Offre de stage en informatique 2017
de 5 mois minimum

Sujet

Big Data & Data Science : veille active et évaluation d'outils

Contexte

La R&D d'EDF (2000 chercheurs) a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du groupe EDF, d'identifier et de préparer les relais de croissance à moyen et long termes. Avec l'accroissement du volume et la diversification des données à analyser, les entreprises doivent être outillées et prêtes à analyser ces données rapidement et de réagir en conséquence. Documents, emails, mesures de capteurs, logs de serveur Web sont autant de sources hétérogènes qu'il faut savoir intégrer et valoriser à grande échelle. Le « **Big Data** » désigne davantage un défi à relever qu'un type de données particulier. Dans le même esprit, l'**Open Data** (données ouvertes) vise à obtenir la mise à disposition des données numériques. En ce sens, les entreprises peuvent considérer le mouvement Open Data comme une opportunité pour développer un écosystème autour de la donnée, augmenter leur capacité d'innovation et renforcer la culture digitale. Afin de préparer le Groupe EDF aux enjeux stratégiques du Big Data et de l'Open Data, une équipe de la R&D travaille activement sur ces sujets. A partir de janvier 2017, EDF R&D met en place un nouveau dispositif appelé Data Innovation Lab (DIL) pour accélérer la valorisation des données par les techniques de Data Analytics et par une approche agile, en forte interaction avec les métiers d'EDF. Ce dispositif réunit sur un plateau projet des Data scientists et des Data analysts autour de cas concrets fournis par les métiers d'EDF. Il s'appuie sur une infrastructure de calcul performante (cluster Hadoop, serveurs dédiés, calculateur haute performance, etc.) et sur un socle d'outils et d'environnements collaboratifs en open source pour la plupart (R, Python, Gitlab, etc.). Nous recherchons un(e) stagiaire pour contribuer aux travaux réalisés dans le DIL R&D.

Objectifs

Dans le cadre du suivi des évolutions des technologies du Big Data et de l'Open Data, nous souhaitons élargir des campagnes ponctuelles à une évaluation active et continue des outils consolidés mais aussi en émergence. L'objectif est d'évaluer un ou plusieurs outils/solutions de Data Science afin d'identifier le potentiel existant en fonction des besoins métiers actuels d'EDF. De plus, il s'agira de mener des tests permettant d'estimer la performance de l'outil, sa prise en main, son intégration dans le système d'information de l'entreprise, etc. Des briques plus spécifiques pourront également être testées (i.e. connecteurs spécifiques, modules et méthodes avancées, etc.). Il s'agira enfin de mener une veille active et régulière sur ces sujets. Nous souhaitons également être accompagnés sur les problématiques liées à l'infogérance des ressources internes et externes de la R&D et à leurs bonnes utilisations.

Profil recherché :

- Programmation (python, R, java, javascript, php, etc.)
- Administration systèmes et réseau
- Bases de données et langage SQL
- La connaissance de l'Open Data, de l'écosystème Hadoop et des solutions compatibles (bases de données NoSQL, Spark, etc.) serait un plus
- Curieux(/se), ingénieux(/se) et motivé(e) pour le domaine de la recherche appliquée

Informations pratiques

Unité d'accueil : Groupe SOAD (Statistique et Outils d'Aide à la Décision), département ICAME d'EDF Lab Paris-Saclay, 7 boulevard Gaspard Monge, 91120 Palaiseau.

Transmettre par mail un CV, une lettre de motivation et les bulletins de notes à :
Geoffrey Aldebert – e-mail : geoffrey.aldebert@edf.fr

Ci-dessous des exemples de travaux en Data Science publiés par notre équipe :

- **A Data Lake and a Data Lab to Optimize Operations and Safety Within a Nuclear Fleet.** Marie-Luce Picard, Jean-Marc Rangod, Christophe Salperwyck. *Hadoop Summit 2016*, California, USA, June 2016: http://fr.slideshare.net/HadoopSummit/a-data-lake-and-a-data-lab-to-optimize-operations-and-safety-within-a-nuclear-fleet
- **Exploring Titan and Spark GraphX for Analyzing Time-Varying Electrical Networks.** Guillaume GERMAINE, Thomas Vial, *Hadoop Summit 2016*, Dublin. http://fr.slideshare.net/HadoopSummit/exploring-titan-and-spark-graphx-for-analyzing-timevarying-electrical-networks
- Vidéo: https://www.youtube.com/watch?v=Xk8UPECiMSw
- **CourboSpark: Decision Tree for Time-series on Spark.** Christophe Salperwyck, Simon Maby, Jérôme Cubillé, Matthieu Lagacherie, *Hadoop Summit 2015*, Dublin, https://speakerdeck.com/simonmaby/courbospark-decision-tree-for-time-series-on-spark
- Vidéo: https://www.youtube.com/watch?v=GNtU-kVL5xI
- **Computing Data Quality Indicators on Big Data Stream Using a CEP.** Wenlu Yang, Alzennyr Gomes Da Silva, Marie-Luce Picard, *IEEE Xplore - IWCIM 2015,* Prague, Novembre 2015. https://tel.archives-ouvertes.fr/LIP6/hal-01367862v1
- **Real-time energy data-analytics with Storm.** Rémy Saissy, Marie-Luce Picard, Charles Bernard, Bruno Jacquin, Simon Maby, Benoît Grossin, *Hadoop Summit 2014*, Californie, USA, 2014. http://fr.slideshare.net/Hadoop_Summit/t-525p212picard
- **HETA: Hadoop environment for text analysis.** Vincent Nicolas, Alzennyr Gomes da Silva, Marie-Luce Picard*, IWCIM (International Workshop on Computational Intelligence for Multimedia Understanding*), IEEExplorer, 2014, 10.1109/IWCIM.2014.7008803
- **Smart Metering x Hadoop x Frost: A Smart Elephant Enabling Massive Time Series Analysis.** Benoît Grossin, Marie-Luce Picard, *Hadoop Summit Europe 2013*, Amsterdam, Mars 2013. http://hadoopsummit.org/amsterdam/
- **Searching time-series with Hadoop in an electric power company.** Alice Bérard, Georges Hébrail, *BigMine Workshop*, KDD2013, Chicago, August 2013. http://bigdata-mining.org/
- **Simulation and forecasting electricity demand at scale.** Alexis Bondu, Yannig Goude, Marie-Luce Picard, Pascal Pompey, Mathieu Sinn, *European Utility Week*, Amsterdam, October 2013.http://www.european-utility-week.com/
- **Empower agile BI & analytics for utilities with a total data approach.** Marie-Luce Picard, Bruno Jacquin, *Teradata Partners Conference*, Dallas, October 2013. http://www.teradata-partners.com
- **A proof of concept with Hadoop: storage and analytics of electrical time-series.** Marie-Luce Picard, Bruno Jacquin, *Hadoop Summit 2012*, Californie, USA, 2012. http://www.slideshare.net/Hadoop_Summit/proof-of-concent-with-hadoop
- **Massive Smart Meter Data Storage and Processing on top of Hadoop**. Leeley D. P. dos Santos, Alzennyr G. da Silva, Bruno Jacquin, Marie-Luce Picard, David Worms,Charles Bernard. *Workshop Big Data 2012*, Conférence VLDB (Very Large Data Bases), Istanbul, Turquie, 2012. http://www.cse.buffalo.edu/faculty/tkosar/bigdata2012/program.php