17/12/23

Analyse du Secteur Culturel Américain

FERNANDES THOMAS
KENNICHE VANESSA

TABLE DES MATIERES

I- CRÉATION DES TABLES 2
A. Description du projet 2
B. Traitement des données et création des tables 2
C. Peuplement des tables 4
II- Requêtes SQL 5
A. Requêtes LID simples, avec jointures et sous-requêtes 5
B. Requêtes LMD 5
C. Requêtes de synthèse 5
D. Requêtes complexes 6
III- Mise à disposition de la base de données et représentation des requêtes sur une carte 6

I- CRÉATION DES TABLES

A. Description du projet

Notre étude se concentre sur l'analyse des données du secteur culturel aux États-Unis, basée sur un ensemble de données fourni par l'"Institute of Museum and Library Services". Ce jeu de données offre des informations détaillées sur les musées ainsi que les organisations connexes telles que les zoos et les aquariums. Notre analyse porte principalement sur la répartition géographique des établissements culturels, la diversité des types de musées et autres organisations culturelles à travers les différentes régions, états et villes. Nous analyserons également la dimension financière de ces établissements.

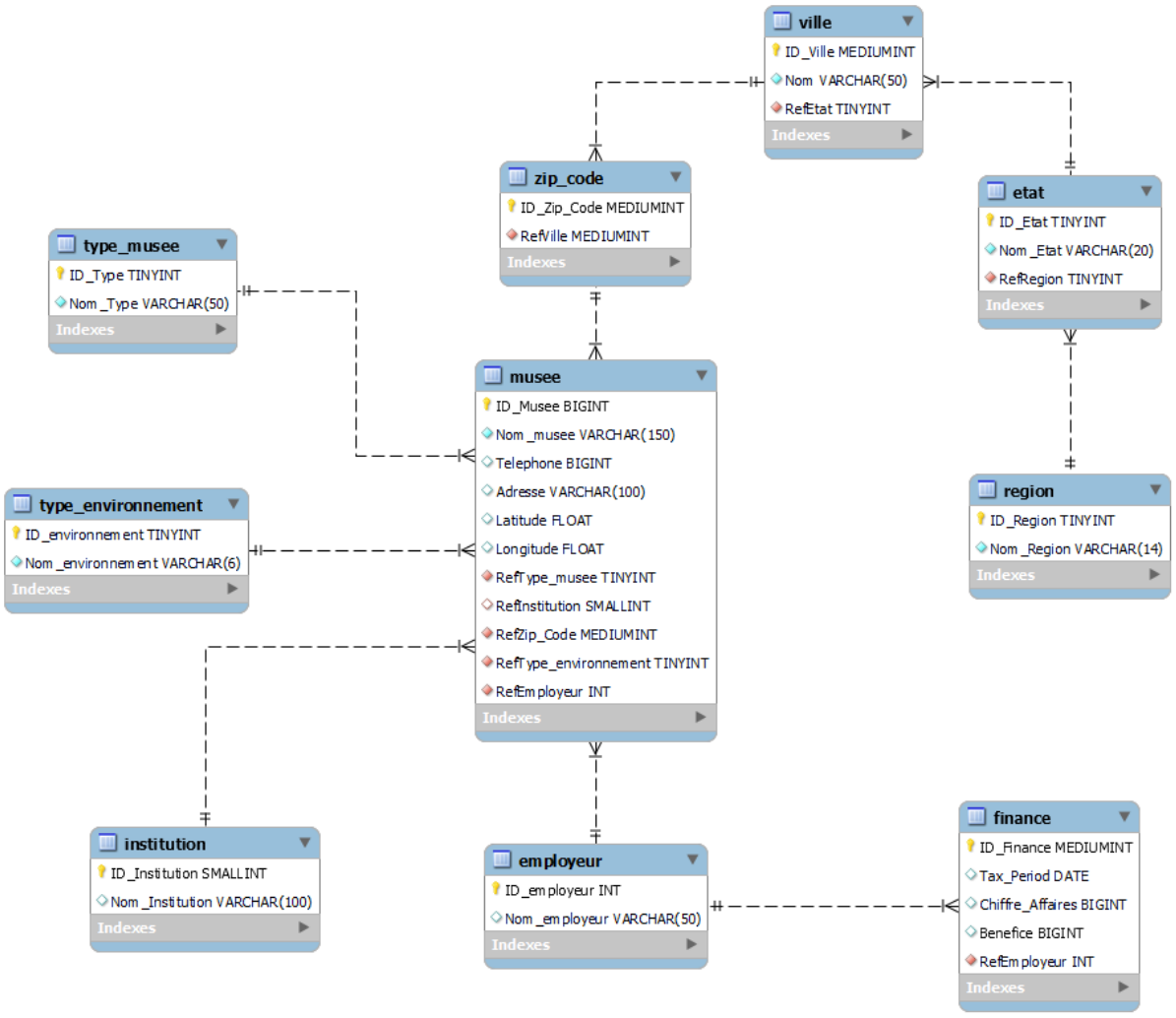
B. Traitement des données et création des tables

