Omniscience dans la Conception des Entrepôts de Données Parallèles sur un Cluster

Soumia Benkrid*,**, Ladjel Bellatreche** , Alfredo Cuzzocrea ***

* Ecole nationale Supérieure d'Informatique (ESI), Alger, Algérie
s_benkrid@esi.dz
**LIAS/ISAE-ENSMA, Futuroscope, Poitiers, France
(soumia.benkrid, bellatreche)@ensma.fr
***ICAR-CNR - University of Calabria, Italy
cuzzocrea@si.deis.unical.it

Résumé. Généralement, le processus de conception d'un entrepôt de données parallèle passe principalement par deux étapes : (1) la fragmentation des données et (2) l'allocation des fragments générés sur les différents nœuds de traitement. Le principal inconvénient d'une telle approche de conception est le coût élevé de communication pour équilibrer la charge entre les nœuds de traitement, ainsi le nœud coordinateur peut devenir un goulot d'étranglement dans le système. Pour remédier à ces problèmes, la réplication de données (RD) est utilisée. Fréquemment, la fragmentation des données, l'allocation des fragments et la réplication de données sont effectuées de manière isolée. En effet, l'interaction entre ces phases est ignorée. Dans cet article, nous proposons une nouvelle approche de conception d'un entrepôt de données parallèle qui traite conjointement la fragmentation, l'allocation et la réplication. Un algorithme d'allocation redondant basé sur l'algorithme de classification floue "Fuzzy k-means" est proposé. Nous avons également formalisé le problème du traitement parallèle des requêtes comme un Dual Bin Packing, un algorithme glouton est proposé pour la résolution du problème. Enfin, une validation de nos propositions en utilisant le banc d'essai "Star Schema Benchmark" (SSB) est proposée.

1 Introduction

Les nouveaux fournisseurs de données, comme les réseaux sociaux (facebook, linkedin, twitter) les médias numériques, les capteurs, systèmes de commerce électronique, etc. ont largement contribué à la naissance d'une nouvelle ère autour de la gestion des données extrêmement volumineuses. Des conférences et des ateliers autour de cette thématique se sont créés. Nous pouvons ainsi citer Extremely Large Databases Workshop [1], régulièrement sponsorisé par eBay, Facebook et Greenplum. Des applications appuyées par des entreprises se sont créées proposant des solutions d'entreposage, de collection, de traitement, d'analyse de cette mine d'information. De nouveaux métiers ont vu le jour : data analyst, data architect, etc. Un

1. http://www.xldb.org/