Data: 18/01/2023. **Nom i Cognoms:** .....

EXAMEN DE CBDE

Instruccions: Respon cada pregunta i subapartat a l'espai habilitat.

PREGUNTA 1. [3p] ***(Temps de referència: 30')***

Considera l'esquema del benchmark TPC-H, i les tres consultes de l'Annex. Discuteix els **avantatges i inconvenients** de **cada un** dels següents models de documents alternatius per a MongoDB, tenint en compte que cal optimitzar el rendiment de les tres consultes alhora. Per a la discussió, tingues en compte aspectes com l'eficiència de les consultes, de les insercions/modificacions, la mantenibilitat, l'espai ocupat, ...

Model 1: Col·lecció LineItems

```
{ "id": 123456,
  "l_orderkey": 654321,
  "l_quantity": 3,
  "l_extendedprice": 100,
  "l_discount": 15,
  "l_returnflag": "a",
  "l_linestatus": "b",
  "l_shipdate": "2023-01-18",
  "nation_supplier": "Italy",
  "region_supplier": "Europe",
  "o_orderdate": "2022-12-28",
  "o_shippriority": 1,
  "customer: {
      "o_custkey": 123,
      "c_mktsegment": "young male",
      "nation": "Italy"}
}
```

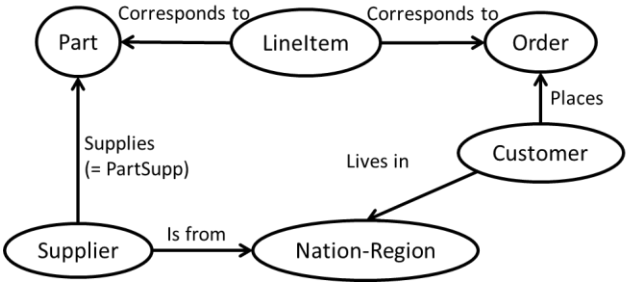
Model 2: Col·lecció Orders

```
{  "id": 342672,
   "o_orderkey": 654321,
   "o_orderdate": "2022-12-28",
   "o_shippriority": 1,
   "lineItems": [{
      "l_orderkey": 654321,
      "l_quantity": 3,
      "l_extendedprice": 100,
      "l_discount": 15,
      "l_returnflag": "a",
      "l_linestatus": "b",
      "l_shipdate": "2023-01-18",
      "nation_supplier": "Italy",
      "region_supplier": "Europe"
   }],
   "c_mktsegment": "young male"
}
```

Data: 18/01/2023. **Nom i Cognoms:** .....

PREGUNTA 2. [3p] ***(Temps de referència: 30')***

Considera de nou el benchmark TPC-H de l'Annex. El següent graf mostra una possible representació d'aquestes dades per a Neo4j, utilitzant una notació simplificada que mostra els tipus dels nodes i les arestes, en lloc de les instàncies:



NOTA: Els atributs de la taula PartSupp es troben a l'aresta "Supplies". Els atributs de la taula Region estan inclosos al node "Nation-Region". La resta d'atributs es troben al node corresponent a cada taula.

Respon les següents preguntes:

a) Digues si aquest graf **permet representar o no** la mateixa informació que el model relacional original, assumint que tant els nodes com les arestes tenen tots els atributs de les taules corresponents. Justifica la resposta.

b) **Proposa un graf** que, a més de representar la mateixa informació que el model relacional original, permeti optimitzar les tres consultes. Raona la resposta en termes de rendiment de les consultes, de les insercions/modificacions, mantenibilitat, espai ocupat, ...

Graf **Justificació**

Data: 18/01/2023. **Nom i Cognoms:** .....

PREGUNTA 3. [4p] ***(Temps de referència: 30')***

Considera la següent taula relacional:

Customer_id	Name	Surname	LastConnection	Registration	EndSubscription
1	Emma	Miralbau	Barcelona	GOLD	12/02/2023
2	Anna	Castellat	Lleida	GOLD	31/12/2023
...	...	...	...	...	...

Aquesta taula emmagatzema els usuaris d'un servei per subscripció. Donat el seu volum (milions de clients) es vol migrar a HBase. A més, suposa la següent consulta que aquest sistema ha de respondre (versió SQL):

```
SELECT Registration, Endsubscription, COUNT(*)
FROM TAULA WHERE LastConnection = [CITY]
GROUP BY Registration, Endsubscription
```

Respon a les següents preguntes:

1) Proposa una clau per a la taula HBase on migrar aquesta taula. Justifica la teva resposta.

