

Raja - Bases de données distribuées

12 décembre 2010

Audrey NOVAK
Romain MANESCHI
John FHAL
Aloys URBAIN

Table des matières

1	Introduction	3
1.1	Projet	3
1.2	Equipe	3
1.3	Organisation	3
1.4	Matériel	3
2	Modèle d’analyse	4
2.1	Besoins	4
2.2	Requêtes	4
2.3	Bases de données	4
2.4	Dérivation puis Intégration	4
2.5	Intégration	5
2.5.1	Schéma RDF Maladies	6
2.5.2	Schéma RDF Facteurs	6
2.5.3	Schéma RDF CIM	6
2.5.4	Schéma RDF Recencement	6
2.5.5	Schéma RDF Global	6
2.6	Insertion, Modification et Suppression	6
3	Modèle de conception	7
3.1	Serveur pseudos-médiateurs	7
3.2	Adaptateurs	7
3.3	Traducteurs	7
4	Glossaire	8
5	Annexes	9
5.1	Schéma des tables des bases de données	9

1 Introduction

1.1 Projet

Ce projet permettra d'accéder à différentes informations concernant les maladies dans le monde. Ceci de la façon la plus simple possible. Dans un premier temps nous devrons interroger le système par le biais de requêtes sql simplifiées. Mais à terme une interface permettrait de simplifier la construction de ces requêtes.

Les données sur les maladies sont stockées sur différentes bases de données. Ces bases de données sont stockées sur des SGBD hétérogènes non seulement par leur implantations mais aussi et surtout par leur langage de communication. Toute la difficulté du projet sera de rassembler toutes les données, pour offrir à l'utilisateur une vue d'ensemble sur ce qu'il cherche.

1.2 Equipe

Pour réaliser ce projet nous sommes une équipe de quatre étudiants en master deuxième année à la faculté des sciences de Montpellier. Ce projet intervient à l'issue de l'unité d'enseignement "Gestion de données distribuées - Intégration - Médiation".

1.3 Organisation

Tout le monde a participé à une réflexion préliminaire pendant les deux premières semaines. Ayant chacun nos propres affinités l'organisation du travail c'est fait tout naturellement par la suite.

Romain Maneschi : chef de projet, rédaction du rapport, organisation générale

Audrey Novak : modèles ontologiques et traduction D2RQ par spécification N3

Aloys Urbain : installation d'un environnement de test

John Fhal : création des scripts d'installation et de tests des bases de données

Cette organisation nous permet de pouvoir tous commencer en même temps. Puis au fur et à mesure que chacun aura fini sa partie, nous nous rejoindrons sur le développement des classes Java pour réaliser le serveur.

1.4 Matériel

Afin de simplifier nos échanges nous avons ouvert un site internet sur google code. Ce site nous permet d'avoir un gestionnaire de versions centralisé, un wiki et un compte ftp. Tout ceci nous permet donc de centraliser en un seul endroit tous nos travaux.

Pour ne pas perdre de temps, un seul membre du groupe, sera en charge de créer une machine virtuelle sur laquelle tournera les différentes bases de données. Cette machine virtuelle pourra ensuite être transmise aux autres membres, ceux-ci auront donc directement un système opérationnel.

Les SGBD utilisées seront Oracle, MySQL et PostgreSQL. Nous les avons choisis car ce sont trois logiciels qui ont fait leurs preuves et sont utilisés tous les jours par différents professionnels.

2 Modèle d'analyse

2.1 Besoins

Notre serveur offrira à un utilisateur la possibilité de chercher des renseignements sur des maladies. Plus exactement il pourra demander au système des informations sur une maladie : son nom, son facteur de transmission, le virus associé à cette maladie et un nombre de cas recensés pour une zone (voir le schéma des tables des bases de données pour de plus amples informations). Pour cela nous avons mis en place un système de requêtes basées sur sql. Ces informations seront récupérées par nos soins auprès de l'OMS.

