ea-4b4c-aac1-c011448106db): describe dados_demograficos
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name:comment, type:string,
deserializer)], properties:null)
```

* Foram aproveitadas as colunas estados_do_brasil, população_residente, rendimento_mensal_domiciliar_per_capita e total_de_veiculos.

Tabela.3: Quantidade de Domicilios		
Nome	**Descrição**	**Tipo de Dado**
sigla	Unidades Federativas	string
Ano_2000	Quantidade de domicílios por Estados	float

```
0: jdbc:hive2://localhost:10000/default> select * from quantidade_domicilios;
INFO  : Compiling command(queryId=hive_20230904193717_0871adf7-3498-4d48-a5d5-2838e07127a8): select * from q
uantidade_domicilios
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:quantidade_domicilios.sigla, type:strin
g, comment:null), FieldSchema(name:quantidade_domicilios.ano_2000, type:float, comment:null)], properties:nu
ll)
INFO  : Completed compiling command(queryId=hive_20230904193717_0871adf7-3498-4d48-a5d5-2838e07127a8); Time
taken: 0.159 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904193717_0871adf7-3498-4d48-a5d5-2838e07127a8): select * from q
uantidade_domicilios
INFO  : Completed executing command(queryId=hive_20230904193717_0871adf7-3498-4d48-a5d5-2838e07127a8); Time
taken: 0.001 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------------------------+------------------------------+
| quantidade_domicilios.sigla  | quantidade_domicilios.ano_2000  |
+------------------------------+------------------------------+
| Sigla                        | NULL                         |
| AC                           | 131580.0                     |
| AL                           | 658873.0                     |
| AM                           | 580900.0                     |
| AP                           | 100765.0                     |
| BA                           | 3214292.0                    |
| CE                           | 1773393.0                    |
| DF                           | 556762.0                     |
| ES                           | 851014.0                     |
| GO                           | 1420822.0                    |
| MA                           | 1246715.0                    |
| MG                           | 4837296.0                    |
| MS                           | 577362.0                     |
| MT                           | 669676.0                     |
| PA                           | 1332248.0                    |
| PB                           | 857989.0                     |
| PE                           | 1994041.0                    |
| PI                           | 665808.0                     |
| PR                           | 2709523.0                    |
| RJ                           | 4315737.0                    |
| RN                           | 678652.0                     |
| RO                           | 354391.0                     |
| RR                           | 76681.0                      |
| RS                           | 3091643.0                    |
| SC                           | 1518651.0                    |
| SE                           | 442256.0                     |
| SP                           | 1.0564745E7                  |
| TO                           | 285701.0                     |
+------------------------------+------------------------------+
28 rows selected (0.202 seconds)
0: jdbc:hive2://localhost:10000/default> LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_municip
ios.csv' OVERWRITE INTO TABLE qtd_num;
INFO  : Compiling command(queryId=hive_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a): LOAD DATA INPAT
H '/user/leonardo_moraes/datasets/quantidade_municípios.csv' OVERWRITE INTO TABLE qtd_num
INFO  : Concurrency mode is disabled, not creating a lock manager
```

Tabela.4: Quantidade Numero de Municípios		
Nome	**Descrição**	**Tipo de Dado**
uf	Unidades Federativas	string
qtd_num	Quantidade de municípios por Estados	float

