Vous êtes Data Analyst chez Laplace Immo, un réseau national d'agences immobilières. Le directeur général est sensible depuis quelque temps à l'importance des données, et il pense que l'agence doit se démarquer de la concurrence en créant un modèle pour mieux prévoir le prix de vente des biens immobiliers.

Ce projet stratégique est appelé en interne le projet "DATAImmo". La CTO de l'entreprise, Clara Daucourt, a la responsabilité de conduire le projet.

Dans ce cadre, elle vous a confié la modification de la base de données permettant de collecter les transactions immobilières et foncières en France. Vous utiliserez ensuite cette base pour analyser le marché et aider les différentes agences régionales à mieux accompagner leurs clients.

À votre arrivée ce matin, vous avez reçu un e-mail de la part de Clara, qui donne plus de détails sur ce qui est attendu de votre part.

Hello,

Afin d'avancer sur le projet DATAImmo, je prévois une première réunion pour valider la modification de la base de données.

Tu trouveras en pièce jointe un fichier zip avec les données suivantes :

- Des données extraites du site open data des Demandes de valeurs foncières (DVF)
- Des données de l'INSEE avec les résultats des recensements de la population
- Des données de data.gouv sur les régions, avec le référentiel géographique français, communes, unités urbaines, aires urbaines, départements, académies, régions

D'ici la réunion, j'aurais besoin que tu travailles sur ces sujets :

1. Il faut que tu prépares le dictionnaire des données en respectant le template en PJ pour répertorier et décrire les données importantes à stocker car nous avons omis de le faire. Il faut le remplir pour les trois fichiers de données.
2. Ensuite, peux-tu modifier le schéma relationnel (aussi en PJ) pour qu'il prenne en compte les nouvelles données région et population ? Ça nous permettra de bien visualiser les différentes entités, associations et cardinalités de la base de données. Enfin, il faudra que tu me présentes ce nouveau schéma relationnel normalisé (il doit suivre la norme 3NF) de la base de données qui donnera lieu à la création des nouvelles tables.
3. Peux-tu nous assurer que nos fichiers soient bien conformes au RGPD ?

Tu peux préparer une présentation avec les différentes informations qu'on validera ensemble lors de notre réunion, afin que tu puisses avancer sur la création de la base de données.

Je reste à ta disposition en cas de besoin.

Bien à toi,

Clara Daucourt
CTO

Étape 1 :

Avant de démarrer cette étape, je dois avoir :

- téléchargé et analysé les données. Cela signifie que vous devez avoir identifié les données non-renseignées, leur type et leur signification.
- identifié les colonnes qui devront absolument être présentes et celles qui devront être supprimées dans la base de données pour répondre aux demandes et aux contraintes RGPD.

Une fois cette étape terminée, je devrais avoir :

- le dictionnaire de données avec les données importantes, conforme à la réglementation RGPD.

Recommandations :

- Souvenez-vous que chaque ligne du dictionnaire doit être définie par :
  - un code
  - une signification
  - un type
  - une longueur
  - une nature
  - une règle de gestion
  - une règle de calcul (si nécessaire)

Ressources :

- Pour en savoir plus sur le dictionnaire de données :
  - Une ressource d'une autre formation qui présente ce qu'est un dictionnaire de données
- Pour vous aider dans la compréhension des variables :
  - Notice descriptive des fichiers « Demande de valeurs foncières»

Étape 2 :

Avant de démarrer cette étape, je dois avoir :

- compris la 3eme forme normale ;
- savoir définir une clé primaire, une clé étrangère et la notion de contrainte.
- pour chaque table du schéma relationnel, je dois avoir identifié :
    - la clé primaire ;
    - la/les clés étrangères ;
    - la nature des relations avec les autres tables.

Une fois cette étape terminée, je devrais avoir :

- le schéma relationnel mis à jour avec les informations supplémentaires. Vous pouvez choisir l'outil de modélisation que vous voulez comme SQL Power Architect, Draw.io ou Looping par exemple.

Recommandations :

- Le choix de vos clés primaires et de vos clés étrangères devra être justifié, ces notions étant fondamentales pour la cohérence de votre base de données. Toutes les données devront être présentes dans ce nouveau modèle.

