L'opérateur CUBE pour les entrepôts de données NoSQL orientés colonnes

Khaled Dehdouh, Fadila Bentayeb, Nadia Kabachi, Omar Boussaid

Laboratoire ERIC, Université de Lyon 2
5 avenue Pierre Mendes-France, 69676 Bron Cedex, France
prénom.nom@univ-lyon2.fr

Résumé. L'émergence de grands volumes de données, imposée par les grands acteurs du web, nécessite de nouveaux modèles de gestion de données et des nouvelles architectures de stockage et de traitement capables de trouver rapidement une information dans une volumétrie considérable de données. Les bases de données NoSQL (Not Only SQL) orientées colonnes offrent pour les *big data*, un modèle approprié aux entrepôts de données et à une structuration multidimensionnelles sous forme de cube OLAP(On-Line Analytical Processing). Cependant, en l'absence d'opérateur de calcul de cube OLAP, nous proposons dans cet article, un nouvel opérateur d'agrégation, baptisé CN-CUBE (*Columnar NoSQL CUBE*), qui permet de calculer des cubes de données à partir d'entrepôts de données stockés dans un système de gestion de base de données NoSQL orientées colonnes. Nous avons implémenté l'opérateur CN-CUBE sous l'interface SQL (*Phoenix* [1]) du SGBD orienté colonnes *Hbase* [2], et réalisé des expérimentations sur un entrepôt de données publiques dans un environnement distribué réalisé avec *Hadoop* [3]. Nous avons pu montrer ainsi que notre opérateur CN-CUBE présente des temps de calcul de cubes OLAP intéressants pour les entrepôts de *big data*.

1 Introduction

Un entrepôt de données est une base de données dédiée à l'analyse en ligne (OLAP) pour l'aide à la décision (Inmon, 1992). Il est souvent implémenté sous un système de gestion de base de données relationnelles (SGBDR) (Codd, 1970). Cependant, dans un monde qui est constamment connecté, les sources de données produisent des données de plus en plus massives, appelées *big data*. Les modèles relationnels classiques ont montré leurs limites quant au stockage et à la gestion des *big data* (Leavitt, 2010). En effet, ce sont les grands acteurs du web tels que Yahoo, Google, Facebook, Twitter et LinkedIn qui ont signalé en premier les limites du modèle relationnel. Le constat était que les SGBDR ne sont pas adaptés aux environnements distribués requis par les volumes gigantesques de données. Pour répondre aux besoins

1. http://phoenix.incubator.apache.org/
2. http://hbase.apache.org/
3. http://hadoop.apache.org/