Laboratoire : Laboratoire d'Informatique de Bourgogne (LIB) – Équipe Science des Données

Durée : 3 ans

Contexte : Cette thèse est financée par la région Bourgogne, Franche-Comté par le dispositif intitulé «Itinéraire Chercheurs Entrepreneurs» (I.C.E). Ce parcours intègre une double compétence recherche et entrepreneuriat/management est Il vise à promouvoir l'émergence d'entreprises à forte valeur ajoutée sur le territoire régional et passe par l'identification et la professionnalisation de chercheurs ayant la volonté de s'inscrire dans ce type de projet.

Le but de cette thèse est de faciliter l'usage d'un data lake et de proposer des méthodes permettant d'automatiser la création de méta-données, accompagnées de solutions techniques pour les mettre en place, afin de faciliter l'exploitation des données et leur analyse. Une approche consistant à automatiser l'annotation des données lors de leur intégration dans le data lake puis à mettre en relation des données avec un graphe de connaissance pour créer des méta-données fiables, est prometteuse. Cet enrichissement peut être incrémental en bénéficiant des résultats produits par les analyses et il peut également être piloté par des ontologies de domaines, par exemple en combinant des techniques de machine learning pour des données textuelles (ou peu structurées) avec des outils de l'IA symbolique (ontologies et logiques de description).

Sujet : L'analyse des données massives est une discipline en plein essor qui a pour objectif d'extraire de la valeur des données. Les informations ainsi extraites peuvent ensuite servir à expliquer et décrire un événement passé, prédire les événements à venir ou encore prescrire des solutions permettant d'améliorer la situation actuelle.

De nombreuses méthodes d'analyse existent (machine learning, algorithmes d'analyse de réseaux complexes, stream processing, etc.), s'appuyant sur des modèles de données différents (graphes, relations, matrices, tenseurs, etc.), ayant des contraintes d'application variées et n'ayant pas les mêmes capacités d'interprétabilité. La collaboration entre des experts des sciences des données et des experts métiers est essentielle.

Le stockage des données est une phase critique qui doit permettre de pouvoir ensuite les exploiter efficacement lors des analyses. Les data warehouses dont le principe date de plus de 25 ans sont peu adaptés à la réalité des données massives. En effet, ces dernières évoluent rapidement tant au niveau de leur nature que de leur format : un data warehouse est statique et ne peut pas incorporer des modifications des schémas des données facilement et encore moins ingérer des flux de données importants et continus.

Pour compenser ce manque de flexibilité, la notion de data lake a été proposée en 2010, par James Dixon [3], de Pentaho (société spécialisée dans les technologies décisionnelles). La définition a évolué depuis vers le consensus suivant : ce sont des systèmes dans lesquels des données hétérogènes (de par leur format, leur provenance, leur utilité, etc.) sont stockées, et qui offrent des outils pour extraire des jeux de données afin de réaliser des analyses beaucoup plus variées que celles possibles avec les data warehouses. Pour prendre en compte la diversité des formats de données, différents systèmes de stockage distribués ou non peuvent être combinés pour former un polystore [11].

Toutefois, les data lakes peuvent être victimes d'un excès de flexibilité, et se transformer en data swamps, dans lesquels il devient extrêmement difficile de naviguer, de localiser et d'extraire des données pertinentes. Il est donc essentiel de développer des techniques permettant d'organiser et de mettre en forme les data lakes. Cela demande souvent un investissement humain considérable, qui a entraîné la naissance de nouveaux rôles liés aux données, comme les data stewards, chargés de

maintenir un catalogue de méta-données du contenu du data lake, afin de pouvoir identifier le ou les jeux de données pertinents pour des analyses métier. Toutefois, la nécessité de tels rôles rend les data lakes inaccessibles pour la majorité des entreprises, et requiert une cohérence dans le catalogue afin de pouvoir exploiter correctement les données.

Les deux orientations principales des recherches autour des data lakes consistent soit à diviser le data lake en data ponds (qui peuvent regrouper des données ayant la même fonctionnalité, le même format, etc.) [8], soit à ajouter des méta-données [9,10,7]. La première approche, très similaire aux data-marts, ne résout pas réellement les problèmes puisqu'elle segmente uniquement le data lake en unités plus petites avec peu de liens entre elles. Or la valeur extraite des données massives provient souvent de leur analyse conjointe découvrant ainsi des liens cachés entre les données.

