Auto-Évaluation : Questions pour la Préparation à l'Examen

Ce document présente des questions permettant de travailler des sujets abordés en cours ainsi que lors du mini-projet.

Voici les sujets abordés.

- Optimisation de requêtes
- Modélisation: détaillée, Snapshot, Update-Records
- Tables pont (hierarchies, associations N:M)
- Partitionnement
- Techniques de mise à jour des dimensions
- Vues matérialisées
- Indexes Bitmap
- Hadoop et Map/Reduce

Optimisation

Exercice 1. Plans d'exécution

Expliquer le plan d'exécution physique suivant fourni par le SGBD Oracle et donner en SQL la requête exécutée qui conduit à ce plan d'exécution.

```
-----------------------------------------------------------------------------------------------------
| 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)
```

Notes :

- Ne pas tenir compte du mot clé "BATCHED" de l'instruction 6 (il s'agit d'optimisation au niveau bloc de données physique)
- Le schéma relationnel de la base de données interrogée est le suivant :

  Acteur (idA, nom, prénom, nationalité)

  Film (idF, titre, annee, pays, nbspectateurs)

  Jouer (*#idActeur, #idFilm*, salaire) avec l'attribut *idActeur* clé étrangère sur l'attribut *idA* de la relation Acteur et l'attribut *idFilm* clé étrangère sur l'attribut *idF* de la relation Film

Exercice 2. Plans d'exécution (Coûts)

Dessiner l'arbre ou donner l'expression algébrique du plan d'exécution logique correspondant au plan d'exécution de l'exercice 1, puis calculer le coût d'E/S de ce plan d'exécution logique. Vous pouvez poser les hypothèses (sur la sélectivité des différentes conditions de la requête) que vous jugerez utiles au calcul du coût

Proposer des plans d'exécution logique équivalents alternatifs en utilisant les règles de réécriture algébriques et expliquer les avantages et inconvénients par rapport au plan d'exécution logique initial.

Existe-t-il des cas où il est préférable d'exécuter les opérations de jointures avant celles de sélections ? Si oui, lesquels ?

Exercice 3 . Plans d'exécution (Coûts avec Vues et Indexes)

Illustrer comment le coût d'une requête peut être réduit en utilisant :

- un index sur une clé primaire
- une vue matérialisée
- un index de type bitmap
- un index de type join-bitmap

Entrepôts de données

Motivations - Questions ouvertes

1. Quelle est la différence entre une base de données relationnelle et un entrepôt de données ?
2. Pourquoi les bases de données relationnelles ne sont-elles pas adaptées à la gestion des données massives ?
3. Pourquoi a-t-on introduit les plateformes de Big-Data ? Quels sont les avantages et les inconvénients par rapport aux entrepôts de données ?
4. Pourquoi est-il nécessaire d'optimiser l'évaluation des requêtes dans les bases de données relationnelles ? Illustrez l'intérêt de l'optimisation avec un exemple.

Modélisation : schémas en étoile

Exercice 1. Télécommunication (analyse statistiques d'appel)

Un entrepôt de données pour une entreprise de télécommunications

Bouygues télécom (fondée en 1994) est la troisième compagnie de télécommunications française, avec 10 millions d'utilisateurs, une part de marché de 20%, et 5.3 milliards d'euros de profits annuels. En 2007, Bouygues a repensé son système d'aide à la décision. L'analyse des données des clients était un point central. La compagnie souhaite analyser des statistiques d'appel très détaillées.

Travail demandé :

1. Proposer un schéma logique et physique pour l'entrepôt de données.

2. Justifier la (semi, non)-additivité des mesures.

3. Proposer 5 requêtes analytiques pertinentes.

4. Estimer la taille de l'entrepôt sur 10 ans.

Exercice 2. Télécommunication (analyse factures client)

Un entrepôt de données pour une entreprise de télécommunications

Bouygues télécom (fondée en 1994) est la troisième compagnie de télécommunications française, avec 10 millions d'utilisateurs, une part de marché de 20%, et 5.3 milliards d'euros de profits annuels. En 2007, Bouygues a repensé son système d'aide à la décision. La compagnie souhaite mettre en place un système d'analyse des factures client.

