

Big data aplicat

Practica 3



Jorge Osarenkhoe Petro

15 de desembre de 2024

Índex

		Page
1	APARTAT 1	1
1.1	Nombre de gols que ha marcat en Lionel Messi (sense comptar autogols). . .	3
1.2	Llistat dels 5 partits més recents que ha jugat la selecció espanyola.	3
1.3	Nombre de gols que ha marcat Espanya en tota la seva història. Aquesta informació s'ha de treure de results, ja que goalscorers no conté tots els gols. .	3
1.4	Llistat dels 5 màxims golejadors amb la selecció espanyola (sense comptar autogols).	4
1.5	Llistat dels jugadors espanyols que han marcat algun gol de penal en alguna Eurocopa (UEFA Euro), ordenats alfabèticament.	5
1.6	Llistat dels 5 màxims golejadors de les fases finals dels mundials (FIFA World Cup) (sense comptar autogols).	6
2	APARTAT 2	7
2.1	Quina de les plataformes té més pel·lícules a la seva col·lecció? Mostra la plataforma i el nombre de pel·lícules.	8
2.2	Quines són les 5 sèries amb millor valoració a IMDB (imdbAverageRating)? Per a cada sèrie, mostra el títol, la valoració i la plataforma on es troba. . .	9
2.3	Quin és el total de vots en IMDB (imdbNumVotes) de totes les sèries del gènere de ciència-ficció en cada una de les plataformes? Per a cada plataforma, mostra la plataforma i el nombre de vots, ordenats de major a menor nombre de vots.	9
2.4	Quins són els 5 anys en què s'han llançat més pel·lícules? Per a cada any, mostra l'any i el nombre de pel·lícules, ordenats de major a menor nombre de pel·lícules.	10

1 APARTAT 1

Anam a treballar amb el dataset de resultats de tots els partits de futbol disputats entre seleccions nacionals des de 1872 fins a l'actualitat, que podeu trobar a Kaggle, on podràs trobar-ne tots els detalls.

Dels tres fitxers de què consta el dataset, ens interessen només dos:

- results.csv, que conté la informació de tots els partits disputats, incloent-hi els equips, el marcador, el campionat i la seu.
- goalscorers.csv, que conté la informació de tots els gols marcats en aquests partits. Per a cada gol, s'indica el partit (data i equips), l'equip i jugador que fa el gol, el minut i dos flags que indiquen si ha estat en pròpia porteria o de penal.

Nota: No estan registrats els gols de tots els partits que apareixen a resultats.csv. En falten els gols de partits antics. Així mateix, els gols dels partits més recents, tampoc s'han recollit encara.

ALERTA

Com que tenim comes dins les dades, podem tenir problemes quan importem aquests fitxers en Hive. Podríem definir la taula amb format Serde, però així no podem mantenir els tipus de dades originals (totes passarien a Strings). Així doncs, abans de carregar les dades en Hive, cal transformar aquests fitxers, emprant el tabulador com a separador de camps. També pots trobar els arxius en el GitHub del curs (actualitzats fins al 22/11/2024), on ja s'han emprat tabuladors com a separadors de camp: results.csv i goalscorers.csv.

Com a analistes de dades ens han demanat una sèrie de preguntes que hem de respondre, utilitzant Apache Hive. Són aquestes:

```
CREATE DATABASE soccer;
USE soccer;

CREATE TABLE soccer.results (
  date DATE,
  home_team STRING,
  away_team STRING,
  home_score INT,
  away_score INT,
  tournament STRING,
  city STRING,
  country STRING,
  neutral BOOLEAN )
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
TBLPROPERTIES ("skip.header.line.count"="1");

CREATE TABLE soccer.goalscorers (
  date STRING,
  home_team STRING,
```

```

away_team STRING,
team STRING,
scorer STRING,
minute INT,
own_goal BOOLEAN,
penalty BOOLEAN )
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
TBLPROPERTIES ("skip.header.line.count"="1");

```

The screenshot shows the Hive query editor interface. The top query is a CREATE TABLE statement for 'soccer.results'. The bottom query is a CREATE TABLE statement for 'soccer.goalscorers'. Both queries are executed successfully, as indicated by the 'Success.' messages and the 'USE soccer' command in the query history.