```
0: jdbc:hive2://localhost:10000/default> select * from qtd_num;
INFO  : Compiling command(queryId=hive_20230904193828_d198e39d-1093-49a0-a74b-c448cd49fc97): select * from q
td_num
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:qtd_num.uf, type:string, comment:null),
 FieldSchema(name:qtd_num.qtd_num, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230904193828_d198e39d-1093-49a0-a74b-c448cd49fc97); Time
taken: 0.161 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904193828_d198e39d-1093-49a0-a74b-c448cd49fc97): select * from q
td_num
INFO  : Completed executing command(queryId=hive_20230904193828_d198e39d-1093-49a0-a74b-c448cd49fc97); Time
taken: 0.004 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------+------------------+
| qtd_num.uf  | qtd_num.qtd_num  |
+-------------+------------------+
| UF          | NULL             |
| RO          | 52.0             |
| AC          | 22.0             |
| AM          | 62.0             |
| RR          | 15.0             |
| PA          | 144.0            |
| AP          | 16.0             |
| TO          | 139.0            |
| MA          | 217.0            |
| PI          | 224.0            |
| CE          | 184.0            |
| RN          | 167.0            |
| PB          | 223.0            |
| PE          | 185.0            |
| AL          | 102.0            |
| SE          | 75.0             |
| BA          | 417.0            |
| MG          | 853.0            |
| ES          | 78.0             |
| RJ          | 92.0             |
| SP          | 645.0            |
| PR          | 399.0            |
| SC          | 295.0            |
| RS          | 497.0            |
| MS          | 79.0             |
| MT          | 141.0            |
| GO          | 246.0            |
| DF          | 1.0              |
+-------------+------------------+
28 rows selected (0.227 seconds)
0: jdbc:hive2://localhost:10000/default>
```

2. Identificação das fontes de onde os dados foram retirados:

Tabela.1: Pagamento Royalties:
    Fonte: IBGE
    url: https://www.ibge.gov.br/cidades-e-estados

Tabela.2: Dados Demográficos:
    Fonte: IBGE
    url: https://www.kaggle.com/datasets/luizph21/pagamento-de-royalties-do-petrleo-no-brasil

Tabela.3: Quantidade Municípios:
    Fonte: Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome
    url: http://blog.mds.gov.br/redesuas/lista-de-municipios-brasileiros/

Tabela.4: Quantidade Domicilios:
    Fonte: Ipea Data
    url: http://www.ipeadata.gov.br/Default.aspx

3. Scripts de criação do database no Hive

Show tables

```
0: jdbc:hive2://localhost:10000/default> show tables;
INFO  : Compiling command(queryId=hive_20230904180344_898394f2-f0ca-4968-b159-ed67e4e728fe): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from des
erializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230904180344_898394f2-f0ca-4968-b159-ed67e4e728fe); Time
taken: 0.02 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904180344_898394f2-f0ca-4968-b159-ed67e4e728fe): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230904180344_898394f2-f0ca-4968-b159-ed67e4e728fe); Time
taken: 0.016 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-----------------------+
|        tab_name       |
+-----------------------+
| dados_demograficos    |
| pagamento_royalties   |
| qtd_num               |
| quantidade_domicilios |
+-----------------------+
4 rows selected (0.067 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Tabela.1: Pagamento Royalties:

CREATE TABLE pagamento_royalties (
    UF STRING,
    Total FLOAT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

describe pagamento_royalties

```
0: jdbc:hive2://localhost:10000/default> describe dados_demograficos;
INFO  : Compiling command(queryId=hive_20230904154303_09abc63b-48ea-4b4c-aac1-c011448106db): describe dados_demograficos
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:col_name, type:string, comment:from deserializer), FieldSchema(name:data_type, type:string, comment:from deserializer), FieldSchema(name
:comment, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230904154303_09abc63b-48ea-4b4c-aac1-c011448106db); Time taken: 0.05 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904154303_09abc63b-48ea-4b4c-aac1-c011448106db): describe dados_demograficos
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20230904154303_09abc63b-48ea-4b4c-aac1-c011448106db); Time taken: 0.021 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------------------------+------------+----------+
|               col_name               |  data_type |  comment |
+--------------------------------------+------------+----------+
| estados_do_brasil                    | string     |          |
| area_territorial                     | float      |          |
| populacao_residente                  | float      |          |
| densidade_demografica                | float      |          |
| idh                                  | float      |          |
| rendimento_mensal_domiciliar_per_capita | float   |          |
| total_de_veiculos                    | float      |          |
+--------------------------------------+------------+----------+
7 rows selected (0.116 seconds)
0: jdbc:hive2://localhost:10000/default> Closing: 0: jdbc:hive2://localhost:10000/default
leonardo_moraes@cluster-4c01-m:~$
```

Tabela.2: Dados Demográficos:

