**EXAMEN SESSION 1 – HAI708I**
**Entrepôt de Données et Big Data**

| | |
|---|---|
| Session : 1<br>Date : 09-janvier-2023<br>Mention Informatique<br>Master 1ère année : EDBD (HAI708I) | Durée de l'épreuve : 2 heures<br>Documents autorisés : tous<br>Matériel utilisé : aucun |

**NUMERO ÉTUDIANT :**

***ATTENTION :***

- ***pour la question 5 de la partie Optimisation et pour la partie Map/Reduce vous devez répondre sur le sujet***
- ***pensez à bien indiquer votre numéro étudiant (ci-dessus).***

## Partie Optimisation

Vous disposez d'une base de données d'une entreprise de vente de produits électroménagers, contenant des données sur des commandes de produits, passées par des clients.
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.

```
PLAN_TABLE_OUTPUT
------------------------------------------------------------------------------------------------
Plan hash value: 1197142960

-------------------------------------------------------------------------------------------
| Id  | Operation                          | Name         | Rows  | Bytes | Cost (%CPU)| Time     |
-------------------------------------------------------------------------------------------
|   0 | SELECT STATEMENT                   |              |     2 |   400 |     3   (0)| 00:00:01 |
|   1 |  NESTED LOOPS                      |              |     2 |   400 |     3   (0)| 00:00:01 |
|   2 |   NESTED LOOPS                     |              |     2 |   400 |     3   (0)| 00:00:01 |
|   3 |    NESTED LOOPS                    |              |     2 |   200 |     3   (0)| 00:00:01 |
|   4 |     TABLE ACCESS FULL              | COMMANDES    |     5 |   260 |     3   (0)| 00:00:01 |
|*  5 |     TABLE ACCESS BY INDEX ROWID    | PRODUITS     |     1 |    48 |     0   (0)| 00:00:01 |
|*  6 |      INDEX UNIQUE SCAN             | PK_PRODUITS  |     1 |       |     0   (0)| 00:00:01 |
|*  7 |    INDEX UNIQUE SCAN               | PK_CLIENTS   |     1 |       |     0   (0)| 00:00:01 |
|   8 |   TABLE ACCESS BY INDEX ROWID      | CLIENTS      |     1 |   100 |     0   (0)| 00:00:01 |
-------------------------------------------------------------------------------------------

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

   5 - filter("PRODUITS"."NOMP"='sable')
   6 - access("PRODUITS"."IDP"="COMMANDES"."IDP")
   7 - access("CLIENTS"."IDCLIENT"="COMMANDES"."IDCLIENT")
```

Plan d'exécution fourni par Oracle

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

produits (idp, nomp, cout)
clients (idclient, nomc, adrc, solde)
commandes (numcom, #idclient, #idp, qte)

Dans la relation « commandes », l'attribut « idclient » est clé étrangère référençant l'attribut « idA » de la relation « clients » et l'attribut « idp » est clé étrangère référençant l'attribut « idp » de la relation « produits ».

Par ailleurs, les contraintes de clé primaire des tables « produits », « clients » et « commandes » sont nommées respectivement « pk_produits », « pk_clients » et « pk_commandes ».

**Question 1 : Donner en SQL la requête exécutée qui conduit au plan d'exécution présenté en page 2.**

**Question 2 : Dessiner l'arbre ou donner l'expression algébrique du plan d'exécution logique correspondant au plan d'exécution physique fourni par Oracle.**

**Question 3 : Proposer un autre plan d'exécution logique que celui d'Oracle (arbre ou expression algébrique).**

**Question 4 : Indiquer quel plan d'exécution logique, entre celui de la question 2 et celui de la question 3, est optimal en argumentant.**

