Faculdade Infnet

Atividade de Infraestrutura Hadoop

Modulo: Armazenamento e Processamento Distribuído em Big Data

Disciplina: Infraestrutura Hadoop

Professor: Andre Ormastroni Victor

Aluno: Wellington da Silva Leal

E-mail: wellington.leal@al.infnet.edu.br

Introdução

Este documento descreve a parte conceitual e prática dos conhecimentos adquiridos durante a disciplina Infraestrutura Hadoop. Neste projeto, exploraremos tanto os aspectos teóricos quanto as habilidades práticas fundamentais para trabalhar com esta plataforma de computação distribuída.

Esse relatório, script de criação do database e tabelas, ingestão e consultas estão disponíveis no repositório público do github que pode ser acessado pelo link abaixo:

https://github.com/wellington-infnet/hadoop_atividade

1. Definição do Dataset

Para esse trabalho utilizarei a base pública do kaggle “F1 GrandPrix Dataset” encontrado no link: https://www.kaggle.com/datasets/harshitstark/f1-grandprix-datavault?select=status.csv

Essa base de dados possui os seguintes arquivos:

- circuits.csv: Detalhes de todos os circuitos de corrida.
- constructor_results.csv: Resultados dos construtores em cada corrida.
- constructor_standings.csv: Classificação dos construtores por temporada.
- constructors.csv: Informações sobre todos os construtores.
- driver_standings.csv: Classificação dos pilotos por temporada.
- drivers.csv: Informações sobre todos os pilotos.
- lap_times.csv: Tempos de volta para cada corrida.
- pit_stops.csv: Detalhes das paradas nos boxes durante as corridas.
- qualifying.csv: Resultados das sessões de qualificação.
- races.csv: Informações sobre todas as corridas.
- results.csv: Resultados detalhados das corridas.
- seasons.csv: Visão geral de todas as temporadas de corridas.
- sprint_results.csv: Resultados das corridas sprint.
- status.csv: Códigos de status para eventos de corrida.

2. Arquitetura

Abaixo, seguem as tabelas que serão criadas e o dicionário de dados

- Database: db_formula1
- Tabelas:
    - races
    - results
    - drivers
    - constructors
    - circuits
    - qualifying

Dicionário de dados de cada tabela:

Tabela: races

Contém informações sobre as corridas de Fórmula 1.

Coluna	Tipo	Decrição
raceId	INT	Identificador único da corrida.
year	INT	Ano em que a corrida ocorreu.
round	INT	Número da rodada da corrida na temporada.
circuitId	INT	Identificador do circuito onde a corrida foi realizada.

name	STRING	Nome oficial da corrida.
race_date	DATE	Data da corrida.
race_time	STRING	Hora de início da corrida.
url	STRING	URL para informações adicionais sobre a corrida.
fp1_date	STRING	Data da primeira sessão de treinos livres.
fp1_time	STRING	Hora da primeira sessão de treinos livres.
fp2_date	STRING	Data da segunda sessão de treinos livres.
fp2_time	STRING	Hora da segunda sessão de treinos livres.
fp3_date	STRING	Data da terceira sessão de treinos livres.
fp3_time	STRING	Hora da terceira sessão de treinos livres.
quali_date	STRING	Data da sessão de qualificação.
quali_time	STRING	Hora da sessão de qualificação.
sprint_date	STRING	Data da corrida Sprint.
sprint_time	STRING	Hora da corrida Sprint.

