# Lacs de Données : Tendances et Perspectives

**Franck RAVAT[1], Yan ZHAO[1,2]**

*1. Institut de Recherche en Informatique de Toulouse, IRIT-CNRS (UMR 5505), Université Toulouse 1 Capitole, Toulouse, France*
*franck.ravat@irit.fr yan.zhao@irit.fr*

*2. Centre Hospitalier Universitaire (CHU) de Toulouse, Toulouse, France*

RESUME. *Le lac de données est actuellement présenté comme le composant essentiel pour l'analyse de mégadonnées. Dans cet article, nous résumons la définition, l'architecture fonctionnelle et les différents axes de recherche associés aux lacs de données de (Ravat, 2019) .*

*Mots-clés : Lac de données, Architecture, Métadonnées.*

*Keywords: data lake, architecture, metadata*

À l'ère des mégadonnées (*Big Data*), l'analyse de données volumineuses, véloces et variées nécessitent des architectures adaptées pour l'intégration, le stockage et la restitution. Les entrepôts de données (ED) définis dans le cadre de la *Business Intelligence* (BI), ne sont plus adaptés pour les raisons suivantes : (i) seuls les besoins exprimés dès le début de la phase de conception peuvent être satisfaits ; (ii) toutes les données sources ne sont pas intégrées ; (iii) le coût d'implantation et de maintenance d'un ED peut croître de façon exponentielle pour assurer de bonnes performances d'interrogation et d'analyse. Pour relever le défi de l'analyse de mégadonnées, (Dixon, 2010) a été le premier à proposer le concept de Lac de Données (LD) dont l'objectif est de stocker toutes les données dans leur format natif. Cette définition imprécise n'est pas suffisante pour comprendre tous les enjeux de ce nouveau concept. Notre objectif est donc d'apporter une définition précise, de définir les composants d'une architecture générique de LD et d'aborder les futurs axes de recherche associés à ce nouveau concept.

Mêmes si les définitions ont évolué au fil du temps et à partir des retours d'expérience, ces dernières sont imprécises voire contradictoires. Pour être aussi complet que possible, nous définissons un LD au travers de ses entrées, ses processus de transformations, ses sorties et la gouvernance associée. Un LD est donc une solution d'analyse de mégadonnées qui (i) ingère et stocke des données brutes hétérogènes provenant de sources diverses dans leur format natif, (ii) permet de traiter et transformer ces données afin de répondre aux besoins d'analyse, (iii) fournit des accès aux données à un grand nombre d'utilisateurs pour des restitutions et/ou analyses

diverses (statistique, tableaux de bord interactif, analyse décisionnelle, exploration de données, apprentissage automatique etc.), et (iv) assure la qualité, la sécurité et le cycle de vie des données.

Afin de répondre aux lacunes et spécificités des différentes propositions, nous proposons une architecture fonctionnelle générique de LD basée quatre zones. Chaque zone contient des processus de traitement et un espace de stockage de données. La zone des données brutes permet d'ingérer en temps réel ou différé tous types de données dans leur format natif. La zone de traitements permet d'appliquer tous les processus de transformations et de calculs avec stockage de données intermédiaires afin de répondre à tout type d'analyse. La zone d'accès permet de préparer et stocker les données pour pouvoir appliquer une interrogation ou une analyse spécifique. Parallèlement, la zone de gouvernance, appliquée à toutes les autres zones, est chargée d'assurer la sécurité, la qualité, le cycle de vie, l'accès et la gestion des données à l'aide d'un système des métadonnées spécifique.

En complément de ces propositions, nous avons identifié différents axes de recherche prioritaires. Le premier axe de recherche est l'intégration d'un LD dans le système d'information d'une organisation pouvant déjà contenir un ED. La problématique de recherche est donc de savoir faire coexister deux systèmes ayant des objectifs distincts, comment faire circuler les données entre ces deux systèmes et quelles analyses complémentaires il est possible d'extraire. Le second axe de recherche est relatif à la définition et à la gestion des métadonnées (Ravat, 2019) afin de ne pas rendre les données d'un LD invisibles, incompréhensibles et inaccessibles. Ces métadonnées doivent informer de manière aussi complète que possible aussi bien sur les données que les processus de transformations (ingestion, traitement, diffusion) de ces données. Enfin, le troisième axe de recherche est relatif à la gouvernance des LD. Ces travaux de recherche doivent permettre de définir des politiques, des normes et des pratiques pour gérer les données de sources hétérogènes et les processus associés (transformation et analyse) afin d'assurer une utilisation efficace et sécurisée et une qualité fiable des résultats d'analyse. La gouvernance doit non seulement traiter des données mais également de tous les systèmes informatiques associés aux LD.

## Bibliographies

Dixon, J. (2010) *Hadoop, and Data Lakes,* https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Ravat, F., Zhao, Y. (2019) *Metadata management for data lakes,* East European Conference on Advances in Databases and Information Systems, Communications in Computer and Information Science. pp. 37–44. Springer International Publishing (2019)

Ravat, F., Zhao, Y. (2019) *Data lakes: Trends and perspectives*. Database and Expert Systems Applications - 30th International Conference, DEXA, Lecture Notes in Computer Science. pp. 304–313. Springer International Publishing (2019)