SUJET STAGE MASTER

Titre : Architectures de lacs de données

Mots-clés : Data lakes, Big data, Metadata

Noms des encadrants :

Jérôme Darmont (laboratoire ERIC), Sabine Loudcher (laboratoire ERIC), Franck Ravat (laboratoire IRIT)

Lieu :

Laboratoire ERIC (Université Lyon 2, Campus Porte des Alpes, Bron) / journées ou courts séjours à l'IRIT (Toulouse)

Durée :

5 mois à partir de mars ou avril 2020

Rémunération :

Indemnités de stage légales (3,60 euros par heure pour 35 heures de travail par semaine)

Sujet :

Le concept de lac de données (*data lake*) a été introduit comme une alternative aux entrepôts et magasins de données pour le stockage et l'analyse des mégadonnées (*big data*). Le lac de données est un vaste dépôt de données brutes de structures hétérogènes, alimenté par des sources de données externes et à partir duquel des analyses diverses peuvent être réalisées. Un lac de données propose un stockage intégré des données sans schéma prédéfini. En l'absence de schéma de données, un système de métadonnées efficace est essentiel pour rendre les données interrogeables et empêcher ainsi le lac de se transformer en « marécage » (*data swamp*) inexploitable.

Les premiers travaux sur les lacs de données ont rapidement associé ce nouveau concept à la technologie Hadoop en le considérant comme une méthodologie consistant à utiliser des technologies libres ou peu coûteuses, typiquement Hadoop, pour assurer le stockage, le traitement et l'exploration des données brutes au sein d'une entreprise. Cependant, cette vision est de plus en plus minoritaire dans la littérature, le concept de lac de données est désormais également associé à des solutions propriétaires comme Azure ou IBM ou encore les multistores.

Par ailleurs, les lacs de données sont le plus souvent considérés comme des bacs à sable au sein desquels les *data scientists* mènent des travaux exploratoires. En revanche, les laboratoires ERIC (Université de Lyon) et IRIT (Université de Toulouse) travaillent de concert à rendre les lacs de données accessibles à un plus large panel d'acteurs, par exemple des *business users* au fait des outils décisionnels ou des chercheur·es. Il s'agit ainsi d'industrialiser les processus de science des données pour étayer le nouveau concept de *business intelligence and analytics* (BI&A).

Dans ce contexte, les objectifs du stage sont :

- sur la base de l'état de l'art :
    - de définir une architecture fonctionnelle de référence pour les lacs de données,
    - de recenser les grands scénarios d'utilisation des lacs de données ;
- de proposer et de tester des architectures techniques (piles technologiques) alternatives relatives à ces scénarios ;
- de développer un outil de génération automatique d'architectures physiques répondant aux différents scénarios.

Compétences requises :

Le sujet de stage s'adresse à des étudiant·es en 1ère ou 2e année de master (ou équivalent) en informatique décisionnelle ou en sciences des données. Des compétences en bases de données, en entrepôts de données, en traitement des données massives ou en technologies liées aux *big data* seront particulièrement appréciées.

Contact :

Merci d'adresser, avant le 15 janvier 2020, 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 jerome.darmont@univ-lyon2.fr, sabine.loudcher@univ-lyon2.fr et Franck.Ravat@irit.fr
Les candidat·es retenus seront convoqué·es pour un entretien fin janvier.