Tabela: results

Coluna	Tipo	Decrição
resultId	INT	Identificador único do resultado.
raceId	INT	Identificador da corrida
driverId	INT	Identificador do piloto
constructorId	INT	Identificador da equipe construtora
number	INT	Número do carro do piloto.
grid	INT	Posição inicial do piloto no grid de largada.
position	INT	Posição final do piloto na corrida.
positionText	STRING	Representação textual da posição final
positionOrder	INT	Ordem final de chegada do piloto
points	FLOAT	Pontos obtidos pelo piloto na corrida
laps	INT	Número de voltas completadas pelo piloto
time	STRING	Tempo total da corrida para o piloto
milliseconds	BIGINT	Tempo total em milissegundos
fastestLap	INT	Número da volta mais rápida do piloto
rank	INT	Classificação da volta mais rápida
fastestLapTime	STRING	Tempo da volta mais rápida
fastestLapSpeed	FLOAT	Velocidade da volta mais rápida
statusId	INT	Identificador do status final do piloto

Tabela: drivers

Coluna	Tipo	Decrição
driverId	INT	Identificador único do piloto.
driverRef	STRING	Referência única do piloto
number	INT	Número do carro do piloto

code	STRING	Código de 3 letras do piloto
forename	STRING	Nome do piloto.
surname	STRING	Sobrenome do piloto.
dob	DATE	Data de nascimento do piloto.
nationality	STRING	Nacionalidade do piloto
url	STRING	URL para mais informações sobre o piloto

Tabela: constructors

Coluna	Tipo	Decrição
constructorId	INT	Identificador único da equipe.
constructorRef	STRING	Referência única da equipe
name	STRING	Nome oficial da equipe
nationality	STRING	Nacionalidade da equipe
url	STRING	URL para mais informações sobre a equipe

Tabela: circuits

Coluna	Tipo	Decrição
circuitId	INT	Identificador único do circuito.
circuitRef	STRING	Referência única do circuito
name	STRING	Nome oficial do circuito
location	STRING	Localização do circuito
country	STRING	País onde o circuito está localizado
lat	FLOAT	Latitude da localização do circuito
lng	FLOAT	Longitude da localização do circuito
alt	INT	Altitude do circuito
url	STRING	URL para mais informações sobre o circuito

Tabela: qualifying

Coluna	Tipo	Decrição
qualifyId	INT	Identificador único da qualificação
raceId	INT	Identificador da corrida
driverId	INT	Identificador do piloto
constructorId	INT	Identificador da equipe
number	INT	Número do carro do piloto
position	INT	Posição de largada do piloto
q1	STRING	Tempo do piloto na primeira sessão de qualificação
q2	STRING	Tempo do piloto na segunda sessão de qualificação
q3	STRING	Tempo do piloto na terceira sessão de qualificação

3. Criação do database e tabelas.

Abaixo conterá a criação do database e tabelas no hive.

- Criação do database

```
hive> create database db_formula1
    > ;
OK
Time taken: 1.595 seconds
hive> show databases;
OK
db_formula1
default
teste_f1
Time taken: 0.125 seconds, Fetched: 3 row(s)
hive>
```

- Criação de tabelas

```
hive> CREATE TABLE races (
    >     raceId INT,
    >     year INT,
    >     round INT,
    >     circuitId INT,
    >     name STRING,
    >     races_date DATE,
    >     races_time STRING,
    >     url STRING,
    >     fp1_date STRING,
    >     fp1_time STRING,
    >     fp2_date STRING,
    >     fp2_time STRING,
    >     fp3_date STRING,
    >     fp3_time STRING,
    >     quali_date STRING,
    >     quali_time STRING,
    >     sprint_date STRING,
    >     sprint_time STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.383 seconds
hive>
```

```
hive> CREATE TABLE results (
    >     resultId INT,
    >     raceId INT,
    >     driverId INT,
    >     constructorId INT,
    >     number STRING,
    >     grid INT,
    >     position INT,
    >     positionText STRING,
    >     positionOrder INT,
    >     points FLOAT,
    >     laps INT,
    >     results_time STRING,
    >     milliseconds INT,
    >     fastestLap INT,
    >     rank STRING,
    >     fastestLapTime STRING,
    >     fastestLapSpeed STRING,
    >     statusId INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.089 seconds
hive>
```

```
hive> CREATE TABLE drivers (
    >     driverId INT,
    >     driverRef STRING,
    >     number STRING,
    >     code STRING,
    >     forename STRING,
    >     surname STRING,
    >     dob DATE,
    >     nationality STRING,
    >     url STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.096 seconds
hive>
```

```
hive> CREATE TABLE constructors (
    >     constructorId INT,
    >     constructorRef STRING,
    >     name STRING,
    >     nationality STRING,
    >     url STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.073 seconds
hive>
```

```
hive> CREATE TABLE circuits (
    >     circuitId INT,
    >     circuitRef STRING,
    >     name STRING,
    >     location STRING,
    >     country STRING,
    >     lat FLOAT,
    >     lng FLOAT,
    >     alt FLOAT,
    >     url STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.082 seconds
hive>
```

```
hive> CREATE TABLE qualifying (
    >     qualifyId INT,
    >     raceId INT,
    >     driverId INT,
    >     constructorId INT,
    >     number STRING,
    >     position INT,
    >     q1 STRING,
    >     q2 STRING,
    >     q3 STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.082 seconds
hive>
```

```
hive> show tables;
OK
circuits
constructors
drivers
qualifying
races
results
Time taken: 0.099 seconds, Fetched: 6 row(s)
hive>
```

