Offre de stage 2022-2023

LISTIC

Titre	**Benchmarking de middleware pour le Big Data**
Niveau du stage	BUT, L3, M1 ou M2 recherche Ingénieur année 1, 2 ou 3
Date de début et durée	De février – mars 2023 à juillet 2023 4 à 6 mois
Ville, Pays	Annecy-le-Vieux, *France*
Laboratoire	LISTIC - Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance https://univ-smb.fr/listic/
Description du sujet de stage	**Multidisciplinaire :** développement, recherche, test **Mots clés :** Processus ETL, Redis, Kafka, Hadoop. **Contexte :** Le sujet de ce stage se situe dans le cadre d'un projet de recherche visant à proposer une approche de modélisation personnalisable d'un pipeline Big Data pour l'acquisition, le traitement et le stockage de données pour une analyse future. En effet, de nos jours les sources et les types de données se multiplient au sein de l'entreprise : fichiers plats, données opérationnelles, nouveaux services internet, différents réseaux sociaux, nouvelles applications de l'internet des objets (IOT), etc. Cette révolution informationnelle a généré une grande masse de données, dite « Big Data ». Le Big Data est caractérisé par le grand « volume » de données collectées par l'entreprise, la « variété » de ces données, qui peuvent être structurées, semi-structurées ou non structurées et aussi par la fréquence de l'arrivée des données « vitesse » qui devrait être prise en considération. Pour faire face aux challenges de Big Data une bonne variété de technologies dédiées est apparue, tels que l'écosystème d'Hadoop (HDFS, Map Reduce, Yarn, etc), Flink, Kafka, Elasticserach, Kibana, etc. Dans la littérature de différentes solutions architecturales Big Data ont été proposées. Dans ces architectures se trouvent une ou plusieurs technologies pour répondre à un besoin spécifique. Par ailleurs, le choix de ces technologies n'est pas toujours suffisamment justifié. **Objectif du stage :** L'objectif de ce stage sera le déploiement de trois différentes architectures Big Data pour l'extraction, le traitement, le chargement (ETL) des données. Dans chacune de ces architectures, l'étudiant teste le déploiement des technologies selon des critères à définir (RAM, réseau, stockage, etc). L'étudiant est appelé aussi à étudier la compatibilité entre les technologies mises en œuvre au sein d'une même architecture. De plus, tout au long du stage, il serait utile de prendre note de tous les problèmes rencontrés, en particulier celle de configuration et de préciser comment sont-ils surmontés. À la fin du stage, l'étudiant est appelé à synthétiser toutes les étapes menées et relever les résultats du travail de benchmarking. Pour la mise en œuvre des architectures proposées le candidat pourra avoir accès durant la période du stage à la plateforme MUST, mésocentre de stockage et de calcul scientifique mutualisée ouverte sur la grille de recherche européenne utilisée par les chercheurs des différents laboratoires de l'USMB ainsi qu'a des machines de calcul internes au laboratoire.

	Travail à effectuer : Les résultats attendus de ce stage sont les suivants : 1. Acquérir des compétences sur les technologies Big Data. 2. Mettre en œuvre trois différentes architectures pour l'extraction, le traitement et le chargement de données. 3. Valider les implémentations par différents exemples. 4. Synthétiser tous les résultats des comparaisons faites. Par exemple, une comparaison des approches de Kafka et Redis pour la gestion des flux de données. **Candidature :** La candidature se fait via une première prise de contact par mail en nous fournissant un CV, une lettre de motivation et le dernier relevé de notes. **Références :** • Asma Dhaouadi, Khadija Bousselmi, Mohamed Mohsen Gammoudi, Sébastien Monnet, Slimane Hammoudi, Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons. Data 7(8): 113 (2022) • Asma Dhaouadi, Khadija Bousselmi, Sébastien Monnet, Mohamed Mohsen Gammoudi, Slimane Hammoudi, A Multi-layer Modeling for the Generation of New Architectures for Big Data Warehousing. AINA (2) 2022: 204-218 • Tardio, R., Mate, A., & Trujillo, J., An iterative methodology for defining big data analytics architectures. *IEEE Access*, *8*, 210597-210616 (2020).
Compétences requises	- Connaissance et compréhension des phases d'acquisition, de traitement, de stockage de données. - Connaissances relativement bonnes de l'écosystème Hadoop et d'autres technologies : Redis, Spark, etc. - Configuration des technologies Big Data. - Comprendre, analyser et rédiger des documents scientifiques et techniques.
Gratification	Selon législation en vigueur (25,20 € / journée de travail)
Tuteurs / Contacts	Asma Dhaouadi, Khadija Bousselmi et Sébastien Monnet E-mail : {prenom.nom}@univ-smb.fr