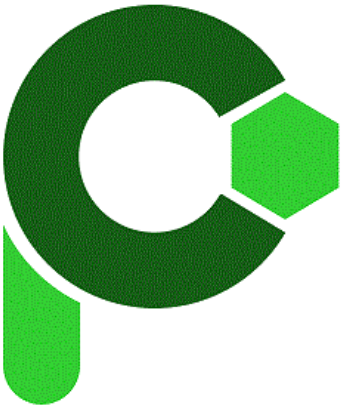

Apache Pig / Apache Hive

1. Comparar el valor d'opinió obtinguda amb el camp label

a> Fitxer CSV amb els següents camps: (text, label, opinio_optinguda, comparació) A on comparació serà un camp booleà, aquest serà true si l'opinio_obtinguda es correspon amb l'etiqueta.

De la taula generada amb l'script feim el següent select on a la columna score interpretam els valors majors a 0 com a positius (1) i els menors com a negatius (0), també incloem una columna on comparam l'opinio_optinguda amb el label.

```
SELECT id, text, label,
CASE WHEN score > 0 THEN 1 ELSE 0 END AS opinio_optinguda,
CASE
   WHEN label = (CASE WHEN score > 0 THEN 1 ELSE 0 END) THEN 1
   ELSE 0
END AS correcte
FROM practica_pig_hive
```

ilian life is the local bank manager is a fussy little man peering at the world through a pair of thick spectacles. It is he who takes the initiative in forming the Home Guard unit and who appoints himself its c
ommander. He is pompous officious with an exaggerated sense of his own importance and of his own powers of leadership the sort of man who does not suffer fools gladly. (And in George Mainwarings world-view the t
erm fool covers most of the rest of the human race). He does however have his good qualities. He is motivated by a genuine patriotic idealism and is capable of great physical courage shown in his encounter with
the Germans.

Wilson is Mainwarings deputy at the bank. The two men are very different in character something emphasised by a difference in appearance Wilson being tall and thin whereas Mainwaring is s
hort and stout. He comes across as being both more intelligent and better educated than his boss. (His accent suggests he may be a former public schoolboy). Nevertheless he has ended up playing second fiddle bot
h in civilian and military life probably because he has the sort of passive personality which leads to pessimism and defeatism and an inability to take anything altogether seriously. Jones is an old soldier who
now runs the local butchers shop. (His promotion to Corporal is due mainly to his ability to bribe Mainwaring with black market sausages). His enthusiasm for his new role is matched only by his incompetence and
ability to cause chaos. Although his catchphrase is Dont panic! he is prone to panicking at any given opportunity.

Several other members of the platoon are featured. Private Fraser the dour Scottish u
ndertaker is even more of a pessimist than Wilson. (Catchphrase: Were doomed man DOOMED!). Private Godfrey is a gentle old man whose main concern is the whereabouts of the nearest lavatory. Private Walker is a s
harp Cockney spiv and Private Pike (another bank employee) a spoilt mummys boy. (Pikes mother is Wilsons mistress although Wilson tries to keep this liaison secret from the disapproving Mainwaring). Two signific
ant outsiders are the mild-mannered Vicar and the ARP warden Mainwarings detested enemy and quite his equal in pompousness and officiousness.

There are occasional bawdy doubles entendres (Keep your ha
nds off my privates- Mainwaring is ostensibly referring to those soldiers who hold that rank) more so than in the television show which was surprisingly free of innuendo. (Its creators David Croft and Jimmy Perr
y would later go on to create comedy shows such as Are You Being Served? and Hi-de-hi which were notorious for suggestive humour). The film does however preserve much of the mixture of gentle wit nostalgia and s
harp characterisation which made the TV series so successful. 7/10,1,1,1

	text	label	opinio_optinguda	correcte
190				
191	ad with this one.	0	0	1
192	with either one.	1	1	1
193	ı it was shot on.	0	0	1
194	FPS games more than i do.7/10 STARS	1	1	1
195	focused on what might have been.	0	1	0
196	ı was well worth it.	0	1	0
197	e this as one of the best movie of the year	1	1	1

b> Fitxer CSV a on es farà un recompta a on la comparació sigui true i un recompta a on la comparació sigui false.

