EXAMEN DE CBDE

17 de Gener del 2017

Instruccions: *Respon cada pregunta al full corresponent. L'examen dura 1h*

Nom i Cognoms: .....

Pregunta 1. [5p]

Has decidit emprar HBase per emmagatzemar les dades de la teva organització. Els PUTs que has escrit són (el format és KEY, FAMILY, ATTRIBUTE, VALUE):

```
PUT 1, Personal, Age, "45"
PUT 1, Personal, Street, "Villarroel"
PUT 1, Personal, Name, "Joan"
PUT 1, Financial, Salary, "30000"
PUT 1, Financial, Currency, "Euro"
PUT 2, Personal, Name, "Anna"
PUT 2, Financial, Salary, "40000"
PUT 2, Financial, Currency, "Euro"
PUT 3, Personal, Name, "Josep"
PUT 3, Personal, Street, "Numancia"
```

Assumeix que el teu clúster HBase té dos region servers (#1 i #2) i el B+ conté les següents entrades ({KEY, REGION SERVER}): {1, #1}, {2, #2}, {3, #1}

1.1. Dibuixa com quedaria els *storefiles* (és a dir, fitxers a disc) dels dos *region servers* després de fer aquests inputs i tenint en compte la distribució de les dades que es faria tenint en compte el B+. Cal indicar, clarament, el nombre de fitxers físics generats i el contingut de cadascun d'ells per a cada un dels regions servers.

1.2. Ara, es fa el següent PUT: PUT 1, Personal, Age, "46". Que canvia a nivell físic? Indica clarament com es guardaria a nivell físic aquest PUT i els canvis en els fitxers ja existents (indica els canvis modificant el dibuix fet a l'exercici anterior).

1.3. Suposa que la única consulta del sistema és la següent:

Q1: "*Per cada persona, retorna la seva edat i el seu salari*"

El disseny que has fet, és correcte? Justifica la teva resposta.

……
……
……
……
……

Pregunta 2. [2p]

1 Identifica les dues estructures vistes a classe per implementar el catàleg global d'una base de dades distribuïda.

……
……
……
……
……

2 Dona un exemple de gestor que implementi cada una d'aquestes estructures.

……
……
……
……
……

3 Durant el curs hem fet pràctiques amb *HBase / MapReduce, MongoDB / Aggregation Framework* i *Neo4J / Cypher*. Ordena de major a menor aquestes parelles *gestor / llenguatge de consultes* segons el grau de paral·lelisme que poden assolir en el millor cas. **Justifica la teva resposta**.

……
……
……
……
……
……
……

Pregunta 3. [3p]

Considera la següent estructura de document que es proposa per modelar l'esquema del TPC-H (pots refrescar-lo a la part de darrere del full) en una base de dades *document-oriented*. **Justifica** (en termes d'espai ocupat, temps d'execució de les consultes i modificació de les dades), fent el **supòsits quantitatius** que calgui, la tria de model (opció A o B), **donada la consulta Query1**:

Query 1:

```
SELECT l_orderkey, sum(l_extendedprice*(1-l_discount)) as revenue,
o_orderdate, o_shippriority
FROM customer, orders, lineitem
WHERE c_mktsegment = '[SEGMENT]' AND c_custkey = o_custkey
AND l_orderkey = o_orderkey AND o_orderdate < '[DATE1]' AND
l_shipdate > '[DATE2]'
GROUP BY l_orderkey, o_orderdate, o_shippriority;
```

Opció A: Cada document representa un *lineitem*. La key del document és doncs l'identificador de *lineitem*. A la resta del document es guarda, a nivell d'arrel del JSON, la informació de l'*order* i *customer* necessària per la consulta.

Opció B: Cada document representa un *order*. Així, la key del document és l'identificador d'*order*. A la resta del document es guarda la informació necessària de *customer* a nivell arrel del JSON, mentre que els *lineitems* d'aquesta *order* es guarden en un *embedded* JSON amb key "*lineitems*".

Per cada aspecte, de forma independent, justifica quina és la millor opció:

Criteri 1. Espai ocupat:

……
……
……
……
……

Criteri 2. Temps d'execució de les consultes:

……
……
……
……
……

Criteri 3. Modificació de les dades:

……
……
……
……
……

Un cop analitzat els tres aspectes, quina és la teva **elecció final**? **Justifica** la resposta
……
……
……
……

Figure 2: The TPC-H Schema