Traitement efficace de requêtes SPARQL avec extensions OLAP pour entrepôts RDF.

Thèse encadrée par Ioana Manolescu (DR INRIA, HDR)
Collaboration avec : Dario Colazzo (MCF, Univ. Paris-Sud) et François Goasdoué (MCF, Univ. Paris-Sud)

21 mars 2011

Résumé

Ce sujet vise à lever des verrous scientifiques à la mise en œuvre d'une chaîne décisionnelle pour un modèle de données du W3C de plus en plus utilisé : RDF. Une telle chaîne doit permettre d'analyser un ensemble de données stockées dans un entrepôt RDF, selon différents critéres d'analyse et à des niveaux de granularité variables [1], afin de faire émerger des informations aidant à la prise de décision dans le pilotage d'une organisation publique ou privée. Ce sujet consiste à concevoir des méthodes d'*exploration* des données selon les axes d'analyse (et aux granularités) voulus par une technologie OLAP pour RDF. L'objectif est l'analyse de données du Linking Open Data Project [3] : la mise à disposition sur le Web, en RDF, des données publiques.

1 Description détaillée

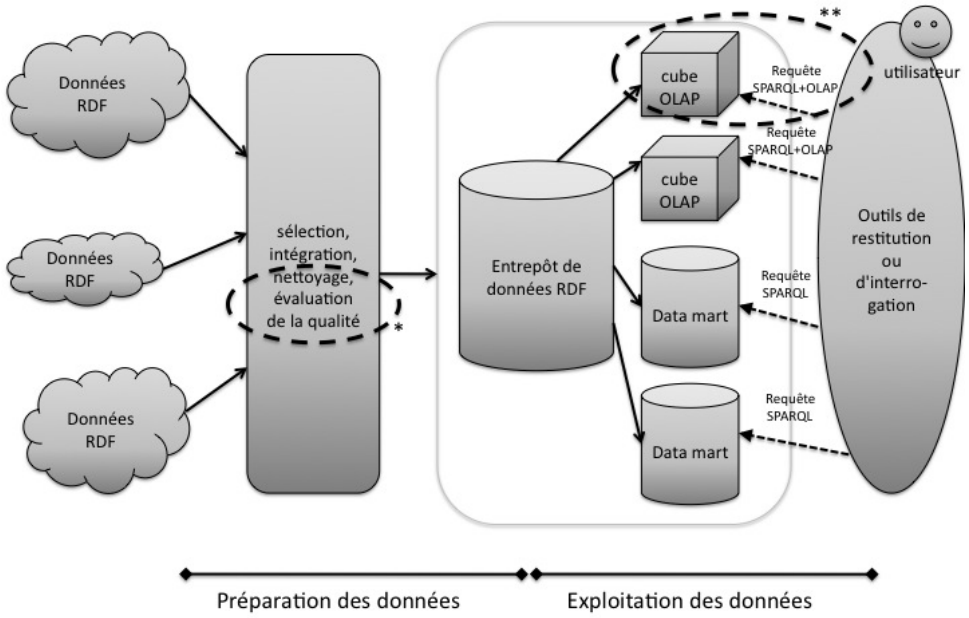
Le Web regorge aujourd'hui de données RDF gratuites, disponibles, et de bonnes qualité (car publiées par des organismes connus et bien identifiés). La quantité de données RDF disponibles sur le Web est en croissance fulgurante, avec un espace global de données de l'ordre de plusieurs milliards d'assertions (le nombre d'assertions RDF a triplé entre mai 2009 et mai 2010 passant de 4 milliards à 13 milliards). Ceci est notamment dû à l'initiative Linking Open Data (LOD, [3]) visant à publier sur le Web – en RDF – les données publiques. Le choix de RDF n'est pas anodin, car ce modèle de données est ouvert, simple, et particulièrement adapté à la description de données non RDF préexistantes du Web. L'exemple emblématique de données non RDF ayant rejoint LOD est Wikipedia, via sa version DBPedia en RDF.

1.1 Positionnement global du travail à réaliser

Cette thèse vise à valoriser des données RDF d'intérêt en permettant leur analyse multi-dimensionnelle, c'est-à-dire selon différents critères d'analyse et à des niveaux de granularité variables. Ce type d'analyse permet de faire émerger certaines caractéristiques à forte valeur ajoutée dans les données. D'ailleurs, l'analyse multidimensionnelle est devenue incontournable pour le pilotage stratégique et politique dans les organisations publiques ou privées.