A partir del select del punt anterior sumam els valors de la columna comparació per obtenir un recompte dels correctes. Per al recompte d'incorrectes restam al total en nombre de valoracions correctes.

```
SELECT
SUM(label) AS n_correctes,
COUNT(label)-SUM(label) AS n_incorrectes
FROM(
   SELECT id, text, label,
   CASE WHEN score > 0 THEN 1 ELSE 0 END AS opinio_optinguda,
   CASE
      WHEN label = (CASE WHEN score > 0 THEN 1 ELSE 0 END) THEN 1
      ELSE 0
   END AS correcte
   FROM practica_pig_hive
) AS opinions
```

```
2473,2442
```

	n_correctes	n_incorrectes
1	2473	2442

2. Relacionar les pel·lícules amb les opinions

a> Descarregar el fitxer d'un repositori de dades.

El fitxer amb les pel·lícules que es descarregarà és el següent:
https://raw.githubusercontent.com/JoanBF20/Pelis_csv/main/pelis.csv

Descarregam el fitxer:

```
wget https://raw.githubusercontent.com/JoanBF20/Pelis_csv/main/pelis.csv
```

```
--2023-03-31 12:40:09--  https://raw.githubusercontent.com/JoanBF20/Pelis_csv/main/pelis.csv
Resolving raw.githubusercontent.com... 185.199.109.133, 185.199.110.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com|185.199.109.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 388 [text/plain]
Saving to: "pelis.csv"

100%[=============================================================================================================================================================>] 388         --.-K/s   in 0s

2023-03-31 12:40:09 (7.91 MB/s) - "pelis.csv" saved [388/388]
```

Visualitzam el contingut del fitxer:

```
cat pelis.csv
```

```
0,Avatar
1,Avengers: Endgame
2,Avatar: The Way of Water
3,Titanic
4,Star Wars: Episodio VII - El despertar de la Fuerza
5,Avengers: Infinity War
6,Spider-Man: No Way Home
7,Jurassic World
8,El rey león
9,The Avengers
10,Fast and Furious 7
11,Top Gun: Maverick
12,Frozen II
13,Avengers: Age of Ultron
14,Black Panther
15,Harry Potter y las reliquias de la muerte: parte 2
```

b> Carregar aquest fitxer el hdfs.

```
hdfs dfs -put pelis.csv /user/cloudera/pig_analisis_opinions
```

Modificam els permisos del fitxer:

```
hdfs dfs -chmod 777 /user/cloudera/pig_analisis_opinions/pelis.csv
```

Per importar les dades a una taula. Primer cream una taula nova:

```
CREATE TABLE IF NOT EXISTS default.pelis (
 id int,
 film string)
 ROW FORMAT DELIMITED
 FIELDS TERMINATED BY ',';
```

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
OK
Time taken: 0.515 seconds
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
```

Carregam les dades del fitxer a la taula creada prèviament:

```
LOAD DATA INPATH '/user/cloudera/pig_analisis_opinions/pelis.csv' INTO TABLE default.pelis;
```

```
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
Loading data to table default.pelis
Table default.pelis stats: [numFiles=10, totalSize=4168]
OK
Time taken: 1.099 seconds
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
```

c> Crear un fitxer CSV amb els següents camps (nom_pelicula, numero_opinions, opinió_optinguda, nlabels_positius, nlabels_negatius, opinio_obtinguda_text) A on opinio_obtinguda_text serà el valor que creguis més adequat per cada pel·lícula. Obtingut a partir de l'anàlisi del text de les crítiques cinematogràfiques.

Realitzam un subselect on feim el join de les 2 taules i a partir d'aquest agrupam per pel·lícula i n'extreim els camps demanats, el nom, el recompte de labels positius i negatius, el nombre d'opinions, l'opinió obtinguda i l'opinió obtinguda en text.

