Conception physique d'un entrepôt de données distribuées basée sur K-means équilibré

Yassine Ramdane*, Omar Boussaid*
Nadia Kabachi**, Fadila Bentayeb *

*Université de Lyon, Lyon 2, ERIC EA 3083, 5, avenue Pierre Mendês 69676 Bron-France,
{Yassine.Ramdane, Omar.Boussaid, Fadila.Bentayeb}@univ-lyon2.fr
**Université de Lyon, université Claude Bernard Lyon 1, ERIC EA 3083, 43 boulevard du 11 novembre 1918, 69100, Villeurbanne-France
Nadia.Kabachi@univ-lyon1.fr

Résumé. Le partitionnement horizontal est l'une des techniques les plus performantes pour améliorer l'exploitation de données sur les plateformes de traitements parallèles comme Hadoop et Spark. Dans les entrepôts de données distribués (EDD), l'opération la plus coûteuse est la jointure en étoile qui nécessite plusieurs cycles MapReduce lors de son exécution. Dans ce papier, nous proposons une nouvelle stratégie de placement des données d'un entrepôt volumineux dans Hadoop, en se basant sur l'algorithme K-means équilibré (*K-means balanced*). Ce schéma de placement permet d'exécuter des opérations de certaines requêtes OLAP, dont la jointure en étoile, en une seule étape de Spark. Dans notre approche, nous prenons en compte les caractéristiques physiques du *cluster* et le volume des données. Pour évaluer notre proposition, nous avons effectué des expérimentations sur un *cluster* de 5 nœuds avec un entrepôt de données issu du banc d'essai TPC-DS. Les résultats obtenus montrent un gain de temps d'exécution, de certaines requêtes OLAP, allant jusqu'à 60% par rapport à d'autres approches existantes.

1 Introduction

Un entrepôt de données (ED) est une grande base de données conçue pour analyser les données. La taille d'un ED peut atteindre des dizaines de teraoctets (To). Il est modélisé à l'aide d'un schéma en étoile ou en flocons de neige, comprenant une ou plusieurs tables de faits et plusieurs dimensions.

Plusieurs techniques de partitionnement horizontal ont été utilisées pour améliorer les performances des entrepôts de données distribuées (EDD), comme l'équilibrage des charges de données ou les stratégies de placement et de distribution des bases de données (Zamanian et al., 2015; Lu et al., 2017). On peut distinguer deux types de partitionnement : statique et dynamique. Dans les techniques statiques, on effectue le placement et la distribution des données avant de traiter une requête en se basant soit sur le schéma de l'entrepôt (Eltabakh et al., 2011; Dittrich et al., 2010), soit sur une charge de requêtes stable (Arres et al., 2015). Dans