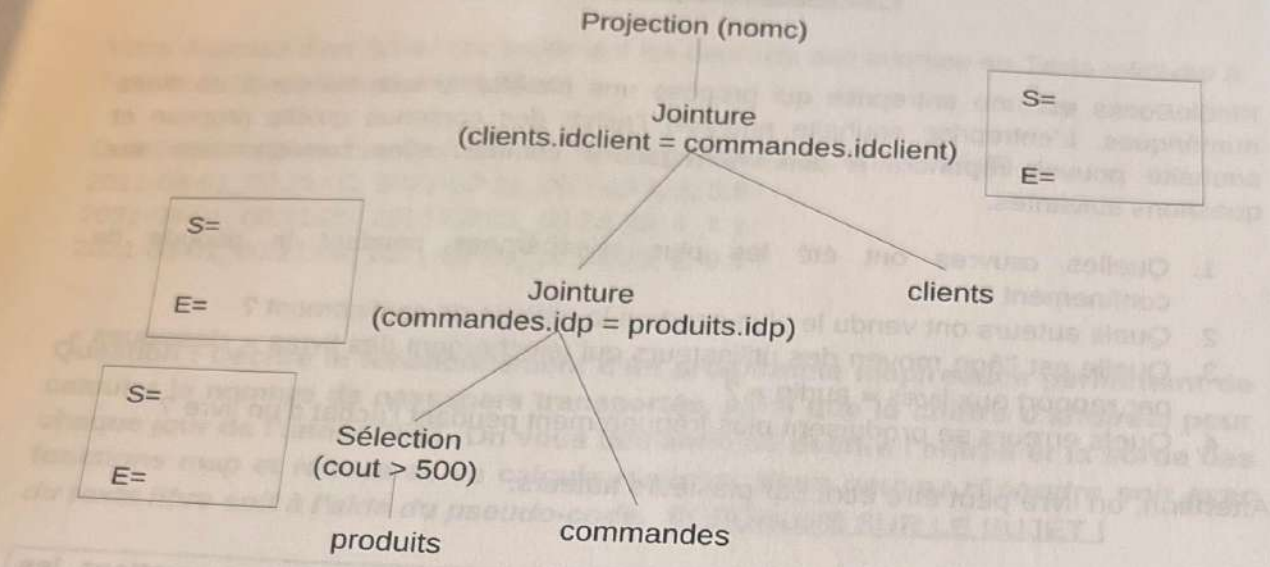
Vous souhaitez maintenant exécuter la requête suivante : « La liste des noms de clients ayant commandé des produits coûtant plus de 500 euros. »

```
SELECT nomc
FROM clients, commandes, produits
WHERE commandes.idp = produits.idp AND clients.idclient = commandes.idclient AND
AND cout > 500;
```

Vous disposez des hypothèses suivantes : 200 clients (espace mémoire = 200 lignes) ; 50 produits (espace mémoire = 50 lignes) dont 50 % avec un coût supérieur à 500€ ; 1000 commandes (espace mémoire = 1000 lignes) dont 60% des commandes concernant des produits coûtant plus de 500€.

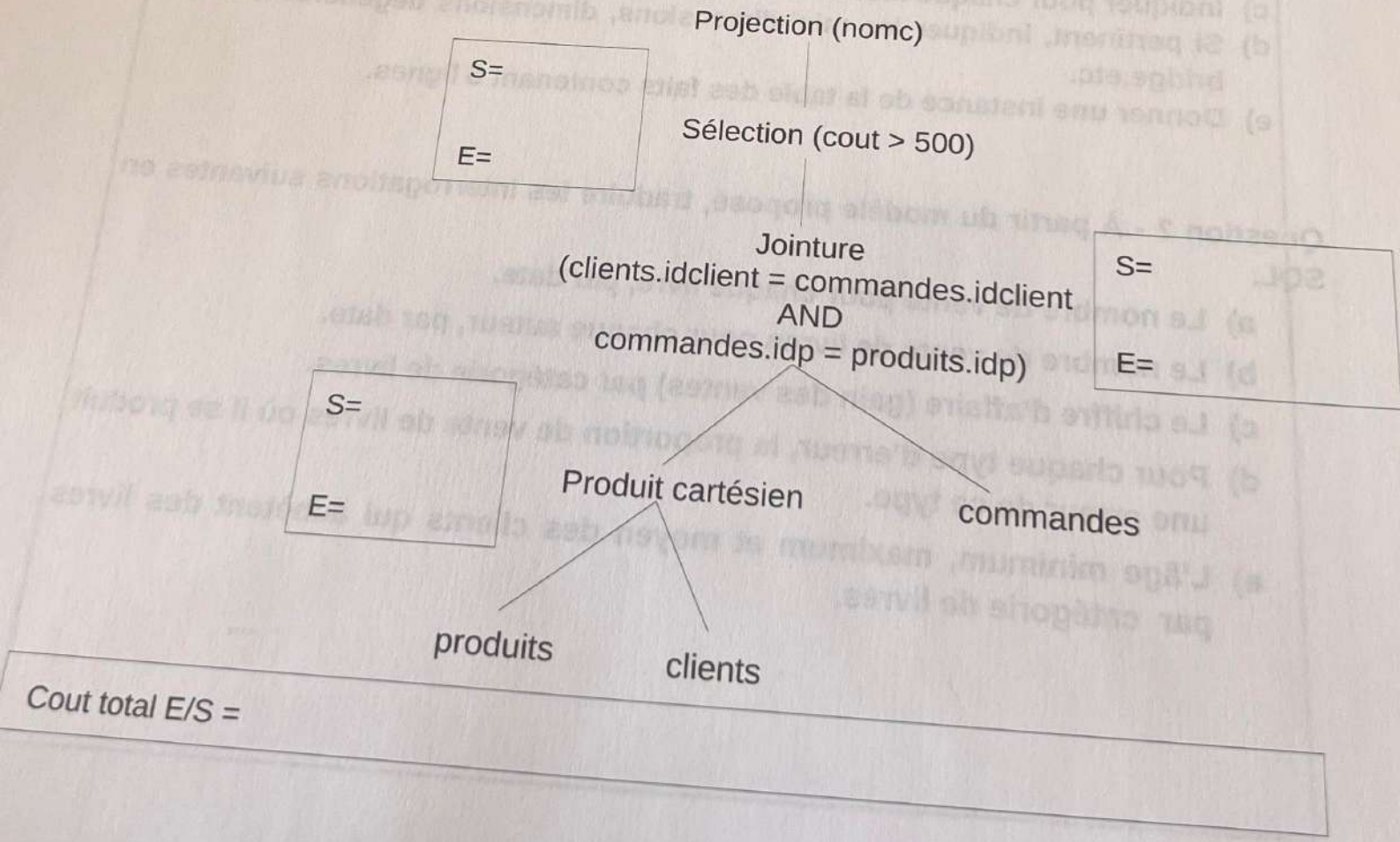
**Question 5 : Pour chacun des plans d'exécution logiques ci-dessous, calculer le coût E/S en remplissant les cadres E/S. <u>RÉPONDRE SUR LE SUJET !</u>**

