yProjet sur les bases de données du Big Data :
Etudes des bases de Données NoSql à partir de critères

Préambule

Le travail ici demandé fait suite au cours « Les Bases de Données du Big Data ». Ces bases de données sont aussi souvent appelées Bases de Données NoSql. L'objectif est, à partir de critères et d'un schéma de données prédéfinis, d'évaluer une base de données Nosql par **groupe de trois étudiants**. Ce TP permet, en plus de l'examen, d'évaluer le cours « Les Bases de Données du Big Data » (50% de la note). Tous les livrables doivent être rendus le 15 Juin 2022 au plus tard.

Table des matières

1. SCHEMA ET LES DONNEES DE L'ETUDE 4

1.1 Schéma de données 4

1.2 Les données de l'étude 5

1.3 Liste des moteurs NoSql à choisir 5

2. IDENTITE DU MOTEUR NOSQL 5

2.1 Editeur 6

2.2 Versions initial (Nr. De version et date) 6

2.3 Versions actuelle (Nr. De version et date) 6

2.4 Modèles de données supportés(Clé/Valeur, Orienté document, Orienté Colonnes, Oriénté Graphe) 6

2.5 Gestion du schéma (sans schéma, dynamique, statique, mixte) 6

2.6 Support de SQL (DDL, DML) 6

2.7 Support d'indexes secondaires 6

2.8 Langage de développement du sgbd nosql 6

2.9 Support / Pérénité (communauté , etc.…) 6

2.10 API Supportés 6

2.11 Théorème CAP (CP, AP, AC) 6

2.12 Méthode de partitionnement 6

2.13 Méthode de réplication 7

2.14 Concept de consistance 7

2.15 Concept de durabilité 7

2.16 Clés étrangères 7

2.17 Support de références (REF) 7

2.18 Licences et prix 7

2.19 Différents types de versions (communautaire, entreprise, …) 7

2.20 Audience dans le marché 7

2.21 Tables et tables filles 7