Query 1: CREATE TABLE soccer.results

```

1 CREATE TABLE soccer.results (
2   date DATE,
3   home_team STRING,
4   away_team STRING,
5   home_score INT,
6   away_score INT,
7   tournament STRING,
8   city STRING,
9   country STRING,
10  neutral BOOLEAN
11 )
12 ROW FORMAT DELIMITED
13 FIELDS TERMINATED BY '\t'
14 TBLPROPERTIES ("skip.header.line.count"="1");

```

Query 2: CREATE TABLE soccer.goalscorers

```

1 CREATE TABLE soccer.goalscorers (
2   date STRING,
3   home_team STRING,
4   away_team STRING,
5   team STRING,
6   scorer STRING,
7   minute INT,
8   own goal BOOLEAN,
9   penalty BOOLEAN
10 )
11 ROW FORMAT DELIMITED
12 FIELDS TERMINATED BY '\t'
13 TBLPROPERTIES ("skip.header.line.count"="1");
14

```

Query History:

Query	Time	Status	Command
CREATE TABLE soccer.results (date DATE, home_team STRING, away_team STRING, home_score INT, away_score INT, tournament STRING, city STRING, country STRING, neutral BOOLEAN) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' TBLPROPERTIES ("skip.header.line.count"="1")	a few seconds ago	✓	USE soccer
USE soccer	a minute ago	✓	

```

LOAD DATA LOCAL INPATH "/home/cloudera/Desktop/goalscorers.csv" INTO TABLE soccer.goalscorers;
LOAD DATA LOCAL INPATH "/home/cloudera/Desktop/results.csv" INTO TABLE soccer.results;

```

- 1.1 Nombre de gols que ha marcat en Lionel Messi (sense comptar autogols).

```
SELECT count(*) FROM soccer.goalscorers WHERE scorer = 'Lionel Messi'
AND own_goal = FALSE;
```

The screenshot shows a SQL query editor with the following query:

```
1 | SELECT count(*) FROM soccer.goalscorers WHERE scorer = 'Lionel Messi' AND own_goal = FALSE;
```

Below the query editor, the results are displayed in a table with one column labeled `_c0`. The result is 55.

_c0
55

1.2 Llistat dels 5 partits més recents que ha jugat la selecció espanyola.

- ```
SELECT * FROM soccer.results
WHERE home_team = 'Spain' OR away_team = 'Spain'
ORDER BY date DESC
LIMIT 5;
```

The screenshot shows a SQL query editor with the following query:

```
1 | SELECT *
2 | FROM soccer.results
3 | WHERE home_team = 'Spain' OR away_team = 'Spain'
4 | ORDER BY date DESC
5 | LIMIT 5;
```

Below the query editor, the results are displayed in a table with 9 columns: `results.date`, `results.home_team`, `results.away_team`, `results.home_score`, `results.away_score`, `results.tournament`, `results.city`, `results.country`, and `results.neutral`. The results are ordered by date in descending order, showing the 5 most recent matches.

|   | results.date | results.home_team | results.away_team | results.home_score | results.away_score | results.tournament  | results.city           | results.country | results.neutral |
|---|--------------|-------------------|-------------------|--------------------|--------------------|---------------------|------------------------|-----------------|-----------------|
| 1 | 2024-11-18   | Spain             | Switzerland       | 3                  | 2                  | UEFA Nations League | Santa Cruz de Tenerife | Spain           | false           |
| 2 | 2024-11-15   | Denmark           | Spain             | 1                  | 2                  | UEFA Nations League | Copenhagen             | Denmark         | false           |
| 3 | 2024-10-15   | Spain             | Serbia            | 3                  | 0                  | UEFA Nations League | Cordoba                | Spain           | false           |
| 4 | 2024-10-12   | Spain             | Denmark           | 1                  | 0                  | UEFA Nations League | Murcia                 | Spain           | false           |
| 5 | 2024-09-08   | Switzerland       | Spain             | 1                  | 4                  | UEFA Nations League | Geneva                 | Switzerland     | false           |