4. Ingestão dos dados.

Abaixo conterá a ingestão dos dados nas tabelas no hive.

Utilizei um bucket no Cloud Storage para armazenar todos os arquivos do dataset e realizei o mapeamento do bucket em um diretório do hadoop.

Após mapeamento, abaixo segue a carga dos registros para cada tabela.

- races

```
hive> use db_formula1
    > ;
OK
Time taken: 0.071 seconds
hive> LOAD DATA INPATH '/user/wellington_leal/datasets/races.csv' OVERWRITE INTO TABLE races;
Loading data to table db_formula1.races
OK
Time taken: 0.571 seconds
hive>
```

- results

```
hive> LOAD DATA INPATH '/user/wellington_leal/datasets/results.csv' OVERWRITE INTO TABLE results;
Loading data to table db_formula1.results
OK
Time taken: 0.24 seconds
hive>
```

- drivers

```
hive> LOAD DATA INPATH '/user/wellington_leal/datasets/drivers.csv' OVERWRITE INTO TABLE drivers;
Loading data to table db_formula1.drivers
OK
Time taken: 0.25 seconds
hive>
```

- constructors

```
hive> LOAD DATA INPATH '/user/wellington_leal/datasets/constructors.csv' OVERWRITE INTO TABLE constructors;
Loading data to table db_formula1.constructors
OK
Time taken: 0.233 seconds
hive>
```

- circuits

```
hive> LOAD DATA INPATH '/user/wellington_leal/datasets/circuits.csv' OVERWRITE INTO TABLE circuits;
Loading data to table db_formula1.circuits
OK
Time taken: 0.231 seconds
hive>
```

- qualifying

```
hive> LOAD DATA INPATH '/user/wellington_leal/datasets/qualifying.csv' OVERWRITE INTO TABLE qualifying;
Loading data to table db_formula1.qualifying
OK
Time taken: 0.22 seconds
hive>
```

5. Análises sobre a base.

1. Qual equipe teve o maior número de vitórias em cada temporada nos últimos 5 anos?

```
hive> SELECT year, c.name, COUNT(*) as wins
    > FROM races r
    > JOIN results res ON r.raceId = res.raceId
    > JOIN constructors c ON res.constructorId = c.constructorId
    > WHERE res.position = 1 AND year >= YEAR(CURRENT_DATE) - 5
    > GROUP BY year, c.name
    > ORDER BY year ASC, wins DESC;
No Stats for db_formula1@races, Columns: raceid, year
No Stats for db_formula1@results, Columns: raceid, constructorid, position
No Stats for db_formula1@constructors, Columns: name, constructorid
Query ID = wellington_leal_20240903004725_2f14738a-376b-40f1-b226-853f9b1220fc
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1725322104894_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 4 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 14.64 s
--------------------------------------------------------------------------------
OK
2019    "Mercedes"      15
2019    "Ferrari"       3
2019    "Red Bull"      3
2020    "Mercedes"      13
2020    "Red Bull"      2
2020    "AlphaTauri"    1
2020    "Racing Point"  1
2021    "Red Bull"      11
2021    "Mercedes"      9
2021    "McLaren"       1
2021    "Alpine F1 Team"        1
2022    "Red Bull"      17
2022    "Ferrari"       4
2022    "Mercedes"      1
2023    "Red Bull"      21
2023    "Ferrari"       1
2024    "Red Bull"      7
2024    "Ferrari"       2
2024    "McLaren"       2
2024    "Mercedes"      2
Time taken: 25.433 seconds, Fetched: 20 row(s)
```

