

Clé de partition multi-attributs pour un partitionnement horizontal optimal des entrepôts de données NoSQL en colonnes

Mohamed Boussahoua*, Fadila Bentayeb *, Omar Boussaid *, Nadia Kabachi **

*Université Lumière Lyon 2, ERIC EA 3083
5, avenue Pierre Mendès 69676 Bron-France
{Mohamed.Boussahoua, Fadila.Bentayeb, Omar.Boussaid}@univ-lyon2.fr

**Université Claude Bernard Lyon 1
43, boulevard du 11 novembre 1918, 69100, Villeurbanne-France
Nadia.Kabachi@univ-lyon1.fr

Résumé. Les systèmes NoSQL en colonnes offrent des techniques de stockage adaptées à la construction d'entrepôts de données. Plusieurs scénarios sont possibles pour stocker des entrepôts de données sur ces systèmes. Dans cet article, nous étudions une nouvelle approche de placement des données d'un entrepôt sur un cluster dans un système NoSQL en colonnes. Notre approche s'appuie sur une méthode basée sur une stratégie de regroupement d'attributs pour définir les clés de partition *RowKey* d'un schéma de partitionnement horizontal des données. Nous obtenons ainsi un modèle physique de données qui permet de garantir une localisation et une distribution plus homogène des données dans les différents nœuds du cluster. Nous utilisons la méthode des règles d'association pour obtenir les partitions sus-mentionnées. Pour évaluer notre méthode, nous avons effectué plusieurs tests sur le benchmark TPC-DS au sein du SGBD NoSQL HBase. Les résultats obtenus montrent que notre stratégie de placement des données augmente les performances des entrepôts NoSQL en colonnes de l'ordre de 48%.

1 Introduction

Le stockage et la gestion de données volumineuses, complexes, structurées ou non structurées constituent une tâche difficile dans les environnements informatiques actuels. Pour faire face à l'essor des volumes et de la diversité des données, de nouveaux systèmes de stockage et de traitement de données robustes et efficaces ont été mis en place par les institutions scientifiques et industrielles tels que MapReduce (Dean et al., 2001), C-Store (Stonebraker et al., 2005), Apache Hadoop¹ ainsi que les systèmes

1. <http://hadoop.apache.org/>