Travail demandé :

1. Proposer un schéma logique et physique pour l'entrepôt de données.

2. Justifier la (semi, non)-additivité des mesures.

3. Proposer 5 requêtes analytiques pertinentes.

4. Estimer la taille de l'entrepôt sur 10 ans.

Exercice 3. Modélisation - Questions ouvertes

- Que signifié dimension corrélée ? Comment gère-t-on cela dans la modélisation de l'entrepôt de données. Illustrer avec un exemple.

- Est-il conseillé d'avoir beaucoup de dimensions dans l'entrepôt de données ? Dans quels cas est-il conseillé d'utiliser une mini-dimension ? Illustrer avec deux exemples différents.

- Qu'est une dimension dégénérée? Quelle est son utilité? Illustrer avec un exemple.

- Quels sont les avantages et les inconvénients des modèles en flocon de neige dans le cadre des entrepôts de données? Dans quel contexte particulier pourrait-on envisager leur utilisation ? Illustrer avec un exemple.

- Quel est le rôle d'une table pont (bridge table) ? Illustrer avec un exemple.

Exercice 4. Modèles Snapshot - Deliveroo

Deliveroo, leader de la livraison de repas, souhaite mettre en place un système d'analyse périodique pour suivre l'état des commandes au cours de la journée. Chaque heure, un snapshot est pris pour enregistrer le nombre total de commandes, les commandes en cours, et les commandes annulées. Deliveroo s'intéresse notamment à l'analyse des différents secteurs géographiques et des types de commandes (eg., repas selon catégorie de restaurant, boissons etc).

Travail demandé :

1. Proposer un schéma logique et physique pour l'entrepôt de données.

2. Justifier la (semi, non)-additivité des mesures.

3. Proposer 5 requêtes analytiques pertinentes.

4. Estimer la taille de l'entrepôt sur 10 ans.

Exercice 5. Modèles Snapshot - Audible

Audible, une plateforme de livres audio et de podcasts, souhaite analyser les tendances mensuelles d'écoute de ses utilisateurs. Les données collectées doivent inclure le nombre total d'écoutes, la durée totale d'écoute, et les évaluations des utilisateurs sur une échelle de 1 à 5. La granularité des données est hebdomadaire.

Travail demandé :

1. Proposer un schéma logique et physique pour l'entrepôt de données.

2. Justifier la (semi, non)-additivité des mesures.

3. Proposer 5 requêtes analytiques pertinentes.

4. Estimer la taille de l'entrepôt sur 10 ans.Exercice 5. Updated-Records (DHL)

DHL, leader mondial de la logistique, traite des millions de colis chaque jour et souhaite suivre leur cycle de vie complet. Les étapes clés incluent :

1. Réception dans le réseau : enregistrement et entrée dans le système.
2. Tri dans un centre de distribution : acheminement vers un hub logistique.
3. Expédition : envoi vers la région cible via des hubs intermédiaires.
4. Livraison finale : remise au destinataire ou à un point de collecte.
5. Statut final : livré, en attente (adresse incorrecte), ou retourné.

Travail demandé :

1. Proposer un schéma logique et physique pour l'entrepôt de données.

2. Justifier la (semi, non)-additivité des mesures.

3. Proposer 5 requêtes analytiques pertinentes.

4. Estimer la taille de l'entrepôt sur 10 ans.

Exercice 6. Updated-Records (Projets informatiques)

Une entreprise de services informatiques souhaite suivre le cycle de vie des projets de développement logiciel ou des prestations qu'elle fournit. Ce processus comprend plusieurs étapes clés :

1. Phase initiale : cadrage du projet, collecte des besoins et validation.
2. Développement et tests : réalisation technique et vérifications fonctionnelles.
3. Livraison au client : transfert des livrables finalisés.
4. Support et maintenance post-livraison : résolution des problèmes et évolutions.

L'objectif est d'assurer le suivi complet du cycle de vie d'un projet, d'évaluer la performance, et d'identifier les points d'amélioration : on s'intéresse notamment aux retards liés aux différentes parties.

Travail demandé :

1. Proposer un schéma logique et physique pour l'entrepôt de données.

2. Justifier la (semi, non)-additivité des mesures.

3. Proposer 5 requêtes analytiques pertinentes.

4. Estimer la taille de l'entrepôt sur 10 ans.

Tables-Pont (bridge table) pour Hierarchies

Exercice 1. Hiérarchie de produits

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).