CREATE TABLE dados_demograficos (
  Estados_do_Brasil STRING,
  Area_Territorial FLOAT,
  Populacao_Residente FLOAT,
  Densidade_Demografica FLOAT,
  IDH FLOAT,
  Rendimento_Mensal_Domiciliar_per_Capita FLOAT,
  Total_de_Veiculos FLOAT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

describe dados_demograficos

Tabela.3: Quantidade Municípios:

CREATE TABLE qtd_num (
UF STRING,
qtd_num FLOAT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';';

Tabela.4: Quantidade Domicilios:

CREATE TABLE quantidade_domicilios (
Sigla STRING,
Ano_2000 FLOAT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';';

4. Scripts de ingestão de dados no Hive

ls -la

Tabela.1: Pagamento Royalties:

LOAD DATA INPATH '/user/leonardo_moraes/datasets/Pagamento_Royalties_Petroleo_Soma_por_UF.csv' OVERWRITE INTO TABLE pagamento_royalties;

```
0: jdbc:hive2://localhost:10000/default> LOAD DATA INPATH '/user/leonardo_moraes/datasets/Pagamento_Royalties_Petroleo_Soma_por_UF.csv' OVERWRITE INTO TABLE pagamento_royalties;
INFO  : Compiling command(queryId=hive_20230904132244_27862da9-f7e6-41b0-b3b7-aceeaa7c3550): LOAD DATA INPATH '/user/leonardo_moraes/datasets/Pagamento_Royalties_Petroleo_Soma_por_UF.csv' OVERWRITE INTO TA
BLE pagamento_royalties
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_20230904132244_27862da9-f7e6-41b0-b3b7-aceeaa7c3550); Time taken: 0.05 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904132244_27862da9-f7e6-41b0-b3b7-aceeaa7c3550): LOAD DATA INPATH '/user/leonardo_moraes/datasets/Pagamento_Royalties_Petroleo_Soma_por_UF.csv' OVERWRITE INTO TA
BLE pagamento_royalties
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.pagamento_royalties from hdfs://cluster-4c01-m/user/leonardo_moraes/datasets/Pagamento_Royalties_Petroleo_Soma_por_UF.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20230904132244_27862da9-f7e6-41b0-b3b7-aceeaa7c3550); Time taken: 0.188 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.257 seconds)
```

Tabela.2: Dados Demográficos:

LOAD DATA INPATH '/user/leonardo_moraes/datasets/dados_demograficos_IBGE.csv' OVERWRITE INTO TABLE dados_demograficos;

```
0: jdbc:hive2://localhost:10000/default> LOAD DATA INPATH '/user/leonardo_moraes/datasets/dados_demograficos_IBGE.csv' OVERWRITE INTO TABLE dados_demograficos;
INFO  : Compiling command(queryId=hive_20230904153414_b43fa3af-fa3d-4ddb-b404-5a6a4ac48761): LOAD DATA INPATH '/user/leonardo_moraes/datasets/dados_demograficos_IBGE.csv' OVERWRITE INTO TABLE dados_demogra
ficos
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_20230904153414_b43fa3af-fa3d-4ddb-b404-5a6a4ac48761); Time taken: 0.048 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904153414_b43fa3af-fa3d-4ddb-b404-5a6a4ac48761): LOAD DATA INPATH '/user/leonardo_moraes/datasets/dados_demograficos_IBGE.csv' OVERWRITE INTO TABLE dados_demogra
ficos
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.dados_demograficos from hdfs://cluster-4c01-m/user/leonardo_moraes/datasets/dados_demograficos_IBGE.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20230904153414_b43fa3af-fa3d-4ddb-b404-5a6a4ac48761); Time taken: 0.224 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.282 seconds)
```

Tabela.3: Quantidade Municípios:

LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_municipios.csv' OVERWRITE INTO TABLE qtd_num;

```
0: jdbc:hive2://localhost:10000/default> LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_municipios.csv' OVERWRITE INTO TABLE qtd_num;
INFO  : Compiling command(queryId=hive_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a): LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_municipios.csv' OVERWRITE INTO TABLE qtd_num
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_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a); Time taken: 0.044 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a): LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_municipios.csv' OVERWRITE INTO TABLE qtd_num
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.qtd_num from hdfs://cluster-4c01-m/user/leonardo_moraes/datasets/quantidade_municipios.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a); Time taken: 0.152 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.208 seconds)
```

