EXAMEN SESSION 1 – HAI708I
Entrepôt de Données et Big Data

Session : 1	Durée de l'épreuve : 2 heures
Date : 12-janvier-2022	Documents autorisés : tous
Mention Informatique	Matériel utilisé : aucun
Master 1ère année : EDBD (HAI708I)	

NUMERO ÉTUDIANT :

ATTENTION, pour la question 6 de la partie Optimisation vous devez répondre sur le sujet .
Pensez à bien indiquer votre numéro étudiant (ci-dessus).

Partie Optimisation

Vous disposez d'une base de données contenant des données sur des acteurs, des films et le salaire que touchent les acteurs pour un film.

Le schéma relationnel de la base de données implémentée sous Oracle est le suivant :
`Acteur (idA, nom, prénom, nationalité)`
`Film (idF, titre, annee, pays, nbspectateurs)`
`Jouer (#idActeur, #idFilm, salaire)`
avec l'attribut *idActeur* clé étrangère référençant l'attribut *idA* de la relation Acteur et l'attribut *idFilm* clé étrangère référençant l'attribut *idF* de la relation Film.

Vous demandez au SGBD Oracle le plan d'exécution physique de la requête que vous venez de concevoir. Voilà ci-dessous la sortie que vous fournit Oracle (le mot clé "BATCHED" de l'instruction 6 correspond à une optimisation au niveau bloc de données physique, n'en tenez pas compte)

```
-----------------------------------------------------------------------------------------------
| Id  | Operation                                | Name       | Rows  | Bytes | Cost (%CPU)| Time     |
-----------------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT                         |            |     1 |   212 |     2   (0)| 00:00:01 |
|   1 |  NESTED LOOPS                            |            |     1 |   212 |     2   (0)| 00:00:01 |
|   2 |   NESTED LOOPS                           |            |     1 |   212 |     2   (0)| 00:00:01 |
|   3 |    NESTED LOOPS                          |            |     1 |   113 |     1   (0)| 00:00:01 |
|   4 |     TABLE ACCESS BY INDEX ROWID          | ACTEUR     |     1 |    74 |     1   (0)| 00:00:01 |
|*  5 |      INDEX UNIQUE SCAN                   | PK_AUTEUR  |     1 |       |     1   (0)| 00:00:01 |
|   6 |     TABLE ACCESS BY INDEX ROWID BATCHED  | JOUER      |     1 |    39 |     0   (0)| 00:00:01 |
|*  7 |      INDEX RANGE SCAN                    | PK_JOUER   |     1 |       |     0   (0)| 00:00:01 |
|*  8 |    INDEX UNIQUE SCAN                     | PK_FILM    |     1 |       |     0   (0)| 00:00:01 |
|*  9 |   TABLE ACCESS BY INDEX ROWID            | FILM       |     1 |    99 |     1   (0)| 00:00:01 |
-----------------------------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------

   5 - access("IDA"=1)
   7 - access("IDACTEUR"=1)
   8 - access("IDF"="IDFILM")
   9 - filter("ANNEE">2000)
```

Question 1 : Expliquer le plan d'exécution physique suivant fourni par le SGBD Oracle

Question 2 : Reconstruire la requête exécutée.

Question 3 : Dessiner l'arbre ou donner l'expression algébrique du plan d'exécution logique correspondant au plan d'exécution physique fourni par Oracle ci-dessus

Question 4 : Proposer un autre plan d'exécution logique que celui d'Oracle (arbre ou expression algébrique).

Question 5 : Indiquer quel plan d'exécution logique, entre celui de la question 3 et celui de la question 4, est optimal en argumentant.

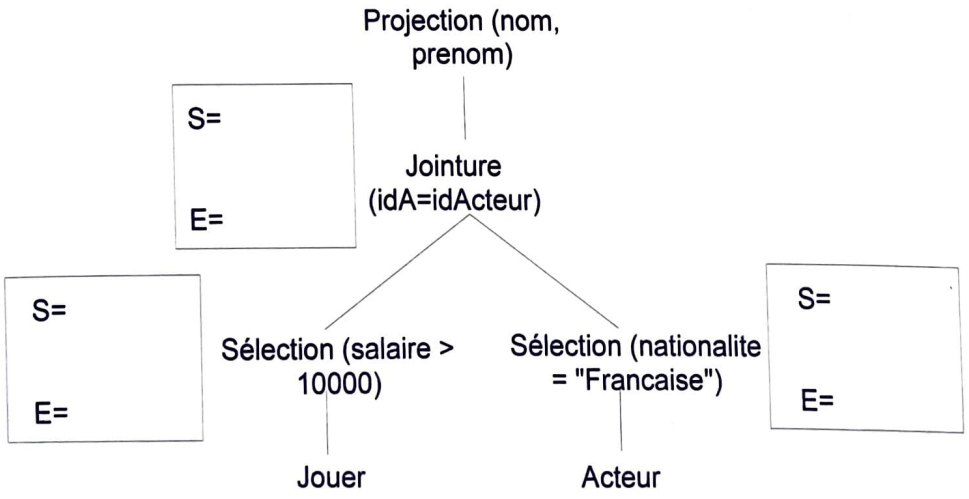
Vous souhaitez maintenant exécuter la requête suivante : « La liste des noms et prénoms des acteurs français ayant déjà touché un salaire supérieur à 10 000 euros pour un film. »

```
SELECT nom, prenom
FROM Acteur, Jouer
WHERE idA = idActeur AND nationalite = "Francaise" AND salaire >
10000 ;
```

Vous disposez des hypothèses suivantes : 1000 films (100 blocs), 200 acteurs (20 blocs) dont 10 % d'acteurs français, 3500 « jeux » d'acteurs dans des films (3500 lignes dans la table Jouer, 350 blocs) dont 20 % des jeux d'acteurs avec un salaire > 10 000 euros et 5% des jeux d'acteurs correspondant à des acteurs français.

