Politique d'utilisation des IA pour le Projet

L'utilisation des IA est permise pour ce qui concerne l'automatisation de tâches répétitives ou de mise en forme liées au projet, rapport et transparents, ou encore pour aider à mieux formuler les besoins analytiques ou corriger une embauche de modélisation. L'idée est que tout informaticien "moderne" soit capable d'utiliser ces outils pour extraire le plein potentiel de son travail ; en même temps l'objectif de ce projet et du module est d'accroître vos connaissance, et cela passe nécessairement par un travail personnel ainsi que par les interactions du travail de groupe. La seule condition imposée dans ce cours est que l'utilisation des IA soit déclarée. Par exemple, si des transparents ont été faits en premier temps, et puis mis en forme avec un assistant d'IA, il faut indiquer "utilisation IA pour création / mise en forme des transparent". Même chose, si on fait rédiger à une IA une partie de notre document, ou réécrire proprement une requête SQL, il faut l'indiquer. Important : utiliser des IA n'a pas vocation à entraîner de pénalisations, au contraire en indiquer l'utilisation permet aux enseignants de mieux évaluer votre travail et à récompenser l'effort et l'originalité des différents projets.

Organisation du projet

Timeline

- 02/10 : validation des groupes et du sujet, création des dépôts git pour déposer les documents de projet et le code (1 groupe = 1 dépôt)
- 09/10 : travail en séance de TD sur Analyse, Traitements et et Schéma
- 16/10 et 23/10 : travail personnel sur Analyse, Traitements et et Schéma
- 03/11 23h59 : rendu intermédiaire = Analyse, Traitements et et Schéma (question 1 à 19 inclus)
- 06/11: retour sur le rendu intermédiaire + travail en séance sur Implémentation
- 13/11 : travail en séance sur Implémentation + rendu sur fiche auto-évaluation
- 18/11 23h59 - rendu rapport final + code + transparents
- 20/11 - Soutenance Mini-projet Entrepôts de Données

Travail à rendre (par groupe de 2 ou 3 personnes) :

- Rapport intermédiaire : analyse des traitements et schéma : question 1 à 19 (3 novembre). Le
- Rapport final
    - questions 20-27 (18 novembre)
    - Révision questions 1 - 19 suite au retour des enseignants
- Transparents de la présentation + code (20 novembre)
- Exposé oral de 10 minutes + 5 minutes de questions (21 novembre)

Bibliographie

- [DW1] The Datawarehouse Toolkit. Kimball, Ross.
- [Git-EDBD] Dépôt git du module avec scripts et exemples. https://gitlab.etu.umontpellier.fr/p00000013857/edbd

Choix du sujet

L'objectif du projet est de concevoir, implémenter et interroger un entrepôt de données. Plus précisément, l'objectif est d'arriver à **mettre en œuvre toutes les techniques de modélisation et d'optimisation vues en cours**.

Le choix du sujet est libre, mais cela doit être lié aux thèmes suivants. Il est bien évidemment possible de proposer d'autres sujets - dont la réalisation d'un entrepôt de données est pertinente.

- Véhicules électriques et infrastructures de recharge (Tesla)
- Intelligence artificielle et apprentissage automatique (OpenAI, DeepMind)
- Santé numérique et télémédecine (Doctolib)
- Cybersécurité (Kaspersky, NordVPN)
- Services de cloud computing (Amazon Web Services (AWS), Microsoft Azure)
- Impression 3D (Stratasys, 3D Systems)
- Drones et véhicules autonomes (DJI, Waymo)
- Plateformes d'économie collaborative (Uber, Bolt)
- Plateformes de travail à distance et collaboration en ligne (Zoom, Slack)
- E-learning et formations en ligne (Coursera, OpenClassrooms)
- Industrie spatiale privée (SpaceX, Blue Origin)
- Services de livraison de repas à domicile (Uber Eats, Deliveroo)
- Services de streaming en direct (Twitch, Facebook Live)
- Plateformes de podcasts et livres audio (Audible)
- Réseaux Sociaux (Instagram, Twitter, Facebook, LinkedIn...)
- Journalisme Numérique (LeMonde, ...)
- Vidéo ou musique à la demande (Youtube, Netflix, Amazon, Spotify...)
- Commerce électronique (LDLC...)
- Expéditions et livraisons (DHL...)
- Sites d'annonces commerciales (LeBonCoin...)
- Infrastructures et services : télécommunications, eau, électricité EDF, SFR ...)
- Énergies renouvelables (solaire, éolien, hydroélectricité) (Vestas, First Solar)

- Réseaux de transport : aérien, terrestre, nautique (SNCF, ...)
- Jeux videos en-ligne (Poker...)
- Gestion de ressources humaines, recrutements, inscriptions (Université, Entreprise)
- Chaînes de magasins ou cinémas (Carrefour, Gaumont...)
- Industrie (secteur automobile, pharmaceutique, ...)
- FinTech (services bancaires en ligne, paiements mobiles, cryptomonnaies) (Revolut)
- Internet des objets (IoT) (Philips Hue, Nest)

Attention : votre sujet doit impérativement

1. être validé par les enseignants
2. être unique (différent de celui des autres groupes)

Liste de sujets “banned” qu’on ne peut pas choisir car trop proches du cours : Amazon.