2.22 Clé avec major et minor key 7

2.23 Gestion des utilisateurs 7

2.24 Gestion des droits 7

2.25 Gestion des namespaces ou databases 8

2.26 Systèmes d'exploitations supportés 8

2.27 Disponible en mode DBaas 8

2.28 Support du Map/Reduce 8

2.29 Lien vers la documentation technique y compris les API 8

2.30 Typage (none, static, dynamique) 8

2.31 Applications communautaires l'utilisant 8

2.32 Domaines d'applications 8

2.33 Architecture du moteur NoSql 8

2.34 Montée en charge 8

2.35 Gestion de la disponibilité 8

2.36 Procédure d'installation 8

3. MODELISATION ET CHARGEMENT DES DONNEES 9

3.1 Modélisation et chargement de données partie Key/value 9

3.2 Modélisation et chargement de données partie JSON 9

3.3 Modélisation et chargement de données partie Orientée colonnes 9

3.4 Modélisation et chargement de données partie Relationnelles 9

3.5 Modélisation et chargement de données partie Graphes 9

4. MISES A JOURS DES DONNEES (INSERT, UPDATE, DELETE) 9

4.1 Maj de données partie Key/value 9

4.2 Maj de données partie JSON 9

4.3 Maj de données partie Orientée colonnes 10

4.4 Maj de données partie Relationnelles 10

4.5 Maj de données partie Graphes 10

5. INTERROGATION DES DONNEES 10

6. LES RESULTATS DE L'ETUDE 11

7. REPARTITION DU TRAVAIL 11

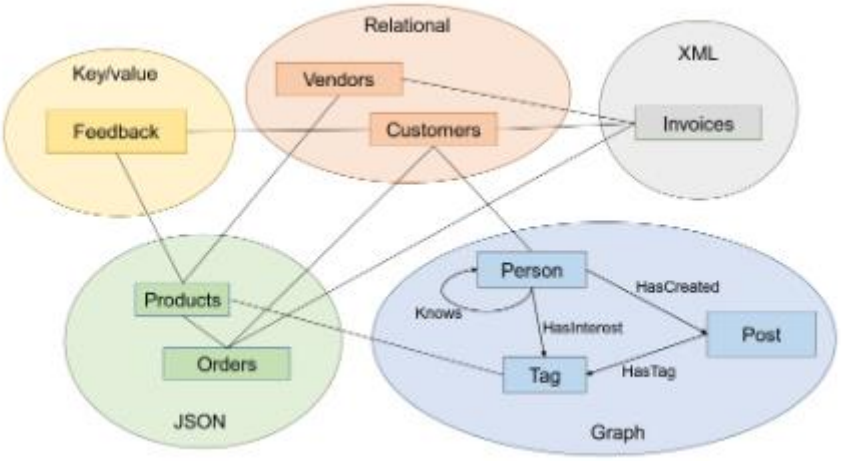
8. OU S'INSCRIRE POUR CONSTITUER SON GROUPE DE 5 ET CHOISIR SON SGBD NOSQL 11

1. Schéma et les données de l'étude

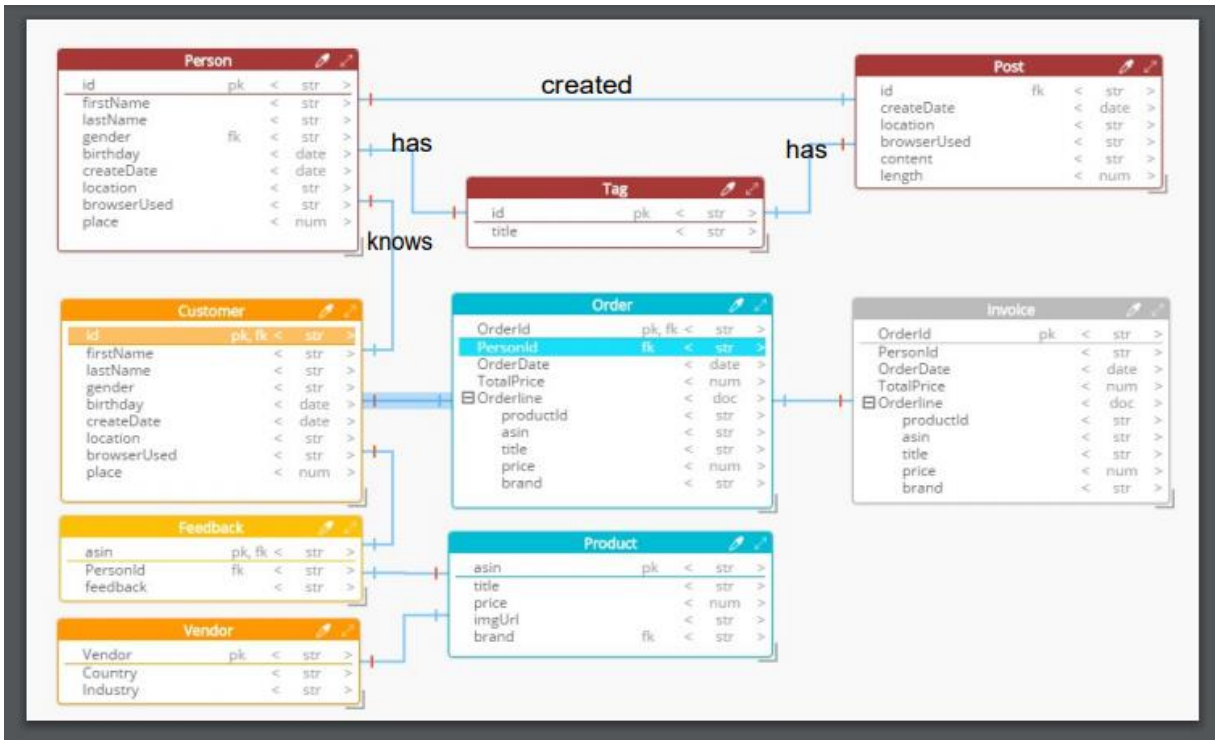
1.1 Schéma de données

Le schéma de données ci-dessous de vente de produits est un schéma multi-modèle. Il y a là les modèles suivants :