Plan d'exécution logique 1 :




Cout total E/S =

Plan d'exécution logique 2 :




Cout total E/S =

# Partie Entrepôts de données

KindleBooks est une entreprise qui propose une plateforme web de vente de livres numériques. L'entreprise souhaite analyser l'achat des contenus qu'elle propose et souhaite pouvoir répondre à des interrogations comme celles correspondant aux questions suivantes.

1. Quelles œuvres ont été les plus téléchargées pendant la période de confinement ?
2. Quels auteurs ont vendu le plus pendant la période de confinement ?
3. Quelle est l'âge moyen des utilisateurs qui téléchargent des livres « classiques » par rapport aux lives « audio » ?
4. Quels erreurs se produisent plus fréquemment pendant l'achat d'un livre ?

Attention, un livre peut être écrit par plusieurs auteurs.

**Question 1 : Proposer le modèle d'un entrepôt de données permettant les analyses ci-dessus. Pour cela, vous devrez :**

a) **Indiquer les tables de faits et de dimensions de l'entrepôt et préciser leurs attributs**
b) **Indiquer s'il s'agit d'un modèle transactionnel ou snapshot en justifiant**
c) **Indiquer pour chaque mesures leur additivité/semi-additivité/non-additivité**
d) **Si pertinent, indiquer les mini-dimensions, dimensions dégénérées, tables bridge,etc.**
e) **Donner une instance de la table des faits contenant 5 lignes.**

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

a) **Le nombre de vente pour chaque livre, par date.**
b) **Le nombre de vente de livres pour chaque auteur, par date.**
c) **Le chiffre d'affaire (gain des ventes) par catégorie de livres.**
d) **Pour chaque type d'erreur, la proportion de vente de livres où il se produit une erreur de ce type.**
e) **L'âge minimum, maximum et moyen des clients qui achètent des livres par catégorie de livres.**

## Partie Map/Reduce

Vous disposez d'un fichier csv contenant les données des courses en Taxis relatives à l'année 2021, dont voici un extrait.

Date/horaire début, Date/horaire fin, Nombre passagers, Prix Total ($)
2021-03-01_00:21:05, 2021-03-01_00:24:23, 3, 5.8
2021-03-01_00:21:05, 2021-03-01_00:24:23, 4, 7.1
2021-03-01_00:21:05, 2021-03-01_00:24:23, 1, 9.3
...

**Question : Décrire le fonctionnement d'un programme map/reduce permettant de calculer le nombre de passagers transportés, ainsi que le chiffre d'affaires, pour chaque jour de l'année 2021. On vous demande de décrire l'entrée et la sortie des fonctions map et reduce et les calculs réalisés. *Vous pouvez répondre soit avec du texte libre soit à l'aide du pseudo-code.* RÉPONDRE SUR LE SUJET !**