Entrepôts de données multidimensionnelles NoSQL

Max Chevalier*, Mohammed El Malki*,**, Arlind Kopliku*,
Olivier Teste*, Ronan Tournier*

*Université de Toulouse, IRIT UMR 5505, Toulouse, France
http://www.irit.fr Prénom.Nom@irit.fr
**Capgemini, 109, avenue du Général Eisenhower,
BP 53655- 31036 Toulouse, France
http://www.capgemini.com

Résumé. Les données des systèmes d'analyse en ligne (OLAP, On-Line Analytical Processing) sont traditionnellement gérées par des bases de données relationnelles. Malheureusement, il devient difficile de gérer des mégadonnées (de gros volumes de données, « Big Data »). Dans un tel contexte, comme alternative, les environnements « Not-Only SQL » (NoSQL) peuvent fournir un passage à l'échelle tout en gardant une certaine flexibilité pour un système OLAP. Nous définissons ainsi des règles pour convertir un schéma en étoile, ainsi que son optimisation, le treillis d'agrégats pré-calculés, en deux modèles logiques NoSQL : orienté-colonnes ou orienté-documents. En utilisant ces règles, nous implémentons et analysons deux systèmes décisionnels, un par modèle, avec MongoDB et HBase. Nous comparons ces derniers sur les phases de chargement des données (générées avec le benchmark TPC-DS), de calcul d'un treillis et d'interrogation.

1 Introduction

Pour faciliter le processus d'aide à la prise de décision, les données à analyser sont centralisées de manière uniforme dans un entrepôt de données Kimball et Ross (2013). Au sein de l'entrepôt, une analyse interactive des données est effectuée via un processus d'analyse en ligne (OLAP On-Line Analytical Processing), Colliat (1996), Chaudhuri et Dayal (1997). Les données sont souvent décrites au moyen d'un modèle multidimensionnel tel qu'un schéma en étoile, Chaudhuri et Dayal (1997), basé sur des sujets d'analyse (appelés faits) et des axes d'analyses (appelés dimensions). Les faits regroupent de manière conceptuelle des indicateurs d'analyse (des mesures). Ces mesures sont associées à des dimensions qui sont composées de différents niveaux de détails (niveau d'agrégation ou paramètres) permettant de constituer des perspectives (hiérarchies) d'analyse. Ces hiérarchies sont des structures employées pour faciliter le pré-calcul des agrégations induites par les analyses OLAP (par exemple, calculer des ventes annuelles à partir des valeurs mensuelles). Ces pré-calculs sont souvent modélisés via un treillis d'agrégats pré-calculés, Gray et al. (1996). Ainsi, un treillis représente l'ensemble des pré-calculs d'un schéma multidimensionnel où chaque noeud du treillis représente un agrégat et chaque arc représente le chemin pour calculer l'agrégat à partir d'autres agrégats. De nos jours, le volume des données d'analyses atteint des tailles critiques, Jacobs (2009), défiant les