

Projet GMIN332 : Gestion de données complexes

ALIJATE Mehdi - COUSOT Kevin - NEGROS Hadrien

27 janvier 2014

Abstract

Résumé

Accès et consultation de données provenant de différentes solutions de persistance (gros volumes de données distribuées et hétérogènes) au travers d'un démonstrateur.

Sommaire

- 1 Introduction et problématique
- 2 Jeu de données
 - Description
 - Schéma
- 3 Différents modèles de représentation des données
 - Modèle relationnel
 - Triple Store
 - Base de donnée orientée graphe
 - Base de donnée orientée colonne
- 4 Application
 - L'API Jena
 - Requêtage SPARQL
 - Architecture de l'application
- 5 Démonstration
- 6 Discussion et conclusion

Introduction

Les systèmes NoSQL et les technologies du Web Sémantique sont une alternative aux SGBD classiques.

Problématique

- Données hétérogènes
- Différents paradigmes de représentation

Comment interconnecter ces données ?

Nous avons fait appel à trois sources :

- **INSEE** : COG : les régions, les départements, les arrondissements, les cantons et les communes, ISF.
- **Geonames** : de nombreuses informations telles que : noms de lieux géographiques, population, code postal...
- **data.gouv.fr** : résidences de tourisme classées en France.

Modèle relationnel

- C'est largement le paradigme le plus utilisé.
- Les données sont représentées par des tuples.
- Les contraintes assurent la cohérence et les liens sur les données.
- Le SGBD que nous avons choisi est **MySQL**

D2RQ

D2RQ nous permet de manipuler des données relationnelles comme des triplets RDF.

Triple Store

- Un Triple Store est une base de donnée NoSQL conçue pour stocker des triplets RDFs.
- La consultation des données se fait avec le langage de requête **Sparql**

TDB

TDB est une solution de persistance pour les données RDFs.

BDD orientée graphe

- Une base de données orientée graphe est une BDD NoSQL utilisant la théorie des graphes.
- Pas besoin de schéma, les relations sont des objets de premier ordre.

Neo4j

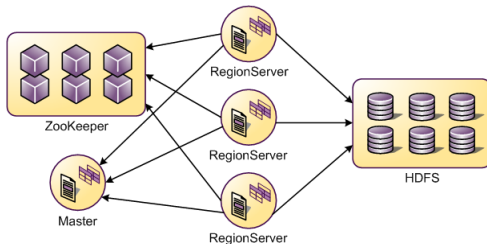
La base de donnée orientée graphe que nous avons décidé d'intégrer est **Neo4j**

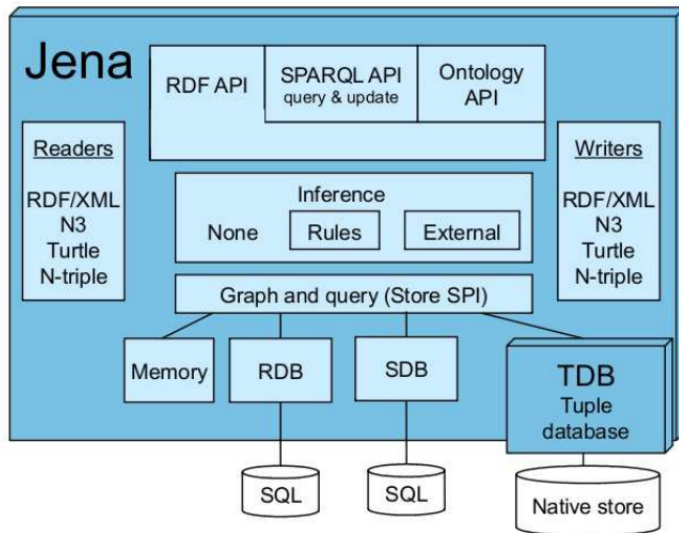
BDD orientée colonne

- C'est une base de données qui stocke les données sous forme de colonnes.
- L'intérêt est de sérialiser les colonnes les unes après les autres

Hbase

La base de donnée orientée colonne que nous avons décidé d'intégrer(à la base) est **Hbase**





SPARQL

- C'est le langage de requêtage sur les données RDFs.
- SPARQL fonctionne par recherche d'homomorphismes du sous-graphe défini dans la requête vers la BDD.
- ARQ est le moteur de requêtes SPARQL de JENA, celui-ci y ajoute diverses fonctionnalités telles que les agrégateurs.

Architecture

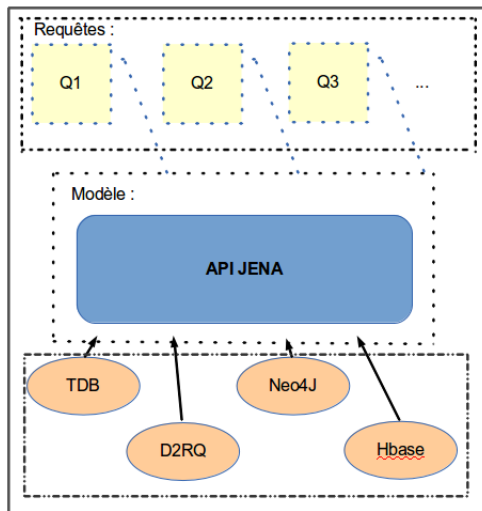
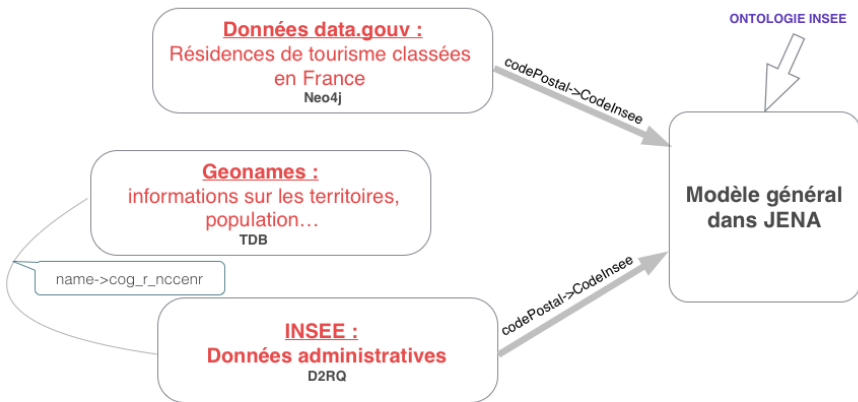


Schéma d'interconnexion des données



Démonstration

Discussion

- Distribution des données sur des sites différents.
- Ne pas utiliser une bdd embarquée mais l'API REST.
- Choisir des données plus facilement interconnectables.

Difficultés

- Quelques difficultés techniques.
- Choix des données : interconnection.

Conclusion

Ce projet nous a permis :

- Explorer les différentes solutions NoSql.
- Familiariser avec l'API Jena.
- Comprendre l'intérêt d'un modèle commun pour des données hétérogènes.