Analyse assistée de documents textuels et tabulaires dans un lac de données avec AUDAL

Pegdwendé N. Sawadogo*, Jérôme Darmont*

*Université de Lyon, Lyon 2, UR ERIC
{pegdwende.sawadogo, jerome.darmont}@univ-lyon2.fr

Résumé. Nous résumons ici un article publié en 2021 dans la conférence internationale ADBIS. Nous y proposons AUDAL, une implémentation de lac de données permettant d'organiser et d'analyser des documents textuels et tabulaires (Sawadogo et al., 2021).

Au cours de la dernière decennie, le concept de lac de données a émergé comme une solution pour le stockage et l'analyse des mégadonnées. Mais le concept de lac de données reste toujours en cours de maturation. De ce fait, il existe encore peu d'approches méthodologiques de conception de lacs de données. Les approches existantes intègrent en effet pour la plupart des données structurées et semi-structurées et excluent donc les données non structurées qui constituent pourtant la majorité des mégadonnées (Diamantini et al., 2018). Pour y remédier, nous proposons AUDAL, une implémentation de lac de données permettant d'organiser et d'analyser des documents textuels et tabulaires.

L'un des composants essentiels dans un lac de données est le système de métadonnées. C'est en effet sur le système de métadonnées que s'appuient les analyses dans un lac de données (Diamantini et al., 2018). Dans AUDAL le système de métadonnées est basé sur le principe de polymorphisme des données (Sawadogo et al., 2019) qui consiste à conserver les données à la fois sous leur forme brute et à travers des représentations raffinées.

Plus concrètement, nous générons systématiquement pour chaque document stocké dans le lac de données une représentation raffinée pouvant servir en l'état pour des analyses automatisées. Dans le cas des documents textuels, cette représentation raffinée prend la forme d'une vectorisation de document (*embedding*). Dans le cas des documents tabulaires, la représentation raffinée correspond à une matérialisation du document à travers une table de base de données relationnelle.

Cette approche est assimilable à l'organisation en zones, communément adoptée dans les lacs de données. En effet, il est d'usage dans les implémentations de lacs de données d'organiser les données à travers plusieurs zones en fonction de leur niveau de maturation (Ravat et Zhao, 2019). Cela permet d'enrichir continuellement les données du lac avec leurs formes transformées, et ainsi faciliter les analyses futures.

Grâce à son système de métadonnées et au principe de polymorphisme des données, le système AUDAL propose à la fois des services de type « recherche de données » et des analyses portant sur le contenu des données. La recherche de données consiste à retrouver des documents stockés dans le lac, sur la base de caractéristiques précises. Ces services, qui sont au nombre de trois, peuvent être utilisés séparément ou conjointement.