CBDE - Laboratori 5

GRUP: Oriol Muñoz i Albert Suàrez

Document explicatiu

Introducció

En aquest laboratori de CBDE se'ns presenta un nou tipus de bases de dades no relacionals (NoSQL) anomenat *document stores*. L'objectiu d'aquesta pràctica és entendre aquest tipus i saber dissenyar els documents de la manera més optima possible.

A l'enunciat d'aquesta pràctica tenim quatre *queries* que accedeixen a un seguit de taules donades, les quals hem de decidir com estructurar els documents per realitzar accessos de manera òptima.

Després d'haver analitzat una per una les quatre consultes donades, hem arribat a la conclusió que la millor solució és emmagatzemar cada **LineItem** com un document separat. Llavors, d'ara cap endavant, explicarem la justificació analitzant consulta per consulta, finalment extraient la conclusió final.

A més a més, s'afegirà a l'explicació quins índexs s'han utilitzat amb la seva respectiva justificació.

Query 1

En aquesta primera consulta disposem d'una *query* que selecciona molts elements, tots ells de l'entitat *LineItem*, agrupats i ordenats per dos atributs de la mateixa entitat, a més a més de filtrar comparant una data donada amb la data dels *lineitems*. Per tant, hem arribat a la conclusió que tota aquesta consulta només depèn de l'entitat *LineItem*.

En conseqüència, dissenyar en aquest cas els documents per *lineitems* aportaria grans avantatges en conceptes d'eficiència i optimització en l'execució de la consulta. Llavors doncs elegiríem l'entitat **LineItem** com a element base per al disseny d'aquesta *query*.

Hem decidit implementar un índex sobre l'atribut *shipdate* situat a l'entitat *lineitem* pel simple fet que aquest atribut és el que condiciona el *where* de la query. Per tant, si indexem per aquest atribut, aconseguirem millor performance en la consulta i millor eficiència i rendiment.

Query 2

Si parlem de la segona consulta donada, veiem com se seleccionen i es filtra per atributs de les entitats *PartSupp*, *Part*, *Supplier*, *Nation* i *Region*. Totes elles, que d'un bon principi hem pensat que no tenien res a veure, tenen un gran tret característic en comú. Si partim de la idea que tenim una *lineitem*, podem veure com des d'allà tenim només una *partsupplier*, la qual disposa d'una *part* i d'una *supplier*. Llavors des d'aquesta última entitat, podem veure que disposem només d'una *nation*, la qual a més a més disposa d'una *region*.

És a dir, dit d'una altra manera, per cada *lineitem* disposarem sempre d'una entitat de les comentades anteriorment. Per tant, partint des d'aquesta idea, hem vist que el disseny més òptim dels documents seria agafar com a entitat base l'entitat **LineItem**.

Hem decidit indexar la segona consulta per l'atribut *name* de l'entitat *region* corresponent al *supplier* per la simple raó que s'utilitza com a condició en el *where*. Per tant, obtindrem millors resultats en l'execució de la consulta.

Query 3

Si ens referim a la tercera consulta, ens adonem que agrupa per tres valors dels quals pertanyen a l'entitat d'*order*, però que tot i així, partint des de l'entitat *lineitem*, només pot tenir un pertinent valor dels tres mencionats, ja que un *lineitem* té només una *order*. A més a més, aquesta consulta s'ordena per dos elements. El primer atribut per al qual s'ordena, *revenue*, és una suma d'un conjunt d'atributs de l'entitat *lineitem*. Per altra banda, el segon d'ells, *orderdate*, tot i pertànyer a l'entitat *order*, per cada *lineitem* només tindrà una *orderdate* donada la seva cardinalitat. I com a últim punt, veure que també es consulten atributs de l'entitat *customer*, però que si partim de *lineitem*, tenim novament que només n'hi ha un.

En conclusió, hem vist que l'entitat base per dissenyar els documents d'aquesta consulta és el **LineItem**.

Per optimitzar encara més la query, podem crear un índex sobre l'atribut del *where* que sigui més selectiu. Com que hi ha 150.000 *customer*, 6.000.000 *linitem* i 1.500.000 *orders*, la millor opció és crear l'índex sobre l'atribut mktsegment de *customer*.

Query 4

Si per últim, si decidim analitzar la quarta consulta, veiem com accedim fins a sis entitats diferents: *customer*, *order*, *lineitem*, *supplier*, *nation* i *region*. Tot i així, si parem atenció als atributs seleccionats i per quins atributs s'agrupa i s'ordena, veiem com tots ells depenen o bé de l'entitat *order* o *lineitem*. És a dir, que partint d'alguna d'aquestes entitats qualsevol, la cardinalitat a les altres entitats sempre seria 1. En conseqüència, aporta gran valor en ser candidata com a entitat base.

Tot i així, hem acabat decidint que el més òptim per elegir com entitat base ha estat el **LineItem**, considerant que les altres tres consultes hem decidit *lineitem* com a entitat. Així doncs, les quatre queries coincidirien en la decisió.

En aquest cas, aquesta query es pot optimitzar creant un índex sobre l'atribut *name* de *region*, donat que apareix com a condició del *where* i que, en existir només 5 regions, és la condició més selectiva.

Conclusió

Finalment, després d'haver analitzat consulta per consulta, hem pogut veure com hem coincidit els

quatre cops en la decisió de quina seria l'entitat base mes optima. Així doncs, elegim **LineItem** com l'entitat per dissenyar la nostra base de dades no relacional de tipus *document stores*.

Considerem, donades totes les justificacions realitzades, que aquesta opció és la que aporta més valor i millor anirà en la performance de les consultes, ja que obtindrem més eficiència a l'hora de rebre els resultats de l'execució de les diferents queries donades.

Nota

Per legibilitat, el document creat *no* té materialitzats a l'arrel els camps rellevants per a les queries, si no que es mantenen en els seus respectius objectes interns. Som conscients que això afecta negativament al rendiment de les *queries*, donat que per cada accés a un atribut anidat (entenent per accés l'obtenció del valor en el JSON que es troba en memòria, no en disc), s'han de fer tants accessos com nivells estigui anidat l'atribut. No obstant això, donat que això no és un cas d'ús crític on aquest detall resulti particularment rellevant, hem preferit deixar-ho, per legibilitat, amb objectes anidats.