Tabela.4: Quantidade Domicilios:

LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_domicilios.csv' OVERWRITE INTO TABLE quantidade_domicilios;

```
0: jdbc:hive2://localhost:10000/default> LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_municipios.csv' OVERWRITE INTO TABLE qtd_num;
INFO  : Compiling command(queryId=hive_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a): LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_municipios.csv' OVERWRITE INTO TABLE qtd_num
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_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a); Time taken: 0.044 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a): LOAD DATA INPATH '/user/leonardo_moraes/datasets/quantidade_municipios.csv' OVERWRITE INTO TABLE qtd_num
INFO  : Starting task [Stage-0:MOVE] in serial mode
INFO  : Loading data to table default.qtd_num from hdfs://cluster-4c01-m/user/leonardo_moraes/datasets/quantidade_municipios.csv
INFO  : Starting task [Stage-1:STATS] in serial mode
INFO  : Completed executing command(queryId=hive_20230904193810_b151a45e-7117-43da-b679-0700aafa6c5a); Time taken: 0.152 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.208 seconds)
```

5. Scripts de carga dos dados para o HDFS

Tabela.1: Pagamento Royalties:

```
leonardo_moraes@cluster-4c01-m:~$ hdfs dfs -put
Pagamento_Royalties_Petroleo_Soma_por_UF.csv /user/leonardo_moraes/datasets
```

cat Pagamento_Royalties_Petroleo_Soma_por_UF.csv

```
leonardo_moraes@cluster-4c01-m:~$ cat Pagamento_Royalties_Petroleo_Soma_por_UF.csv
UF;Total
DF;32011704373
RJ;31807476068
ES;7740799070
SP;6152401070
RN;2364487067
BA;2358126668
SE;1587916179
AM;1457279878
AL;599557919
RS;469761128
CE;337681373
MA;329627650
PE;299824032
SC;281323599
PB;166319910
MG;87406390
PR;46649298
PA;8291489
AP;1463204
leonardo_moraes@cluster-4c01-m:~$
```

Tabela.2: Dados Demográficos:

leonardo_moraes@cluster-4c01-m:~$ hdfs dfs -put dados_demograficos_IBGE.csv /user/leonardo_moraes/datasets

cat dados_demograficos_IBGE.csv

```
leonardo_moraes@cluster-4c01-m:~$ cat dados_demograficos_IBGE.csv
Estados_do_Brasil; Area_Territorial; Populacao_Residente; Densidade_Demografica; IDH ; Rendimento_Mensal_Domiciliar_per_Capita;Total_de_Veiculos
AL;27830;3127511;112;0.6;935;1034187
AP;142470;733508;5;0.6;1177;232691
AM;1559255;3941175;2;0.7;965;1069794
BA;564760;14136417;25;0.6;1010;4887673
CE;148894;8791688;59;0.7;1050;3625994
DF;5760;2817068;489;0.8;2913;2021627
ES;46074;3833486;83;0.7;1723;2248960
MA;329651;6775152;20;0.6;814;2031236
MG;586513;20538718;35;0.7;1529;13028883
PA;1245870;8116132;6;0.6;1061;2478988
PB;56467;3974495;70,39;0.6;1096;1523167
PR;199298;11443208;57;0.7;1846;8575905
PE;98067;9058155;92;0.7;1010;3439164
RJ;43750;16054524;366;0.762;1971;7475503
RN;52809;3302406;62;0.7;1267;1495094
RS;281707;10880506;38;0.7;2087;7869630
SC;95730;7609601;79;0.7;2018;5974106
SP;248219;44420459;178;0.8;2148;32293191
SE;21938;2209558;100;0.7;1187;907388
```