1) Proposer un schéma d'entrepôt de données permettant les analyses suivantes, 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)
5. Pour quelles requêtes avez-vous utilisé la table-pont ?

Exercice 1. Hiérarchie des entreprises et filiales

On souhaite analyser la hiérarchie de propriété entre les entreprises et leurs filiales. Une société mère peut posséder plusieurs filiales, qui à leur tour peuvent posséder d'autres entreprises, créant ainsi une hiérarchie. Le groupe souhaite analyser les revenus des filiales. Ces analyses permettent de prendre des décisions sur leur restructuration. Par exemple, la société mère Alphabet Inc a comme filiales Google LLC, DeepMind, X Development, Verily et Fitbit. À son tour, Google LLC a comme filiales YouTube, Google Cloud, Waymo. Notez que dans certains cas on pourrait avoir des participations croisées aux entreprises, c'est-à-dire, une filiale pourrait être détenue par deux sociétés mères différentes.

Objectif :

1) Proposer un schéma d'entrepôt de données permettant les analyses suivantes, et donner les requêtes SQL correspondantes.

1. Identifier toutes les entreprises détenues directement ou indirectement par une société mère dont le nom est "Alphabet Inc".
2. Donner la liste des sociétés mères aus sommet de la hiérarchie
3. Calculer le chiffre d'affaires de toutes les filiales sous l'entreprise "Google LLC".
4. Donner la liste de toutes les filiales impliquées dans des participations croisées.

Exercice 3 : Hiérarchie des employés

Une entreprise multinationale souhaite analyser la structure de son organisation, en particulier les relations hiérarchiques entre les employés. Chaque employé peut avoir un manager, et les managers peuvent rapporter à des responsables de niveau supérieur, formant ainsi une hiérarchie. L'entreprise souhaite également exploiter cette hiérarchie pour évaluer des indicateurs métier tels que l'efficacité des managers et la productivité des équipes.

Objectif :

Proposer un schéma d'entrepôt de données permettant les analyses suivantes, et donner les requêtes SQL correspondantes.

1. Calculer le salaire moyen des employés sous la responsabilité de chaque manager.
2. Déterminer le chiffre d'affaire des équipes qui dépendent directement ou indirectement de chaque manager
3. Pour quelles requêtes avez-vous utilisé la table-pont ?

Tables-Pont (bridge table) pour Relations N:M

Exercice 1. Groupes d'agents

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 notamment les ventes par secteur (Boutonnet, Facultés, Gare, etc…), le type d'appartements (T2, T3, etc) et la période. On se concentre ici sur les performances des agents de l'agence qui réalisent les ventes. Notamment pour les gros biens, une vente peut être réalisée par N agents, travaillant en équipe. Chaque agent réalise une marge individuelle sur une vente correspondant à un certain pourcentage du montant de la vente.

Proposer un schéma d'entrepôt de données permettant les analyses suivantes, et donner les requêtes SQL correspondantes.

1. Pour chaque vente supérieure au million d'euros, le nombre d'agents impliqués dans la vente.
2. Pour chaque agent, le montant des marges réalisées en 2019.
3. Pour chaque vente, la somme des marges réalisées par les agents impliqués.
4. Pour quelles requêtes avez-vous utilisé la table-pont ?

Exercice 2. Projets et Compétences

Une entreprise de développement logiciel souhaite analyser les compétences mobilisées pour l'analyse des projets réalisés. Chaque projet nécessite plusieurs compétences, et chaque compétence est utilisée dans une certaine proportion pour le projet (ex, 40% programmation java, 30% test réseau). Pour chaque projet on a également des mesures telles que le budget.

Proposer un schéma d'entrepôt de données permettant les analyses suivantes et donner les requêtes SQL correspondantes :

1. Pour chaque projet, la liste des compétences utilisées et leur proportion
2. Pour chaque projet dont le budget est inférieur à 100K euros, la liste des compétences associées
3. Pour chaque compétence, la liste des projets dans lesquels elle est mobilisée, la somme des proportions de coûts de cette compétence en fonction du pourcentage et du budget total associé.
4. Identifier les compétences les plus mobilisées en termes de proportions cumulées sur des projets dépassant le budget de 100K euros.
5. Pour quelles requêtes avez-vous utilisé la table-pont ?