Points de vigilance :

- Attention aux redondances d'information (Par exemple : redondance entre la voie et le code_voie ou departement et Code_departement).
- Parfois les clés de votre base de données peuvent être une concaténation des différentes données. Dans le cas de ce projet, la clé id_codedep_codecommune est la concaténation de deux variables (Code département et Code commune). Par exemple, pour la commune de Montpellier, nous avons les informations suivantes :
    - Commune : Montpellier
    - Code Département : 34
    - Code commune : 172

La clé id_codedep_codecommune sera donc 34172 (quand le code département est 34000).

Voici pourquoi une nouvelle clé au lieu du code département a été utilisée :

1. L'information du code département n'est pas disponible dans les fichiers de data.gouv et de l'INSEE.

2. Un code département peut contenir plusieurs communes, ce n'est donc pas une clé unique (vous pouvez faire le test en regardant le code postal 05100 qui a pour communes Briançon, Nevache et Montgenevre).

Ressources :
Pour comprendre simplement les étapes de création d'un schéma relationnel :

- Fiche sur les MPD sur le site base-données.com

Étape 3 :

Avant de démarrer cette étape, je dois avoir :

- créé la base de données dans l'outil système de gestion de base de données (SGBD).

Une fois cette étape terminée, je devrais avoir :

- une base de données opérationnelle ;
- des tables qui sont en adéquation avec le schéma relationnel normalisé (nom des tables, des colonnes, des types de variables, etc.) ;
- des liaisons entre les tables avec des clés primaires / étrangères.

Recommandations :

- La création des tables dans la base de données peut se faire soit :
    - avec un script SQL utilisant CREATE TABLE, le script peut être :
        - écrit à la main ;
        - généré automatiquement via un outil comme SQL Power Architect.
    - en important les fichiers CSV directement, fait à l'étape suivante.
- Les différentes clés et contraintes devront être implémentées sur vos tables quelque soit la méthode de création choisie.

Points de vigilance :
Il est important que les types de vos colonnes soient bien définis pour éviter le rejet pendant le chargement. Par exemple, si une donnée est définie comme numérique dans votre table, elle devra être numérique dans le fichier .csv également.

Ressources :

- Tutoriel pour créer une base de données SQLite avec SQLite studio réalisé par un mentor OpenClassrooms

Deux semaines plus tard

Une fois cette première partie de votre travail effectuée, vous présentez votre travail à Clara qui valide votre schéma relationnel ainsi que le dictionnaire des données.
À ce stade du projet, vous pouvez valider avec votre mentor que votre conception est correcte avant d'aller plus loin.
Félicitations ! Elle vous envoie un e-mail à la suite de cette réunion.

Hello,

Félicitations pour cette première étape de conception ! Comme on en a parlé en réunion ce matin, on est bons pour partir sur la modification de la base de données. Il faut maintenant que tu implémentes les tables dans la base de données en respectant ce qu'on s'est dit ce matin (cf. compte rendu de réunion en pièce jointe).

On a échangé en interne, et on pense qu'un outil comme SQLite est pertinent pour ce type d'implémentation. Si tu es plus à l'aise, tu peux essayer d'implémenter un outil comme mySQL ou postgreSQL. N'hésite pas à contrôler ta BDD avec une requête pour vérifier le nombre de lignes par exemple, il faut qu'elle soit opérationnelle et conforme au RGPD.

Une fois que tu auras fait ça, tu pourras faire les requêtes pour extraire les données dont nous avons besoin. Pour ça, tu peux compléter le document de présentation que tu trouveras en pièce jointe. Tu pourras rajouter tous les résultats des requêtes ainsi que la requête associée. Utilise des alias pour que ça soit plus lisible. Tu trouveras toutes les demandes auxquelles il faudra répondre dans le CR de réunion.

Bon courage !