Tabela.3: Quantidade Municípios:

leonardo_moraes@cluster-4c01-m:~$ hdfs dfs -put quantidade_municipios.csv /user/leonardo_moraes/datasets

cat quantidade_municipios.csv

```
leonardo_moraes@cluster-4c01-m:~$ cat quantidade_municipios.csv
UF;qtd_mun
RO;52
AC;22
AM;62
RR;15
PA;144
AP;16
TO;139
MA;217
PI;224
CE;184
RN;167
PB;223
PE;185
AL;102
SE;75
BA;417
MG;853
ES;78
RJ;92
SP;645
PR;399
SC;295
RS;497
MS;79
MT;141
GO;246
DF;1
leonardo_moraes@cluster-4c01-m:~$
```

<u>Tabela.4: Quantidade Domicilios:</u>

leonardo_moraes@cluster-4c01-m:~$ hdfs dfs -put quantidade_domicilios.csv /user/leonardo_moraes/datasets

cat quantidade_domicilios.csv

```
leonardo_moraes@cluster-4c01-m:~$ cat quantidade_domicilios.csv
Sigla;Ano_2000
AC;131580
AL;658873
AM;580900
AP;100765
BA;3214292
CE;1773393
DF;556762
ES;851014
GO;1420822
MA;1246715
MG;4837296
MS;577362
MT;669676
PA;1332248
PB;857989
PE;1994041
PI;665808
PR;2709523
RJ;4315737
RN;678652
RO;354391
RR;76681
RS;3091643
SC;1518651
SE;442256
SP;10564745
TO;285701
leonardo_moraes@cluster-4c01-m:~$
```

6. Formulação de perguntas analíticas a serem respondidas com os dados:

Consulta.1:
- Quais são as três UF que mais recebem royalties no Brasil?
DF, RJ, ES.

Consulta.2:
- Quais são as três UF (siglas) com maior quantidade de domicilios no Brasil?
SP, MG, RJ.

Consulta.3:
- Quais são as cinco UF que mais possuem municípios no Brasil?
MG, SP, RS, BA, PR.

Consulta.4
- Quais são as três UF que mais possuem veículos no Brasil?
SP, MG, PR.

Consulta.5
- Quais são as três UF com o maior número de população residente no Brasil?
SP, MG, RJ.

7. Scripts com as consultas SQL que foram utilizadas nas análises:

Consulta.1
- select uf, total from pagamento_royalties order by total desc limit 3;

```
0: jdbc:hive2://localhost:10000/default> select uf, total from pagamento_royalties order by total desc limit 3;
INFO  : Compiling command(queryId=hive_20230904230318_daba2327-3520-4c75-bce9-d71b3de0c315): select uf, total from pagamento_royalties order by total desc limit 3
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:uf, type:string, comment:null), FieldSchema(name:total, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230904230318_daba2327-3520-4c75-bce9-d71b3de0c315); Time taken: 0.441 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904230318_daba2327-3520-4c75-bce9-d71b3de0c315): select uf, total from pagamento_royalties order by total desc limit 3
INFO  : Query ID = hive_20230904230318_daba2327-3520-4c75-bce9-d71b3de0c315
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_20230904230318_daba2327-3520-4c75-bce9-d71b3de0c315
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: select uf, total from pagamento_royaltie...3 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1693759391598_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container       SUCCEEDED     1          1        0        0       0       0
Reducer 2 ...... container       SUCCEEDED     1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.18 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230904230318_daba2327-3520-4c75-bce9-d71b3de0c315); Time taken: 35.63 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------+---------------+
| uf   |    total      |
+------+---------------+
| DF   | 3.20117043E10 |
| RJ   | 3.18074757E10 |
| ES   | 7.740799E9    |
+------+---------------+
3 rows selected (36.217 seconds)
0: jdbc:hive2://localhost:10000/default>
```

