Détection de Singularités en temps-réel par combinaison d'apprentissage automatique et web sémantique basés sur Spark

Badre Belabbess[*,**] Musab Bairat [**] Jeremy Lhez [*]
Olivier Curé [**]

[*]Innovation Lab, ATOS, F-95870, Bezons, France
prénom.nom@atos.net,
[**]LIGM (UMR 8049), CNRS, F-77454, MLV, France.
prénom.nom@univ-paris-est.fr

1 Introduction

L'apprentissage automatique contient un ensemble puissant d'approches qui peut aider à détecter des anomalies de manière efficace. Cependant, il représente un processus lourd avec des règles strictes et une multitude de tâches telles que l'analyse et le nettoyage des données, la réduction de dimension, l'échantillonnage, la sélection d'algorithmes appropriés, le réglage précis des hyper-paramètres, etc. Notre système a été spécifiquement conçu pour simplifier ce processus lourd et accélérer le déploiement d'une solution en peu de temps. Notre système vise à identifier les anomalies dans un grand réseau d'eau potable géré par un leader national expert dans le domaine de l'eau. En fait, la découverte de telles irrégularités dans le réseau d'eau est une préoccupation critique tant sur le plan écologique que financier. Le volume réel d'eau perdue dans le monde a généré une perte de 32 milliards de m3 / an (soit 14 milliards d'euros par an) dont 90 % reste difficilement identifiable en raison de la nature souterraine du réseau. Sur la base de recherches approfondies menées par les experts, les anomalies peuvent être identifiées en utilisant des mesures de pression et de débit envoyées par des capteurs spécifiques dispersés sur tout le réseau de canalisations.

2 Architecture

Le système a été conçu pour traiter à la fois des données massives dynamiques et statiques à l'aide d'une architecture distribuée tolérante aux pannes. L'objectif principal est de pouvoir traiter des flux massifs de données en temps réel et de lancer des modèles intensifs d'apprentissage automatique. Pour répondre aux besoins d'un système distribué robuste, scalable et à faible latence, nous avons basé notre conception sur une architecture Lambda. Ce type d'architecture Big Data résout le problème des fonctions de calcul lourdes sur des données en temps réel en décomposant le problème en trois couches : une couche batch, une couche vitesse et couche service. Un scénario général de bout en bout commence par le stockage des données