SGIA : Stratégie Intelligente de Groupement pour Améliorer le Traitement des Requêtes OLAP en MapReduce

Yassine Ramdane*, Nadia Kabachi**
Omar Boussaid*, 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. L'amélioration des performances d'une requête OLAP dans un système distribué tel que Hadoop ou Spark est une tâche ardue. Une requête OLAP est composée de plusieurs opérations, tels que le filtrage, la jointure et le *Group By*. Chaque opération peut être exécutée dans la phase *map* ou la phase *reduce* avec un ou plusieurs cycles *MapReduce*. Étant donné qu'il est possible de collecter au préalable quelques connaissances sur le système distribué, certaines opérations, comme la jointure en étoile et le filtrage, peuvent être optimisées en utilisant une technique statique de partitionnement. Cependant, l'optimisation du *Group By* nécessite généralement l'utilisation d'une technique dynamique de partitionnement et de distribution qui permet d'équilibrer à la volée les charges des *reducers*, car nous ne pouvons pas collecter les informations pertinentes qui aident le système à établir le bon schéma qu'au moment de l'exécution de la requête. Dans cet article, nous proposons une méthode intelligente, appelée SGIA, permettant d'équilibrer les données d'entrées des *reducers*. Nous avons utilisé un système multi-agents qui permet d'équilibrer à la volée les charges des *reducers*. Les expérimentations révèlent que notre approche est plus performante que celles existantes en termes de temps d'exécution des requêtes.

1 Introduction

Les requêtes OLAP sont généralement des requêtes coûteuses qui nécessitent beaucoup de temps pour être exécutées sur des Entrepôts de Données Distribuées (EDD) massives. Dans les EDD massives, l'amélioration du traitement d'une requête OLAP est une tâche ardue. Une requête OLAP est composée de plusieurs clauses et prend généralement la forme "*Select... Function ()... From...Where... Join-Predicates ...Filters... GROUP BY...*". Chaque clause peut être exécutée dans la phase *map* ou dans la phase *reduce*, et chaque opération peut être exécutée avec un ou plusieurs cycles MapReduce ou stages de Spark, avec une quantité considérable de données transférée entre les nœuds. La jointure en étoile n'est pas la seule opération coûteuse