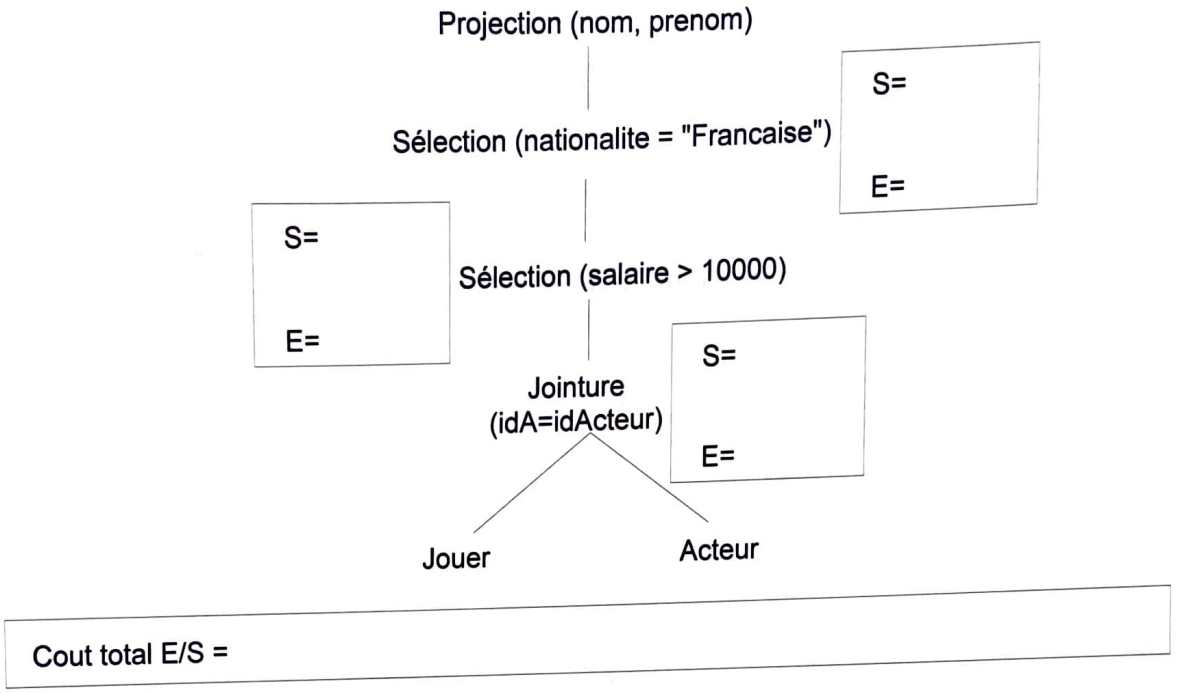
Question 6 : Pour chacun des plans d'exécution logiques ci-dessous, calculer le coût E/S en remplissant les cadres E/S. RÉPONDRE SUR LE SUJET !

Plan d'exécution logique 1 :



Cout total E/S =

Plan d'exécution logique 2 :



Entrepôts de données

Une agence immobilière nationale souhaite mettre en place un entrepôt de données pour mesurer ses performances au niveau du marché Montpelliérain. Par simplicité, nous nous concentrerons sur un entrepôt de données restreint à un certain nombre d'analyses qui seront traduites par les requêtes suivantes.

1. Pour chaque secteur (Boutonnet, Facultés, Gare, etc...), le nombre d'appartements vendus par type (T2, T3, etc) et par mois en 2019.
2. La moyenne du prix de vente des maisons à 2 étages dans l'Écusson en 2019.
3. Le nombre d'acheteurs étrangers ayant acquis un bien en 2019.
4. Pour chaque vente supérieure au million d'euros, le nombre d'agents impliqués dans la vente.
5. Pour chaque agent, le montant des marges réalisées en 2019.
6. Pour chaque vente, la somme des marges réalisées par les agents impliqués.

À travers ces analyses, on veut mesurer à la fois la demande du marché et les performances des agents de l'agence qui réalisent les ventes. Notamment pour les gros biens, une vente peut être réalisée par un ou plusieurs agents, travaillant dans ce dernier cas en équipe. Dans le cas d'une vente en équipe, chaque agent contribue à la vente avec un pourcentage de participation. Par ailleurs, chaque agent réalise une marge individuelle sur une vente correspondant à un certain pourcentage du montant de la vente.

Question 1 : Définir le schéma d'un entrepôt de données permettant les analyses ci-dessus.

1.1 Donner la liste des tables des faits et des dimensions de l'entrepôt

1.2 S'agit-il d'un modèle transactionnel ou snapshot ? Justifier.

1.3 Indiquer les mesures et décrire leur additivité/semi-additivité/non-additivité.

1.4 Si pertinent, indiquer les mini-dimensions, dimensions dégénérées, tables bridge,etc.

1.5 Donner une instance de la table des faits contenant 5 lignes.

Question 2 - À partir du modèle proposé, traduire les interrogations suivantes en SQL.

Attention : il n'est pas demandé de donner le code SQL pour la création des tables.

2.1 Le chiffre d'affaire concernant la vente des maison par secteur.

2.2 Pour chaque agent, le montant total des marges qu'il a réalisées sur l'ensemble de ses ventes.

2.3 Pour chaque vente, la somme des marges réalisées par les agents.

2.4 Pour chaque agent, la moyenne des participations aux ventes d'appartements.

Question 3 - Expliquer avec vos mots l'intérêt d'un entrepôt de données par rapport à une base de données transactionnelle.