Optimisation d'architecture de lacs de données basée sur les chaînes d'approvisionnement

Marzieh Derakhshannia, Anne Laurent, Dickson Owuor

LIRMM, Université de Montpellier, CNRS, Montpellier, France
{prenom.nom}@umontpellier.fr
http://www.lirmm.fr/

Résumé. Les lacs de données constituent une nouvelle génération de dépôts de données. Dans cet article, nous nous appuyons sur une modélisation mathématique de problèmes joints de "location-allocation" utilisés dans la conception de réseau de chaîne d'approvisionnement afin d'améliorer l'architecture des lacs de données et leur performance. Un lac de données est alors considéré comme étant une chaîne d'approvisionnement et les données du lac sont considérées comme des produits avec une durée de vie déterminée. Nous faisons l'hypothèse d'un lac géré avec la paradigme MapReduce et nous résolvons le modèle mathématique à l'aide d'algorithmes gloutons pour déterminer les optimaux de tâches à exécuter pour optimiser les performances tout en minimisant les coûts totaux.

1 Introduction

Au cours de la dernière décennie, de nombreuses organisations ont décidé d'améliorer leurs plates-formes de stockage de données afin de traiter le volume important de données (organisationnelles, de capteurs, ...) produites de manière continue Llave (2018). Les défis principaux sont liés à la préparation d'environnements efficaces pour recueillir et maintenir les données pertinentes à un haut niveau de qualité. La problématique du stockage de données n'est pas nouvelle, et les entrepôts avaient émergé dans les années 1990 pour répondre en partie. Cependant, ils supposent la définition a priori des informations et connaissances à extraire des données, ce qui n'est pas toujours le cas. On parle de *lacs de données* qui sont vus comme une nouvelle génération de systèmes de stockage de données évolutifs pour répondre à l'exigence des plates-formes de stockage de données flexibles et agiles pour les organisations Giudice et al. (2019); Fang (2015). Ils constituent une nouvelle manière de charger, stocker, traiter et visualiser les données Pasupuleti et Purra (2015).

Le lac de données contient d'énormes données structurées et non-structurées dans leurs formats natifs, non pas pour une utilisation immédiate, mais pour de futures interrogations Gorelik (2019). Les architectures de lacs de données sont créées de manière différente de celles des entrepôts de données traditionnels, et tirent parti des outils et des structures pour réduire ou éliminer les processus de préparation de données inefficaces et ingérables Loshin (2013); LaPlante et Sharma (2016); Giebler et al. (2019). Par exemple, Inmon (2016) a défini un lac de données sous la forme d'une structure constituée de différentes sous-structures (appelés