Entrepots de Données et Big Data : Omptimisation de requête
TD Optimisation 1 - PEL

Rendu du TD optimisation 1 - PEL obligatoire avant le 24/09 (attention : soigné, clair, et synthétique - max 5MB) à déposer dans l'espace Moodle dédié au cours.

1 Coût de plans d'exécution logiques

Soit le modèle relationnel composé des relations suivantes :
ETUDIANTS(IDE, NOM, AGE) – la relation contenant tous les étudiants
MODULES(IDM, RESPONSABLE, INTITULE) – la relation contenant tous les modules
IP(#IDE, #IDM) – la relation contenant la liste des inscripions pédagogiques (inscription d'un étudiants à un module)
FORMATION(IDF, NOMF) – la relation contenant toutes les formations
IA (#IDE, #IDF) – la relation contenant la liste des inscripions administratives (inscription d'un étudiants à une formation)

Hypothèses sur les données :
- 200 étudiants (200 lignes)
- 70 modules (70 lignes) dont un module dont l'intitulé est "EDBD".
- 4200 IP i.e. inscriptions pédagogiques (4200 lignes) dont 10% concernent le module EDBD.
- 50 formations (50 lignes).
- 250 IA i.e. inscription administrative, sachant que des étudiants peuvent être inscrits à plusieurs formations (250 lignes).

Nous souhaitons exécuter la requête suivante :

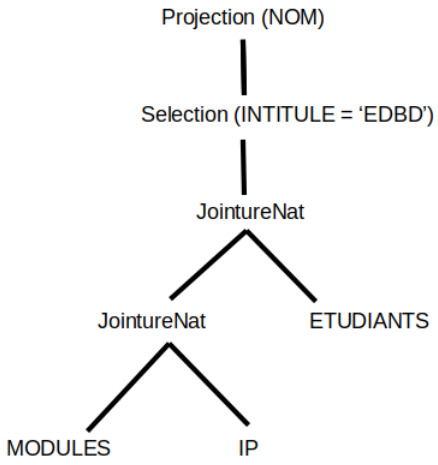
```
SELECT NOM
FROM ETUDIANTS E ,MODULES M ,IP I
WHERE E.IDE = I.IDE AND M.IDM=I.IDM
AND INTITULE = "EDBD" ;
```

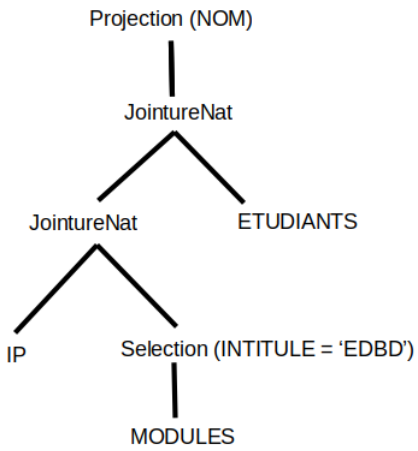
Question 1 Que permet d'obtenir la requête ci-dessus ?

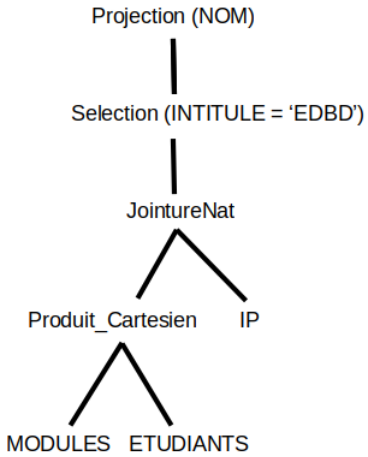
Pour cette requête, nous proposons 3 plans d'exécution logiques représentés par des arbres algébriques ci-dessous.

Question 2 Pour chaque plan d'exécution logique, calculer le coût E/S (en terme de nombre de lignes).

Question 3 Quel est le plan d'exécution logique optimal parmi les plans proposés ? Pourquoi ?







2 Définition de plans d'exécution logiques

Question Pour la requête ci-dessous :

- indiquer ce qu'elle permet d'obtenir,
- donner 2 plans d'exécution logique,
- indiquer le plan qui vous semble optimal parmi les plans proposés en justifiant.

Hypothèses complémentaires sur les données :

— il n'y a que des modules de master informatique (les intitulés commencent tous par "HAI").
— il y a 40% des IP qui concernent le Master GL.

Requête :

```
SELECT NOM
FROM ETUDIANTS E ,FORMATION F ,IA A, IP I, MODULE M
WHERE E.IDE = A.IDE AND F.IDF=A.IDF AND I.IDE=E.IDE AND M.IDM=I.IDM,
AND NOMF = "MASTER GL" AND M.IDM = "HAI708I" ;
```

3 Réécriture de plans d'exécution logiques

Soit le schéma relationnel suivant :
JOURNALISTE (IDJ, NOM, PRENOM) – La relation contenant tous les journalistes
JOURNAL (TITRE, REDACTION, #REDACTEUR_ID) – La relation contenant tous les journaux rédigés par des journalistes

On considère la requête suivante :

```
SELECT NOM
FROM JOURNAL, JOURNALISTE
WHERE TITRE='Le Monde' AND IDJ=REDACTEUR_ID AND PRENOM='Jean' ;
```

Voici deux expressions algébriques :

$$\pi_{nom}(\sigma_{titre='Le\ Monde' \wedge prenom='Jean'}(Journaliste \bowtie_{jid=redacteur\_id} Journal))$$

et

$$\pi_{nom}(\sigma_{prenom='Jean'}(Journaliste) \bowtie_{jid=redacteur\_id} \sigma_{titre='Le\ Monde'}(Journal))$$

Question 1 Les deux expressions retournent-elles le même résultat (sont-elles équivalentes) ? Justifiez votre réponse en indiquant les règles de réécriture que l'on peut appliquer.

Question 2 Une expression vous semble-t-elle meilleure que l'autre si on les considère comme des plans d'exécution ?

4 Préparation pour les prochaines séances

Assurez vous d'avoir un compte ORACLE sur l'instance MASTER du SGBD ORACLE de l'Université de Montpellier. Pour cela, vous pouvez vous référer à la partie "Instructions pour les Travaux Pratiques" du Moodle du module.
Penser à tester votre compte ORACLE en utilisant X2GO.