

Explicació de l'Estructura dels Models de Dades

En aquest document explicarem les decisions que hem pres per tal de fer que l'estructura dels models de dades sigui òptima segons les 4 queries donades. Hem dividit l'estructura en dues col·leccions: "Orders" (dissenyada per optimitzar les queries 1, 3 i 4) i "Partsupp" (dissenyada per optimitzar la query 2).

ORDERS

```
{
  "o_orderkey": 123,
  "o_orderdate": 2023-11-29,
  "o_shippriority": "Primary",
  "c_mktsegment": "Cars",
  "nation": {
    "n_name": "Spain",
    "n_regionkey": 1,
    "r_name": "Europe",
    "supplier": [
      {
        "s_suppkey": 1093
      }
    ]
  },
  "lineitems": [
    {
      "l_suppkey": "Nissan",
      "l_extendprice": 1000,
      "l_discount": 50,
      "l_shipdate": 2023-11-14,
      "l_returnflag": 'Y',
      "l_linestatus": 'S',
      "l_quantity": 100,
      "l_tax": 133
    }
  ]
}
```

(En aquest cas, sabem que posar *nation* com a embedded empitjorarà l'eficiència, però millora molt la comprensió del codi)

En primer lloc, vam començar a fixar-nos en la query 3, que accedeix a *orders*, *customer* i *lineitems*. Aquí, ens vam trobar amb el nostre primer dilema a l'hora de tractar la relació entre aquestes tres classes:

Si fem que *lineitems* englobi les altres dues, s'hauria de cercar la informació en molts documents diferents, ja que per buscar les dades d'una *order* hauríem de cercar en varis *lineitems* i, tenint en compte que *lineitems* té la multiplicitat més alta de totes, fa que aquesta no sigui l'opció ideal.

Si fem que *customer* englobi les altres dues, tampoc seria una bona idea. Al cap i a la fi, el que més es crea en aquesta database són *orders*, llavors estant englobades dins de *customer*, es podrien generar problemes de padding, ja que el document podria créixer sovint i MongoDB no deixaria l'espai reservat suficient.

Per tant, creiem que la millor idea és que *orders* englobi les altres dues classes, evitant aquests problemes i, a més, fent que l'agrupació de la query sigui òptima, ja que tots els paràmetres del GROUP BY estan dins d'*orders*. A més, hem posat un índex a *c_mktsegment* per tal de millorar l'eficiència de la cerca a l'hora de fer el WHERE de la query.

En segon lloc, vam fixar-nos en la query 1, que només vol accedir a informació de *lineitems*, cosa que fa senzilla la seva integració amb l'esquema explicat anteriorment. A més, hem posat un índex a *l_shipdate* per a millorar l'eficiència del WHERE de la query.

Finalment, en tercer lloc, vam fixar-nos en la query 4, per a nosaltres la més difícil d'integrar. En aquesta query es vol accedir a informació de *customer*, *orders*, *lineitems*, *supplier*, *nation* i *region*; per tant, havíem d'integrar les tres últimes amb l'esquema anterior.

Per a fer-ho, vam decidir agrupar *supplier* i *region* dins de *nation*, ja que aquesta era el nexa entre les classes i, així, es facilitava l'agrupament de la query (el GROUP BY és per *n_name*).

A més, hem posat un array amb tots els *suppliers* de cada *nation*, per tal de tenir tots els *suppliers* agrupats per *nations*, cosa que fa que el join amb *customer*, *nation* i *supplier* sigui més ràpid, eficient i més fàcil. També hem posat indexos a *n_regionkey* i a *r_name* per tal de millorar la rapidesa del WHERE de la query.

PARTSUPP

```
{
  "ps_partkey": 20201102112,
  "ps_suppkey": 1093,
  "ps_supplycost": 15000,
  "s_acctbal": 100,
  "s_name": "Nissan",
  "s_address": "17-1 Ginza 6-Chome Chuo-Ku Tokyo, Japan",
  "s_phone": "0120-315-232",
  "s_comment": "Somos un grupo empresarial de tradición y líder del mercado
que aporta bienestar económico y social a todos nuestros clientes,
empleados y proveedores.",
  "n_name": "Japan",
  "r_name": "Asia",
  "p_mfgr": "MICHELIN PRIMACY 4 195/65 R15 91 V",
  "p_size": 125,
  "p_type": "Tyres",
}
```

En aquest cas, vam decidir fer una altra col·lecció dissenyada per optimitzar la query 2, ja que trobàvem que aquesta era molt disjunta a les de més i, per tant, no hi havia sentit intentar ajuntar-ho tot en una.

En la query 2 s'accedeix a les classes *part*, *supplier*, *partsupp*, *nation* i *region*. Vam decidir englobar les altres classes a *partsupp*, ja que des de *partsupp* pots accedir a les altres classes (a *part* i a *supplier* mitjançant *ps_partKey* i *ps_suppkey*, a *nation* mitjançant *s_nationkey* i a *region* mitjançant *n_regionkey*). Així, tot el document pot ser un document pla amb els atributs necessaris, millorant l'eficiència. Finalment, hem posat un índex en *r_name* per tal que la condició del WHERE de la query sigui més ràpida.