EXAMEN DE CBDE

22 de Juny del 2018

Instruccions: Respon cada pregunta al full corresponent.

L'examen dura 1 hora.

Nom i Cognoms: .....

Pregunta 1. [4p]

Considera una base de dades *column-oriented* que empra ***Run-Length Encoding*** **amb diccionari** com a mecanisme de compressió per l'emmagatzemament de les columnes. Donada la columna (és a dir, atribut) *activity* d'una taula qualsevol, representa com s'emmagatzemaria aquesta columna en aquesta base de dades. No oblidis detallar totes les estructures internes necessàries, inclòs el diccionari resultant.

Column Activity
Karate
Karate
Running
null
null
Running
Running
Tai-txi
Tai-txi
Tai-txi

En què consisteix la tècnica *block iteration* que apliquen les bases de dades *column-oriented*? Exemplifica la teva explicació amb el resultat de la pregunta anterior.

……
……
……
……
……

Quines altres condicions s'han de complir per a poder dir que una base de dades *column-oriented* aplica *vectorization*?

……

PREGUNTA 2. [2p]

1 Justifica una avantatja i un inconvenient de l'arquitectura d'HBase, basada en un B+ distribuït, i la de MongoDB MMAPv1, basada en Consistent Hashing.

.....

.....

.....

.....

2 Aplica HBase fragmentació horitzontal i/o vertical? Justifica la teva resposta.

.....

.....

.....

.....

Nom i Cognoms: .....

PREGUNTA 3. [4p] Considera els dos sistemes que es mostren a continuació:

Sistema 1: Base de dades centralitzada

- Consta d'un únic node
- La base de dades ocupa 12TB. Per simplicitat, assumeix que conté una única taula T amb totes les dades i n'hi ha 100.000.000 de tuples
- La latència de llegir de disc és de 5ms
- L'ample de banda màxim que pot aconseguir el disc és de 100Mb/s

Sistema 2: Base de dades distribuïda

- Consta de 3 nodes connectats per una LAN. Qualsevol d'ells pot llençar queries
- La base de dades ocupa 12Tb. Pots assumir que la taula T s'ha distribuït de forma uniforme entre els nodes mitjançant una fragmentació horitzontal. No hi ha replicació
- La latència de llegir de disc és de 5ms i la de la xarxa 1 ms
- L'ample de banda màxim que pot aconseguir el disc és de 100Mb/s
- Els nodes estan connectats a través d'una LAN amb ample de banda màxim de 10Mb/s

Suposa que l'única *query* és: `SELECT SUM(a) FROM T, on a és un atribut de T`

No hi ha índexs ni cap altra estructura definida en el sistema.

En el millor cas, quant trigarà (en segons) en fer un accés seqüencial de T en el sistema 1?

Latència .....

Lectura seqüencial .....

Total .....

I en el sistema 2 en el millor cas?

Latència .....

Lectura seqüencial .....

Total .....

Suposa ara que la única query del sistema fos:

`SELECT * FROM T WHERE pk = 1, on pk és la clau primària de T i té un B+ associat (pots suposar que el B+ està a memòria)`

Que canviaria respecte a la pregunta anterior?

.....
.....
.....
.....