Concrètement, ce type d'analyse est réalisée grâce à une *chaîne décisionnelle*. Celle-ci est composée de deux grandes phases de traitement des données [2, 5] : la *préparation des données* (sélection, intégration, nettoyage, évaluation de la qualité) issues de sources multiples afin de les mettre à disposition dans un entrepôt de données et l'*exploitation des données* de l'entrepôt via leur analyse mutidimensionnelle fondée sur la technologie OLAP (OnLine Analitical Processing

1. Par exemple, temps : année, mois, jours ; espace : région, département, ville ; *etc*



** Étape concernée par le sujet de thèse *Traitement efficace de requêtes SPARQL avec extensions OLAP pour entrepôts RDF.*

FIGURE 1 – Chaîne décisionnelle

[4]). La figure 1 illustre (de façon simplifiée) le fonctionnement d'une telle chaîne décisionnelle pour des données RDF comme nous la voyons cette thèse.

Dans ce sujet de thèse, nous nous intéressons à l'exploitation des données de l'entrepôt RDF, en particulier l'exploration de *cubes* de données qui sont des vues multidimensionnelles.

Un cube est dédié à *une* analyse multidimentionnelle de données (dite *mesure*) de l'entrepôt principal en fonction des critères d'analyses voulus (dits *dimensions*) et à leurs différents niveaux de granularité (dits *niveaux de hiérarchie*). Une telle structure nécessite des opérations particulières pour être explorée (dites *opérations OLAP*). De telles opérations ont été définies en termes de structures multidimensionnelles, c'est-à-dire indépendamment d'un modèle de données (`http://www.olapcouncil.org/`). Dans le cadre du modèle de données relationnel, le langage de requêtes SQL a déjà été étendu avec ces opérations [5] et implanté dans, ou "au-dessus", de nombreux SGBD (`http://www.bi-verdict.com/`).

Concernant le sujet de thèse proposé, le langage recommandé par le W3C pour manipuler des données RDF est SPARQL (`http://www.w3.org/TR/rdf-sparql-query/`). Le princial objectif sera donc d'une part d'étendre SPARQL avec les opérations OLAP, d'autre part de founir des mécanismes d'évaluation de requêtes efficaces et effectifs pour ce nouveau langage.

1.2 Positionnement des travaux par rapport à l'état de l'art du domaine de recherche

La notion de chaîne décisionnelle a donné lieu, dans le cadre de données relationnelles, à de nombreux travaux de recherche [2, 5], ainsi qu'à de nombreux outils utilisés quotidiennement dans le pilotage des organisations (ORACLE Business Intelligence Suite, MICROSOFT SQL Server Analysis Services, TM1, Pentaho Analysis Services, etc.)

Récemment, des travaux se sont intéressés à l'utilisation de RDF dans la mise en œuvre de chaînes décisionnelles pour des données relationnelles [6]. Notamment RDF est utilisé comme modèle pivot pour masquer l'hétérogénéité des sources de données alimentant l'entrepôt princi-

pal (relationnel) et pour exprimer simplement et élégamment les critères d'analyse multidimensionnelle à partir desquels sera généré le cube de données (relationnelles).

À notre connaissance, aucun travail n'a porté sur une chaîne décisionnelle dédiée aux données RDF. On peut raisonnablement penser que ceci est dû à la version actuellement standardisée de SPARQL (1.0, `http://www.w3.org/TR/rdf-sparql-query/`) qui ne permet pas le calcul d'aggrégats, prérequis pour la mise en œuvre et l'exploration OLAP de cubes de données RDF.

Notre sujet de recherche se projette dans la révision à venir de SPARQL 1.0 : SPARQL 1.1 (`http://www.w3.org/TR/sparql11-query/`). Cette prochaine version est à l'état de Working Draft au W3C et devrait obtenir le status de recommandation (c-à-d de norme du W3C) sous peu. Parmi les nouveautés apportées, nous trouvons la notion de calcul d'aggrégats permettant d'envisager une extension OLAP de SPARQL.

2 Concepts, outils, défis et enjeux, pistes bibliographiques

L'enjeu principal de cette thèse est l'extention de SPARQL avec les principaux opérateurs OLAP [2], à savoir :

- *Roll up* : zoomer sur un critère d'analyse en passant à un niveau de granularité plus fin (par ex. pour le critère de temps, passer d'une année à un mois),
- *Drill down* : dézoomer sur un critère d'analyse en passant à un niveau de granularité plus grossier (par ex. pour le critère spatial, passer d'un département à une région), et
- *Slice* et *Dice* : considérer le sous-cube obtenu en selectionnant une valeur pour un (*slice*) ou plus (*dice*) critères d'analyse (par ex. en fixant le mois et la région).