Scéma Use case très simple!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

2.2 Requêtes

Notre serveur prendra en entrée une requête dont la syntaxe correspondra au langage SQL. Mais notre système ne pourra gérer toute la complexité de ce langage. Nous le limiterons donc à quelques instructions : select, insert, update, delete

De plus, et toujours dans un soucis d'efficacité, nous n'accepterons dans un premier temps, que des instructions dites basiques (toutes les instructions entre crochets sont optionnelles) :

- select column_name [, column_name2...] from table_name [, table_name2] [where column_name = value...];
- insert into tabel_name values(value_table, value_table2...) [where column_name = value and column_name2 = value2...];
- update table_name set column_name = value1 [, column_name2 = value2] [where column_name = value...];
- delete from table_name [, table_name2] [where column_name = value...];

Nous avons donc mis en place un type requête particulier : SChéma UML des IQuery (avec la hiérarchie)

2.3 Bases de données

Pour la distribution de nos données, nous avons décidé de partager nos données verticalement :

- les facteurs de transmissions d'une maladie
- les virus à l'origine d'une maladie
- le nom des maladies
- le nombre de cas recensés d'une maladie par zone, cette base de donnée sera une base virtuelle et sera donc découpée horizontalement par zone OMS (ici nous en prendrons trois : ???????, ???????, ???????)

Ce qui donne le shéma de tables suivant :

2.4 Dérivation puis Intégration

Maintenant que nous avons réparties nos données sur différents SGBD avec une approche qui est dite par dérivation. Nous devons nous demander comment nous allons rendre possibles la recherche sur ces différentes bases. Deux méthodes s'offrent à nous : nous continuons à dériver, ou nous intégrons nos différents schémas de bases.

La première solution revient à découper notre schéma global de façon à ce qu'il colle à nos

schémas locaux : donc à chaque base de données. Cette solution est la plus simple des deux mais nécessite de gros efforts d'implémentation puisque le schéma global doit être découpé en schémas locaux "à la main" pour chaque base de données. De plus cette solution devient immaintenable lors du changement de schémas de base : en d'autres termes il faut se replonger dans la programmation.

La deuxième revient à prendre tous les schémas locaux pour faire émerger un schéma global. Cette solution est bien plus compliquée au niveau de la réflexion et de la logique qu'il y a derrière. Mais la généricité de l'implémentation permet de perdre moins de temps en cas de changement des schémas des bases de données. En effet nous avons imaginé une solution entièrement basée sur les méta-tables, ce qui nous permet d'abstraire tout notre code afin de le rendre réutilisable et modulable. Nous avons donc en toute logique retenu cette deuxième solution.

Schéma explicatif!!!!!!!!!!!!!!!!!!!!!!!!!!!!

2.5 Intégration

Puisque notre système repose sur les méta-informations des schémas locaux, le problème suivant est de pouvoir les rassembler pour faire émerger un schéma global correspondant à toutes les informations accessibles par un utilisateur.

Pour réaliser cela nous avons immédiatement choisis la technologie RDF. En effet elle nous permet de représenter nos données sous forme de triplets : deux entités et une relation entre eux. Ce système nous permet d'ajouter à volonté des relations de tous types entre toutes nos entités. Vous comprenez donc qu'il devient facile de regrouper des entités de différentes bases de données pour faire émerger notre fameux schéma global, en précisant que les données d'une base correspondent aux mêmes données d'une autre.

De plus RDF nous est fourni avec plusieurs outils très utiles :

- D2RQ : traducteur d'objet relationnel en objet rdf. - N3 : cette notation qui va de pair avec D2RQ permet de lier des bases de données avec des objets rdf. En décrivant le mappage entre les tables et columns d'un côté et les types RDF de l'autre, D2RQ peut construire tout le schéma RDF correspondant à une base de données. Nous utiliserons donc D2RQ comme traducteur pour toutes nos bases de données.

- SparQL : ce langage permet de réaliser très simplement des requêtes complexes dans nos triplets RDF. Et notamment de faire émerger des connaissances plus subtilement que par des recherches "classiques" basées sur la comparaison de valeurs.

