Modélisation Mathématique du problème d'Ordonnancement dans Hadoop

Aymen Jlassi*[1], Patrick Martineau[1], and Vincent Tkindt[1]

[1]Laboratoire Informatique de Tours (LI) – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, Université François-Rabelais de Tours, CNRS, LI EA 6300, OC ERL CNRS 6305, Tours, France – France

Résumé

Modélisation Mathématique du problème d'Ordonnancement dans Hadoop

JLASSI Aymen

Jlassi.aymen@etu.univ-tours.fr

Tkindt Vincent

tkindt@univ-tours.fr

Patrick Martineau

patrick.martineau@univ-tours.fr

Université François-Rabelais de Tours, CNRS, LI EA 6300, OC ERL CNRS 6305, Tours, France

Notre objectif est l'optimisation de l'exécution de travaux sur un cluster Hadoop.

Hadoop est un logiciel libre de gestion de gros volumes de données, basé sur le calcul distribué. Il est fondé sur le paradigme map-reduce introduit par Google et sur un système de fichiers distribué nommée HDFS. Malgré son adoption par des entreprises de grande envergure, des études tel que (Palvo, et al. 2009) montrent que la configuration par défaut de Hadoop ne fournit ni les meilleures performances ni la meilleure exploitation du cluster physique. Afin de remédier à cette problématique, plusieurs travaux de recherches sont apparus : (Bogdan Nicolae 2010) qui propose un nouveau système de fichier adapté au paradigme map-reduce et (Zhao, et al. 2012) qui contribue sur la localisation des données et la gestion des flux sur le cluster.

Le travail présenté vise à optimiser l'affectation de travaux, décomposés en tâches map et

*Intervenant

reduce, sur un ensemble de machines du réseaux de sorte à réduire la durée de traitement et les migrations de données sur le réseau. Nous introduirons un modèle mathématique indexé sur le temps, dans le but non seulement de définir le problème d'optimisation mais également de proposer par la suite des heuristiques d'ordonnancement plus performantes que celles existantes au sein du système Hadoop.

Bibliographie

Bogdan Nicolae, Gabriel Antoniu, Luc Bouge, Diana Moise, Alexandra Carpen-Amarie. "BlobSeer: Next Generation Data Management for Large Scale Infrastrutures." Journal of Parallel and Distributed Computing, aout 2010: 168-184.

Palvo, Andrew, et al. "A comparison of Approaches to Large-scale Data Analysis." (ACM) 09 2009.Zhao, Yanrong, Weiping Wang, Meng Dan, Shubin Zhang, et Gan Guan. "A Data Locality Optimization Algorithm for large-scale Data Processing in Hadoop." IEEE, 2012: 655-661.

Mots-Clés: Cloud computing, Hadoop, HDFS, MapReduce, ordonnancement