- Le modèle clé-Valeur pour gérer les FeedBack,
- Le modèle relationnel pour gérer Vendors et Customers
- Le modèle XML pour gérer Invoices (ce sera pour vous un modèle orienté colonnes)
- Le modèle JSON (documents) pour gérer Products et Orders
- Le modèle Graph pour gérer Person, Tag et Post



Le but initial de cette application était de gérer chaque modèle avec le moteur BD qui va. Le but qui vous est assigné ici est de gérer chaque modèle avec un seul moteur NoSQL que vous aurez choisi. Si le SGBD NoSQL supporte tous ces modèles, alors utilisés les. Si un modèle n'est pas possible avec avec votre SGBD NoSQL alors convertissez en ce qui est possible avec votre SGBD.



Le schéma ci-dessus représente le détail des structures par modèle. Faites attention, les cardinalités n'apparaissent pas. Vous devez en tenir compte.

L'ensemble du projet est disponible sur jalon :

Nom du fichier à télécharger : ?

1.2 Les données de l'étude

Les données pour chaque modèle sont disponibles dans le dossier NoSQL\data du zip à télécharger sur jalon.

1.3 Liste des moteurs NoSql à choisir

Moteurs Clé-Valeur : REDIS, Amazon DynamoDB, Microsoft Azur CosmosDB, RIAK

Moteurs orientés colonnes : Cassandra, Hbase, Big Table de Google, Microsoft Azur Table store

Moteurs orientés documents : CouchBase, CouchDB, OrientDB, MarkLogic

Moteurs orientés graphes :Neo4J, ArangoDB, Virtuoso, Giraph

2. Identité du moteur NoSql

Vous devez présenter sur quelques lignes chaque item ci-dessous du moteur nosql que vous aurez à évaluer.

2.1 Editeur

2.2 Versions initial (Nr. De version et date)

2.3 Versions actuelle (Nr. De version et date)

2.4 Modèles de données supportés(Clé/Valeur, Orienté document, Orienté Colonnes, Oriénté Graphe)

2.5 Gestion du schéma (sans schéma, dynamique, statique, mixte)

2.6 Support de SQL (DDL, DML)

2.7 Support d'indexes secondaires

2.8 Langage de développement du sgbd nosql

2.9 Support / Pérénité (communauté , etc.…)

2.10 API Supportés

2.11 Théorème CAP (CP, AP, AC)

2.12 Méthode de partitionnement

2.13 Méthode de réplication

2.14 Concept de consistance

2.15 Concept de durabilité

2.16 Clés étrangères

2.17 Support de références (REF)

2.18 Licences et prix

2.19 Différents types de versions (communautaire, entreprise, …)

2.20 Audience dans le marché

2.21 Tables et tables filles

2.22 Clé avec major et minor key

2.23 Gestion des utilisateurs

2.24 Gestion des droits

2.25 Gestion des namespaces ou databases

2.26 Systèmes d’exploitations supportés

2.27 Disponible en mode DBaas

2.28 Support du Map/Reduce

2.29 Lien vers la documentation technique y compris les API

2.30 Typage (none, static, dynamique)

2.31 Applications communautaires l’utilisant

2.32 Domaines d'applications

2.33 Architecture du moteur NoSql

2.34 Montée en charge

2.35 Gestion de la disponibilité

2.36 Procédure d’installation

3. Modélisation et chargement des données

Vous devez ici procéder au chargement des données dans votre BD Nosql. Trois solutions possibles :

- Ecrire un programme Java avec l'api fournit par le sgbd nosql
- Utiliser l'interface ligne de commandes du sgbd nosql
- Utiliser un utilitaire de chargement de données

Vous avez le droit de choisir une des solutions ou de mixer.

