

## Contextualização:

Uma empresa de transportes particular que atua em Nova Iorque reuniu dados de corridas dos meses de maio e junho de 2020 para fazer o levantamento das características principais das corridas realizadas no período. O setor financeiro da empresa solicitou algumas informações:

Quantidade de passageiros por corrida  
Valores máximo e mínimo de gorjeta no período  
Valor médio da corrida  
Média de distância percorrida nos meses de maio e junho

## Fonte de Dados:

Os dados foram extraídos do Kaggle (<https://www.kaggle.com/datasets/microize/newyork-yellow-taxi-trip-data-2020-2019>), em fomato .csv.

## Dicionário de Dados:

### Descrições de colunas

- VendorID : Identificador do provedor de TPEP que fornece o registro.
  - 1 = Tecnologias Móveis Criativas, LLC
  - 2 = VeriFone Inc.
- tpep\_pickup\_datetime : A data e a hora em que o medidor foi ativado.
- tpep\_dropoff\_datetime : A data e a hora em que o medidor foi desligado.
- Passenger\_count : O número de passageiros no veículo, conforme inserido pelo motorista.
- Trip\_distance : A distância da viagem em milhas, conforme registrada pelo taxímetro.
- PULocationID : Zona de táxi TLC onde o taxímetro foi acionado.
- DOLocationID : Zona de táxi TLC onde o taxímetro foi desativado.
- RateCodeID : O código de tarifa aplicável no final da viagem.
  - 1 = Taxa padrão
  - 2 = JFK
  - 3 = Nova Iorque
  - 4 = Nassau ou Westchester
  - 5 = Tarifa negociada
  - 6 = Passeio em grupo
- Store\_and\_fwd\_flag : indica se o registro da viagem foi armazenado na memória do veículo antes da transmissão ao fornecedor devido à falta de conexão com o servidor.
  - Y = Viagem de armazenamento e encaminhamento
  - N = Não é uma viagem de armazenamento e encaminhamento
- Payment\_type : Como o passageiro pagou pela viagem, representado por um código numérico.
  - 1 = Cartão de crédito
  - 2 = Dinheiro

- 3 = Sem custo
- 4 = Disputa
- 5 = Desconhecido
- 6 = Viagem anulada
- Fare\_amount : A tarifa calculada pelo taxímetro com base no tempo e na distância.
- Extra : Taxas adicionais, atualmente incluindo apenas as taxas de US\$ 0,50 e US\$ 1 para o horário de pico e pernoite.
- MTA\_tax : Um imposto de US\$ 0,50 adicionado automaticamente com base na taxa medida.
- Improvement\_surcharge : Uma sobretaxa de US\$ 0,30 adicionada no início da viagem, implementada desde 2015.
- Tip\_amount : valores de gorjetas de cartão de crédito. (Observação: gorjetas em dinheiro não são registradas aqui.)
- Tolls\_amount : Total de pedágios pagos durante a viagem.
- Total\_amount : O valor total cobrado dos passageiros, excluindo gorjetas em dinheiro.

## Scripts

### Script de carga de dados para o HDFS

Os dados foram baixados e carregados no HDFS através do comando `hdfs dfs -put` na pasta 'projeto'

```
mariana_cmenezes@projhadoop-m:~$ ls
movies.dat projeto yellow_tripdata_2019-06.csv.csupload yellow_tripdata_2020-06.csv
mariana_cmenezes@projhadoop-m:~$ hdfs dfs -ls
Found 1 items
drwxr-xr-x - mariana_cmenezes hadoop 0 2024-09-11 21:22 projeto
mariana_cmenezes@projhadoop-m:~$ hdfs dfs -put yellow_tripdata_2020-06.csv projeto
mariana_cmenezes@projhadoop-m:~$
```