Tâches à réaliser

(prenez le temps de lire chaque section attentivement)

1. Analyse des besoins métier

La mise en place d'un entrepôt de données dépend exclusivement des besoins métier. C'est à dire, des requêtes analytiques permettant à une entreprise d'améliorer sa position dans le marché (e.g., la vente d'un produit) ou encore la logistique et la gestion de son fonctionnement interne (e.g., les inventaires, la chaîne de production). Il est donc impératif de ne pas stocker dans un entrepôt toutes les données disponibles dans les différentes sources (sauf pour des cas rares), et de l'alimenter seulement avec les données nécessaires aux requêtes analytiques. En outre, vu le coût de mise en place d'un système d'aide à la décision, il est obligatoire de modéliser avec une grande précision les actions / opérations dont l'analyse est la plus rentable (e.g., augmenter les ventes en proposant des promotions de produit pour une population ciblée).

L'analyse des besoins métier impacte les requêtes analytiques, et les requêtes analytiques impactent les schémas en étoile de l'entrepôt de données. Ainsi, sans une bonne analyse des besoins métier, on peut être amenés à revenir sur notre modélisation ou implémentation et à refaire une partie importante du travail.

L'analyse des besoins métier aura un poids très important dans l'évaluation du projet.

En cas de manque d'inspiration, allez à la recherche d'éléments pour le cas à étudier dans le Web ou encore dans [DW1] !

Travail demandé pour cette partie.

1. Réaliser une analyse complète du cas considéré (au moins 1 page). Quels sont les objectifs de l'entreprise (ou institution) considérée dans votre sujet ? Quelle est sa position sur le marché ? Quels services ou produits propose-t-elle ? Quelles sont ses formes de revenu ? Quelles informations seraient utiles pour la prise de décision au sein de l'entreprise ?
2. Indiquez les actions / opérations (e.g., ventes, livraisons) à tracer pour récupérer ces informations.
3. Pour chaque action / opération, proposez au moins trois traitements possibles (i.e., requêtes analytiques). Montrer comment chaque requête permet d'aider à la prise de décision sur le sujet, c'est-à-dire, pointer le besoin métier auquel contribue cette requête.
4. Ordonnez les actions par ordre d'importance / rentabilité potentielle (e.g., augmentation des ventes vs. utilisation optimale de l'espace de stockage dans le magasin).
5. Identifiez les deux actions / opérations les plus importantes à analyser.
6. La conception de l'entrepôt dépendra exclusivement des traitements que vous avez indiqués comme les plus importants..

2. Conception du premier modèle détaillé

Il s'agit maintenant de proposer un premier datamart permettant d'analyser le besoin métier le plus important. Dans cette partie on met en œuvre les notions de modèle en étoile, faits, dimensions, attributs et mesures.

Travail demandé pour cette partie.

7. Concevez un data-mart (c'est-à-dire, un modèle en étoile) pour l'action / opération la plus importante. Prévoyez une modélisation détaillée dans la table des faits.
8. Pour le modèle, prévoir au moins 5 dimensions. Pour chaque dimension, prévoir une dizaine d'attributs.
9. Pour la table des faits, indiquer la liste des mesures. Pour chaque mesure, indiquer s'il s'agit d'une mesure additive, semi-additive, ou non-additive. S'il s'agit d'une mesure semi-additive indiquer la dimension qui fait que l'addition des valeurs n'ait pas de sens.
10. Est-il possible de répondre au principal besoin métier que vous avez indiqué avec le modèle que vous avez mis en place ? Expliquez pourquoi et comment.
11. Pour tester la pertinence de votre modèle, donner un exemple d'instance de l'entrepôt de données (2 ou 3 lignes par table suffisent).
    Estimez la taille du datamart (en terme du nombre de lignes) sur 12 mois. Est ce que cette taille justifie la mise en place d'un entrepôt de données ou par exemple, pourrait-on tout gérer avec un fichier excel ? Justifiez votre réponse.
12. Vérifier toutes les mesures indiquées dans le datamart.