Étape 4 :

Avant de démarrer cette étape, je dois avoir :

- la base de données dans l'outil système de gestion de base de données (SGBD) ;
- préparé les données en utilisant :

    - soit Excel : vous devez alors créer les différents fichiers .csv, en prenant soin d'avoir incrémenté les clés étrangères et supprimé les éventuels doublons. Petite astuce pour les clés étrangères, si votre ordinateur a suffisamment de mémoire, vous pouvez utiliser la fonction excel RECHERCHEV.
    - soit Power query : vous pouvez dans ce cas vous aider de cette ressource de Microsoft Learn.
    - il est aussi possible de créer l'ensemble de la base en SQL uniquement. Cette solution est plus complexe car il vous faudra importer tout le fichier Excel dans votre SGBD et faire toutes les requêtes nécessaires pour créer l'ensemble des tables. Il ne vous restera ensuite qu'à importer les fichiers .csv dans les tables.

Une fois cette étape terminée, je devrais avoir :

- la base de données opérationnelle avec l'ensemble des données ;
- vérifié l'intégrité de ma base de données en m'assurant que l'ensemble des données est bien présent.

Recommandations :

- Vos fichiers Excel devront correspondre parfaitement aux tables de votre modèle physique de données.
- Il ne faut pas oublier de créer les clés de votre base de données. Dans notre projet, il y a plusieurs façons de créer les clés :
    - Auto increment (automatique ou manuel) pour les clés (id_bien et id_vente)
    - Concaténation de plusieurs variables pour id_codedep_codecommune
    - Variable déjà présente que nous transformons en clé avec id_region
- Vous devez vérifier que l'intégralité des données est bien chargée, pour cela vous pouvez :
    - comparer le nombre de lignes chargées lors de l'intégration de vos données pour toutes les tables ;
    - lire le nombre de lignes dans votre base de données (si votre SGBD vous le permet) ;
    - faire une requête SQL pour donner le nombre de lignes dans la base de données.

Points de vigilance :

Il est important de vérifier que l'ensemble des données est présent. Dans le cas contraire, il sera important d'en comprendre les raisons afin de pouvoir le justifier (attention cela peut motiver un refus lors de la soutenance).

Ressources :
Pour vous aider à importer et exporter un .csv avec MySQL et PostgreSQL :

- Un tutoriel rédigé par un mentor

Étape 5 :

Avant de démarrer cette étape, je dois avoir :

- la base de données opérationnelle et chargée avec l'ensemble des données ;
- les clés primaires et étrangères implémentées également.

Une fois cette étape terminée, je devrais avoir :

- un document avec une requête par demande et son résultat.

Recommandations :

- Relisez bien les différentes questions afin de décomposer les demandes.
- Utilisez dans votre code SQL :
    - des alias pour rendre la lecture plus facile ;
    - des sous requêtes ou des tables temporaires.
- Sauvegardez systématiquement les requêtes qui fonctionnent et les résultats associés aux requêtes.

Ressources :

- Pour connaître les principales commandes SQL : le site SQL.sh

Étape 6 :

Avant de démarrer cette étape, je dois avoir :

- terminé l'intégralité des requêtes SQL.

Résultats attendus :

- un support de présentation contenant les différentes demandes :
    - Le contexte du projet
    - La stratégie de sauvegarde et la vérification de la conformité RGPD
    - Les données initiales
    - Un extrait du dictionnaire des données
    - Le schéma relationnel normalisé
    - La base de données avec les tables créées et les données chargées

    - Le code SQL et le résultat des différentes requêtes permettant de répondre aux questions

Vérifications :

- Complétez cette fiche d'autoévaluation**.**

Recommandations pour préparer votre soutenance :

- Préparez-vous bien à la soutenance en vous **entraînant** (chronométrez la soutenance pour être sûr de ne pas dépasser le temps imparti).
- Soyez capable d'expliquer tout le cheminement du projet en commençant par les demandes du projet, en passant par les fichiers Excel, les étapes de création de la base de données et les différents résultats.
- N'oubliez pas que vous êtes un analyste, vous devez donc être capable de commenter les différents résultats des requêtes (vous ne devez pas vous contenter de commenter le code SQL).
- Votre base de données doit être fonctionnelle afin de vous préparer à tester les 3 requêtes que votre évaluateur va vous demander en live.
- Adaptez votre vocabulaire et votre posture à votre interlocuteur.