

Entrepôts de Données et Big Data : Optimisation de requête - partie 1

TD/TP en 2 parties (22/09 et 29/09). Rendu obligatoire avant le 08/10 (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 inscriptions 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 inscriptions administratives (inscription d'un étudiants à une formation)

Hypothèses sur les données :

- 200 étudiants (200 lignes)
- 70 modules (70 lignes)
- 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 :

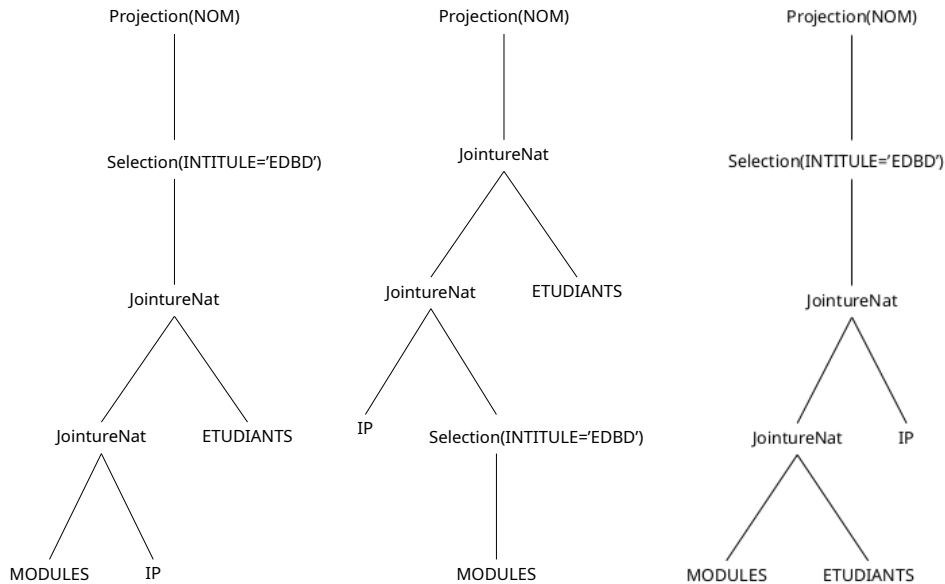
```
SELECT NOM
FROM ETUDIANTS E ,MODULES M ,IP I
WHERE E.IDE = I.IDE AND M.IDM=I.IDM
AND INTITULE = "EDBD";
```

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 une des requêtes ci-dessous :
- indiquer ce qu'elle permet d'obtenir,
 - donner 3 plans d'exécution logique,
 - indiquer le plan optimal parmi les plans proposés.

Requête 1 :
 SELECT RESPONSABLE
 FROM ETUDIANTS E ,MODULES M ,IP I
 WHERE E.IDE = I.IDE AND M.IDM=I.IDM
 AND NOM = "DUPOND" AND INTITULE LIKE "HMIN%";

Requête 2 :
 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 INTITULE = "EDBD";

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=IDJOURNALISTE 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 Tous les plans d'exécution logiques

Soit le modèle relationnel suivant :

ACTEUR (idA, nom, prenom, nationalite) – la relation contenant tous les acteurs

FILM (idF, titre, annee, nbspectateurs, #idRealisateurs, #idGenre) – la relation contenant tous les films

JOUER (#idActeur,#idFilm, salaire) – la relation contenant la listes des acteurs et des films dans lesquels ils jouent

REALISATEUR (idR, nom, prenom, nationalite) – la relation contenant tous les réalisateurs

GENRE (idG, description) – la relation contenant tous les genres des films (horeur, comédie,)

Hypothèses sur les données :

- 100 acteurs (100 lignes), dont 30% d'acteurs français et 10% d'acteurs de film
- 1000 films (1000 lignes)
- 25 réalisateurs (25 lignes)
- 10 genres (10 lignes)
- 2500 acteurs jouant dans un film (couples acteur-film) (2500 lignes)

Soit la requête suivante :

SELECT acteur.nom,acteur.prenom

FROM acteur,jouer,film,genre, realisateur

WHERE (idA=idActeur) AND (idFilm=idF) AND (idGenre=idG) AND (idRealisateur=idR)

AND (acteur.nationalite='France') AND (description='comédie') AND (realisateur.nom = "Les frères Coen");

Question 1 Pour la requête ci-dessus, donner 2 plans d'exécution logiques : un premier plan fera intervenir les jointures en premier et un deuxième commencera par les sélections.

Question 2 Parmi les plans d'exécution proposés, quel est le plan optimal? Justifier votre choix.