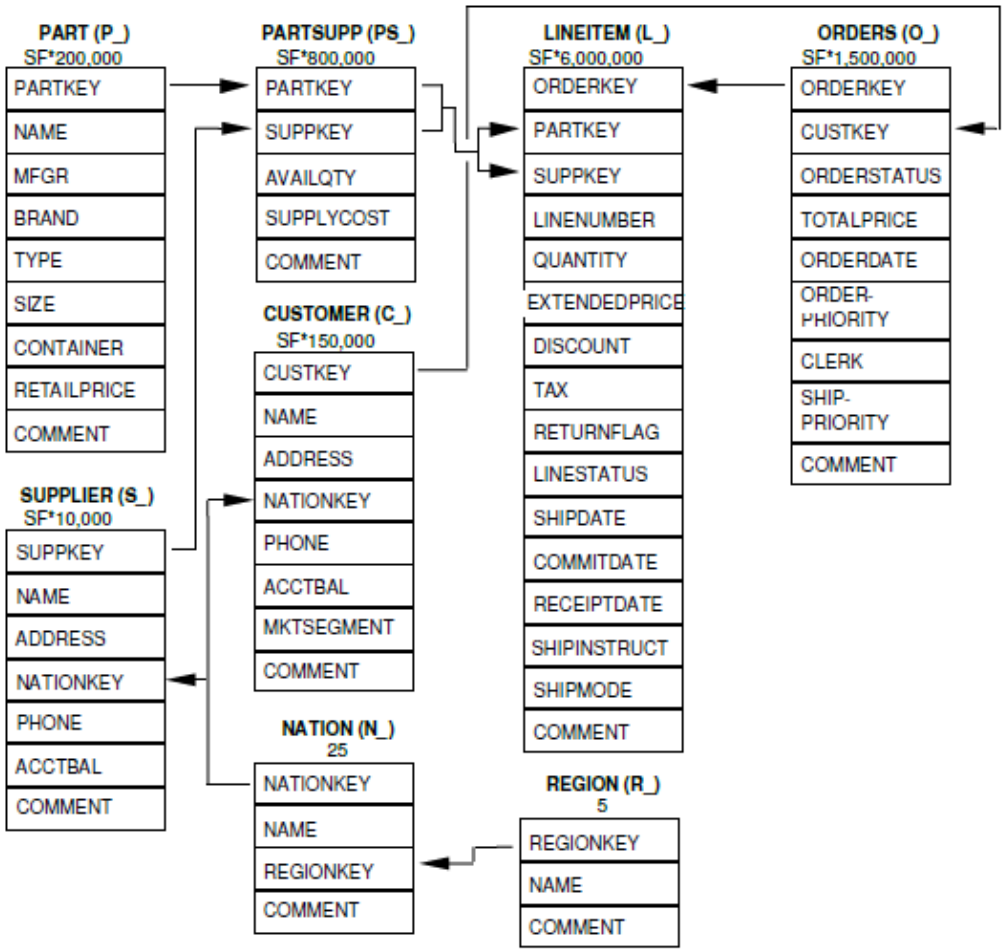
2) Considerant la clau proposada, proposa ara un disseny de la resta de la taula. Aquest disseny ha d'especificar les famílies proposades i els atributs a cada família. Pots assumir que no hi ha cap altra consulta al sistema (ni n'haurà cap altra). Justifica la teva resposta.

Data: 18/01/2023. **Nom i Cognoms:** .....

3) Suposa un sistema amb 5 servidors (RegionServers) i que les dues instàncies que es mostren a la taula van a parar al mateix RegionServer. Tenint en compte la clau i el disseny proposat, dibuixa els fitxers que es crearien en aquest RegionServer assumint que només hi ha aquestes dues instàncies. Els fitxers de la teva resposta han de contenir tots els elements físics d'un HFile (o fitxer HBase físic).

Data: 18/01/2023. **Nom i Cognoms:** .....

ANNEX: TPC-H Benchmark



Legend:

- The parentheses following each table name contain the prefix of the column names for that table;
- The arrows point in the direction of the one-to-many relationships between tables;
- The number/formula below each table name represents the cardinality (number of rows) of the table. Some are factored by SF, the Scale Factor, to obtain the chosen database size. The cardinality for the LINEITEM table is approximate (see Clause 4.2.5).

Claus Primàries

- **PART:** P_PARTKEY
- **SUPPLIER:** S_SUPPKEY
- **PARTSUPP:** PS_PARTKEY + PS_SUPPKEY
- **CUSTOMER:** C_CUSTKEY
- **NATION:** N_NATIONKEY
- **REGION:** R_REGIONKEY
- **LINEITEM:** L_ORDERKEY+L_PARTKEY+L_SUPPKEY
- **ORDERS:** O_ORDERKEY

Data: 18/01/2023. **Nom i Cognoms:** .....

Consultes

Query 1

```
SELECT l_returnflag, l_linestatus, sum(l_quantity) as sum_qty,
 sum(l_extendedprice) as sum_base_price,
 sum(l_extendedprice*(1-l_discount)) as sum_disc_price,
 sum(l_extendedprice*(1-l_discount)*(1+l_tax)) as sum_charge, avg(l_quantity) as avg_qty,
 avg(l_extendedprice) as avg_price, avg(l_discount) as avg_disc, count(*) as count_order
FROM lineitem
WHERE l_shipdate <= '[date]'
GROUP BY l_returnflag, l_linestatus
ORDER BY l_returnflag, l_linestatus;
```

Where [date] is a constant that may vary between executions of the query.

Query 3

```
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
ORDER BY revenue desc, o_orderdate;
```

Where [segment], [date1] and [date2] are constants that may vary between executions of the query.

Query 4

```
SELECT n_name, sum(l_extendedprice * (1 - l_discount)) as revenue
FROM customer, orders, lineitem, supplier, nation, region
WHERE c_custkey = o_custkey AND l_orderkey = o_orderkey
 AND l_suppkey = s_suppkey AND c_nationkey = s_nationkey
 AND s_nationkey = n_nationkey AND n_regionkey = r_regionkey
 AND r_name = '[REGION]' AND o_orderdate >= date '[DATE]'
 AND o_orderdate < date '[DATE]' + interval '1' year
GROUP BY n_name
ORDER BY revenue desc;
```

Where [date] and [region] are constants that may vary between executions of the query.