Consulta.2

- select sigla, ano_2000 from quantidade_domicilios order by ano_2000 desc limit 3;

```
0: jdbc:hive2://localhost:10000/default> select sigla, ano_2000 from quantidade_domicilios order by ano_2000 desc limit 3;
INFO  : Compiling command(queryId=hive_20230904231116_bc535c6-034d-4401-a7da-888e5c8bf865): select sigla, ano_2000 from quantidade_domicilios order by ano_2000 desc limit 3
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:sigla, type:string, comment:null), FieldSchema(name:ano_2000, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230904231116_bc535c6-034d-4401-a7da-888e5c8bf865); Time taken: 0.146 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230904231116_bc535c6-034d-4401-a7da-888e5c8bf865): select sigla, ano_2000 from quantidade_domicilios order by ano_2000 desc limit 3
INFO  : Query ID = hive_20230904231116_bc535c6-034d-4401-a7da-888e5c8bf865
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_20230904231116_bc535c6-034d-4401-a7da-888e5c8bf865
INFO  : Session is already open
INFO  : Dag name: select sigla, ano_2000 from quantidade_d...3 (Stage-1)
INFO  : Tez session was closed. Reopening...
INFO  : Session re-established.
INFO  : Session re-established.
INFO  : Status: Running (Executing on YARN cluster with App id application_1693759391598_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container       SUCCEEDED     1          1        0        0       0       0
Reducer 2 ...... container       SUCCEEDED     1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.46 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230904231116_bc535c6-034d-4401-a7da-888e5c8bf865); Time taken: 23.349 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------+------------+
| sigla  |  ano_2000  |
+--------+------------+
| SP     | 1.0564745E7 |
| MG     | 4837296.0  |
| RJ     | 4315737.0  |
+--------+------------+
```

Consulta.3

- select uf, qtd_num from qtd_num order by qtd_num desc limit 5;

```
0: jdbc:hive2://localhost:10000/default> select uf, qtd_num from qtd_num order by qtd_num desc limit 5;
INFO  : Compiling command(queryId=hive_20230905000930_eb15c593-8b8d-40cf-b019-6774d6cea9ac): select uf, qtd_num from qtd_num order by qtd_num desc limit 5
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:uf, type:string, comment:null), FieldSchema(name:qtd_num, type:float, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20230905000930_eb15c593-8b8d-40cf-b019-6774d6cea9ac); Time taken: 0.224 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20230905000930_eb15c593-8b8d-40cf-b019-6774d6cea9ac): select uf, qtd_num from qtd_num order by qtd_num desc limit 5
INFO  : Query ID = hive_20230905000930_eb15c593-8b8d-40cf-b019-6774d6cea9ac
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_20230905000930_eb15c593-8b8d-40cf-b019-6774d6cea9ac
INFO  : Session is already open
INFO  : Dag name: select uf, qtd_num from qtd_num order by...5 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1693759391598_0003)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container       SUCCEEDED     1          1        0        0       0       0
Reducer 2 ...... container       SUCCEEDED     1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.80 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20230905000930_eb15c593-8b8d-40cf-b019-6774d6cea9ac); Time taken: 13.16 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------+-----------+
| uf   | qtd_num   |
+------+-----------+
| MG   | 853.0     |
| SP   | 645.0     |
| RS   | 497.0     |
| BA   | 417.0     |
| PR   | 399.0     |
+------+-----------+
5 rows selected (13.426 seconds)
0: jdbc:hive2://localhost:10000/default> Closing: 0: jdbc:hive2://localhost:10000/default
leonardo_moraes@cluster-4c01-m:~$
```

Consulta.4

- select estados_do_brasil, total_de_veiculos from dados_demograficos order by total_de_veiculos desc limit 3;

Consulta.5

- select estados_do_brasil, populacao_residente from dados_demograficos order by populacao_residente desc limit 3;

Relatório:

Como análise das consultas feitas é possível afirmar que dos principais Estados que recebem royalties de petróleo (DF, RJ, ES), o Estado do RJ é o que apresenta maior incidência quando observados outros parâmetros como a quantidade de domicílios e a quantidade de população residente. Esta pode ser uma observação preliminar a ser considerada para a distribuição dos royalties, apesar de que ainda restariam muitas outras análises a serem feitas para agregar ao presente trabalho.

Proposta de evolução do trabalho:

Como proposta de evolução do trabalho, poderão ser incluídos outros parâmetros considerados importantes como, por exemplo, Estados que possuem maior produção de petróleo em seus territórios ou aqueles que carecem de maiores investimentos em serviços de saúde e de infraestrutura.