TD5 : Optimisation des opérateurs

Exercice 1 (question cours)

1. Lorsque l'on calcule un coût pour un chemin d'accès aux données quelle est l'unité que l'on utilise ?
2. Quels sont les deux éléments principaux que l'on considère dans un calcul de coût qui met en jeu un index ?
3. Qu'est-ce que la séléctivité d'un attribut ?
4. Dans quels cas le SGBD a-t-il la possibilité d'utiliser un index mettant en jeu plusieurs attributs (considérez par exemple un index structuré en arbre B ayant pour clé d'index (A,B)) ? Quelle séléctivité utilise-t-on ?

Exercice 2

On considère la relation suivante : Employees(<u>eid</u>: integer, ename: string, sal: integer, title: string, age: integer). On suppose que l'on dispose des index suivants qui sont tous de type 2 :

- un index non groupant avec table de hachage sur eid
- un index arbre B+ non groupant sur sal
- un index non groupant avec table de hachage sur age
- un index arbre B+ groupant sur (age, sal)

Chaque n-uplet de la table Employees occupe un espace de 100 octets, et on suppose (quelque soit l'index) que chaque entrée d'index occupe un espace de 20 octets. La relation Employees est stockée sur 10 000 pages, chaque page contenant 20 n-uplets de la relation. On suppose que la séléctivité moyenne est de 0.1 (quelque soit la condition et l'attribut).

Calculer le coût du chemin d'accès le plus sélectif pour chacune des requêtes suivantes (on fait varier les conditions et les attributs attendus en réponse à la requête).

1. Requête R1 : On recherche l'ensemble des n-uplets de la table Employees qui satisfont les conditions suivantes :
   (a) sal $\succ$ 100
   (b) age = 25
   (c) age $\succ$ 20
   (d) eid = 1 000
   (e) sal $\succ$ 200 AND age $\succ$ 30
   (f) sal $\succ$ 200 AND age = 20

(g) sal $\succ$ 200 AND title ="CFO"

(h) sal $\succ$ 200 AND age $\succ$ 30 AND title ="CFO"

2. Requête R2 : On recherche le salaire moyen des n-uplets de la table Employees qui satisfont la condition (a). Même question en considerant la condition (b), ..., Même question en considérant la condition (g).

3. Requête R3 : On recherche le salaire moyen pour chaque groupe d'age des n-uplets de la table Employees qui satisfont la condition (a). Même question en considerant la condition (b), ..., Même question en considérant la condition (g).

Exercice 3 (Exam 2017)

On considre les relations suivantes

```
Produit(pid:integer,nom:varchar(50),prix:integer,qtite:integer)-- pid est la cl
Client(cid:integer,email:varchar(50),pointfidelite:integer)-- cid est la cl
Achat(aid:integer,pid:integer,cid:integer,qtite:integer,date:integer)-- aid est la cl
```

On suppose une distribution uniforme des donnes. Il y a 10 000 produits (pid compris entre 0 et 9999), 200 clients (cid compris entre 0 et 199), 2 000 achats (aid compris entre 0 et 1999). Les emails des clients sont uniques, les points de fidlit vont de 0 400, les prix sont compris entre 0 et 1000 euros et les quantits entre 1 et 20. L'attribut `qtite` dans `Produit` indique la quantit d'un produit (identifié par `pid`) en stock alors que l'attribut `qtite` dans `Achat` indique la quantite de produit (identifié par `pid`) qui est acheté (lors d'un achat identifié par `aid`). Les relations `Produit`, `Client` et `Achat` occupent respectivement 500, 40 et 200 pages. En plus des clés primaires, on dispose d'un index B+ sur `Produit.prix`, non-groupant de type 2 dont l'ensemble des feuilles occupe 100 pages.

On considre la requte SQL suivante :

```
SELECT email
FROM Produit P, Client C, Achat A
WHERE A.pid = P.pid
     AND A.cid = C.cid
     AND A.qtite = P.qtite
     AND P.prix < 50
     AND C.pointfidelite < 200
ORDER BY email
```

1. Que renvoie cette requte (exprimez-la en francais) ?

2. Proposer un plan de requtes (ou arbre) qu'un optimiseur classique pourrait considrer (soyez prcis et justifiez votre rponse). Vous pouvez utiliser un noeud étiqueté `GroupBy` dans votre arbre sans autre précision.

3. Donner une estimation du cot de l'évaluation de la condition `P.prix` $\prec$ `50`.

4. Mme question pour `C.pointfidelite` $\prec$ `200`.

Exercice 4 (non corrigé)

Considérez deux schémas de relation R(A,C) et S(A,B), et leurs instances ci-dessous. Considérez la jointure spécifiée par : R $\bowtie$ S.

R	A	C
	a2	c4
	a1	c3
	a3	c5
	a1	c1
	a1	c2

S	A	B
	a5	b6
	a6	b7
	a1	b1
	a2	b3
	a4	b2
	a2	b4

Supposez que les n-uplets des instances sont stockés dans des fichiers dans l'ordre ci-dessus et que chaque page de fichier (pour R comme pour S) ne contient pas plus de deux enregistrements.

- Pour chacune des méthodes ci-dessous, donnez le principe de la méthode et la formule générale du coût en entrées/sorties en fonction de M (le nombre de pages de R), pR (le nombre de n-uplets par page dans R), N (le nombre de pages de S), et pS (le nombre de n-uplets par page dans S).
- Lister les couples de n-uplets de R et S dans l'ordre dans lequel ils sont "examinés" par la méthode et donnez le coût correspondant à l'opération de jointure.
  1. jointure itérative "brute"
  2. jointure itérative page à page
  3. jointure itérative par bloc (en supposant que le buffer disponible est de taille 4)
  4. jointure itérative avec index
  5. jointure par tri (en supposant que la phase de tri de R et S soit effectuée)