## Script de ingestão de dados no HIVE

- 1) Foi realizado o comando create database taxi para a criação do database
- 2) Foram criadas duas tabelas conforme o script abaixo:

```
CREATE EXTERNAL TABLE tb_dados_taxis (VendorID VARCHAR(3), tpep_pickup_datetime TIMESTAMP,
tpep_dropoff_datetime TIMESTAMP, passenger_count SMALLINT, trip_distance DECIMAL(6,3),Rate_code_id SMALLINT,
store_and_fwd_flag VARCHAR(1), PULocationID SMALLINT, DOLocationID SMALLINT, payment_type VARCHAR(3),
fare_amount DECIMAL(6,2), extra DECIMAL(6,2), mta_tax DECIMAL (6,2), tip_amount decimal(6,2), tolls_amount
DECIMAL(6,2), improvement_surcharge DECIMAL(6,2), total_amount DECIMAL(6,2), congestion_surcharge DECIMAL (6,2))
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
```

```
CREATE EXTERNAL TABLE tb_dados_taxis_maio (VendorID VARCHAR(3), tpep_pickup_datetime TIMESTAMP,
tpep_dropoff_datetime TIMESTAMP, passenger_count SMALLINT, trip_distance DECIMAL(6,3),Rate_code_id SMALLINT,
store_and_fwd_flag VARCHAR(1), PULocationID SMALLINT, DOLocationID SMALLINT, payment_type VARCHAR(3),
fare_amount DECIMAL(6,2), extra DECIMAL(6,2), mta_tax DECIMAL (6,2), tip_amount decimal(6,2), tolls_amount
DECIMAL(6,2), improvement_surcharge DECIMAL(6,2), total_amount DECIMAL(6,2), congestion_surcharge DECIMAL (6,2))
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
```

```
INFO : Concurrency
+-----+
| database_name |
+-----+
| default      |
| taxi         |
+-----+
2 rows selected (0.0591 seconds)
0: jdbc:hive2://localhost:10000/default>
```

## Comando de carga dos dados

```
overwrite into table tb_dados_taxis;
LOAD DATA INPATH '/user/mariana_cmenezes/projeto' overwrite into table tb_dados_taxis
```

## Consultas HQL

Valor máximo de gorjetas em cartão de crédito em maio e junho:

```
select max(tb_dados_taxis_maio.tip_amount) from tb_dados_taxis_maio;  
select max(tb_dados_taxis.tip_amount) from tb_dados_taxis;
```

```
-----  
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: 6.83 s  
-----  
INFO : Completed executing command(queryId=hive_20240911235519_284f153a-81ee-4ce1-b3b8-f2822bffc004); Time taken: 14.173 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+  
| _c0 |  
+-----+  
| 442.18 |  
+-----+  
1 row selected (14.326 seconds)  
0: jdbc:hive2://localhost:10000/default>
```

Valor máximo em maio: 422.2

```
-----  
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: 6.57 s  
-----  
INFO : Completed executing command(queryId=hive_20240912000008_8b7dd43c-db55-45fd-99c8-47d96b65a9d0); Time taken: 6.769 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+  
| _c0 |  
+-----+  
| 422.68 |  
+-----+  
1 row selected (6.94 seconds)  
0: jdbc:hive2://localhost:10000/default>
```

Valor máximo em junho: 422.7

Valor mínimo de gorjetas em cartão de crédito em maio e junho:

```
select min(tb_dados_taxis_maio.tip_amount) from tb_dados_taxis_maio;  
select min(tb_dados_taxis.tip_amount) from tb_dados_taxis;
```

```
-----  
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: 7.14 s  
-----  
INFO : Completed executing command(queryId=hive_20240912000203_592e4e23-a602-41fb-9f20-1be3d11ca89f); Time taken: 7.282 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+  
|  _c0 |  
+-----+  
| -11.06 |  
+-----+  
1 row selected (7.431 seconds)  
0: jdbc:hive2://localhost:10000/default>
```

Valor mínimo em maio: 11.06

```
-----  
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: 6.47 s  
-----  
INFO : Completed executing command(queryId=hive_20240912000405_aa80a913-6595-4715-a79f-b389a1d03da4); Time taken: 6.629 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+  
|  _c0 |  
+-----+  
| -36.30 |  
+-----+  
1 row selected (6.794 seconds)  
0: jdbc:hive2://localhost:10000/default>
```

Valor mínimo em junho: 36.3

Valor médio das corridas em maio e junho:

query (maio): select avg(tb\_dados\_taxis\_maio.total\_amount) from tb\_dados\_taxis\_maio

query (junho): select avg(tb\_dados\_taxis.total\_amount) from tb\_dados\_taxis

```
-----  
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: 6.58 s  
-----  
INFO : Completed executing command(queryId=hive_20240912000757_369e4e65-e215-4ec5-86b2-055c1a2afe6c); Time taken: 6.761 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+  
|          _c0          |  
+-----+  
| 18.4416174756724172575136 |  
+-----+
```

