

Exercice n°1 : SQL et plan d'exécution

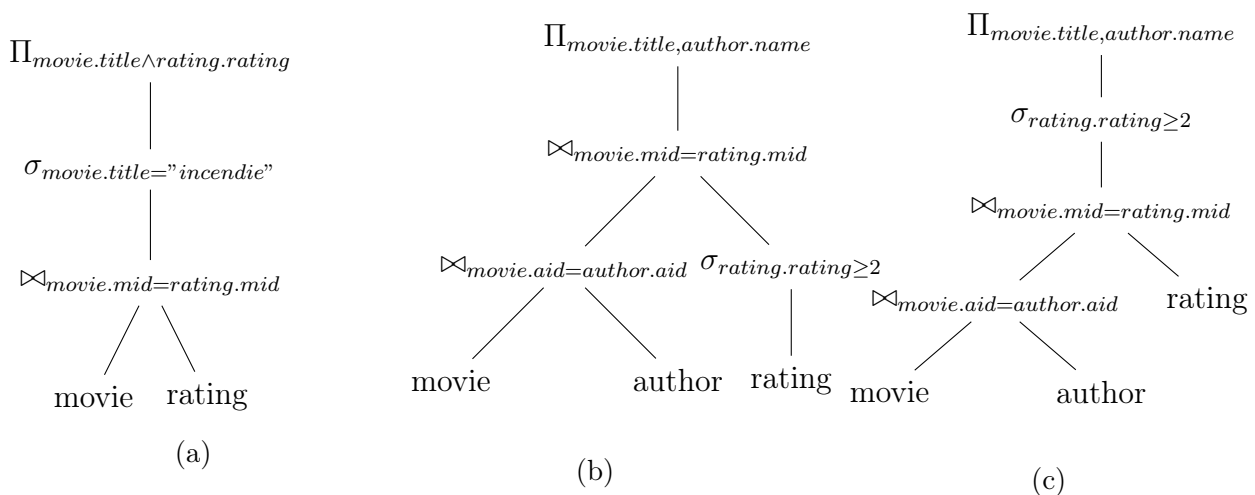


FIGURE 1 – Trois arbres de requête algébrique

Question n°1 : Proposez pour les trois arbres de requêtes algébriques (figure 1) une requête SQL possible.

Question n°2 : Pour les trois arbres de la figure 1, dites si oui ou non des matérialisations sont nécessaires. On rappelle que toutes les boucles externes de jointure correspondent au sous arbre gauche (si jointure par itération).

Question n°3 : Dessinez un arbre de requête algébrique pour :

SELECT movie.title, rating.rating **FROM** movie **INNER JOIN** rating **ON** movie.mid=rating.mid **WHERE** movie.year = 2000

Exercice n°2 : Estimation du coût d'un plan

On considère dans cet exercice une base de données contenant trois relations :

- library (lid, name, location)
- book (bid, title, author, publisher, publication_year)
- stock (bid, lid, quantity) (bid et lid sont des clefs étrangères)

La base de données contient 250.000 livres, 55 bibliothèques et la relation **stock** contient 1.000.000 d'entrées. Une page peut contenir 250 livres, 500 bibliothèques et 1000 entrées de la relation stock. Il existe un index B+ de type 2 sur les clefs primaires de la table. On dispose aussi d'un index B+ (de type 2) pour les champs *publication_year* et *quantity*, chaque page de l'index peut contenir (pour les deux index) 2000 entrées et la profondeur des arbres vaut respectivement 2 et 3.

Question n°1 : On cherche les bibliothèques et le nom des livres tel que la bibliothèque contient au moins 3 exemplaires où le livre a été publié après 2020. La requête qui peut s'exprimer en SQL comme

