Nouvelle stratégie pour le traitement distribué des processus décisionnels massifs dans un Big Data Warehouse

Rado Ratsimbazafy*, Fadila Bentayeb*, Omar Boussaid*

*Université de Lyon, Université Lumière Lyon 2, Laboratoire ERIC
5 Avenue Pierre Mendes France, 69676 Bron Cedex
prenom.nom@univ-lyon2.fr, http://eric.ish-lyon.cnrs.fr

Résumé. Cet article traite du problème de l'optimisation de l'exécution des charges de requêtes massives dans le cadre des entrepôts de données (ED) distribués où le nombre de processus simultanés à traiter se compte par milliers. En nous inspirant des techniques d'optimisation utilisées dans le contexte centralisé, nous proposons dans cet article une nouvelle stratégie de sélection et de stockage de vues matérialisées (MV) basée sur système de fichiers distribués ; puis nous abordons le traitement des charges de requêtes décisionnelles massives en utilisant les MV. Notre approche joue un rôle de médiateur entre les utilisateurs et l'entrepôt de données pour proposer de meilleurs plans d'exécution à leurs requêtes. Les premiers résultats que nous avons obtenus, à partir de nos expérimentations montrent que dans un environnement distribué notre approche améliore de plus de 50% le coût d'exécution d'une charge de requêtes par rapport au système fourni par défaut.

1 Introduction

L'entreposage et l'analyse en ligne des données massives (*big data*) sont devenus en quelques années l'activité principale de beaucoup d'entreprises et de chercheurs (Ahuja et al. (2009); Thusoo et al. (2010b)). L'intérêt porté à l'avènement des données massives a fait évoluer le système d'information décisionnel (SID), et par conséquent les entrepôts de données et l'OLAP (Online Analytical Processing) (Chaudhuri et al. (2011)). De nouveaux modèles d'entrepôts de données sont alors apparus (Figure 1) : les systèmes basés sur des systèmes de gestion de bases de données relationnelles (SGBDR) tels que *Teradata* [1], *Greenplum* [2], et les systèmes basés sur le paradigme *MapReduce* (Dean et Ghemawat (2004)), comme *Hive* (Thusoo et al. (2010a)), où les données sont stockées sur un système de fichiers distribués tels que GFS de *Google* ou HDFS de *Hadoop*.

Par ailleurs, l'intérêt grandissant autour du "*big data analytics*" a fait naître plusieurs techniques, stratégies et approches, mais aussi de nouveaux profils métiers (*data scientist, big data engineer,* ...). Cependant, l'analyse de telles quantités de données, pour récupérer les informations pertinentes, doit être réalisée dans une durée acceptable (Cohen et al. (2009)). Les

1. http ://www.teradata.com
2. http ://greenplum.org/