Valor médio em maio: 18.44

```
-----  
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: 5.99 s  
-----  
INFO : Completed executing command(queryId=hive_20240912001219_22cf290a-c959-4179-9f9e-41e7e44b95b7); Time taken: 6.135 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+  
|          _c0          |  
+-----+  
| 18.7689117614959254947614 |  
+-----+  
1 row selected (6.288 seconds)  
0: jdbc:hive2://localhost:10000/default>
```

Valor médio em junho: 18.80

Média de distância das corridas em maio e junho

query(maio): select avg(tb\_dados\_taxis\_maiو.trip\_distance) from tb\_dados\_taxis\_maiو  
query(junho): select avg(tb\_dados\_taxis.trip\_distance) from tb\_dados\_taxis

```
-----  
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: 7.22 s  
-----  
INFO : Completed executing command(queryId=hive_20240912001431_6d5d1daa-b9e9-4b03-ac32-f53ba97db8a6); Time taken: 7.363 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+  
|          _c0          |  
+-----+  
| 3.69791752547724989234965 |  
+-----+  
1 row selected (7.504 seconds)  
0: jdbc:hive2://localhost:10000/default>
```

Maio: média de 3,7 milhas

```
-----  
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: 7.03 s  
-----  
INFO : Completed executing command(queryId=hive_20240912001647_bdaaf734-259a-4c9e-91f1-d3b8ac5a823b); Time taken: 7.153 seconds  
INFO : OK  
INFO : Concurrency mode is disabled, not creating a lock manager  
+-----+  
|          _c0          |  
+-----+  
| 3.35473243535756837136543 |  
+-----+  
1 row selected (7.303 seconds)  
0: jdbc:hive2://localhost:10000/default>
```

Junho: média de 3,5 milhas

Quantidade de passageiros por corrida em maio e junho

```
select tb_dados_taxis_maio.passenger_count, count (*) from tb_dados_taxis_maio
select tb_dados_taxis.passenger_count, count (*) from tb_dados_taxis
```

Observa-se que a maioria das corridas tem apenas um passageiro

```
-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    1         1         0         0         0         0
Reducer 2 ..... container  SUCCEEDED    2         2         0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 6.79 s
-----
INFO : Completed executing command(queryId=hive_20240912004720_d989e39c-b808-4f71-b7cf-80057c302537); Time taken: 6.945 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| tb_dados_taxis_maio.passenger_count | _c1 |
+-----+
| 0 | 9607 |
| 2 | 29564 |
| 3 | 7525 |
| 4 | 2655 |
| 9 | 1 |
| NULL | 58892 |
| 1 | 229321 |
| 5 | 6060 |
| 6 | 4746 |
| 8 | 1 |
+-----+
10 rows selected (7.108 seconds)
0: jdbc:hive2://localhost:10000/default>
```

```
INFO : Status: Running (Executing on YARN cluster with App ID application_1720089463273_0005)
-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    1         1         0         0         0         0
Reducer 2 ..... container  SUCCEEDED    2         2         0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 7.62 s
-----
INFO : Completed executing command(queryId=hive_20240912004501_2cfc5d4d-fded-4372-8a31-8381d0bd0a17); Time taken: 15.536 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| tb_dados_taxis.passenger_count | _c1 |
+-----+
| 0 | 13305 |
| 2 | 57222 |
| 3 | 14333 |
| 4 | 4947 |
| 9 | 2 |
| NULL | 50718 |
| 1 | 388932 |
| 5 | 11219 |
| 6 | 9082 |
| 8 | 1 |
+-----+
10 rows selected (15.7 seconds)
0: jdbc:hive2://localhost:10000/default> select tb_dados_taxis_maio.passenger_count, count(*) from tb_dados_taxis_maio group by pa
```



Proposta de evolução do projeto:

Fazer a integração com Spark para tratamento das bases de dados e utilização da Mllib para aplicação de modelos de machine learning em busca de correlações mais detalhadas entre as características das viagens. É possível também fazer a carga de todo histórico disponível no Kaggle para prever valores de corridas futuras com base na série temporal.

Este projeto está disponível em  
[https://github.com/maricmenezes/hadoop\\_projeto](https://github.com/maricmenezes/hadoop_projeto)