Des solutions telles que Delta Lake [1] ou Lakehouse [2] regroupent différents moyens techniques facilitant l'utilisation de données hétérogènes et les interactions entre les différents acteurs. Ces solutions nécessitent souvent un expert technique pour orchestrer et tirer profit de tels systèmes. En se concentrant sur les aspects stockage, elles ne permettent pas de gérer finement l'organisation et la navigation dans le data lake pour localiser les jeux de données adaptés. De ce fait, cela pénalise les entreprises qui ne peuvent pas avoir d'équipe dédiée à cette tâche, et les empêche donc de bénéficier du gain de compétitivité que peut leur apporter l'exploitation de leurs données.

Le but de cette thèse est de faciliter l'usage d'un data lake et de proposer des méthodes permettant d'automatiser la création de méta-données, accompagnées de solutions techniques pour les mettre en place, afin de faciliter l'exploitation des données et leur analyse. Une approche consistant à automatiser l'annotation des données lors de leur intégration dans le data lake puis à mettre en relation des données avec un graphe de connaissance pour créer des méta-données fiables, est prometteuse. Cet enrichissement peut être incrémental en bénéficiant des résultats produits par les analyses et il peut également être piloté par des ontologies de domaines, par exemple en combinant des techniques de machine learning pour des données textuelles (ou peu structurées) avec des outils de l'IA symbolique (ontologies et logiques de description). En effet, les mécanismes d'annotation manuels ont démontré leurs limites comme le présente Gorelik [6] au travers d'un biais d'annotation qui traduit le fait que les données les mieux annotées et documentées sont celles qui sont le plus accédées, entraînant par la même occasion un ajout d'annotations sur ces mêmes données populaires, et ce au détriment des autres jeux de données indifféremment de leur qualité.

Références

[1] Michael Armbrust, Tathagata Das, Liwen Sun, Burak Yavuz, Shixiong Zhu, Mukul Murthy, Joseph Torres, Herman van Hovell, Adrian Ionescu, Alicja Luszczak, et al. Delta lake : high-performance acid table storage over cloud object stores. Proceedings of the VLDB Endowment,13(12) :3411-3424, 2020.

[2] Michael Armbrust, Ali Ghodsi, Reynold Xin, and Matei Zaharia. Lakehouse : A new generation of open platforms that unify data warehousing and advanced analytics. CIDR, 2021.

[3] James Dixon. Pentaho, Hadoop, and data lakes. blog, Oct, 2010.

[4] Annabelle Gillet, Eric Leclercq, and Nadine Cullot. Evolution et formalisation de la lambda architecture pour des analyses à hautes performances-application aux données de twitter. Revue ouverte d'ingénierie des systèmes d'information, 2021.

[5] Annabelle Gillet, Eric Leclercq, and Nadine Cullot. Lambda+, the renewal of the lambda architecture : Category theory to the rescue. In 33rd International Conference on Advanced Information Systems Engineering (CAISE) (à paraître), page 15, 2021.

[6] Alex Gorelik. The enterprise big data lake : Delivering the promise of big data and data science. O'Reilly Media, 2019.

[7] Moditha Hewasinghage, Jovan Varga, Alberto Abello, and Esteban Zimanyi. Managing polyglot systems metadata with hypergraphs. In International Conference on Conceptual Modeling, pages 463-478. Springer, 2018.

[8] Bill Inmon. Data Lake Architecture : Designing the Data Lake and avoiding the garbage dump. Technics publications, 2016.

[9] Pegdwendé Sawadogo and Jérôme Darmont. On data lake architectures and metadata management. Journal of Intelligent Information Systems, pages 1-24, 2020.

[10] Pegdwendé Sawadogo, Tokio Kibata, and Jérôme Darmont. Metadata management for textual documents in data lakes. International Conference on Enterprise Information Systems (ICEIS), 2019.

[11] Michael Stonebraker and Ugur Cetintemel. "one size ts all" an idea whose time has come and gone. In Making Databases Work : the Pragmatic Wisdom of Michael Stonebraker, pages 441-462. 2018.

Profil du candidat : Le candidat à cette thèse doit avoir un Master 2 en informatique (ou équivalent).

Formation et compétences requises :

Le candidat devra avoir effectué un cursus en informatique et démontré ses compétences en gestion des données et en intelligence artificielle.

De bonnes connaissances en bases de données, web sémantique, ontologies, logiques du premier ordre sont nécessaires. Le candidat devra également avoir une première expérience en analyse de données massives (données de réseaux sociaux par exemple).

Adresse d'emploi :

Laboratoire d'Informatique de Bourgogne (LIB), Université de Bourgogne, UFR Sciences et Techniques, 9, Avenue Alain Savary 21078 Dijon.

Contacts :

Eric Leclercq - Eric.Leclercq@u-bourgogne.Fr

Nadine Cullot – Nadine.Cullot@u-bourgogne.fr