M1 - PLDAC 2018
Requêtes SPARQL réparties pour Wikidata.

Encadrants : H. Naacke, B. Amann, Contact : hubert.naacke@lip6.fr

Mots clés : SPARQL, RDF, parsing et exécution de requêtes

Description :
Ce projet concerne le traitement de requêtes SPARQL[2] sur des grands jeux de données (tel que Wikidata contenant plusieurs milliards d'éléments). On suppose que le volume des données dépasse la capacité de stockage d'une machine. Cela nécessite de répartir le stockage des données sur plusieurs machines. Dans ce contexte « réparti », le problème étudié concerne les performances des requêtes. On constate qu'une requête provoque de nombreux transferts de données entre les machines.
L'objectif du projet est de proposer, pour une requête SPARQL, un plan d'exécution efficace qui minimise la quantité de données transférées entre les machines.

Travail à faire :
Apres avoir parsé une requête SPARQL (en utilisant le parseur de Jena), déterminer quelles sont les opérations de sélection, projection et jointure à exécuter pour obtenir le résultat de la requête.
Proposer une méthode pour collecter des statistiques sur les données afin d'estimer la taille du résultat d'une opération (de sélection, projection ou jointure).
Comprendre les deux algorithmes de jointure détaillés dans [1] et en particulier le coût des transferts associés. Proposer une méthode pour :

- choisir, pour chaque opération de jointure de la requête, l'algorithme qui transfert le moins de données,
- ordonner les opérations de telle sorte que la quantité totale des transferts soit la plus petite possible.

Implémenter votre solution en scala (ou en java).
Tester votre solution sur la plateforme Apache Spark avec les données de Wikidata [3] et les exemples de requêtes proposées dans [4]

Lecture:

[1] Hubert Naacke, Bernd Amann, Olivier Curé: SPARQL Graph Pattern Processing with Apache Spark. Grades 2017. https://event.cwi.nl/grades/2017/01-Naacke.pdf

[2] SPARQL 1.1 Query Language : http://www.w3.org/TR/sparql11-query/

[3] Wikidata: https://www.wikidata.org

[4] Query service https://query.wikidata.org