Notre ensemble de données de départ comprenait 25 variables. Nous avons observé que plusieurs de ces variables étaient redondantes. Par exemple, nous avions à la fois 'Museum Name' et 'Legal Name', qui, bien qu'étant techniquement distinctes, ne présentaient pas d'intérêt significatif pour notre analyse en étant séparées. Par conséquent, nous avons décidé de conserver uniquement 18 de ces variables, en éliminant celles qui étaient superflues ou se chevauchaient dans le contexte de notre étude.

Variables	Description
Museum ID	ID officiel utilisé par le gouvernement américain pour identifier les musées
Museum Name	Nom du musée
Museum Type	Type ou discipline auquel le musée se rattache (Histoire, Sciences, etc.)
Institution Name	Institution auquel appartient le musée. Ce sont en grande majorité des universités.
Street Address	Adresse
City	Ville du musée
State	Etat du musée
Zip Code	Adresse postale du musée
Phone Number	Numéro de téléphone
Latitude	Coordonnées géographiques
Longitude	Coordonnées géographiques
Locale Code	Code représentant le type d'environnement dans lequel le musée est situé (ville, banlieue, village ou campagne)
State Code	ID de l'etat selon la nomenclature américaine
Region Code	ID de la région selon la nomenclature américaine. Une région représente un ensemble d'etat géographiquement voisins. (Midwest, Western, etc.)
Employer ID	Numéro officiel de l'employeur possédant le musée.
Tax Period	Date à laquelle les informations fiscales ont été enregistrées.
Income	Chiffre d'affaires
Revenue	Bénéfices

Pour créer les tables constituant notre base, nous avons fait en sorte que ces dernières respectent les 3 premières formes normales afin d'assurer l'intégrité, la cohérence et l'efficacité de nos données.

- Première forme normale (1NF) : Chaque table a été conçue pour avoir des colonnes atomiques, évitant ainsi les groupes répétitifs et garantissant que chaque champ contient des valeurs uniques et indivisibles. Par exemple, dans la table musee, chaque musée est identifié par un ID_Musee unique, et toutes les informations sont stockées dans des colonnes séparées et bien définies.
- Deuxième forme normale (2NF) : Nous avons éliminé la redondance des données en décomposant les données en tables supplémentaires et en utilisant des clés étrangères. Par exemple, les informations relatives aux régions et aux états sont stockées dans des tables distinctes (region et etat) et sont liées à d'autres tables via des clés étrangères. Cela permet non seulement d'éviter les redondances mais aussi de faciliter les mises à jour des données.
- Troisième forme normale (3NF) : Pour assurer que les données de chaque table soient non seulement reliées à la clé primaire, mais également indépendantes entre elles, nous avons éliminé toute dépendance transitive. Par exemple, pour la table finance, chaque enregistrement est directement associé à l'ID_Finance, qui est la clé primaire. Toutes les autres informations, telles que le Chiffre_Affaires, le Benefice, et la Tax_Period, sont exclusivement dépendantes de cette clé primaire, sans aucune dépendance intermédiaire.



1. type_musee(**ID_Type**, Nom_Type) : Stocke les différents types de musées, avec un identifiant unique et le nom du type.

2. institution(**ID_Institution**, Nom_Institution) : Contient les informations des institutions gérant les musées, avec un identifiant unique et le nom de l'institution. Ils sont NOT NULL car on ne peut pas définir une nouvelle institution sans donner son nom.
3. type_environnement(**ID_Environnement**, Nom_Environnement) : Répertorie les catégories d'environnement où se situent les musées, avec un identifiant unique et le nom de l'environnement.
4. region(**ID_Region**, Nom_Region) : Liste les régions géographiques, avec un identifiant unique et le nom de la région.
5. etat(**ID_Etat**, Nom_Etat, *RefRegion*) : Définit les états avec un identifiant unique, le nom de l'état et une référence à la région dans laquelle il se situe.
6. ville(**ID_Ville**, Nom, *RefEtat*) : Contient les informations sur les villes, avec un identifiant unique, le nom de la ville et une référence à son etat.
7. zip_code(ID_Zip_Code, *RefVille*) : Associe un code postal à une ville, avec un identifiant unique pour le code postal et une référence à sa ville.
8. employeur(ID_Employeur, Nom_Employeur) : Stocke les informations sur les employeurs des musées, avec un identifiant unique et le nom de l'employeur.
9. finance(ID_Finance, Tax_Period, Chiffre_Affaires, Benefice, *RefEmployeur*) : Gère les informations financières, avec un identifiant unique, la période fiscale, le chiffre d'affaires, le bénéfice et une référence à l'employeur.
10. musee(ID_Musee, Nom_Musee, Telephone, Adresse, Latitude, Longitude, *RefType_Musee*, *RefInstitution*, *RefZip_Code*, *RefType_Environnement*, RefEmployeur) : Détaille chaque musée, avec un identifiant unique, le nom, les coordonnées, les adresses, les références aux types de musée, à l'institution, au code postal, au type d'environnement, et à l'employeur.

