Systèmes de métadonnées dans les lacs de données : modélisation et fonctionnalités

Étienne Scholly*,**, Pegdwendé N. Sawadogo*,
Cécile Favre*, Éric Ferey**, Sabine Loudcher*, Jérôme Darmont*

*Université de Lyon, Lyon 2, ERIC EA 3083
{etienne.scholly, pegdwende.sawadogo,
cecile.favre, sabine.loudcher, jerome.darmont}@univ-lyon2.fr
https://eric.ish-lyon.cnrs.fr/
**BIAL-X
{etienne.scholly, eric.ferey}@bial-x.com
https://www.bial-x.com/

Résumé. Au cours de la dernière décennie, le concept de lac de données a émergé comme une alternative aux entrepôts de données pour le stockage et l'analyse des mégadonnées. Le lac de données propose un stockage des données sans schéma prédéfini. En l'absence de schéma, l'interrogation et l'analyse des données dépendent alors d'un système de métadonnées qui se doit d'être efficace et complet. Cependant, la gestion des métadonnées dans les lacs de données demeure une problématique d'actualité et les critères d'évaluation de son efficacité sont peu ou prou inexistants.

Dans cet article, nous proposons MEDAL, un modèle générique pour la gestion des métadonnées d'un lac de données. MEDAL adopte une modélisation du système de métadonnées à base de graphes. Nous proposons aussi des critères d'évaluation du système de métadonnées d'un lac de données à travers une liste de fonctionnalités attendues et montrons que notre approche est plus complète que les systèmes de métadonnées existants.

1 Introduction

Depuis le début du 21e siècle, les usages des organisations dans les processus de prise de décision sont bouleversés par la disponibilité de grands volumes de données appelées *big data*. Ces mégadonnées, principalement issues des médias sociaux (Facebook, Twitter, Wikipédia, Youtube, etc.) et des objets connectés (*Internet of Things*), constituent une véritable opportunité pour les organisations. Cependant, elles s'accompagnent entre autres de problématiques de volume, de vélocité et de variété, qui surpassent les capacités des systèmes traditionnels de stockage et de traitement des données (Miloslavskaya et Tolstoy, 2016).

C'est dans ce contexte que Dixon (2010) introduit le concept de lac de données (*data lake*), en guise de solution aux problèmes induits par l'hétérogénéité des mégadonnées. Un lac de données propose un stockage intégré des données sans schéma prédéfini (Hai et al., 2016). En