3.1 Modélisation et chargement de données partie Key/value

3.2 Modélisation et chargement de données partie JSON

3.3 Modélisation et chargement de données partie Orientée colonnes

3.4 Modélisation et chargement de données partie Relationnelles

3.5 Modélisation et chargement de données partie Graphes

4. Mises à jours des données (insert, update, delete)

Afin d'effectuer les mises à jour vous devez :

- Ecrire un programme Java qui permet d'insérer, modifier ou supprimer des enregistrements (le faire pour un enreigstrement et aussi pour plusieurs)
- Effectuer l'activité précédente en utilisant l'interface ligne de commande du sgbd nosql

4.1 Maj de données partie Key/value

4.2 Maj de données partie JSON

4.3 Maj de données partie Orientée colonnes

4.4 Maj de données partie Relationnelles

4.5 Maj de données partie Graphes

5. Interrogation des données

Ecrire les programmes java qui répondent aux questions ci-dessous.

- Query 1. For a given customer, find his/her all related data including profile, orders, invoices, feedback, comments, and posts in the last month, return the category in which he/she has bought the largest number of products, and return the tag which he/she has engaged the greatest times in the posts.

- Query 2. For a given product during a given period, find the people who commented or posted on it, and had bought it.

- Query 3. For a given product during a given period, find people who have undertaken activities related to it, e.g., posts, comments, and review, and return sentences from these texts that contain negative sentiments.

- Query 4. Find the top-2 persons who spend the highest amount of money in orders. Then for each person, traverse her knows-graph with 3-hop to find the friends, and finally return the common friends of these two persons.

- Query 5. Given a start customer and a product category, find persons who are this customer's friends within 3-hop friendships in Knows graph, besides, they have bought products in the given category. Finally, return feedback with the 5-rating review of those bought products.

- Query 6. Given customer 1 and customer 2, find persons in the shortest path between them in the subgraph, and return the TOP 3 best sellers from all these persons' purchases.

- Query 7. For the products of a given vendor with declining sales compare to the former quarter, analyze the reviews for these items to see if there are any negative sentiments.

- Query 8. For all the products of a given category during a given year, compute its total sales amount, and measure its popularity in the social media.

- Query 9. Find top-3 companies who have the largest amount of sales at one country, for each company, compare the number of the male and female customers, and return the most recent posts of them.

- Query 10. Find the top-10 most active persons by aggregating the posts during the last year, then calculate their RFM (Recency, Frequency, Monetary) value in the same period, and return their recent reviews and tags of interest.

6. Les résultats de l'étude

- Rapport d'identification de votre moteur Nosql (remplir pour le cela le chapitre 2 de ce document)
- Les programmes et scripts de chargement de données (remplir chapitre 3). Joindre les scripts et/ou les programmes dans des fichiers textes
- Les programmes et scripts de mise à jour de données (remplir le chapitre 4). Joindre les programmes et les scripts dans des fichiers textes
- Les programmes de consultation des données (remplir le chapitre 5), joindre les programmes dans des fichiers textes

7. Répartition du travail

- 5 Membres par groupe (activités par membre): chap. 2 (7 propriétés), chap. 3 (chargement des données d'1 modèle), chap. 4 (mise à jour des données d'1 modèle), chap. 5 (2 requêtes).
- Vous devez lors de la restitution identifier ce que chaque membre a fait

8. Ou s'inscrire pour constituer son groupe de 5 et choisir son SGBD NoSql

- Dans google drive, voir le mail que vous avez reçu :
- Un moteur nosql ne peut pas être choisi par deux groupes.
- Vous devez vous repartir le travail à part égal par membre du groupe
- Vous devez préciser dans vos rendus qui a fait quoi !!!
- Une présentation de 15 minutes par groupe est attendu !!!
- Deadline 15 Juin 2022