Nous avons choisi de séparer les tables finance et musee en raison de la structure particulière de notre jeu de données. Dans notre base, les informations financières sont associées non pas à chaque musée individuellement, mais à leur employeur respectif. Ce choix est dicté par le fait que dans nos données, un employeur peut être responsable de plusieurs musées. Cependant, les détails financiers ne sont pas fournis pour chaque musée séparément, mais globalement pour l'employeur. Par conséquent, il n'est pas possible de déterminer les performances financières d'un musée spécifique, mais uniquement celles de l'employeur qui gère plusieurs établissements.

C. Peuplement des tables

Une fois nos tables créées, nous nous sommes attaqués à la tâche de les peupler avec les données. Notre base de données provient d'un unique fichier CSV, ce qui a nécessité un traitement préalable avant de pouvoir importer ces données. Pour ce faire, nous avons utilisé le langage R et le package dplyr, spécialisé dans la manipulation de données. Durant cette phase de traitement, nous avons dû éliminer environ 5 000 lignes. La majorité de ces lignes étaient soit erronées, soit insuffisamment détaillées pour être utilisées. Nous avons également pris soin d'uniformiser les données en convertissant par exemple les dates dans un format compatible avec MySQL et en standardisant les noms des états pour éviter les doublons et les incohérences (par exemple, "DC", "District Columbia" et "District of Columbia" ont été unifiés sous un seul nom). En outre, des contrôles de cohérence ont été effectués pour assurer que les codes postaux (zip codes) correspondent bien aux villes et aux Etats appropriés. A la fin, nous avons créé un csv par table pour que l'importation des données sur MySQL soit la plus simple possible.

II- Requêtes SQL

A. Requêtes LID simples, avec jointures et sous-requêtes

- LID Simple :

1 : Sélectionne toutes les stations du 18e arrondissement.

2 : Sélectionne le nom des institutions

3 : Sélectionne les musées qui ont ZOO dans leur nom

4 : Sélectionne les musées qui n'ont pas de numéro de téléphone

- LID Jointure :

1 : Sélectionne les Régions correspondants à chaque Etat

2 : Sélectionne tous les musées du Mississipi (Optimisation avec index)

3 : Sélectionne tous les musées qui ne font pas de bénéfices (Optimisation avec index)

4 : Sélectionne les musées qui ont une renseigné leurs données fiscales en 204 ou 2015 en Louisiane

- LID sous-requête :

1 : Sélectionne les musées de la ville de Baton Rouge

2 : Sélectionne les musées sans données financières en Louisiane pour les années 2014 et 2015

B. Requêtes LMD

Requêtes d'insertion et de mise à jour :

1 : Ajout d'un nouveau type de musée

2 : Ajout d'un nouveau musée

3 : Rachat d'un musée par une institution

4 : Fusion des villes de Baton Rouge et Port Allen

5 : Le THREE NOTCH MUSEUM change de téléphone

6 : Le JUDSON COLLEGE ferme et tous ses musées ferment avec

C. Requêtes de synthèse

Requêtes de synthèses :

1 : Répartition des types de musées

2 : Types de musées les plus représentées par type d'environnement

3 : Classement des institutions par nombre de musées

4 : Nombre de musées par ville

5 : Nombre de musées par Etat

6 : Nombre de musées par Région

7 : Moyenne des chiffres d'affaires par type de musée

8 : Top 5 villes ayant le plus de musées

9 : Pourcentage de musées par type d'environnement
10 : Les musées les plus éloignés de New York

D. Requêtes complexes

- Requêtes avec CTE :

1 : Taux d'évolution des bénéfices par rapport à l'année précédente

2 : Répartition de la santé financière des employeurs de musées

- Requêtes et vues :

3 : Création d'une vue avec la distance moyenne à vol d'oiseau pour chaque musée depuis Baton Rouge et Lafayette

4 : Sélectionne les musées de chaque type qui minimise la somme des distances depuis les deux villes

- Requêtes avec jointures externes :

5 : Sélectionne les musées ainsi que leurs institutions même s'ils n'en ont pas

6 : Sélectionne l'employeur ayant le plus gros chiffre d'affaires et combien de musées il possède

III- Mise à disposition de la base de données et représentation des requêtes sur une carte