Pour cela, nous définirons une spécification formelle et algébrique de l'extension de SPARQL (que nous appelons SPARQL-OLAP dans la suite). Nous implanterons ensuite cette algèbre dans notre système de gestion de données RDF. Celui-ci est issu du projet RDFViewS [1] mené dans le cadre des thèses de Konstantinos Karanasos et Julien Leblay, dont le coencadrement est assuré par François Goasdoué et Ioana Manolescu.

Nous développerons ensuite des techniques d'évaluation efficaces des requêtes SPARQL-OLAP. Ces techniques dépendront du type de cube RDF considéré, à l'instar des cubes relationnels [5]. Un cube peut-être une vue non-matérialisée ou matérialisée sur l'entrepôt [2]. Dans le premier cas, l'optimisation consiste à reformuler efficacement une requête SPARQL-OLAP sur le cube en un ensemble de requêtes SPARQL sur l'entrepôt. Dans le second cas, l'optimisation consiste en l'évaluation d'une requête SPARQL-OLAP sur un cube en ayant recours à une indexation multidimensionnelle de ses données.

Dans la majeure partie des scénarii, les données RDF sont susceptibles de mises à jour fréquentes. Nous développerons des techniques pour la maintenance efficace des vues/index mentionnés ci-dessus. L'objectif principal sera de minimiser l'impact des mises à jour sur les vues/index, par exemple en détectant l'indépendance entre mises à jour et index/vues afin de ne propager que les modifications pertinentes.

3 Résultats attendus et critères pour juger du succès de la thèse

Les résultats attendus rélèvent de l'innovation scientifique et du développement de logiciels. Ils sont présentés dans le tableau 1.

2. Il existe aussi des approches hybrides consistant en une matérialisation partielle de la vue.

Résultats scientifiques	**Critères d'évaluation**
1. Spécification formelle et algébrique d'une extension de SPARQL pour l'analyse OLAP de cubes de données RDF ; en particulier via l'adaptation des opérateurs *roll-up*, *drill-down*, *slice and dice*, etc au contexte RDF **2.** Techniques d'évaluation efficace de requêtes SPARQL avec extensions OLAP sur des cubes de données RDF	Publications et leur visibilité (citations, conférences invitées...) Présentations dans des PMEs travaillant dans le domaine de la gestion de données du Web Sémantique (Nexedi ou DataPublica déjà mentionnées, mais aussi p.ex. Mondeca, intéressée par nos technologies RDF actuelles etc.)
Résultats logiciels	**Critères d'évaluation**
3. Outil permettant la spécification des cubes de données RDF **4.** Outil simple puis optimisé pour la matérialisation et la manipulation des cubes **5.** Etude de passage à l'échelle dans le volume et la complexité des données **6.** Générateur de données et requêtes paramétrées (éventuellement en partant d'un jeu de données existant) pour mesurer la performance des opérations liées aux cubes RDF	Démonstrations dans des conférences, journées d'échange INRIA-PMEs, contrats de collaboration sur ces thématiques. Echelle des données utilisées, niveaux d'aggrégation, temps de réponse rapporté à la taille des données. Nombre d'utilisations (dans des cours et stages de Master, par nos partenaires européens dans des activités EIT ICT Labs etc).

TABLE 1 – Résultats attendus et critères d'évaluation.

Références

[1] François Goasdoué, Konstantinos Karanasos, Julien Leblay, and Ioana Manolescu. RDF-ViewS : A storage tuning wizard for RDF applications. In *ACM International Conference on Information and Knowledge Management*, Canada Toronto, Oct 2010.

[2] Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis. *Fundamentals of Data Warehouses*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2nd edition, 2001.

[3] "linked open data" web site. Available at `http://linkeddata.org/`.

[4] The OLAP council web site. Available at `http://www.olapcouncil.org/`.

[5] Maurizio Rafanelli, editor. *Multidimensional Databases : Problems and Solutions*. Idea Group, 2003.

[6] Stefano Spaccapietra, editor. *Journal on Data Semantics XIII – Special Issue on Semantic Data Warehouses*, volume 5530 of *Lecture Notes in Computer Science*. Springer, 2009.