Le seul problème de D2RQ est qu'il permet de faire tout ce travail qu'en sélection. En effet nous ne pouvons pas insérer, mettre à jour ou supprimer des données d'une base. Il faudra donc trouver un autre moyen.

Nous nous servirons donc de RDF aussi bien pour réaliser les requêtes de sélection de nos utilisateurs, que pour diriger nos requêtes vers telle ou telle base de données en fonction des schémas locaux.

2.5.1 Schéma RDF Maladies

2.5.2 Schéma RDF Facteurs

2.5.3 Schéma RDF CIM

2.5.4 Schéma RDF Recensement

2.5.5 Schéma RDF Global

2.6 Insertion, Modification et Suppression

Le dernier problème restant est donc de savoir comment nous allons insérer, modifier ou supprimer nos données des bases. Pour cela nous utiliserons la technique classique des traducteurs. Il y aura donc un traducteur par SGBD différent afin de faire correspondre la requête entrante aux normes de la base de donnée.

3 Modèle de conception

Schéma global simple : requête -> serveur -> adaptateur -> traducteur -> bd

3.1 Serveur pseudos-médiateurs

Comme nous pouvons le voir sur ce schéma, le serveur sera le premier maillon de la chaîne à recevoir la requête de l'utilisateur. Il aura alors la charge de transformer la requête qui arrive sous forme de chaîne de caractères en type Query. Pour cela nous mettrons en place une fabrique à Query. Ensuite le serveur demandera à l'adaptateur principal si la requête le concerne, si c'est le cas il lui transmettra la partie de la requête qui lui correspond.

Il n'aura plus alors qu'à attendre le résultat, qui lui sera retourné par l'adaptateur, et au besoin transformer celui-ci du format rdf au format souhaité.

Puisque notre serveur est au centre de la recherche : il prend la recherche et retourne le résultat, nous avons décidé de le caractériser par le nom pseudo-médiateur. En effet dans une architecture classique le médiateur est celui qui fractionne les requêtes pour que celles-ci correspondent au schéma local des adaptateurs. C'est ce que fait notre serveur, mais un médiateur au sens propre du terme, à d'autres exigences que nous ne traiterons pas dans ce projet, d'où le nom de pseudo-médiateur.

3.2 Adaptateurs

Nos adaptateurs sont les pièces principales du projet. Ils devront publier le schéma local des bases de données dont ils ont la responsabilité. Ces schémas seront sous la forme RDF pour faciliter la recherche. Ils contiendront le nom des tables et des colonnes des bases.

Pour faciliter l'implémentation nous avons imaginé deux types d'adaptateur : les composites et les finaux. Les adaptateurs composites, sont des adaptateurs qui ont à leur charge non pas une base de données mais des adaptateurs qui peuvent être eux aussi composite ou finaux. Ces adaptateurs devront donc pouvoir faire émerger un schéma rassemblant l'ensemble de leur sous-adaptateurs.

Puisqu'un adaptateur permet de faire tout cela, nous avons décidé d'utiliser un adaptateur composite dans notre serveur. Celui-ci devra alors faire émerger le schéma global de nos bases de données.

3.3 Traducteurs

Comme son nom l'indique un traducteur permet tout simplement de convertir une requête de notre système en requête compréhensible pour le SGBD. Dans le cas de la sélection nous utiliserons D2RQ comme traducteur. Pour tout le reste nous l'implémenterons nous même.

3.4 Diagramme de classes

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

4 Glossaire

- **SGBD** : Système de gestion de base de données.
- **schéma global** : le schéma de nos tables et de nos colonnes comme s’il ne s’agissait que d’une seule base de données.
- **schéma local** : le schéma physiquement stocké sur une base de donnée particulière.
- **méta-informations** : les informations concernant le nom des tables et des colonnes d’une base de donnée.
- **requête** : dans notre système nous parlons de deux sortes de requêtes différentes, les requêtes du client donc qui arrivent dans notre serveur sous forme de chaîne de caractères et les requêtes internes qui sont ces mêmes requêtes mais transformées en objet.

5 Annexes

5.1 Schéma des tables des bases de données