Contrat Ingénieur de Recherche (Post-Doc)
Laboratoire L3i, La Rochelle

Sujet

Analyse des données à l'échelle de l'entreprise : étendre le DataWarehouse au Data Lake.

Projet

Projet FEDER-FSE PLAIBDE (Région Nouvelle Aquitaine, Programme 2014-2020) : « Plateforme Intégrée Big-data pour les Données Entreprise ».

Informations complémentaires

Lieu de travail : Université de La Rochelle, Laboratoire L3i
Encadrant(s) : Jamal Malki (2), Alain Bouju (2), Jean-Philippe Baron (1)
Date de début du contrat : 1er janvier 2018
Durée du contrat : 12 mois
Rémunération : 2300 €/mois Net

Contexte

Le contrat 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 aYaline[1] : partenaire industriel et chef de fil du projet.
2. Le laboratoire L3i[2] : partenaire scientifique, le L3i (laboratoire Informatique, Image, Interaction) fait partie de l'université de La Rochelle.
3. Le laboratoire LIAS[3] : 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 PLAIBDE est le développement d'un écosystème Big-Data métier dans les domaines d'activités relevant de l'expertise de aYaline : E-Commerce, E-Tourisme, E-Collectivé, …

[1] http://www.ayaline.com
[2] http://l3i.univ-larochelle.fr
[3] https://www.lias-lab.fr

Objectifs

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 (entrepôts de données) 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 pris 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 sur l'ensemble des données de l'entreprise. Le Data Lake (ou lac de données) fait son apparition pour répondre à ces besoins. C'est un système informatique capable de stocker en un seul endroit toutes les données présentes dans une entreprise. Elle tend à se substituer peu à peu à son ancêtre, le DataWarehouse. Toutefois, adapter l'analytique aux lacs de données présente des verrous scientifiques et technologiques liés à l'état de l'art des solutions actuelles.

Dernièrement, de nombreuses entreprises ont investi dans le développement de nouvelles technologies capables de répondre à ces nouvelles problématiques. Elles commencent à adopter et à intégrer l'écosystème Hadoop pour améliorer leurs capacités de traitement des masses de données. Dans un tel scénario, les données de l'entreprise sont d'abord chargées sur la plateforme Hadoop, puis on leur applique des outils d'exploration de données et d'analytique, à l'emplacement qu'elles occupent sur les noeuds d'ordinateurs génériques du cluster Hadoop.

Cependant, l'adoption des solution basées sur l'écosystème Hadoop apporte de nouveaux défis pour l'architecture entreprise en matière de stockage, de persistance, de traitement, d'analyse et de visualisation des données, mais aussi en matière de gouvernance. Si Hadoop apparaît comme une évidence pour construire un Data Lake d'ampleur, il serait assez réducteur de penser qu'il soit l'unique solution à implémenter. De ce fait aujourd'hui, on trouve des possibilités, avec Kafka, Storm, Spark-Streaming et récemment le projet Kylo de Teradata.

Les missions principales de ce travail 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. Les technologies Data Lake basées sur Hadoop étant relativement nouvelles, 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 Data Lake Hadoop dans l'environnement d'entrepôt de données existant de d'entreprise. Cette étude doit comprendre l'explication de l'intégration des données dans Hadoop et les plans de communication entre les sources de données opérationnelles métiers de l'entreprise et le Data Lake ;

4. la mise en œuvre du processus d'extraction, de chargement et de transformation (ELT : Extract-Load-Transform), l'utilisation des outils de business intelligence et de reporting, puis la mise en œuvre physique de Hadoop et l'emplacement du cluster Hadoop (comparaison entre une architecture Cloud et On-Premise).

Profile recherché

Le ou la candidate à ce poste doit être titulaire d'un doctorat en informatique, de préférence dans le domaine de la gestion des grandes masses de données.

Vous êtes rigoureux dans votre travail mais aussi créatif avec une forte envie d'apprendre et de vous investir dans un projet Big-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 détaillant les activités de recherche et les publications
- ✓ la lettre de motivation
- ✓ tout autres documents pouvant appuyer la candidature