Modélisation des métadonnées d'un *data lake* en *data vault*

Iuri D. Nogueira, Maram Romdhane, Jérôme Darmont

Université de Lyon, Lyon 2, ERIC EA 3083
5 avenue Pierre Mendès France, F69676 Bron Cedex
iuri.deolindonogueira@univ-lyon2.fr, maram.romdhane@univ-lyon2.fr,
jerome.darmont@univ-lyon2.fr

Résumé. Avec l'avènement des mégadonnées, l'informatique décisionnelle a dû trouver des solutions pour gérer des données de très grands volume et variété. Les lacs de données (*data lakes*) répondent à ces besoins du point du vue du stockage, mais nécessitent la gestion de métadonnées adéquates pour garantir un accès efficace aux données. Sur la base d'un modèle multidimensionnel de métadonnées conçu pour un lac de données présentant un défaut d'évolutivité de schéma, nous proposons l'utilisation d'un *data vault* pour traiter ce problème. Pour montrer la faisabilité de cette approche, nous instancions notre modèle conceptuel de métadonnées en modèles logiques et physiques relationnel et orienté document. Nous comparons également les modèles physiques en termes de stockage et de temps de réponse aux requêtes sur les métadonnées.

1 Introduction

Les lacs de données (*data lakes*) ont été introduits par Dixon (2010). Ils proposent une manière, née avec les mégadonnées (*big data*), de stocker dans leur format natif des données volumineuses, variées et diversement structurées, en vue de les analyser (*reporting*, visualisation, fouille de données...).Ce concept s'oppose à celui des entrepôts de données, très intégrés et orientés sujet, mais qui ont l'inconvénient de diviser les données en silos étanches (Stein et Morrison, 2014). Toutefois, tout le monde s'accorde pour dire qu'un lac de données doit être bien conçu sous peine de devenir un marécage (*data swamp*) inexploitable (Alrehamy et Walker, 2015) ; c'est à dire qu'il doit permettre le requêtage des données (sélection/restriction) avec un bon temps de réponse et pas seulement leur stockage et leur accès « clé-valeur ». En revanche, les solutions pour y parvenir sont peu ou prou inexistantes dans la littérature et relèvent à l'heure actuelle de pratiques industrielles peu divulguées.

C'est pourquoi Pathirana (2015) a proposé un modèle conceptuel de métadonnées permettant l'indexation et l'interrogation efficace d'un lac de données patrimoniales. Ce modèle multidimensionnel est proche des modèles en flocons en usage dans les entrepôts de données, mais ne concerne que les métadonnées et non le corpus de documents lui-même. Il a été instancié au niveau physique dans différents systèmes de gestion de bases de données (SGBD) NoSQL. Cependant, le type de schéma employé est très difficile à faire évoluer lorsque ceux des sources de données évoluent ou que de nouvelles sources sont à prendre en compte, alors que c'est un point crucial dans la gestion d'un lac de données.