Mises à jour - Évolution de l'entrepôt de données

Exercice 1. Mise à jour des Dimensions (le cas de la dimension Client)

Une entreprise de services en ligne souhaite concevoir une dimension client dans son entrepôt de données pour suivre à la fois les informations de base des clients, comme , le nom, la date d'inscription, et l'adresse actuelle, ainsi que des attributs analytiques spécifiques tels que la tranche d'âge (catégorisation basée sur l'âge du client, par exemple 18-24, 25-34, 35-44), le profil (ex. : Acheteur fréquent ou Acheteur occasionnel) et un indice de confiance (score entre 0 et 100 évaluant la fiabilité, basé sur l'historique des paiements ou des retours). Les tranches d'âge sont recalculées périodiquement en fonction de la date de naissance, le profil évolue en fonction des comportements d'achat, et l'indice de confiance est mis à jour régulièrement.

1. Pour chaque attribut de la dimension, indiquer quelle stratégie utiliser pour gérer les changements des attributs dynamiques (type 1, type 2, type 3).

Exprimer les requêtes suivantes en SQL.

- Chiffre d'affaire des clients dont l'indice de confiance est supérieur à 90% en 2024
- Identifier les clients qui ont changé de profil plus de deux fois au cours des 12 derniers mois.
- Donner un exemple d'instance de données pour lequel en choisissant la stratégie 1 (écrasement de valeur) pour mettre à jour l'indice de confiance du client, on pourrait avoir des fausses réponses à la requête.
- Donner un exemple d'instance de données pour lequel en choisissant la stratégie 3 (historique partiel) pour mettre à jour l'indice de confiance du client, on pourrait avoir des fausses réponses à la requête.

Partitionnement

Exercice 1. Partitionnement par lignes et colonnes pour la dimension Clients

China Mobile, le plus grand opérateur de télécommunications au monde, gère plus d'un milliard de clients répartis dans plusieurs régions géographiques. La dimension Clients de son entrepôt de données contient des informations variées, dont certaines sont mises à jour fréquemment, et d'autres rarement. Pour optimiser les performances des requêtes sur ce volume massif de données, on souhaite mettre en place un partitionnement des données.

1) À l'aide d'SQL mettre en place un partitionnement hybride selon les consignes suivantes.

1. Partitionner par colonnes : Divisez les attributs en deux groupes :
    - Les attributs statiques : rarement mis à jour (pays/region)
    - Les attributs dynamiques : fréquemment mis à jour (indice de confiance)
2. Partitionner par lignes : Divisez les données en fonction de la région géographique des clients (par exemple : Nord de la Chine, Sud de la Chine, etc.).

2) Donner la requête SQL qui sélectionne les utilisateurs avec indice de confiance supérieur à 90% qui habitant dans la région "Nord de la Chine".

Évolution de l'entrepôt de données

Exercice 1. Questions ouvertes.

- Une nouvelle source de données a une granularité différente de celle des données existantes. Comment gérer cette différence ? Illustrer cela à l'aide d'un exemple.

- Si la nouvelle source contient des dimensions supplémentaires non présentes dans l'entrepôt actuel, comment intégrer ces dimensions tout en conservant la cohérence avec les dimensions existantes ? Illustrer cela à l'aide d'un exemple.

- Vous ajoutez un nouvel attribut dans une table dimensionnelle. Comment gérer les lignes existantes pour lesquelles cet attribut n'est pas disponible ? Illustrer avec un exemple.

- Vous ajoutez une nouvelle mesure dans une table de faits. Comment gérer les lignes existantes pour lesquelles cette mesure n'est pas disponible ? Illustrer avec un exemple.

