EXAMEN DE CBDE

13 de Gener de 2021

Instruccions: Respon cada pregunta i subapartat a l'espai habilitat.

L'examen dura 1h 45'.

Nom i Cognoms: Júlia Gasull i Navarro

PREGUNTA 1. [4p] ***(Temps de referència: 30')***

Considera la següent taula que guarda un *log* d'accions a una web de compres online (p.e., *Amazon*). La taula emmagatzema l'usuari, el moment, l'acció feta i sobre quin element l'ha feta.

UserID	Timestamp	Action	Element
Jmercade	010120210930	Click	Menu_Products
Jmercade	010120210931	Free_Text	Search_Bar
Jmercade	010120210932	Click	Product_Item
Jmercade	010120210933	Click	Product_Description
Jmercade	010120210934	Click	Buy_Product
Jmercade	010120210935	Exit	*NULL*
Acanyelles	010120210936	Click	Contact_Page
Ogarcia	010120210936	Click	Menu_Products
Ogarcia	010120210938	Back_Button	*NULL*
Acanyelles	010120210944	Free_Text	Contact_Text
Acanyelles	010120210950	Click	Send_Contact_Info
Acanyelles	010120210951	Exit	*NULL*
Ogarcia	010120210952	Click	Menu_About
Jmercade	010120210955	Exit	*NULL*
Ogarcia	010120210956	Exit	*NULL*

1) Dibuixa les estructures físiques que es crearien per emmagatzemar aquesta taula en una base de dades columnar emprant *Run-Length Encoding (RLE),* diccionaris i *Ending Row Indexing (ERI)*. Només cal dibuixar la solució per l'atribut *action,* però heu de dibuixar i identificar clarament les estructures finals (si dibuixes cap estructura temporal, distingeix-la de la resta).

	Action
0	NULL
1	Click
2	Free_Text
3	Exit
4	Back_Button

Action
1
2
1
1
1
3
1
1
4
2
1
3
1
3
3

ERI	Dict position
0	1
1	2
4	1
5	3
7	1
8	4

9	2
10	1
11	3
12	1
14	3

Les temporals son la columna d'action i dictionary position i les no temporals ERI, dict.position, i dictionary.

2) Suposa que s'han creat totes les estructures físiques amb RLE, diccionari i ERI per tots els atributs de la taula. La següent afirmació, és certa?

"*No hi ha cap atribut del qual les estructures físiques creades a la base de dades columnar ocupin MÉS espai que el que ocuparia aquest atribut a nivell físic en una taula relacional tradicional*".

Nota: Per la base de dades relacional, pots assumir el *heap file* com a estructura física de referència.

Justifica la teva resposta:

Cert. Com que compta nombre de repeticions, mai podrà haver-n'hi més. Com a màxim igual si estan molt desordenades.

3) Suposa la següent consulta sobre la taula un cop creada i emmagatzemada en una base de dades columnar.

```
SELECT COUNT(*) FROM T WHERE Action = 'Click' AND Element = 'Menu_Products'
```

Considerant el pla d'accés òptim que puguis pensar, aquest pla d'accés ha de:

- Aplicar una operació per reconstruir les tuples originals? NO
- Desfer el *Run Length Encoding* per resoldre la consulta? NO

Justifica la teva resposta:

No és necessari reconstruïr donat que sabem en quines posicions tenim tots els Action="click". Per altra banda, tampoc hem de desfer (definint desfer com accedir a un lloc que no sigui ERI) el Run Length Encoding, donat que només accedint a l'ERI ja podem saber les posicions a les que Action="click".

Nom i Cognoms: Júlia Gasull i Navarro

PREGUNTA 2. [2p] ***(Temps de referència: 15')***

Descriu l'efecte que tindria habilitar l'execució de la funció *combine* en aquest *MapReduce job*. Considera que la funció *combine* té el mateix codi que la funció *reduce*.

```
public void map(LongWritable key, Text value) {
    String line = value.toString();
    StringTokenizer tokenizer = new StringTokenizer(line);
    while (tokenizer.hasMoreTokens()) {
        write(new Text(tokenizer.nextToken()), new IntWritable(1));
    }
}

public void reduce(Text key, Iterable<IntWritable> values) {
    int count = 0;
    for (IntWritable val : values) {
       count += 1;
    }
    write(key, new IntWritable(count));
}
```

Justifica la teva resposta en termes del marc d'execució *MapReduce* (és a dir, cal fer referència a les fases i impacte en les entrades / sortides d'aquestes, si n'hi ha cap). Empra l'espai restant en aquest full per respondre-hi:

Combine:
- Permet optimitzar (menys lectures de disc)
- Redueix trànsit de xarxa
- Important: commutativa i associativa --> sempre i quant sigui igual que el reducer i el reducer ho sigui
- Tot això sempre i quant: |input| / |output| >> #CPU

En aquest cas, el que fa el combine és sumar nombre de paraules. Si no hi hagués el combine, s'hauria de fer: 1+1+1+1+1+…..+1+1+1+1+1. En canvi, amb el combine, es podria arribar a tenir: 2+4+2+1+…+5+1, per exemple.