```
SELECT film,
SUM(label) AS nlabels_positius,
COUNT(label)-SUM(label) AS nlabels_negatius,
COUNT(label) AS n_opinions,
CASE WHEN AVG(score) > 0 THEN 1 ELSE 0 END AS opinio_optinguda,
CASE WHEN AVG(score) > 0 THEN 'Positiu' ELSE 'Negatiu' END AS opinio_optinguda_text
FROM (
   SELECT film, practica_pig_hive.* FROM practica_pig_hive
   INNER JOIN pelis ON practica_pig_hive.id=pelis.id
) AS joined_tables
GROUP BY film
```

```
"Avatar",576,747,1323,1,Positiu
"Avatar: The Way of Water",1368,1548,2916,1,Positiu
"Avengers: Age of Ultron",1647,1521,3168,1,Positiu
"Avengers: Endgame",1530,1440,2970,1,Positiu
"Avengers: Infinity War",1584,1404,2988,1,Positiu
"Black Panther",1575,1593,3168,1,Positiu
"El rey león",1512,1602,3114,1,Positiu
"Fast and Furious 7",1647,1386,3033,1,Positiu
"Frozen II",1530,1440,2970,1,Positiu
"Harry Potter y las reliquias de la muerte: parte 2",702,648,1350,1,Positiu
"Jurassic World",1368,1485,2853,1,Positiu
"Spider-Man: No Way Home",1377,1476,2853,1,Positiu
"Star Wars: Episodio VII - El despertar de la Fuerza",1476,1458,2934,1,Positiu
"The Avengers",1314,1467,2781,1,Positiu
"Titanic",1521,1287,2808,1,Positiu
"Top Gun: Maverick",1530,1476,3006,1,Positiu
```

	film	nlabels_positius	nlabels_negatius	n_opinions	opinio_opt
1	"Frozen II"	170	160	330	1
2	"Avengers: Age of Ultron"	183	169	352	1
3	"El rey león"	168	178	346	1
4	"The Avengers"	146	163	309	1

d> Crear un fitxer CSV amb els següents camps (nom_pelicula, opinio_labels, opinio_text) A on opinio_labels serà un valor obtingut a partir del camp labels que podrà tenir els següents valors de 0,1,2,3,4. A on opinio_text serà un valor obtingut a partir del camp opinio_obtinguda_text que podrà tenir els següents valors de 0,1,2,3,4.
Realitzam un subselect on feim el join de les 2 taules i a partir d'aquest feim la mitjana de la columna label que ens retornarà un valor entre 0 i 1 i li assignem un valor de 0 a 4, a partir de la mitja de la columna score que ens retorna valors entre -1125 fins a 2375 i li assignem un valor de 0 a 4.

```
SELECT film,
CASE
   WHEN AVG(label) > 0.80 THEN 4
   WHEN AVG(label) > 0.60 THEN 3
   WHEN AVG(label) > 0.40 THEN 2
   WHEN AVG(label) > 0.20 THEN 1
   ELSE 0
END AS opinio_labels,
CASE
   WHEN AVG(score) > 1675 THEN 4
   WHEN AVG(score) > 975 THEN 3
   WHEN AVG(score) > 275 THEN 2
   WHEN AVG(score) > -425 THEN 1
   ELSE 0
END AS opinio_text
FROM (
   SELECT film, practica_pig_hive.* FROM practica_pig_hive
   INNER JOIN pelis ON practica_pig_hive.id=pelis.id
) AS joined_tables
GROUP BY film
```

```
"Avatar",2,1
"Avatar: The Way of Water",2,1
"Avengers: Age of Ultron",2,1
"Avengers: Endgame",2,1
"Avengers: Infinity War",2,1
"Black Panther",2,1
"El rey león",2,1
"Fast and Furious 7",2,1
"Frozen II",2,1
"Harry Potter y las reliquias de la muerte: parte 2",2,1
"Jurassic World",2,1
"Spider-Man: No Way Home",2,1
"Star Wars: Episodio VII - El despertar de la Fuerza",2,1
"The Avengers",2,1
"Titanic",2,1
"Top Gun: Maverick",2,1
```

	film	opinio_labels	opinio_text
1	"Frozen II"	2	1
2	"Avengers: Age of Ultron"	2	1
3	"El rey león"	2	1
4	"The Avengers"	2	1
5	"Avengers: Endgame"	2	1