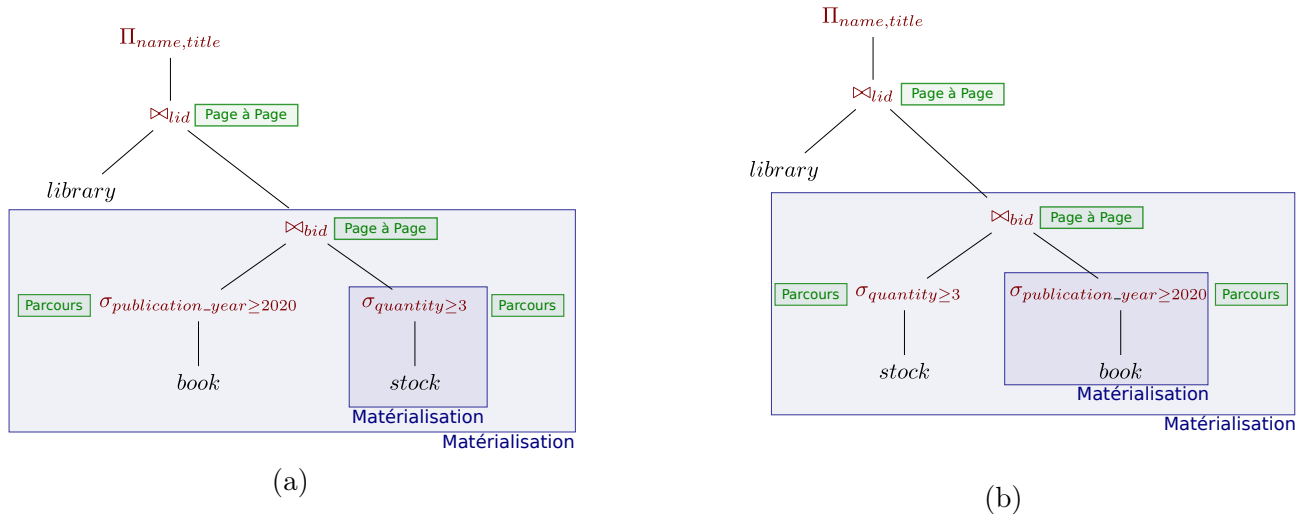
SELECT name, title **FROM** library **INNER JOIN** stock **ON** lid **INNER JOIN** book **ON** bid **WHERE** publication_year \geq 2020 **AND** quantity \geq 3

Expliquez quelles sont les données dont vous auriez besoins pour calculer le nombre d'E/S du plan d'exécution.

Question n°2 : En considérant que l'évaluation de $c_1 = \sigma_{publication_year \geq 2020}$ produit 5000

enregistrements, $c_2 = \sigma_{quantity \geq 3}$ 500 enregistrements et $c_1 \wedge c_2$ 100 enregistrements. De plus en pipelining la jointure page à page attend de remplir une cellule du buffer/cache (l'équivalent d'une page) pour chaque itération externe et que la taille du buffer est de 3 cellules.

Calculez et comparez le temps d'exécution des plans suivants :



Question n°3 : Si maintenant on matérialise tous les résultats des conditions, quel est le coût en terme d'E/S des deux plans de la question précédente si le cache contient 3 cellules ?

Question n°4 : Même question que la précédente, mais cette fois-ci le cache disponible pour la requêtes contient 100 cellules et l'on considère donc des jointures par paquets de pages.

Question n°5 : Dans les deux cas précédents, deux matérialisations sont nécessaires, proposez un plan d'exécution ne nécessitant qu'une seule matérialisation en gardant les méthodes d'évaluation des opérateurs de la question 3. Calculez le nombre d'E/S si 3 cellules sont disponibles dans le cache.

Question n°6 : On se propose pour le plan d'exécution précédent d'utiliser la sélection via les index B+ disponibles. Donner le coût minimum possible en utilisant ou non ces index (les fichiers de données ne sont pas triés selon ces index).

Exercice n°3 : Un exemple complet (Difficile - Type sujet examen)

Pour l'examen il vous sera demandé de rechercher un plan d'exécution "performant" étant donnée une requête. Pour ce faire, on vous demandera d'abord d'évaluer tous les plans "mono-relation" (sélection), puis d'estimer le coût des plans multi-relation (c'est à dire les jointures).

Dans cet exercice on considérera que toutes les sélections sont effectués avant les jointures, pour chaque sous-plan, il faudra calculer la taille le nombre d'enregistrements résultant de l'évaluation du sous-plan, le nombre de page (page écrites), mais aussi le coût (en terme de page) pour l'exécution du sous-plan.

