EXAMEN SESSION 1 – HAI708I
Entrepôt de Données et Big Data

Session : 1	Durée de l'épreuve : 2 heures
Date : janvier-2025	Documents autorisés : documents papier
Mention Informatique	Matériel utilisé : aucun

ATTENTION

Pensez à bien indiquer votre numéro étudiant en haut des pages !

Répondre sur la copie pour la partie 1 (Optimisation)

Répondre sur le sujet pour les parties 2 et 3 (Entrepôt et Map Reduce)

1 – Optimisation [RÉPONDRE SUR LA COPIE]

Vous disposez d'une base de données d'une entreprise spécialisée dans l'évènementiel. Cette entreprise conserve les données concernant les spectacles qu'elle propose et les artistes qui se produisent dans ces spectacles.
Vous demandez au SGBD Oracle d'afficher le plan d'exécution physique de la requête que vous venez de concevoir. Voilà ci-dessous la sortie que vous fournit Oracle.

```
-----------------------------------------------------------------------------------------
| Id  | Operation                              | Name          | Rows  | Bytes | Cost (%CPU)| Time     |
-----------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT                       |               |     3 |   360 |     8   (0)| 00:00:01 |
|   1 |  NESTED LOOPS                          |               |     3 |   360 |     8   (0)| 00:00:01 |
|   2 |   NESTED LOOPS                         |               |     3 |   360 |     8   (0)| 00:00:01 |
|*  3 |    HASH JOIN                           |               |     3 |   219 |     5   (0)| 00:00:01 |
|   4 |     TABLE ACCESS BY INDEX ROWID BATCHED| ARTISTE       |     2 |   104 |     2   (0)| 00:00:01 |
|*  5 |      INDEX RANGE SCAN                  | INDEX_AGE     |     2 |       |     1   (0)| 00:00:01 |
|   6 |     TABLE ACCESS FULL                  | SEPRODUITDANS |    10 |   210 |     3   (0)| 00:00:01 |
|*  7 |    INDEX UNIQUE SCAN                   | SPECTACLE_PK  |     1 |       |     0   (0)| 00:00:01 |
|   8 |   TABLE ACCESS BY INDEX ROWID          | SPECTACLE     |     1 |    47 |     1   (0)| 00:00:01 |
-----------------------------------------------------------------------------------------

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

   3 - access("ARTISTE"."CODEA"="SEPRODUITDANS"."CODEA")
   5 - access("AGE">50)
   7 - access("SPECTACLE"."CODES"="SEPRODUITDANS"."CODES")
```

Plan d'exécution fourni par Oracle

NUMÉRO ÉTUDIANT :

Le schéma relationnel de la base de données implémentée sous Oracle est le suivant :

Artiste (codeA, nomA, prenomA, age, genre, pays)
Spectacle (codeS, titre, localisation, debut, fin)
seProduitDans (#codeA, #codeS, nomRole, categorieRole)

Par ailleurs, voila les contraintes connues :

- Dans la relation «seProduitDans», l'attribut «codeA» est une clé étrangère référençant l'attribut «codeA» de la relation «Artiste» et l'attribut «codeS» est une clé étrangère référençant l'attribut «codeS» de la relation «Spectacle».
- les clés primaires des relations « Artiste », « Spectacle » et « seProduitDans » sont respectivement « codeA », « codeS » et « codeA,codeS,nomRole ».
- La contrainte de clé primaire de la table «Artiste» est nommée «ARTISTE_PK».
- Un index supplémentaire nommé «INDEX_AGE » a été créé sur l'attribut « age » de la relation « Artiste ».

Question 1 : Donner en SQL la requête exécutée qui conduit au plan d'exécution présenté ci-dessus.

Question 2 – optimisation logique : Dessiner l'arbre ou donner l'expression algébrique du plan d'exécution logique correspondant au plan d'exécution fourni par Oracle, puis calculer le coût E/S de ce plan d'exécution logique en utilisant les hypothèses suivantes.

Vous disposez maintenant des hypothèses suivantes :
- 1000 artistes (espace mémoire = 1000 lignes) dont 20 % ont plus de 50 ans;
- 300 spectacles (espace mémoire = 300 lignes) ;
- 2000 productions d'artistes dans des spectacles (espace mémoire = 2000 lignes) et 40 % de ces productions correspondant à des artistes de plus de 50 ans (plusieurs artistes peuvent se produire dans un spectacle et un artiste peut se produire dans différents spectacles).

Question 3 – optimisation physique : Est-ce que la création de l'index « INDEX_AGE » engendre une modification du plan d'exécution proposé par Oracle ? En d'autres termes, si cet index n'avait pas été créé, est-ce que le plan d'exécution proposé par Oracle aurait été différent ? Justifier votre réponse.

NUMÉRO ÉTUDIANT :

2 – Entrepôt de données [RÉPONDRE SUR LE SUJET]

Un leader mondial dans la construction d'appareils photo souhaite faire évoluer sa propre offre de produits sur la base de données de la plateforme Flickr. Précisément, l'objectif est de concevoir un entrepôt de données permettant d'analyser l'utilisation des appareils par le biais des photos publiées sur Flickr. L'entrepôt doit permettre d'étudier les lieux ainsi que les périodes de l'année et les horaires de la journée où les appareils sont utilisés, mais aussi le lignage des appareils photos (par exemple, le modèle Nikon D3200 dérive du modèle Nikon D3100 qui dérive du D3000), et le créateur de chaque modèle (par exemple, le modèle Nikon D3200 à été conçu par Eiji Fumio).

Question 1. Proposer un schéma d'entrepôt de données permettant les analyses suivantes en justifiant vos choix, et donner les requêtes SQL correspondantes.

1. Compter le nombre de photos réalisées pour chaque modèle d'appareil photo ;
2. Compter le nombre de photos prises par un appareil conçu par Eiji Fumio ;
3. Pour tous les modèles dérivés du Nikon D3000, compter le nombre de photos réalisées par jour ;
4. Indiquer les modèles historiquement les plus influents (on considère ici les appareils au sommets des hiérarchies de lignage).

Comment répondre pour le schéma d'entrepôt de données : décrire les faits, les dimensions, et les mesures nécessaires, ainsi que leur additivité. Justifier vos choix de modélisation si nécessaire.

NUMÉRO ÉTUDIANT :

Question 3.

Vous disposez d'un fichier csv contenant les données des courses en Taxis relatives à l'année 2023, dont voici un extrait.

Date/horaire début	Date/horaire fin	Nombre passagers	Prix Total ($)
2023-03-01_00:21:05	2023-03-01_00:23:05	3	5.8
2023-03-01_00:13:15	2023-03-01_00:33:15	4	7.1
2023-03-01_00:11:45	2023-03-01_00:14:45	1	9.3
…	…	…	…

On vous demande de décrire l'entrée et la sortie des fonctions map et reduce permettant de réaliser les analyses suivantes. *Vous pouvez répondre soit avec du texte libre soit à l'aide du pseudo-code.*

1. Calculer le nombre de passagers transportés entre (12h et 13h) et entre (18h et 19h) pour toute l'année 2023.
2. Donner le nombre de courses dont les passagers sont (entre 1 et 2), (entre 3 et 5), (entre 5 et 10) pour toute l'année 2023.