- 1.3 Nombre de gols que ha marcat Espanya en tota la seva història. Aquesta informació s'ha de treure de results, ja que goalscorers no conté tots els gols.

```
SELECT SUM
```

```
(CASE WHEN home_team = 'Spain' THEN home_score ELSE 0 END +
CASE WHEN away_team = 'Spain' THEN away_score ELSE 0 END)
FROM soccer.results;
```

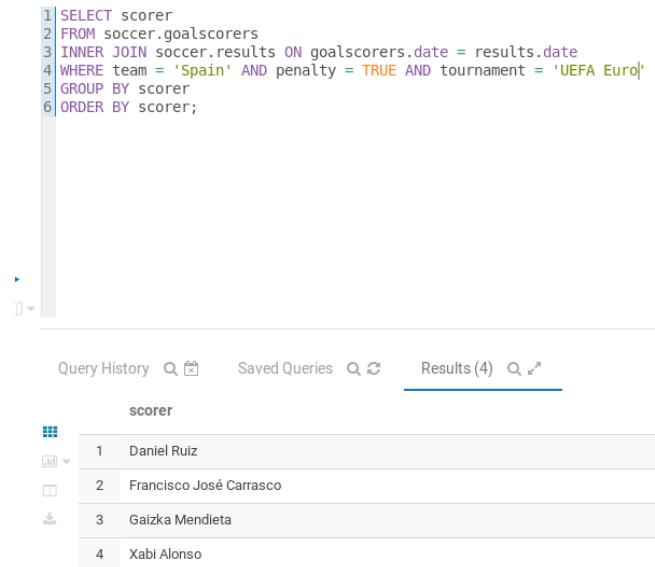
```
1 SELECT sum(home_score) + sum(away_score) FROM soccer.results
2 WHERE home_team = 'Spain' OR away_team = 'Spain';
```

Query History Saved Queries Results (1)

| _c0 |      |
|-----|------|
| 1   | 2236 |

#### 1.4 Llistat dels 5 màxims golejadors amb la selecció espanyola (sense comptar autogols).

- ```
SELECT scorer, count(*) as goals  
FROM soccer.goalscorers  
WHERE team = 'Spain' AND own_goal = FALSE  
GROUP BY scorer  
ORDER BY goals DESC  
LIMIT 5;
```



The screenshot shows a SQL query editor with the following code:

```
1 SELECT scorer
2 FROM soccer.goalscorers
3 INNER JOIN soccer.results ON goalscorers.date = results.date
4 WHERE team = 'Spain' AND penalty = TRUE AND tournament = 'UEFA Euro'
5 GROUP BY scorer
6 ORDER BY scorer;
```

Below the query editor, there is a tab labeled "Results (4)". The results are displayed in a table with the following data:

	scorer
1	Daniel Ruiz
2	Francisco José Carrasco
3	Gaizka Mendieta
4	Xabi Alonso

- 1.5 Llistat dels jugadors espanyols que han marcat algun gol de penal en alguna Eurocopa (UEFA Euro), ordenats alfabèticament.

```
SELECT scorer
FROM soccer.goalscorers
INNER JOIN soccer.results ON goalscorers.date = results.date
WHERE team = 'Spain' AND penalty = TRUE AND tournament = 'UEFA Euro'
GROUP BY scorer
ORDER BY scorer;
```

```
1 SELECT scorer
2 FROM soccer.goalscorers
3 INNER JOIN soccer.results ON goalscorers.date = results.date
4 WHERE team = 'Spain' AND penalty = TRUE AND tournament LIKE '%UEFA Euro%'
5 GROUP BY scorer
6 ORDER BY scorer;
```

Query History Saved Queries Results (13)

	scorer
1	Andrés Iniesta
2	Daniel Ruiz
3	David Villa
4	Fernando Hierro
5	Francisco José Carrasco
6	Gaizka Mendieta
7	José Claramunt
8	Juan Antonio Señor
9	Michel
10	Pirri
11	Sergio Ramos
12	Xabi Alonso
13	Álvaro Morata

