Exercice 4 : Optimisation de requêtes réparties 7 pts

On considère une BD répartie sur 10 sites S_1 à S_{10}, et dont le schéma global est :

Article (num, marque, taille, cat, desc)
cat est la catégorie de l'article : valeur entière de 1 à 100.
desc est la description de l'article.

Client (id, ville, dep, zone, fid, profil)
zone est la zone géographique. Il y a 5 zones distinctes numérotées de 1 à 5.
fid est la fidélité du client valant de 0 à 4 étoiles.

Loue (num, id, date, prix, commentaire)
num fait référence au numéro d'article, *id* fait référence à l'identifiant de client.
date est la date exprimée en nombre de jours depuis le 01/01/2010 (entier positif).
Un nuplet représente la location d'un article par un client.

On suppose que les attributs sont tous indépendants des autres, et qu'ils ont des valeurs uniformément distribuées sur leurs domaines respectifs. La taille d'un attribut numérique (par exemple id ou num) est de 100 octets. Les attributs desc, profil et commentaire sont de grande taille (plusieurs dizaines de Ko).

Les données sont fragmentées sur les sites S_i (ou S_j) ainsi :

Pour $1 \leq i \leq 10$ **Article$_i$** = σ_{ci} (Article) avec le prédicat ci défini par $10*(i-1) < cat \leq 10*i$

Loue$_i$ = Loue $\ltimes_{num}$ Article$_i$

Pour $1 \leq j \leq 5$ **Client$_j$** = $\sigma_{zone = j}$ (Client) rmq : les sites S6 à S10 n'ont pas de fragment de Client.

On donne les cardinalités suivantes

Card(Article$_i$) =10 000 Card(Loue$_i$) = 100 000 Card(Client$_j$) = 20 000 Card(π_{id} (Loue$_i$)) = 50 000

On donne la taille des fragments en Go (pour simplifier les calculs, on pose 1Go = 10^3 Mo = 10^6 Ko = 10^9 octets).

Taille(Article$_i$) = 3 Go Taille(Loue$_i$) = 20 Go Taille(Client$_j$) = 10 Go

Toutes les requêtes sont posées sur le site S_0. Le coût d'une requête est la somme des tailles des données transférées. On ne tient pas compte du coût de calcul des opérations algébriques.

Question 1

1.1) Quelles sont les cardinalités des relations du schéma global

```
card(Article) = ……
card(Loue) = ……
card(Client) = …….
```

1.2) Combien de clients distincts sont référencés dans un fragment Loue$_i$? En moyenne, combien y a-t-il de locations par client dans un site S_i ? Dans combien de sites différents se trouvent les locations d'un client donné?

```
Nombre de clients distincts dans Loue_i : ..............

Nombre de locations par client dans un site S_i : ..........

Nombre de sites pour les locations d'un client : ..............
```

Question 2

Soit la requête R1 : afficher tous les attributs des clients ayant loué au moins un article avant le 5 janvier 2010 (date $\leq$ 5). R1 ne contient aucun double. On suppose que R1 renvoie 1% des clients, et que dans chaque fragment $Client_j$ ou $Loue_i$, 1% des clients référencés satisfont R1.

2.1) Quelle est l'expression algébrique de la requête R1 exprimée sur le schéma global ?

2.2) Pour exécuter R1, on considère le plan P1 suivant :

Etape 1 : Transmettre les fragments $Client_j$ à tous les sites possédant un fragment $Loue_i$.

Etape 2 : Sur chaque site Si, traiter une requête Ti et envoyer le résultat au site S_0.

a) Quelle est l'expression algébrique de Ti, en fonction des données stockées ou reçues sur Si ?

b) Quelle opération algébrique fait-on sur S_0 pour obtenir R1 ?

c) Quelle est la cardinalité de T_i ?

```
card(Ti) = ..............
```

d) Quelle est la taille de Ti (en Go) ?

```
taille(Ti) = ..............
```

e) Quel est le coût de P1 (exprimé en Go) ?

```
coût(P1) = ..............
```

2.3) Pour exécuter R1, proposez un plan P2 de coût minimal. A chaque étape de P2, décrire les transferts de données effectués : quelle donnée est transférée vers quel site, quelle est la taille du transfert. Décrire aussi quelles sont les requêtes traitées sur chaque site.

Question 3

Soit la requête R2 :

```
Select *
From Article a, Loue l
Where a.num = l.num
```

3.1) On suppose que R2 contient tous les attributs de Article et Loue. Que valent card(R2) et taille(R2) ?

```
card(R2) = ..............


taille(R2) = ..............
```

3.2) Pour exécuter R2, on considère le plan P1 où chaque site traite une jointure. Détailler ce plan et donner son coût.

3.3) Pour exécuter R2, on considère le plan P2 où seul S_0 traite une jointure. Détailler ce plan et donner son coût.

Question 4

Soit la requête R3 : Select avg(prix) From Loue

On rappelle que la fonction avg() calcule la moyenne.

4.1) Proposer un plan de cout minimal pour traiter R3.

4.2) Le plan proposé est-il toujours correct si certains fragments $Loue_i$ ont moins de nuplet que d'autres ?