

Proposition de Stage Master II ou projet de fin d'études 2017 - 2018

Sujet Stage

Participer au projet de mise en place d'un OLAP entreprise basé sur l'écosystème Hadoop.

Mots clés

Modélisation multi-décisionnelle, Entrepôt de données (Data Warehouse), Serveur OLAP-ROLAP, Architecture Hadoop (HDFS, Yarn, Hive, Impala, Kylin, etc.)

Informations complémentaires

Encadrant(s) : Jamal Malki (*), Jean-Philippe Baron (**)

Cadre de coopération : L3i (*) – aYaline (**)

Date de début du stage : à partir de janvier 2018

Durée du contrat : 5 mois (minimum)

Stage rémunéré

Contexte

Le stage se déroule dans le cadre du projet PLAIBDE : « Plateforme Intégrée Big-data pour les Données Entreprise ». Ce projet fait partie du programme FEDER-FSE 2014-2020 porté par la région Nouvelle Aquitaine.

Ce projet est dirigé par le consortium suivant :

1. L'entreprise aYaline¹ : partenaire industriel et chef de fil du projet
2. Le laboratoire L3i² : partenaire scientifique, le L3i (laboratoire Informatique, Image, Interaction) fait partie de l'université de La Rochelle
3. Le laboratoire LIAS³ : partenaire scientifique, le LIAS (Laboratoire d'Informatique et d'Automatique pour les Systèmes), fait partie de l'ENSMA (École Nationale Supérieure de Mécanique de d'Aéronautique – Futuroscope), université de Poitiers

L'objectif du projet FEDER-PLAIBDE est le développement d'un écosystème Big-Data métier dans les domaines d'activités relevant de l'expertise de l'entreprise aYaline : E-Commerce, E-Tourisme, E-Collectivité et d'autres.

¹ <http://www.ayaline.com>

² <http://l3i.univ-larochelle.fr>

³ <https://www.lias-lab.fr>

Objectifs du stage

Actuellement, l'entreprise aYaline développe pour ses clients des plateformes pour le traitement analytique en ligne (OLAP). Pour implémenter le concept OLAP dans le cadre de ses projets, aYaline a choisi l'architecture ROALP qui se base sur le principe de gestion des dataWarehouse dans des bases de données relationnelles. L'architecture technique des plateformes en production reposent sur le serveur OLAP Mondrian géré par le système Saiku en mode standalone ou intégré au système Pentaho Community.

En matière de traitement analytique en ligne (en anglais BI : Business Intelligence), les architectes étaient confrontés à un choix relativement simple entre deux technologies de traitement analytique en ligne : multidimensionnelle ou relationnelle. Aujourd'hui, l'intelligence décisionnelle proposée aux entreprises est considérablement plus exhaustive, et les blocs fonctionnels qui font l'architecture des plates-formes BI se sont multipliés, tout comme les systèmes de gestion d'entrepôts de données sous-jacents. Deux points importants sont prises en compte :

1. les données : les volumes de données augmentent rapidement ; les données sont hétérogènes ; les données sont temps réels, etc. ;
2. les usages : les exigences de l'utilisateur final en matière de rapports et de données d'intelligence décisionnelle se développent et se complexifient ...

L'entreprise aYaline souhaite exploiter les nouvelles sources de données émergentes, la volumétrie croissante des données et leurs nouveaux usages pour développer des applications analytiques efficaces répondant aux nouvelles exigences de ses clients. Cependant, les technologies traditionnelles ne sont pas suffisantes pour traiter de gros volumes de données ou diffuser des données en temps réel. Par conséquent, au cours des dernières années, de nombreuses entreprises ont investi dans le développement de nouvelles technologies capables de répondre à ces nouvelles problématiques. Ces technologies de traitement des données sont généralement coûteuses. Par conséquent, Hadoop, un framework open source pour l'informatique distribuée a été développé.

Dernièrement, les entreprises commencent à adopter et à intégrer le framework Hadoop pour améliorer leurs capacités de traitement des données. Cependant, l'adoption de Hadoop apporte de nouveaux défis pour l'architecture entreprise en matière de stockage, de persistance, de traitement et d'analyse des données, et en matière de requêtage et génération de rapport.

Les missions principales de ce travail de stage sont :

1. comprendre et analyser l'architecture des entrepôts de données présente ;
2. étudier les solutions de l'intégration de la technologie Hadoop dans l'environnement d'entreposage de données déjà en cours. La technologie Hadoop étant relativement nouvelle, les cas d'utilisation professionnelle où les implémentations ont réussies et ont été publiées ne sont pas nombreuses. Par conséquent, il n'existe pas de pratiques exemplaires ou de directives existantes ;
3. étudier l'intégration de Hadoop dans l'environnement d'entrepôt de données d'entreprise existant. Cette étude soit comprendre l'explication de l'intégration des données dans Hadoop

suite aux meilleures pratiques de Kimball provenant de l'entreposage général de données, comme la mise en œuvre du schéma de démarrage et les plans des processus individuels de gestion des données, etc.

4. la mise en œuvre du processus d'extraction, de transformation et de chargement, l'utilisation des outils de business intelligence et de reporting, puis la mise en œuvre physique de Hadoop et l'emplacement du cluster Hadoop.

Prérequis

Ce sujet d'adresse aux étudiant(e)s en Master 2 Informatique, ou élève de dernière année d'une école d'ingénieur en informatique.

Vous êtes rigoureux dans votre travail mais aussi créatif avec une forte envie d'apprendre et de vous investir dans un projet décisionnel Blg-Data de taille réelle au sein d'un environnement professionnel regroupant divers acteurs.

Candidature

Merci d'adresser votre dossier de candidature à : jmalki@univ-lr.fr

Le dossier de candidature doit contenir :

- ✓ le CV
- ✓ les relevés de notes des 2 dernières années (M1 et M2)
- ✓ la lettre de motivation
- ✓ tout autres documents pouvant appuyer la candidature