1.6 Llistat dels 5 màxims golejadors de les fases finals dels mundials (FIFA World Cup) (sense comptar autogols).

- ```
SELECT scorer, count(*) as goals
FROM soccer.goalscorers
INNER JOIN soccer.results ON goalscorers.date = results.date
WHERE tournament = 'FIFA World Cup' AND own_goal = FALSE
GROUP BY scorer
ORDER BY goals DESC
LIMIT 5;
```



```

1|SELECT scorer, count(*) as goals
2|FROM soccer.goalscorers
3|INNER JOIN soccer.results ON goalscorers.date = results.date
4|WHERE tournament = 'FIFA World Cup' AND own_goal = FALSE
5|GROUP BY scorer
6|ORDER BY goals DESC
7|LIMIT 5;

```

Query History Q Saved Queries Q Results (5) Q

|   | scorer         | goals |
|---|----------------|-------|
| 1 | Just Fontaine  | 60    |
| 2 | Gerd Müller    | 49    |
| 3 | Helmut Rahn    | 44    |
| 4 | Uwe Seeler     | 44    |
| 5 | Miroslav Klose | 43    |

## 2 APARTAT 2

En la tasca del lliurament 2 varem fer feina amb el dataset de pel·lícules i sèries de la plataforma Amazon Prime, publicat a Kaggle per OctopusTeam. A més d'Amazon Prime, OctopusTeam també publica el dataset de les altres principals plataformes de streaming:

- Netflix
- Apple TV+
- Amazon Prime
- Hulu
- HBO Max

Tots els datasets tenen la mateixa estructura.

ALERTA També aquí tenim comes dins les dades. Així doncs, abans de carregar les dades en Hive, cal transformar aquests fitxers, emprant el tabulador com a separador de camps. També pots trobar els arxius en el GitHub del curs (actualitzats fins al 22/11/2024), on ja s'han emprat tabuladors com a separadors de camp: Netflix, Apple TV+, Amazon Prime, Hulu i HBO Max. Has de descarregar l'arxiu de cada plataforma i importar les dades en una taula Hive, utilitzant 5 particions estàtiques, una per a cada plataforma.

ALERTA És obligatori definir una partició estàtica per a cada plataforma!

Només s'han de tenir en compte aquelles sèries o pel·lícules que estan registrades a IMDB (tenen un imdbId). Pots llevar les altres files directament dels arxius de dades, abans de fer la càrrega. Si una sèrie o pel·lícula està disponible a diverses plataformes, podem considerar-les com a sèries o pel·lícules diferents. Recorda que el camp type ens indica si es tracta d'una pel·lícula (movie) o una sèrie (tv). Has de respondre les següents consultes:

```

CREATE DATABASE soccer;
USE soccer;

```

```

CREATE TABLE streaming.movies (
 title STRING,
 type STRING,
 genres STRING,
 releaseYear FLOAT,
 imdbId STRING,
 imdbAverageRating FLOAT,
 imdbNumVotes INT,
 availableCountries STRING
)
PARTITIONED BY(platform STRING)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t'
TBLPROPERTIES ("skip.header.line.count"="1");

LOAD DATA LOCAL INPATH '/home/cloudera/desktop/streaming/amazon.csv' INTO TABLE streaming.movies
LOAD DATA LOCAL INPATH '/home/cloudera/desktop/streaming/apple.csv' INTO TABLE streaming.movies
LOAD DATA LOCAL INPATH '/home/cloudera/desktop/streaming/hbo.csv' INTO TABLE streaming.movies
LOAD DATA LOCAL INPATH '/home/cloudera/desktop/streaming/hulu.csv' INTO TABLE streaming.movies
LOAD DATA LOCAL INPATH '/home/cloudera/desktop/streaming/netflix.csv' INTO TABLE streaming.movies


```

- **2.1 Quina de les plataformes té més pel·lícules a la seva col·lecció?**  
Mostra la plataforma i el nombre de pel·lícules.

```

SELECT platform, count(*) as num_movies
FROM streaming.movies
WHERE type = 'movie'
GROUP BY platform
ORDER BY num_movies DESC
LIMIT 1;

