Extraction du schéma d'une BD NoSQL orientée documents

Amal Ait Brahim *, Rabah Tighilt Ferhat * et Gilles Zurfluh *

*Institut de Recherche en Informatique de Toulouse (IRIT)
Université Toulouse Capitole - 2 Rue du Doyen-Gabriel-Marty - 31042 Toulouse - France
<prenom.nom>@gmail.com
http://www.ut-capitole.fr/

Résumé. Au cours de ces dernières années, l'utilisation des systèmes NoSQL pour stocker et exploiter les bases de données (BD) massives n'a cessé de s'accentuer. Ces systèmes apportent une grande souplesse d'utilisation notamment grâce à leur propriété « schema less ». Cependant, le schéma de la BD NoSQL est nécessaire pour permettre aux utilisateurs (notamment les décideurs) d'exprimer leurs requêtes. Dans ce qui suit, la notion de « BD NoSQL » est utilisée pour désigner une BD gérée par un SGBD NoSQL. L'objectif de cet article est d'extraire le schéma d'une BD NoSQL orientée documents. Pour cela, nous proposons un processus incrémental d'élaboration du schéma pendant l'exploitation de la BD.

1 Introduction

Au cours de ces dernières années, la nécessité d'exploiter les données qui sont générées et accumulées par les dispositifs informatiques n'a cessé de s'accentuer. Ceci est à cause du volume, qui peut dépasser plusieurs téraoctets, et de la variété de ces données qui sont qualifiées de complexes. De plus, ces données sont souvent saisies à très haute fréquence et doivent donc être filtrées et agrégées en temps réel pour éviter une saturation inutile de l'espace de stockage. Ces problématiques sont désignées par l'expression « Big Data » Chen et Zhang (2014) et caractérisées par la règle dite des « 3V » Laney (2001). Il s'agit du Volume (des masses des données considérables à gérer), de la Variété (des données complexes) et de la Vélocité (en référence à la collecte et au traitement en temps-réel de ces données). Les approches classiques basées principalement sur le paradigme relationnel, ne peuvent pas répondre à ces objectifs et exigent de nouvelles approches de stockage et de manipulation des données.

Regroupées sous le terme NoSQL (Not Only SQL) (Han et al., 2011), ces approches permettent une plus grande adaptabilité dans des contextes fortement distribués, ainsi qu'une gestion performante des données complexes Darmont et al. (2007). Les BD NoSQL proposent une nouvelle façon de gérer des données. Elles sont classées en quatre catégories : orientée clé-valeur, orientée documents, orientée colonnes et orientée graphes.

La plupart des BD NoSQL sont caractérisées par l'absence du schéma de données lors de la création. Dans les SGBD relationnels, tous les éléments du schéma sont fixés à l'avance ;