

Ús de bases de dades NoSQL

PRA1

Exercici 1: Cassandra (20%)

Es vol dissenyar una base de dades en Cassandra d'emmagatzematge dels vídeos dels usuaris així com els comentaris i les puntuacions sobre els vídeos. La informació requerida és la següent:

- Usuaris: identificador d'usuari, correu electrònic, nom, cognoms i data de registre. Els usuaris s'identifiquen per el seu identificador.
- Vídeos: identificador del vídeo, usuari, títol, descripció, tipus de vídeo, URL, data de pujada, puntuació mitjana, etiquetes i llista de gèneres.
- Comentaris: vídeo al qual es fa el comentari, usuari que fa el comentari, data a la qual es va registrar i text del comentari.
- Valoracions: vídeo puntuat i puntuació (de 0 a 10).

Sobre aquesta base de dades, es realitzen les següents consultes freqüents:

- Vídeos per gèneres
- Comentaris per usuari ordenats per data de registre de més recent a més antics.
- Comentaris per vídeo ordenats per data de registre de més recent a més antics.

Es demana:

1. Disseny físic de la base de dades en CQL, explicant raonadament el motiu pel qual es dissenya de la forma proposada.
2. Inserir alguns registres a la base de dades i realitzar les següents consultes, utilitzant CQL, sobre la base de dades dissenyada:
 - a. Trobar un usuari per email especificat
 - b. Trobar l'últim vídeo pujat.
 - c. Trobar els comentaris sobre un vídeo concret començant pels més recents
 - d. Trobar els vídeos d'un determinat gènere.

Exercici 2: MongoDB (30 %)

Considera el document "Implementació base de dades" que descriu com s'ha implementat en MongoDB la base de dades descrita en el document "Disseny d'una

base de dades per a una app de missatgeria instantània" que es troba en els materials del curs. També es necessitarà la màquina virtual "Màquina virtual MongoDB" que conté una instal·lació de MongoDB que té carregada la base de dades descrita anteriorment.

Es demana proporcionar les sentències per al shell de MongoDB i els resultats que s'obtenen per a les següents operacions:

1. Trobar tots els usuaris que hagin bloquejat a usuaris que procedeixen de Pamplona i tenen entre 20 i 40 anys. Mostra el nom i cognoms. (Pista: 2 registres)
2. Trobar quants usuaris són membres alhora dels grups: "Intercanvi de videojocs" i "Jugadors d'escacs"
3. Trobar tots els grups que tinguin membres de Valladolid o que no en tinguin cap de Madrid. Mostra el nom de grup i identificador del grup.
4. Crear la col·lecció "ciutats_usuaris" que contingui per cada ciutat un llistat d'usuaris amb les dades: nom, cognoms, email i nombre de grups als quals pertany. (Pista: 27 registres)
5. Donada la consulta següent:

```
db.Usuarios_desbloqueados.find ({
    'Desbloques.Usuario_desbloqueador.Ciudad': 'Madrid',
    'Desbloques.Usuario_desbloqueador.Edad': {'$gt': 40}},
    {_id: 0,
     "Usuario_desbloqueado.Nombre": 1,
     "Usuario_desbloqueado.Apellidos": 1,
     "Usuario_desbloqueado.Edad": 1, })
.sort ({'Usuario_desbloqueado.Edad': 1} );
```

Indicar quins índexs crearies per millorar el rendiment de la consulta.

(Pista: <https://docs.mongodb.com/manual/reference/method/cursor.explain/>)

Exercici 3: Neo4j (30%)

Considera el document que es troba en els materials del curs "Disseny d'una base de dades per analitzar l'activitat d'usuaris a Twitter" que descriu la implementació d'una base de dades en Neo4j. També es necessitarà la màquina virtual "Màquina virtual Neo4j" que conté una instal·lació de Neo4j que té carregada la base de dades descrita anteriorment.