3. Conception de deux modèles moins détaillées

Il s'agit maintenant de proposer deux autres datamarts permettant d'analyser le besoin métier secondaires. Dans cette partie l'objectif est de mettre en œuvre les notions de modèle snapshot et updated records.

Travail demandé pour cette partie.

13. Concevez un datamart snapshot, pour une (1) des deux actions / opérations les plus importantes à analyser. Il est à vous de choisir l'action la plus pertinente. Dans le rare cas qu'aucune des deux deux actions / opérations les plus importantes se prêtent à cette modélisation, vous pouvez choisir une troisième action moins importante à analyser.
14. Concevez un datamart updated-records, pour une (1) des deux actions / opérations les plus importantes à analyser. Il est à vous de choisir l'action la plus pertinente. Dans le rare cas qu'aucune des deux deux actions / opérations les plus importantes se prêtent à cette modélisation, vous pouvez choisir une troisième action moins importante à analyser.
15. Pour chaque modèle, prévoir au moins 5 dimensions. Pour chaque dimension, prévoir une dizaine d'attributs.
16. Pour chaque table des faits, indiquer la liste des mesures. Pour chaque mesure, indiquer s'il s'agit d'une mesure additive, semi-additive, ou non-additive. S'il s'agit d'une mesure semi-additive indiquer la dimension qui fait que l'addition des valeurs n'ait pas de sens.
17. Est-il possible de répondre aux besoins métier que vous avez indiqué avec le modèle que vous avez mis en place ? Expliquez pourquoi et comment.
18. Pour tester la pertinence de vos modèles, donner un exemple d'instance de l'entrepôt de données (2 ou 3 lignes par table suffisent).
19. Estimez la taille des deux datamarts (en terme du nombre de lignes) sur 12 mois. Est ce que cette taille justifie la mise en place d'un entrepôt de données ou par exemple, pourrait-on tout gérer avec un fichier excel ? Justifiez votre réponse.

4. Conception - Techniques Avancées de Modélisation

Dans cette partie, il s'agit de mettre en œuvre des techniques de modélisation avancées. Plus précisément, on se concentrera sur l'utilisation de tables pont pour modéliser hiérarchies et relations N:M, de techniques pour gérer l'évolution des dimensions, et de techniques de partitionnement.

Travail demandé pour cette partie.

20. Montrer comment une table-pont pourrait aider à modéliser une hiérarchie ou une association N:M dans votre cas d'étude. Concevoir la table-pont et ensuite donner une requête analytique qui utilise la table pont. Hiérarchies et associations N:M sont moins fréquentes mais se présentent régulièrement dans des cas pratiques. En cas de manque d'inspiration, allez à la recherche d'éléments dans le Web ou encore dans [DW1] !
21. Considérer la dimension la plus volumineuse de votre entrepôt de données. Pour cette dimension, indiquer les attributs statiques (qui ne sont jamais mis à jour) et dynamiques (qui peuvent être mis à jour).
22. Pour chaque attribut dynamique (de la dimension plus volumineuse de votre entrepôt de données), donner un exemple de mise à jour possible des valeurs contenues dans l'attribut. Ensuite, pour chaque attribut indiquer quelle est la meilleure stratégie pour la gestion des mises à jour de ses valeurs (type 1, type 2, type 3).
23. Mettez en place un schéma de partitionnement hybride (toujours pour la dimensions la pus volumineuse de votre entrepôt de données). Le partitionnement par colonne doit se faire selon l'évolutivité des attributs. Le partitionnement par colonne doit se faire par un critère qui doit permettre d'améliorer une partie des requêtes analytiques de l'entrepôt. Expliquer en quoi ce partitionnement entraîne moins de lectures de données lors de l'évaluation des requêtes. Pour familiariser avec la création des schémas de partitionnement, vous pouvez consulter le script qui a été mis à disposition dans le dépôt git associé au cours [Git-EDBD].

5. L'implémentation et le requêtage, les Vues Matérialisées

24. Implantez en ORACLE les tables de faits et des dimensions prévues par votre modèle. (En cas de problème avec ORACLE utiliser SQL Live ou installer POSTGRES : www.postgresql.org/download)
    a. Utilisez des vues virtuelles pour les dimensions partagées.
25. Proposez 8 requêtes analytiques correspondant aux traitements que vous avez indiqués.
26. Donnez l'ensemble des vues matérialisées permettant de répondre à l'ensemble de vos requêtes. Utiliser la technique basée sur le treillis d'agrégation présentée en cours. Si possible, éviter les deux scénarios suivants :
    - une vue qui répond à toutes les requêtes : cela signifie que soit la vue contient trop d'attributs (et donc la vue perd d'intérêt), soit les requêtes proposées sont très peu variées (et done les requêtes perdent d'intérêt)
    - chaque requête est adressée par une vue différente : ce cas n'est pas souhaitable en pratique, car lorsqu'on investit (calcul, stockage, maintenance) dans la matérialisation d'une vue, on souhaite que plusieurs requêtes puissent en bénéficier
27. Mettre en place un index de type bitmap join permettant d'optimiser deux (2) requêtes analytiques parmi celles que vous avez proposé. Pour familiariser avec la création de ces indexes, vous pouvez consulter le script qui a été mis à disposition dans le dépôt git associé au cours [Git-EDBD].