Baseado no retorno da consulta, 2019 foi a equipe Mercedes, em 2020 foi a equipe Mercedes, em 2021 foi a equipe Red Bull, em 2022 foi a equipe Red Bull, em 2023 foi a equipe Red Bull, 2024 baseado nos dados é a equipe Red Bull.

2. Qual piloto acumulou o maior número de voltas mais rápidas em corridas desde o ano 2000?

```
hive> SELECT d.driverRef, COUNT(*) as fastest_laps
    > FROM results res
    > JOIN drivers d ON res.driverId = d.driverId
    > JOIN races r ON res.raceId = r.raceId and r.year >= 2000
    > WHERE res.rank = 1
    > GROUP BY d.driverRef
    > ORDER BY fastest_laps DESC
    > LIMIT 1;
No Stats for db_formula1@results, Columns: raceid, driverid, rank
No Stats for db_formula1@drivers, Columns: driverid, driverref
No Stats for db_formula1@races, Columns: raceid, year
Query ID = wellington_leal_20240903005608_b5d8c998-77eb-4a89-8e59-fe29691a5bf4
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1725322104894_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 4 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 14.86 s
--------------------------------------------------------------------------------
OK
"hamilton"      66
Time taken: 24.235 seconds, Fetched: 1 row(s)
hive>
```

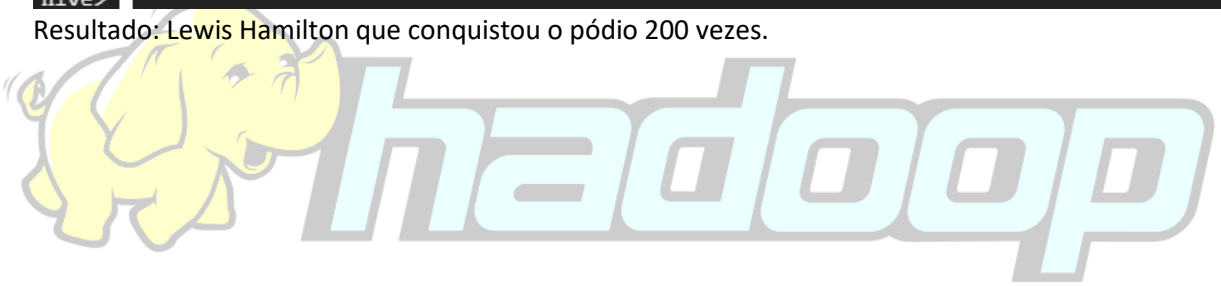
Resultado: Hamilton acumulou o maior número de voltas mais rápidas desde o ano 2000, totalizando 66 voltas nesse período.

3. Qual foi o piloto que mais vezes conquistou uma posição no pódio?

```
hive> SELECT d.forename,d.surname, COUNT(*) as podiums
    > FROM drivers d
    > JOIN results res ON d.driverId = res.driverId and res.position <= 3
    > GROUP BY d.forename,d.surname
    > ORDER BY podiums DESC
    > LIMIT 1;
Query ID = wellington_leal_20240903015010_d1326c36-237e-49bf-b17e-fcab75a979f8
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1725322104894_0009)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 2 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 10.94 s
--------------------------------------------------------------------------------
OK
"Lewis" "Hamilton"      200
Time taken: 19.317 seconds, Fetched: 1 row(s)
hive>
```

Resultado: Lewis Hamilton que conquistou o pódio 200 vezes.