Vues Matérialisées

Exercice 1

1. Considérons un entrepôt de données pour l'analyse des ventes en ligne (Amazon). Traduire les interrogations suivantes en requêtes SQL.
    a. Le nombre de produits par pays achetés après 22h.
    b. Le nombre de produits achetés pour chaque heure de la journée.
    c. Le nombre de produits achetés après 22h par des clients ayant un compte Amazon premium.
    d. Le nombre de produits achetés après 21h par des clients français.
2. Construire le treilli d'agrégation pour le workload {Q1,Q2,Q3,Q4}.
3. Proposer un ensemble de vues matérialisées permettant de répondre aux requêtes.
4. Supposons que la table des faits de l'entrepôt contient 1 milliard de lignes concernant les ventes. Supposons que l'entreprise vend dans 100 pays. Comparez le coût de la requête a (Le nombre de produits par pays achetés après 22h) dans les deux scénarios suivants. 1) Sans vues matérialisées et 2) avec les vues matérialisées.

Index de type "Join"

Exercice 1

Considérons la déclaration de base de données suivante.

- CREATE TABLE
  Dim_table_produit (id VARCHAR2(10), nom VARCHAR2(50), prix NUMBER);
- CREATE TABLE
  Fact_table_ventes (product_id VARCHAR2(10), store_id VARCHAR2(10),
  sale_date DATE, amount NUMBER);
- CREATE BITMAP INDEX idx
  ON Fact_table_ventes(d.product_id)
  FROM Fact_table_ventes f, Dim_table_produit d
  WHERE f.product_id = d.product_id;

Pour chaque requête : (1) donner le plan logique et (2) estimer le coût d'évaluation de la requête

1. SELECT amount FROM Fact_table_sales f, Dim_table_produit d
   WHERE f.product_id = d.product_id and d.product_id = 'p500'
2. SELECT amount FROM Fact_table_sales f, Dim_table_produit d
   WHERE f.product_id = d.product_id and d.price = 50

Hadoop et Map/Reduce

Exercice 1. Questions ouvertes

1. Pourquoi a-t-on besoin de la parallélisation pour exécuter des traitements sur des données massives ?
2. Quels sont les avantages des architectures shared-nothing pour l'exploitation des données massives ? Pourquoi les architectures à mémoire partagée ne sont-elles pas adaptées ?
3. Indiquez trois différences entre les entrepôts de données et Hadoop en termes d'architecture ou de cas d'usage.
4. Illustrer le principe de fonctionnement de HDFS :
    a. Comment se déroule le stockage d'un fichier dans HDFS ? Supposons de devoir stocker un fichier de 1GB avec facteur de réplication 3 et taille du split de 128MB sur un cluster de 10 machines. Comment ce fichier est géré par HDFS?
    b. Comment se déroule l'accès à un fichier dans HDFS ?
5. Donnez un exemple de problème :
    a. Parallélisable avec communication entre nœuds.
    b. Non parallélisable.
6. Expliquez comment les fonctions map et reduce permettent de paralléliser les problèmes. Quel est le rôle du map ? Quel est le rôle du reduce ?
7. Comprendre le fonctionnement de Map/Reduce dans un cluster Hadoop :
8. Dans un cluster de machines exécutant un traitement Map/Reduce, une machine peut-elle exécuter à la fois une tâche map et une tâche reduce ?
    a. Quel module logiciel choisit les machines qui réalisent les tâches map et reduce ?
    b. Quels critères sont utilisés pour réaliser ce choix ?
9. Pratique de Map/Reduce :
    a. Résolvez des problèmes classiques en Map/Reduce (comme ceux vus en travaux pratiques).
10. Quelles sont les limites principales de l'approche Map/Reduce du point de vue de la consommation des ressources dans un cluster ? Comment peut-on atténuer ces problématiques ?

Exercice 2. Programmation Map/Reduce (Taxi)

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
…			

1. Décrire les traitements map/reduce nécessaires pour calculer, pour chaque mois de l'année 2021, le nombre moyen de passagers transportés par jour. 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.*

Exercice 3. Programmation Map/Reduce (Join sur plusieurs tables)

Décrire le fonctionnement d'un programme map/reduce pour calculer la requête
SELECT S.B FROM S, R, T WHERE R.A = S.A AND R.B = T.A

On supposera que R, S, et T sont des relations binaires enregistrées séparément.

	S.csv	
Nom	A	B
S	1	2
S	3	4
	...	

	R.csv	
Nom	A	B
R	3	2
R	2	6
	…	

	T.csv	
Nom	A	B
T	6	3
T	8	7
	…	

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.*