Soit une base de données composée des relations suivantes :

- **movie** (mid : int, title : str, year : int) (10^5 enregistrements)
- **class** (cid : int, categorie : str) (15 enregistrements)
- **movie_info** (#mid : int, #pid : int, #cid : int) (3×10^5 enregistrements)

On considère dans cet base de données qu'un entier occupe 4 octets, une chaîne de caractère 100 octets, un pointeur est composé de 8 octets, la taille des pages est de 8000 octets (pour les indexes aussi). On dispose des index suivants :

Nom de l'index	Relation	Clé	Grouper	Structure	Coût/hauteur
I_1	movie_info	cid	Hachage	Oui	2
I_2	movie_info	pid	B+	Non	$h = 2$
I_3	movie_info	mid	B+	Non	$h = 3$
I_4	movie	mid	B+	Non	$h = 3$
I_5	movie	year	B+	Oui	$h = 3$
I_6	class	cid	B+	Non	$h = 1$

On considérera tout au long de l'exercice la requête suivante :

SELECT title, categorie FROM movie M, class C movie_info MI WHERE M.mid = MI.mid AND C.cid = MI.cid AND categorie = 'Adventure' AND year \geq 2022.

On suppose que les catégories sont uniformément répartis selon les films et qu'un film à en moyenne 3 catégories. Chaque catégorie est unique, donc il y a un enregistrement par catégorie et donc 15 catégories. Pour l'année de sortie des films on a autant de films sur les quatre intervalles suivants :

[1895, 1960], [1960, 1995], [1995, 2013], [2013, 2023]

Dans ces intervalles on pose **l'hypothèse** suivante : *les films sont uniformément répartis par année* (autant de films chaque années).

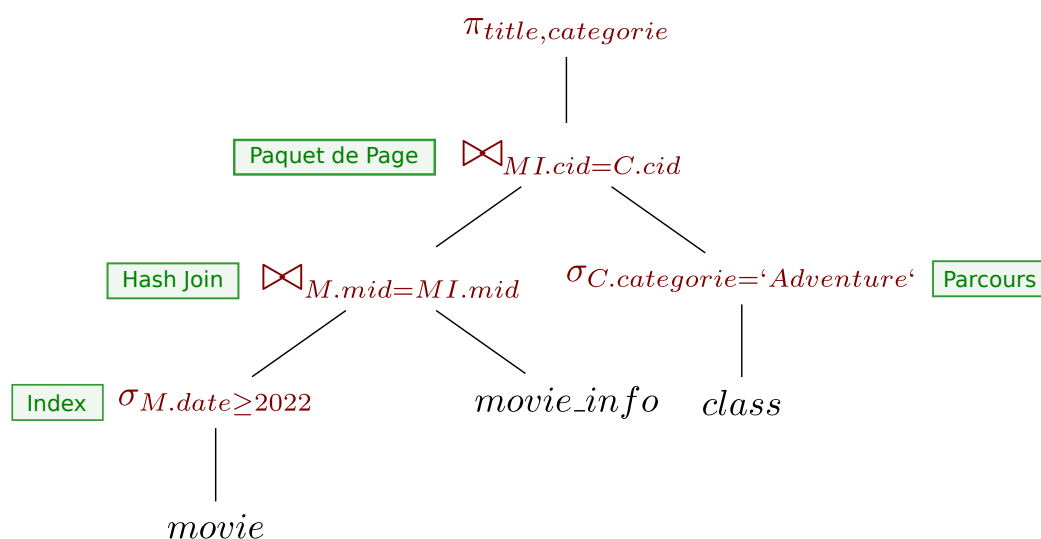
On considère si non renseigné, 3 cellules disponibles dans le cache.

Question n°1 : Combien d'entrées au maximum peut contenir une page d'un index haché sur une chaîne de caractères ?

Question n°2 : Déterminez tous les sous-plans mono relation pour la requête

Question n°3 : Combien de jointures sont possibles ? Lesquelles peuvent être facilement éliminées ? Pourquoi ?

Question n°4 : Sachant que le hachage produit au maximum des paquets de taille 2 pour *mid* de la table *movie* et que l'on dispose d'au moins 4 cellules libres dans le buffer. Calculez le coût du plan d'exécution suivant :



Question n°5 : Si c'est possible, proposez un plan n'effectuant aucune matérialisation des tables (sinon expliquez pourquoi) et calculez son coût.

Question n°6 : Proposez un plan d'exécution effectuant au moins une jointure par index et calculez son coût.