4. Quais foram os circuitos que mais apareceram no calendário da F1 nos últimos 30 anos?

```
hive> SELECT c.name as circuit_name, COUNT(*) as appearances
    > FROM races r
    > JOIN circuits c ON r.circuitId = c.circuitId
    > WHERE year >= YEAR(CURRENT_DATE) - 30
    > GROUP BY c.name
    > ORDER BY appearances DESC;
Query ID = wellington_leal_20240903012915_7b18b5f8-e76c-4f75-9545-a8dc34253b96
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1725322104894_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 4 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 10.62 s
--------------------------------------------------------------------------------
OK
"Silverstone Circuit"   32
"Autodromo Nazionale di Monza"  31
"Hungaroring"   31
"Circuit de Barcelona-Catalunya"        31
"Autódromo José Carlos Pace"    30
"Circuit de Monaco"     30
"Circuit de Spa-Francorchamps"  29
"Circuit Gilles Villeneuve"     28
"Suzuka Circuit"        27
"Albert Park Grand Prix Circuit"        27
"Bahrain International Circuit" 21
"Hockenheimring"        20
"Red Bull Ring" 20
"Sepang International Circuit"  19
"Nürburgring"   17
"Shanghai International Circuit"        17
"Autodromo Enzo e Dino Ferrari" 17
"Yas Marina Circuit"    16
"Circuit de Nevers Magny-Cours" 15
"Marina Bay Street Circuit"     15
"Circuit of the Americas"       12
"Autódromo Hermanos Rodríguez"  9
"Istanbul Park" 9
"Indianapolis Motor Speedway"   8
"Baku City Circuit"     8
"Sochi Autodrom"        8
"Valencia Street Circuit"       5
"Autódromo Juan y Oscar Gálvez" 4
"Jeddah Corniche Circuit"       4
"Korean International Circuit"  4
"Circuit Paul Ricard"   4
"Circuit Park Zandvoort"        4
"Losail International Circuit"  3
"Autódromo do Estoril"  3
"Miami International Autodrome" 3
"Buddh International Circuit"   3
"Adelaide Street Circuit"       2
"Autódromo Internacional do Algarve"    2
```

Resultado: Silverstone é o circuito que mais vezes esteve no calendário da F1.

5. Qual a distribuição de pódios entre as equipes nos últimos 15 anos?

```
hive> SELECT c.name as constructor_name, COUNT(*) as podiums
    > FROM results res
    > JOIN constructors c ON res.constructorId = c.constructorId
    > JOIN races r ON res.raceId = r.raceId AND year >= YEAR(CURRENT_DATE) - 15
    > WHERE res.position <= 3
    > GROUP BY c.name
    > ORDER BY podiums DESC;
No Stats for db_formula1@results, Columns: raceid, constructorid, position
No Stats for db_formula1@constructors, Columns: name, constructorid
No Stats for db_formula1@races, Columns: raceid, year
Query ID = wellington_leal_20240903013600_c2741b1e-52a7-4f66-9eb9-2337fa62159d
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1725322104894_0008)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 4 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 5 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 15.11 s
--------------------------------------------------------------------------------
OK
"Mercedes"      277
"Red Bull"      274
"Ferrari"       195
"McLaren"       83
"Lotus F1"      25
"Williams"      17
"Brawn" 15
"Aston Martin"  9
"Renault"       9
"Force India"   6
"Toyota"        5
"Racing Point"  4
"Alpine F1 Team"        4
"Sauber"        4
"BMW Sauber"    2
"Toro Rosso"    2
"AlphaTauri"    2
Time taken: 24.863 seconds, Fetched: 17 row(s)
```

Resultado: O resultado, considerando os 3 primeiros, Mercedes nos últimos 15 anos esteve 277 vezes no pódio, Red Bull esteve 274 vezes e Ferrari 195 vezes.

6. Evolução do trabalho.

Para evolução e continuidade do trabalho, pode ser realizado algumas transformações na base e a aplicação de um modelo de ML, para realizar previsões ou análises classificatórias.