	Université de Corse - Pasquale PAOLI
	Diplôme : M1 DFS — 2023-2024
	Module : Bases de données NoSQL **Projet Gestion des données de la recherche** **Enseignant : Evelyne VITTORI**

Dans le cadre de ce projet, vous incarnerez le rôle d'ingénieur des données chargé de concevoir et de mettre en œuvre une infrastructure de gestion des données de la recherche pour un laboratoire de recherche multidisciplinaire.

L'objectif principal est de mettre en pratique vos connaissances et compétences en bases de données (relationnelles et nosql) pour accompagner les chercheurs du laboratoire dans l'organisation et le stockage efficaces de leurs données de recherche, ainsi que dans la structuration et l'organisation des métadonnées associées en vue d'un futur dépôt sur une plateforme d'open data.

Les données de la recherche sont de nature très variée selon les champs disciplinaires. Elles peuvent être issues d'observations, d'expériences ou de simulations par exemple. Leur format peut également varier (différents types de fichiers) et les types des données être très différents :

- Données numériques
- Données textuelles : articles, rapports, notes de recherche, transcription d'entretiens.
- Données multimédias : images, vidéos, enregistrements audios.
- Données génomiques : séquences d'ADN, résultats d'analyses.
- Données géospatiales : cartes, coordonnées GPS, données de terrain.
- Données temporelles : séries chronologiques, données d'événements.

On considère que le stockage des publications (articles de conférences, articles de revues, ouvrages, …) est déjà organisé dans le laboratoire et on s'intéresse plus spécifiquement aux jeux de données utilisés dans le cadre des projets de recherche.

On suppose qu'actuellement les chercheurs partagent ces jeux de données au sein des projets de recherche en plaçant leurs fichiers sur un espace de stockage partagé.

Votre tâche sera tout d'abord de les convaincre de l'inefficacité d'un simple espace de stockage partagé pour la gestion de données de recherche diverses et variées. Vous devrez ensuite leur proposer d'utiliser des bases de données pour stocker les jeux de données et les métadonnées associées. Vous devrez concevoir et implémenter ces bases de données et illustrer leur intérêt en définissant des requêtes d'interrogation pertinentes.

Travail à faire

1. **Choix des Champs Disciplinaires**

Sélectionnez deux champs disciplinaires de votre choix, comme par exemple la biologie, la physique, la sociologie, l'informatique, la linguistique, ….

2. **Définition des Jeux de Données et des métadonnées**

a) Pour chaque champ disciplinaire sélectionné, définissez au moins cinq jeux de données fictifs de nature variée. Les jeux de données devront inclure au moins :
   - Une liste de mesures numériques,
   - Des documents de type texte,
   - Des documents vidéo ou audio,
   - Des documents ayant un format spécifique lié à un logiciel spécialisé (ex :cartographie, …).

Il n'est pas nécessaire d'avoir un nombre important de données dans chaque jeu de données : prévoyez environ 5 à 10 données par jeu de données.

b) Définissez les métadonnées associées à chacun de vos jeux de données.

Les métadonnées décrivent les caractéristiques du jeu de données (par exemple le propriétaire, le nombre de données, le format, …) (cf. exemples en annexe).
Il est important de distinguer ces métadonnées des métadonnées associées aux données elle-même (par ex : date de création pour un fichier vidéo) et qui font partie du jeu de données lui-même (cf. exemples en annexe).

3. **Conception et implémentation**

Au niveau du stockage des jeux de données, vous devrez envisager deux solutions : un SGBD No SQL (par exemple MongoDB mais vous pouvez opter pour une autre solution) et un SGBD relationnel de votre choix.
En ce qui concerne les métadonnées associées décrivant les jeux de données, elles pourront être stockées dans la même base de données ou une base de données séparée.

Concevez et implémentez deux versions de la base de données des jeux de données :
- Une version avec le sgbd MongoDB
- Une version avec un sgbd relationnel de votre choix

Pour chacune des deux solutions, proposez une solution pour l'organisation et le stockage des métadonnées et leur association aux jeux de données. Justifiez vos choix.

