Sujet de thèse BIAL-X & ERIC :
Data lakes & Analytics

Lieu : société Bial-X (Limonest, Rhône) et laboratoire ERIC de l'Université de Lyon (Bron, Rhône).

Directeurs de thèse : Sabine LOUDCHER (Professeure en Informatique, laboratoire ERIC), Jérôme DARMONT (Professeur en Informatique, laboratoire ERIC) et Eric FEREY (Président de la société BIAL-X).

Mots-clés : lacs de données, *big data*, science des données.

Contexte

Cette proposition de thèse se place dans le cadre d'une collaboration entre le laboratoire ERIC, qui mène des recherches dans les domaines de la science des données et de l'informatique décisionnelle (*business intelligence*), et l'entreprise Bial-X, cabinet d'experts en *business intelligence* et *big data*. Une première thèse CIFRE entre les deux partenaires, portant sur la conception et l'implémentation d'un premier lac de données destiné à l'habitat social, va être soutenue d'ici décembre 2021.

Sujet

Depuis le début du 21^{e} siècle, les usages des organisations dans les processus de prise de décision sont bouleversés par la disponibilité de grands volumes de données hétérogènes appelées *big data*. Ces mégadonnées constituent une véritable opportunité pour les organisations, mais 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 [6]. C'est dans ce contexte que Dixon introduit le concept de lac de données (*data lake*), en guise de solution aux problèmes induits par l'hétérogénéité des mégadonnées [7].

Un lac de données propose de stocker les données dans leur format d'origine et sans schéma prédéfini [5]. Cette approche, qualifiée de *schema-on-read*, s'oppose à celle des entrepôts de données, appelée *schema-on-write*, où les données sont transformées avant leur stockage. Avec un tel principe, tous types de données peuvent cohabiter dans un lac de données, qu'elles soient structurées ou non. Pour être exploitable, un lac de données a besoin de métadonnées qui permettent de décrire les données stockées dans le lac, ainsi qu'un système efficace de gestion de ces métadonnées. Le laboratoire ERIC a étendu la définition du concept de lacs de données ainsi que les fonctionnalités que le système de métadonnées devait avoir pour être complet et efficace [9]. Récemment, il vient de proposer un modèle de métadonnées, baptisé goldMEDAL, basé sur 4 concepts principaux : entité de données, groupement, lien et processus [11]. Une étude des modèles de métadonnées actuels montre que goldMEDAL permet de généraliser les concepts proposés dans la littérature, faisant de lui le modèle le plus générique [4, 7, 8].

La 1re thèse CIFRE entre le laboratoire ERIC et la société Bial-X est ancrée dans le domaine de l'habitat social, domaine en lien avec les clients de l'entreprise. C'est dans ce contexte qu'un premier prototype de lac de données dédié à l'habitat social vient d'être développé [10].

Après avoir démontré l'intérêt et la faisabilité d'utiliser un lac de données dans le contexte de l'habitat social, les partenaires souhaitent poursuivre avec la conception, la mise en place et l'industrialisation de lacs dans différents domaines liés aux activités des clients de la société Bial-X. De plus les partenaires souhaitent explorer le nouveau concept de *data mesh* pour l'organisation et l'exploitation des données hétérogènes massives [1].

A partir de 2022, dans le cadre du concept de *business intelligence and analytics* (BI&A), l'objectif de la présente thèse sera de permettre :

- la création assistée ou semi-automatique de métadonnées au moment de l'insertion de nouvelles entités de données dans un lac, et ce grâce à l'extraction automatique d'informations depuis les données « primaires » par des méthodes d'intelligence artificielle ;
- l'interrogation des données du lac sur la base de requêtes formulées sur les métadonnées ;
- l'utilisation du lac possible non seulement par des *data scientists*, mais aussi par des experts métier pour extraire et analyser des données hétérogènes ;
- la généralisation et l'industrialisation des lacs de données dans différents projets de la société Bial-X ;
- l'étude les possibilités offertes par le nouveau concept de *data mesh* pour l'industrialisation des processus de science de données.

Cette thèse comprendra trois grand niveaux de réalisation : un niveau conceptuel ou théorique pour concevoir les différentes propositions, un niveau technique pour l'implémentation informatique des solutions et un niveau applicatif avec la mise en œuvre des propositions sur des données réelles et sur des problématiques métiers des clients de la société Bial-X.

D'un point de vue technique, les propositions faites par le/la doctorant(e) seront implémentées et intégrées aux solutions logicielles développées par la société Bial-X. Le/la doctorant(e) intégrera une équipe de spécialistes passionnés, à dimension humaine, où il pourra mettre en action ses propositions, sa créativité et ses compétences sur des cas concrets, réels et passionnants.

Contact

Merci d'adresser, avant le 1er juin 2021, votre candidature avec un CV, une lettre de motivation ainsi que vos notes de l'année universitaire en cours et de l'année dernière à eric.ferey@bial-x.com, jerome.darmont@univ-lyon2.fr et sabine.loudcher@univ-lyon2.fr

Les candidat·es retenu·es seront convoqué·es pour un entretien en visioconférence.

Références

[1] Zhamak Dehghani. 2019. https://martinfowler.com/articles/data-monolith-to-mesh.html

[2] Diamantini C., Giudice P. L., Musarella L., Potena D., Storti E., and Ursino D. 2018. A New Metadata Model to Uniformly Handle Heterogeneous Data Lake Sources. *In European Conference on Advances in Databases and Information Systems (ADBIS 2018)*, Budapest, Hungary, pp. 165–177.

[3] Dixon J. 2010. Pentaho, Hadoop, and Data Lakes. *https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/.*

[4] Eichler R., Giebler C., Gröger C., Schwarz H., and Mitschang B. 2020. HANDLE-A Generic Metadata Model for Data Lakes. *In International Conference on Big Data Analytics and Knowledge Discovery (DaWak 2020)*, Bratislava, Slovakia, pp.73–88.

[5] Hai R., Geisler S. and Quix C. 2016. An Intelligent Data Lake System. *In International Conference on Management of Data (SIGMOD 2016)*, San Francisco, CA, USA, ACM Digital Library, pp. 2097–2100.

[6] Miloslavskaya, N. and A. Tolstoy. 2016. Big Data, Fast Data and Data Lake Concepts. *In Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2016)*, NY, USA, Volume 88 of Procedia Computer Science, pp. 1–6.

[7] Quix C., Hai R., and Vatov I. 2016. Metadata Extraction and Management in Data Lakes With GEMMS. In Complex Systems Informatics and Modeling Quarterly 9 (December 2016), pp. 289–293.

[8] Ravat F. and Zhao Y. 2019. Metadata management for data lakes. *In European Conference on Advances in Databases and Information Systems (ADBIS 2019)*, Bled, Slovenia. Springer, pp. 37–44.

[9] Sawadogo P. N., Scholly E., Favre C., Ferey E., Loudcher S., and Darmont J. 2019. Metadata systems for data lakes: models and features. *In International Workshop on BI and Big Data Applications (BBIGAP@ADBIS 2019)*, Bled, Slovenia, Springer, pp. 440-451.

[10] Scholly E., Favre C., Ferey E., and Loudcher S. 2021. Houdal : A data lake implemented for public housing. *In International Conference on Enterprise Information Systems (ICEIS 2021).* To appear.

[11] Scholly E., Sawadogo P. N., Liu P., Espinosa-Oviedo J. A., Favre C., Loudcher S., Darmont J., and e Noûs. Coining goldmedal: A new contribution to data lake generic metadata modeling. *In International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP@ EDBT 2021)*, Nicosia, Cyprus. pp.31-40.