Le Rapport

L'exposé est un moment de restitution de l’ensemble de votre travail. Voici des lignes guide pour la rédaction de votre rapport.

- Le rapport doit être clair, complet, mais en même temps concis.
- La taille maximale du rapport est de 10 pages. Par contre, il n'y a pas de contrainte pour ce qui concerne la taille de l'annexe ou vous pouvez détailler tout élément le nécessitant. Attention que le rapport est censé être lisible même sans consulter l'annexe ; le rôle de l'annexe est seulement de permettre d'approfondir les informations présentées dans le corps du rapport.
- Éviter de faire référence au code : par exemple, inclure au moins un exemple en SQL pour les requêtes analytiques et les vues (en cas de manque d’espaces ces éléments peuvent aller dans l’annexe)
- Si vous utilisez un assistant d'IA pour la rédaction de votre rapport, vérifiez le texte généré pour supprimer toute information redondante ou erronée.
- Voici quelques points pour vous aider à vérifier que votre rapport soit bien complet.
- Le rapport est censé illustrer comment vous avez répondu à chaque question du sujet ; une section est attendue pour chaque question.
- Ne pas oublier la partie sur la volumétrie de l'entrepôt de données

L'exposé

L'exposé est un moment de restitution de l'ensemble de votre travail. Voici des lignes guide pour la réalisation des transparents.

- pour un exposé de 10 minutes, il faut en général prévoir entre 5 et 7 transparents (titre et bibliographie exclus).
- utiliser un format portable : PDF
- les transparents doivent être lisibles : éviter les polices trop petites et les diagrammes trop denses (dans ce dernier cas pensez à réarranger ou découper vos modèles).
- éviter les transparents "sans contenu" (exemple limite : un transparent vide avec le seul titre "Bilan"), et plus en général éviter de parler trop longtemps sans vous appuyer sur des éléments du transparent projeté.
- éviter toute animation, sauf si vraiment nécessaire (et même dans ce cas, éviter)
- il n'est pas demandé d'inclure un transparent sur la gestion du projet ou les outils collaboratifs que vous avez utilisés.
- vue la contrainte sur la durée de l'exposé, il se peut qu'il soit impossible de présenter l'intégralité du projet. Dans ce cas, indiquer des éléments techniques intéressants qui seront développés plus en détail que les autres.
- en même temps, il est important d'utiliser tout le temps à disposition et de couvrir le maximum d'aspects possibles (analyse du cas, modélisation, requêtes analytiques, estimation de la taille de l'entrepôt, implémentation, vues).
- beamer (latex) est une ressource si bien maîtrisé, mais en même temps il pourrait vous contraindre trop au niveau de l'organisation de vos transparents et vous ralentir dans la préparation de l'exposé.
- éviter de fair référence au rapport : les transparents doivent être autonomes et indépendants de celui ci
- Ne pas oublier d'inclure des exemples (au mois 1) de requêtes analytiques et de vues dans vos transparents (texte + SQL)

Annexe

A - Vérification des Mesures

Les mesures constituent un élément basique d'un entrepôt de données. À l'issue de la phase de modélisation et de discussion du rendu intermédiaire avec l'équipe des enseignants l'objectif visé et attendu est qu'il n'y ait aucune erreur concernant les mesures de l'entrepôt de données. En effet, cela aurait peu de sens de continuer avec la mise en ouvre de l'entrepôt si les mesures ne sont pas pertinentes.

Voici des questions pour vérifier la correction des mesures

Les mesures hors-place :

- dans les faits :
    i. pourquoi la mesure X est incluse dans le fait F ?
    ii. est ce que la mesure est pertinente/compatible avec les dimensions du schéma en étoile ?
    iii. est ce que la mesure est utilisée par les requêtes analytiques ? Si oui, quelle information apporte cette mesure ?
- dans les dimensions :
    iv. Pourquoi l'attribut Y est inclus dans la dimension D ? Ne serait-il pas plutôt une mesure ?

La classification des mesures

28. la mesure X dans le fait F est indiquée comme semi-additive : quelle est la dimension qui détermine la semi-additivité ?