4. **Définition de requêtes**

Afin d'illustrer l'intérêt de vos bases de données auprès des chercheurs, définissez plusieurs (au moins 5) requêtes d'interrogation pertinentes pour extraire des informations utiles sur les jeux de données et leurs métadonnées.
Définissez deux versions de ces requêtes :
- SGBD NoSQL
- SGBD relationnel

5. **Argumentation et Analyse critique des deux solutions**

Proposez une argumentation pour expliquer aux chercheurs les limites de leur mode de stockage et de partage actuel et les convaincre de l'utilité d'une approche basée sur des bases de données en général.
Vous devrez analyser de manière critique les deux solutions (NoSQL et relationnelle) que vous aurez implémentées afin de proposer aux chercheurs celle qui vous semble la mieux adaptée au contexte spécifique de leurs données en justifiant naturellement votre choix.

Rendus

Date rendu : 25 avril 2024

Ce travail doit être réalisé en groupes de 3 à 4 étudiants.
Le rendu sera effectué dans l'onglet travaux sur l'ENT sous la forme d'un rapport synthétique (un **seul fichier pdf)** contenant les éléments suivants :

- La description des champs disciplinaires et des jeux de données choisis
- La description de la structure de votre ou vos bases de données en NoSQL et en relationnel
- La définition des requêtes d'interrogation dans les deux SGBD.
- L'argumentation présentée aux chercheurs pour justifier d'une part l'utilisation d'une base de données et d'autre part le choix du SGBD (relationnel ou nosql) pour les jeux de données et les métadonnées. Le choix s'appuiera sur l'analyse critique de vos deux implémentations.
- Une réflexion sur l'évolution vers le dépôt des jeux de données sur un entrepot spécialisé (exemple : https://recherche.data.gouv.fr/fr)

Soutenance Orale

- Date prévue : 25 avril
- Présentation de 20 minutes décrivant votre démarche, présentant vos bases de données en justifiant vos choix comme si vous vous adressiez aux chercheurs qui devront les utiliser
- Démonstration de vos bases de données
- Bilan et suggestions

Critères d'évaluation

1. Pertinence de votre argumentation
2. Précision de la justification des choix dans le contexte spécifique choisi
3. Pertinence de l'organisation de vos bases de données
4. Maitrise des sgbd manipulés
5. Qualité et adéquation des requêtes proposées pour illustrer l'intérêt de vos bases

ANNEXE : Exemples de jeu de données et de métadonnées associées

Exemple 1 : Données médicales

a) Contenu du jeu de données

Ensemble de fichiers :

1. "patient001_brain_scan.png"
2. "patient002_brain_scan.png"
3. "patient003_brain_scan.png"
4. ...

Pour chaque fichier (Métadonnées associées aux données) :

- Identifiant du patient :
  - Patient 001
  - Patient 002
  - Patient 003
  - ...
- Date de l'examen :
  - 2023-01-15
  - 2023-02-20
  - 2023-03-10
  - ...

b) Métadonnées

- Nom du jeu de données : "Images_Medicales"
- **Description :** Ce jeu de données contient des images médicales représentant des scanners IRM du cerveau de patients.
- **Format des fichiers :** Images au format PNG
- Nombre total d'images : 100
- Date de création du jeu de données : 2023-05-20
- **Auteur/Responsable :** Dr. Jeanne Dupont, Service de radiologie
- **Institution :** Hôpital Universitaire de la Ville

Exemple 2 : Données analyse de l'eau

a) Contenu du jeu de données

Emplacement	Substance A (mg/L)	Substance B (mg/L)	Substance C (mg/L)	pH
Point d'eau 1	2.5	1.8	3.2	7.0
Point d'eau 2	3.1	2.0	2.9	6.8
Point d'eau 3	2.8	1.6	3.5	7.2
...	...	...	...	...

b) Métadonnées

- **Nom du jeu de données :** "Donnees_Scientifiques"
- **Description :** Ce jeu de données contient des mesures expérimentales de concentration de différentes substances chimiques dans des échantillons d'eau prélevés dans différents endroits.
- Nombre total d'échantillons : 50
- Date de collecte des données : 2023-03-15
- **Auteur/Responsable :** Dr. Pierre Lefebvre, Laboratoire de Chimie Analytique
- **Institution :** Université de la Ville
- **Méthode d'analyse :** Spectrophotométrie UV-Visible
- **Unités de mesure :** milligrammes par litre (mg/L) pour les concentrations, unité de pH pour le pH
- **Contexte de l'étude :** Analyse de la qualité de l'eau dans la région métropolitaine
- **Notes :** Les données sont le résultat d'une étude menée sur une période de six mois pour évaluer la qualité de l'eau dans différents points de prélèvement.