Nom i Cognoms: Júlia Gasull

PREGUNTA 3. [4p] ***(Temps de referència: 60')***

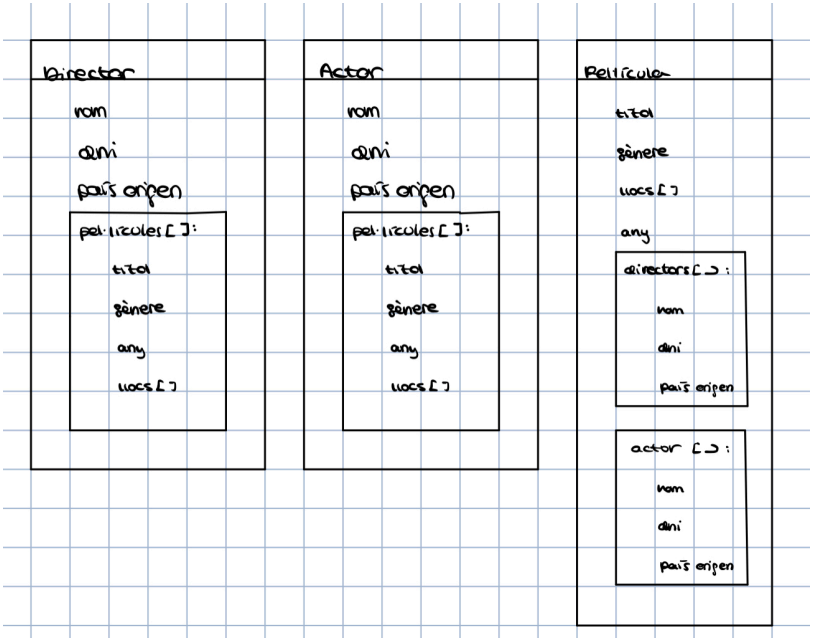
Volem crear una base de dades sobre el domini pel·lícules. Una pel·lícula està dirigida per un o més directors. Un actor pot actuar en una o més pel·lícules. Una pel·lícula es pot filmar en un o més llocs. A més del títol, els directors, els actors, any i localitzacions on es va rodar, s'espera que en el futur altra informació sigui necessària i calgui afegir-la.

El conjunt actual de consultes previstes per ara és el següent:

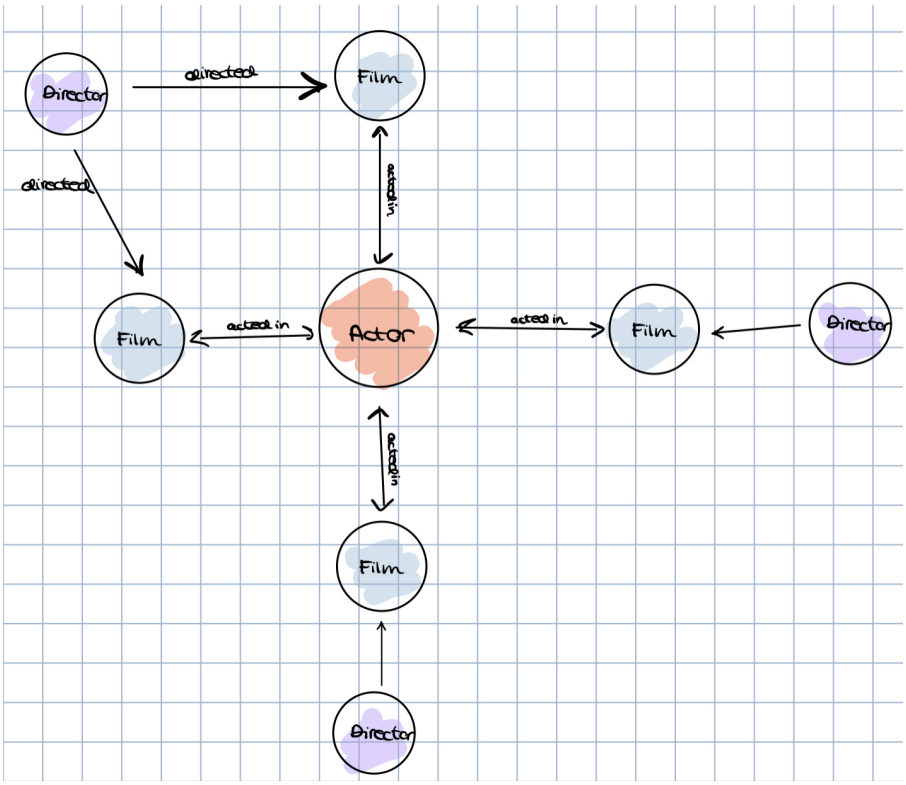
- Retorna els títols de les pel·lícules de terror filmades entre el 2010 i el 2020.
- Retorna el nom de tots els directors que s'han dirigit a sí mateix en una pel·lícula.
- Retorna el títol i l'any de les pel·lícules d'EEUU dirigides per directors estrangers i interpretades per almenys un actor estranger.
- Per cada actor, retorna el nombre de pel·lícules on hagi participat que estiguin filmades íntegrament a Catalunya.

Tenint en compte tota aquesta informació, i la flexibilitat d'esquema requerida:

1) Dibuixa un model de dades document per emmagatzemar aquestes dades a *MongoDB*. Identifica clarament la clau i elements del document. Empra l'espai restant a aquesta pàgina.



2) Dibuixa un model de dades graf per emmagatzemar aquestes dades a Neo4J. Representeu-lo en format graf (nodes i arestes) i identifica clarament els seus elements. Empreu l'espai habilitat entre aquest i el següent punt.



3) Tenint en compte la redundància de dades i el seu impacte en insercions / actualitzacions / esborrats, així com el rendiment de les consultes proposades, justifiqueu quin dels dos models (i per tant, quina base de dades) triaríeu sabent que aquesta base de dades té com objectiu emmagatzemar el màxim nombre de pel·lícules possibles (idealment, totes mai rodades) . Utilitzeu l'espai restant d'aquesta pàgina.

Jo triaria la de MongoDB, donat que tenim un document només per pel·lícula. Els altres documents es poden actualitzar a partir d'aquesta.