```



The screenshot shows a SQL query editor with the following query:

```

1 SELECT platform, count(*) as num_movies
2 FROM streaming.movies
3 WHERE type = 'movie'
4 GROUP BY platform
5 ORDER BY num_movies DESC
6 LIMIT 1;

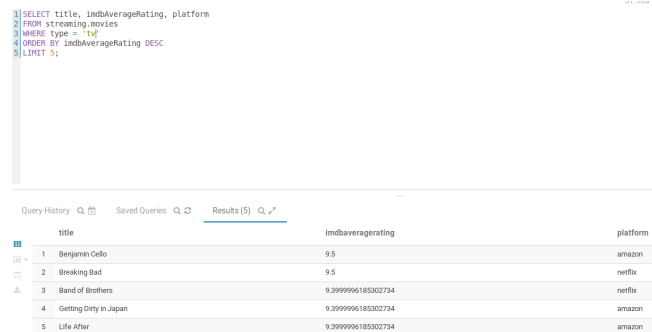
```

Below the query, the results are displayed in a table with the columns 'platform' and 'num\_movies'.

| platform | num_movies |
|----------|------------|
| amazon   | 59244      |

**2.2** Quines són les 5 sèries amb millor valoració a IMDB (imdbAverageRating)? Per a cada sèrie, mostra el títol, la valoració i la plataforma on es troba.

- ```
SELECT title, imdbAverageRating, platform
FROM streaming.movies
WHERE type = 'tv'
ORDER BY imdbAverageRating DESC
LIMIT 5;
```



The screenshot shows a SQL query editor with the following query:

```
1 SELECT title, imdbAverageRating, platform
2 FROM streaming.movies
3 WHERE type = 'tv'
4 ORDER BY imdbAverageRating DESC
5 LIMIT 5;
```

Below the query editor, the results are displayed in a table with the following columns: title, imdbAverageRating, and platform. The results are ordered by the average rating in descending order.

	title	imdbAverageRating	platform
1	Benjamin Cello	9.5	amazon
2	Breaking Bad	9.5	netflix
3	Band of Brothers	9.3999996185302734	netflix
4	Getting Dirty in Japan	9.3999996185302734	amazon
5	Life After	9.3999996185302734	amazon

- **2.3** Quin és el total de vots en IMDB (imdbNumVotes) de totes les sèries del gènere de ciència-ficció en cada una de les plataformes? Per a cada plataforma, mostra la plataforma i el nombre de vots, ordenats de major a menor nombre de vots.

```
SELECT platform, SUM(imdbNumVotes) AS total_votes
FROM streaming.movies
WHERE genres LIKE '%Science Fiction%'
    OR genres LIKE '%Sci-Fi%'
    AND type = 'tv'
GROUP BY platform
ORDER BY total_votes DESC;
```

```
1 SELECT platform, SUM(imdbNumVotes) AS total_votes
2 FROM streaming.movies
3 WHERE genres LIKE '%Science Fiction%'
4 OR genres LIKE '%Sci-Fi%'
5 AND type = 'tv'
6 GROUP BY platform
7 ORDER BY total_votes DESC;
```

Query History Saved Queries Results (5)

	platform	total_votes
1	amazon	3403767
2	apple	2375367
3	netflix	1665835
4	hulu	1551810
5	hbo	745275

2.4 Quins són els 5 anys en què s'han llançat més pel·lícules? Per a cada any, mostra l'any i el nombre de pel·lícules, ordenats de major a menor nombre de pel·lícules.

- ```
SELECT releaseYear, count(*) as num_movies
FROM streaming.movies
WHERE type = 'movie'
GROUP BY releaseYear
ORDER BY num_movies DESC
LIMIT 5;
```

```
1 SELECT releaseYear, count(*) as num_movies
2 FROM streaming.movies
3 WHERE type = 'movie'
4 GROUP BY releaseYear
5 ORDER BY num_movies DESC
6 LIMIT 5;
```

Query History Saved Queries Results (5)

|   | releaseyear | num_movies |
|---|-------------|------------|
| 1 | 2022        | 6068       |
| 2 | 2019        | 6045       |
| 3 | 2018        | 5711       |
| 4 | 2021        | 5522       |
| 5 | 2017        | 5209       |