DISPOSITIF « JEUNES CHERCHEURS ENTREPRENEURS »

Dossier de PROJET DE THESE

Année 2018

Cadre de la recherche	Renforcement de l'axe Sciences des données dans le cadre de la restructuration du laboratoire LE2I et des collaborations avec la société ATOL C&D en recherche
Ecole Doctorale (N° et Nom)	Ecole Doctorale SPIM (Sciences Pour l'Ingénieur et Microtechniques)
Unité de recherche (label, N° et nom) + directeur	**Laboratoire Electronique Informatique et Image (LE2I) – EA 7508** **Directeur : Dominique Ginhac**
Equipe interne	« Combinatoire, Réseaux et Sciences des données »

ENCADREMENT	
Directeur de Thèse (Nom et Prénom)	**CULLOT Nadine, Professeur 27 (PR1)**
Téléphone et mail	Tél : 03 80 39 58 85/06 71 96 62 58 Mail : nadine.cullot-bourgogne.fr
Adresse professionnelle	Laboratoire LE2I 8 Rue Alain Savary - BP 47870 21078 DIJON CEDEX

PROJET	
Sujet de la Thèse	**Modélisation et développement d'un observatoire générique pour la collecte et l'analyse de données massives**
Description du projet de thèse	Les données qualifiées de données à grandes dimensions (Big-Data) en raison de leur volume et du nombre de caractéristiques décrivant chaque donnée, sont issues de domaines très variés. Leur valorisation ne peut se faire qu'avec des algorithmes complexes et souvent couteux à exécuter, qui pris séparément, ne permettent d'éclairer qu'une partie des propriétés des données comme les structures communautaires, les modes de diffusions de messages viraux, les motifs récurrents, etc. L'analyse de ces données à grande dimension peut se faire selon deux objectifs fondamentalement différents dans leur finalité. Il peut s'agir de construire des systèmes logiciels pour effectuer des prédictions, des recommandations ou plus généralement guider des actions ou bien d'analyser ces données dans le but de produire de la connaissance, expliquer ou comprendre des phénomènes et dans ces cas, l'analyse des données doit être intégrée dans une méthodologie itérative et incrémentale de la production de connaissance. Aussi, pour une analyse fine des données à grandes dimensions, il est nécessaire d'avoir recours à plusieurs types d'algorithmes qui reposent sur fondations

	formelles différentes : théorie des graphes, statistiques, algèbre linéaire et multi-linéaire, etc. Une des problématiques est de mettre en adéquation le modèle des données stockées (relationnel, orienté colonne, graphe, document, etc.) avec le modèle de données requis par les algorithmes (séries temporelles, graphes, hypergraphe, graphes multi-couches, matrices d'adjacence, matrices stochastiques, etc.). Cette problématique fait resurgir le problème de l'indépendance logique des données, et concerne l'évolutivité des systèmes logiciels et des systèmes de stockage des données. Actuellement les outils d'analyse de données sont fortement couplés au stockage et le recours aux processus ETL (*extract transform load*) impacte négativement les temps de développement et de mise à disposition des résultats des analyses en réalisant des transformations de données et de modèles complexes à mettre en œuvre. Notre proposition se situe dans le contexte du *data intensive HPC* nouveau champ de recherche issu de l'association du calcul haute performance (*HPC – high performance computing*) avec le stockage et l'analyse des masses de données (*Big Data analytics*). Elle vise à développer une architecture pour le stockage de données reposant sur une approche multi-paradigmes, c'est-à-dire stockant les données dans un ou plusieurs systèmes (SGBDR, graphe, orienté colonne etc.) en fonction de leur nature et de leur utilisation et offrant des services d'analyses ciblés. Ce travail nécessite une collaboration étroite avec des experts métiers pour la collecte des données et la validation des propositions. Les développements menés dans cette thèse viendront enrichir la plateforme de stockage et d'analyse déjà initiée dans le cadre du travail de thèse de Ian Basaille-Gahitte (soutenue en février 2018) pour développer un « Observatoire de collecte stockage et d'analyse de données massives
Connaissances et compétences requises	Le candidat devra être titulaire d'un Master informatique, ou équivalent avec des compétences en Bases de données, Systèmes d'informations, Techniques d'intelligence artificielle, et en Sciences de données pour les aspects architectures distribuées, stockage distribué et méthodes d'analyses. Le profil de thèse pourra être orienté selon les compétences du candidat davantage vers les aspects stockage multI-modèles ou analyse et éventuellement également visualisation pour le rendu des résultats d'analyse.