Optimisation des performances dans les entrepôts de données NoSQL en colonnes

Mohamed Boussahoua*, Omar Boussaid*, Fadila Bentayeb *

*Université de Lyon, Université Lyon 2, ERIC EA 3083
5 avenue Pierre Mendes-France, F-69676 Bron Cedex, France
{mohamed.boussahoua, omar.boussaid, fadila.bentayeb}@univ-lyon2.fr

Résumé. Le modèle NoSQL orienté colonnes propose un schéma de données flexible et hautement dénormalisé. Dans cet article, nous proposons une méthode d'implantation d'un entrepôt de données dans un système NoSQL en colonnes. Notre méthode est basée sur une stratégie de regroupement des attributs issus des tables de faits et de dimensions, sous forme de familles de colonnes. Nous utilisons deux algorithmes *OEP* et *k-means*. Pour évaluer notre méthode, nous avons effectué plusieurs tests sur le benchmark TPC-DS au sein du SGBD NoSQL orienté colonnes *Hbase*, avec une architecture de type *MapReduce* sur une plateforme *Hadoop*.

1 Introduction

Les entrepôts de données jouent un rôle important dans la collecte et l'analyse de grandes masses de données pour l'aide à la décision. Généralement, ils sont souvent implémentés sous les systèmes de gestion de bases de données relationnelles (SGBDR). Ces derniers s'imposent par la richesse de leurs fonctionnalités et les performances de leurs requêtes. Cependant, ils sont peu appropriés pour construire des entrepôts de données distribuées, nécessaires pour faire face à l'augmentation du volume de données et à la scalabilité de l'espace de stockage Leavitt (2010). De plus, l'exécution des requêtes décisionnelles dégrade les performances des entrepôts de données dans un SGBDR. Pour améliorer les performances des entrepôts de données relationnels, différents travaux de recherche existent et portent notamment sur les techniques d'indexation, la fragmentation ou la compression des données. De la même manière, il est nécessaire de recourir à de nouvelles solutions de stockage fiables et à moindre coût dans les systèmes décisionnels distribués. Parmi ces solutions, il y a lieu de citer la plateforme Hadoop [1], qui comprend différents modules, tels que Apache Hive [2], un système d'entreposage de données muni d'une interface de type SQL, Apache Pig et de nouveaux modèles de données dits NoSQL (*Not Only SQL*) [3] apparus ces dernières années sous l'impulsion des grands acteurs du Web (*Google*, *Yahoo*, *Facebook*, *Twitter*, *Amazon*...). Ces SGBD, dits aussi non relationnels, s'appuient sur le théorème de CAP (*Consistency,*

1. http ://hadoop.apache.org
2. https ://hive.apache.org
3. https ://fr.wikipedia.org/wiki/NoSQL