**Explanation document**

Rafael Guardeño | Jacobo Moral

**Query 1**

La primera consulta de todas es la más sencilla de todas, en el sentido que solamente obtiene datos de una entidad, *Lineitem.* Por esto mismo hemos considerado óptimo diseñar los documentos para esta query basándonos en *Lineitem*.

En cuanto a el id del operador $group del pipeline usamos l\_returnflag + l\_linestatus, ya que esta consulta los utiliza para hacer una agrupación, y, posteriormente, una ordenación.

Hemos decidido no compartir el mismo documento que para hacer las queries 3 y 4, ya que la lectura efectiva al realizar esta consulta no sería tan alta. Si bien, de esta manera usamos más espacio en disco y memoria que de la otra forma, podemos asimismo mejorar la eficiencia en la lectura, ya que ahorramos leer datos innecesarios de cada documento (aquellos atributos de las queries 3 y 4 que no corresponden a la 1), la cual cosa consideramos óptimo.

En caso que estuviéramos faltos de espacio y pudiéramos permitirnos perder algo de velocidad en esta query, podríamos optar por juntar la información de las tres consultas en un solo documento.

Otro caso en que podríamos juntar esta query a las otras, es si la frecuencia de ésta es muy inferior al resto.

Podemos comparar el número de bytes han de leerse tanto teniendo un documento propio para la query 1 como compartiéndolo con las queries 3 y 4:

Para el primer caso, donde tenemos un documento exclusivo, tenemos que ocupa 58 bytes cada uno. Como tenemos 6.000.000 line items, esto significa que tendremos que leer 6.000.000\*58bytes = 348 MB para cada consulta.

Para el segundo caso, el documento ocupa estos 58bytes + 72 bytes más de los otros atributos que no son necesarios para esta query, en total ocupa 130bytes. Esto significa que tendríamos que leer 780MB de datos para la consulta.

Con estos argumentos, hemos decidido guardar la consulta número 1 en documentos con la información justa para realizar la consulta.

**Query 2**

Para responder la query 2, creamos un documento por cada Partsupp, cada Partsupp tiene multiplicidad 1 con Part y con Supplier por lo que podemos poner los atributos de ambos como si fuesen un atributo de Partsupp, por la misma razón podemos poner los atributos de Nation y Region como atributos de Partsupp ya que Supplier tiene multiplicidad 1 con Nation y Nation tiene multiplicidad 1 con Region, en este documento incluimos los atributos necesarios para responder la query;

Al guardar las asociaciones implícitamente en el documento, no nos hace falta guardar las key que se usan en el where de la query en SQL para hacer los joins.

En cuanto a la query, la dividimos en dos agregaciones, una para hacer la subquery y obtener el PARTSUPP.supplycost menor para poder utilizar este supplycost como atributo en el “match” de la segunda agregación.

**Query 3**

Para la tercera consulta, creamos un documento para cada *Lineitem*. En este caso necesitamos acceder a la información de *Customer*, *Order* y *Lineitem.* Hemos elegido usar *Lineitem* porque de esta manera podemos acceder directamente a los atributos de las otras dos entidades, ya que la cardinalidad respecto *Lineitem* es 1.

**Query 4**

Para la cuarta y última consulta nos pedían algo muy similar a la tercera, y por tanto hemos hecho lo mismo.

Ya la query 3 y la query 4, ambas usan *Lineitem* como punto de partida para acceder al resto de datos, hemos decidido juntar la información que se requiere para ambas consultas en un documento, maximizando la eficiencia en espacio que se utiliza, ya que ambas consultas utilizan un número elevado de atributos, y, por tanto, de bytes. Por el ratio de lectura efectiva es suficiente para utilizar un documento para ambas queries.

**Estructura de la BD**

**Documento de la colección para la query 1**

|  |
| --- |
| **{  \_id: l\_orderkey + l\_linenumber  l\_linestatus  l\_returnflag  l\_quantity  l\_extendedprice  l\_discount  l\_tax  l\_shipdate }** |

**Documento de la colección para la query 2**

|  |
| --- |
| **{  \_id: p\_partkey + s\_suppkey  ps\_supplycost  p\_mfgr  p\_size  p\_type  s\_name  s\_address  s\_phone  s\_acctbal  s\_comment  n\_name  r\_name }** |

**Documento de la colección para la query 3 y la query 4**

|  |
| --- |
| **{  \_id: l\_orderkey + l\_linenumber  l\_returnflag  l\_linestatus  l\_quantity  l\_extendedprice  l\_discount  l\_tax  l\_shipdate  n\_name  r\_name  o\_orderdate  o\_shippriority  c\_mktsegment }** |