Un benchmark enrichi pour l'évaluation des entrepôts de données noSQL volumineuses et variables

Max Chevalier*, Mohammed El Malki*, Arlind Kopliku *, Olivier Teste*, Ronan Tournier *

*IRIT : Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne, 31062 TOULOUSE
max.chevalier,Mohammed.el.malki,arlind.Kopliku,olivier.teste,ronan.tournier@irit.fr
http://www.irit.fr

Résumé. Avec le développement des données massives (Big Data), de nouveaux besoins émergent dans l'évaluation des systèmes d'information décisionnels. En particulier, les bancs d'essais (benchmarks) dédiés aux entrepôts de données multidimensionnelles doivent être adaptés aux volumes et à la diversité des données massives. Dans ce contexte, nous proposons un nouveau benchmark dédié aux entrepôts des données multidimensionnelles qui supporte plusieurs types de systèmes (relationnel, noSQL) et des modèles de données (flocon, étoile, aplati) structurées et non structurées. Pour tester des volumes très importants, il supporte la génération parallèle sur plusieurs machines (cluster). Il enrichit le processus de génération des données pour évaluer plusieurs niveaux de diversité des données. Dans ce papier, nous présentons ce benchmark, appelé KoalaBench, et les premiers résultats expérimentaux de son utilisation.

1 Introduction

Différents bancs d'essais (benchmarks) ont été proposés pour comparer des systèmes d'informations décisionnels. Ils fournissent des jeux de données et des scenarii d'utilisation permettant d'évaluer le comportement des systèmes, dans des conditions équivalentes, permettant ainsi une évaluation comparative. Dans le contexte des entrepôts de données multidimensionnelles (Chaudhuri et Dayal (1997)), les bancs d'essais les plus utilisés sont TPC-DS et TPC-H [1], (Zhang et al. (2004), Poess et al. (2007)). Ces solutions sont développées et optimisées pour les bases de données relationnelles (R-OLAP) et pour une utilisation sur une seule machine. La généralisation des technologies de l'information au travers de réseaux mondialisés, la diffusion massive de moyens de communications mobiles, et le développement d'objets autonomes connectés, produisent des quantités de données numérisées dans des proportions et avec un rythme sans commune mesure avec le passé. On désigne ce phénomène par le terme de mégadonnées ou "Big Data". Les Big Data remettent en cause bon nombre d'approches classiques dans les systèmes d'informations décisionnels. Ces derniers doivent faire face à

1. COUNCIL, Transaction Processing Performance. TPC-H (ad-hoc, decision support) benchmark. URL : http ://www. tpc. org/tpch, 2004