Es demana proporcionar les següents consultes en Cypher i els resultats que s'obtenen per a les següents operacions:

1. Considerar l'usuari rellevant amb un nombre de seguidors, comptabilitzant els seguidors de primer i segon nivell de forma conjunta, és de 1436. Mostra el nom i el nombre de seguidors.
2. Considerar el tweet que ha tingut més retweets i respostes. Mostra la suma total de retweets i respostes, i els 20 primers caràcters del tweet.
3. Obtenir el percentatge de tweets que estan geolocalitzats a la latitud 42.0355
4. Considerar el sisè usuari que ha enviat més tweets geolocalitzats. Mostra el seu nom i el nombre de tweets.
5. Considerar els tweets geolocalitzats que són rèpliques a altres tweets geolocalitzats. Mostra el nombre de tweets que són rèpliques on la diferència en valor absolut entre les latituds que caracteritzen la rèplica i al tweet replicat és més gran que 20.

Exercici 4: Neo4j (20%)

Es vol organitzar la guia de carrers d'Espanya a nivell de ciutats. Per a això cal tenir en compte que els noms dels carrers es poden repetir entre les diferents unitats de l'estructura territorial excepte a nivell d'un mateix municipi. En un mateix municipi no poden existir dos carrers diferents amb el mateix nom. Si es realitza una simplificació, l'estructura territorial seria la següent: Comunitats Autònomes, Províncies, i municipis. S'ha decidit utilitzar una base de dades en graf per emmagatzemar tota la informació.

Es demana:

- a) Argumentar quina és la forma més eficient d'organitzar la informació en una base de dades orientada a grafs si es volen recuperar els noms dels carrers.
- b) A partir de les següents dades crea, usant Cypher en Neo4J, una base de dades d'acord amb la solució plantejada en a). Has de mostrar les consultes en Cypher així com captures dels grafs creats.

Carrer: Ramón y Cajal

Comunidad	Provincia	Municipio
Madrid	Madrid	Torrejón de Ardoz
Castilla y León	Segovia	Segovia
Andalucía	Granada	Albolote
Valencia	Castellón de la Plana	Onda
Cataluña	Barcelona	Tarrasa

Carrer: Cervantes

Comunidad	Provincia	Municipio
Madrid	Madrid	Alcalá de Henares

Castilla y León	Salamanca	Peñaranda de Bracamonte
Andalucía	Granada	Armillá
Valencia	Alicante	Villajoyosa
Cataluña	Barcelona	Badalona

Criteris de valoració

Els exercicis primer i quart tenen un pes del 20%; el segon i tercer tenen un pes del 30%. Es valorarà, en cada exercici, la validesa de la solució i la claredat de l'argumentació.

Format i data de lliurament

Heu d'enviar la PRA1 a la bústia de Lliurament i registre d'AC disponible a l'aula (apartat Avaluació). El format de l'arxiu que conté la vostra solució pot ser .pdf, .odt, .doc i .docx. Es recomana el format .pdf. Per a altres opcions, si us plau, contacteu prèviament amb el vostre professor col·laborador. El nom del fitxer ha de contenir el codi de l'assignatura, el vostre cognom i el vostre nom, així com el nom de l'activitat (PRA1). Per exemple nomcognom1_nosql_pra1.pdf

La data límit per lliurar la PRA1 és el **9 de juny**.

Propietat intel·lectual

En presentar una pràctica o PAC que faci ús de recursos aliens, s'ha de presentar juntament amb ella un document en què es detallin tots ells, especificant el nom de cada recurs, el seu autor, el lloc on es va obtenir i el seu estatus legal: si l'obra està protegida pel copyright o s'acull a alguna altra llicència d'ús (Creative Commons, llicència GNU, GPL etc.). L'estudiant haurà de assegurar que la llicència que sigui no impedeix específicament seu ús en el marc de la pràctica o PAC. En cas de no trobar la informació corresponent haurà d'assumir que l'obra està protegida pel copyright. Caldrà, a més, adjuntar els fitxers originals quan les obres utilitzades siguin digitals, i el seu codi font, si així correspon.