Gestion de Connaissances en Temps Réel depuis des Flux Massifs de Données et Apprentissage Automatique

Badre Belabbess*, Jérémy Lhez**
Olivier Curé***

*badre.belabbess@atos.net
**jeremy.lhez@u-pem.fr
***olivier.cure@u-pem.fr

Résumé. L'analyse en temps-réel de données massives envoyées par des capteurs a connu ces dernières années un essor important. Du fait de l'hétérogénéité de ces données, l'application de modèles de machine learning spécialement calibrés pour des cas d'usages précis a permis d'extraire et d'inférer des informations de très grandes valeurs. Néanmoins, peu de systèmes proposent une implémentation distribuée sur un vrai cluster industriel permettant de tirer profit de capacités de calcul décuplées. Nous présentons ici une démonstration de détection d'anomalie sur réseau souterrain d'eau potable en île-de-France réalisé avec notre plateforme, dénotée WAVES.

1 Introduction

Les avancées technologiques en termes de communication sans fil et de microélectronique ont mené au développement de capteurs intelligents toujours plus efficients et déployables à large échelle .Les domaines d'application se sont alors rapidement diversifiés avec, entre autres, la surveillance d'habitat (A. Rozyyev et F.Subhan (2011)), la géolocalisation d'objets communicants (S.Chauhdary (2009)) et la gestion d'environnement (L.Lee et C.Chen (2008)). Le recours intensif à ces capteurs a conduit à la génération d'un large volume de mesures dynamiques, hétérogènes et géographiquement distribuées. Si ces informations sont analysées efficacement, cela pourrait aider à inférer automatiquement de nouvelles connaissances à haute valeur ajoutée.

Le système décrit ici s'inscrit dans un projet de recherche pour le déploiement d'une solution industrielle nommée WAVES [1]. Il s'agit d'une plateforme de traitement en temps-réel de flux massifs provenant de capteurs installés sur un large réseau souterrain d'eau potable. Un de ces traitements correspond à la détection d'anomalies dans la consommation d'eau correspondant à une fuite sur le réseau. Ce projet est né de la nécessité de trouver des solutions innovantes pour réduire les déperditions d'eau qui sont estimées en moyenne à 20% du volume d'eau introduit dans le réseau [2]. Dans cette démonstration, nous présenterons l'architecture globale du

1. Détails sur le projet WAVES disponible à l'adresse : `http://waves-rsp.org/`
2. Rapport de l'Observatoire des services publics d'eau et d'assainissement : `www.services.eaufrance.fr/docs/synthese/rapport/Rapport_SISPEA\%202011_resume_DEF.pdf/`

module de raisonnement, puis nous détaillerons le scénario mis en jeu ainsi que les résultats obtenus.

2 Architecture

Le module de raisonnement est basé sur une architecture modulaire, scalable, distribuée sur un large cluster de machines et open-source. Le but principal du module est de créer un outil générique permettant d'extraire et d'inférer des connaissances à partir de flux de données hétérogènes dans un environnement temp-réel. Cette solution est assez flexible pour s'adapter à un large panel de cas d'usages permettant de résoudre des problèmes concrets à fort impact social, économique et environnemental.

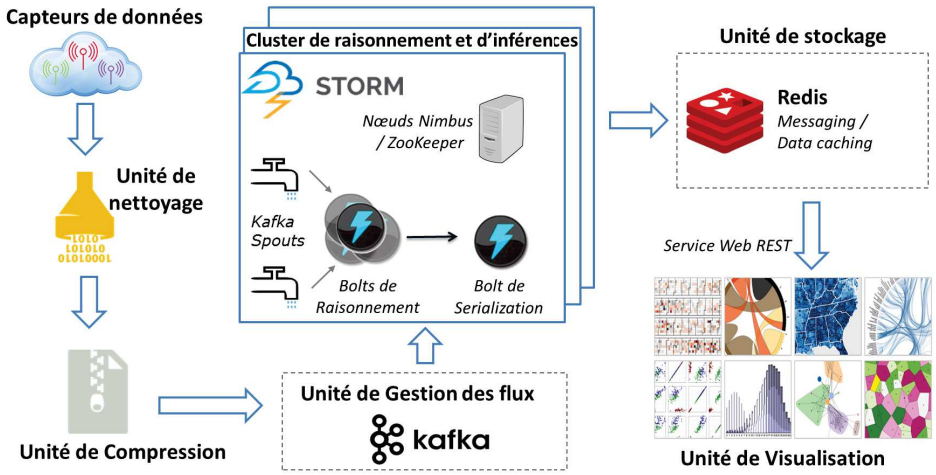

FIG. 1 – *Architecture du module de raisonnement WAVES.*

La figure 1 donne une vue d'ensemble de l'architecture sur laquelle est basé le module de raisonnement et d'inférence. Ce module traite des flux dynamiques organisés sous forme d'évènements et pris en charge par des composants distribués qui ont fait leur preuves dans les déploiements industriels d'envergure : Redis (Red) comme base de stockage en mémoire à haute vélocité, Kafka (Wang et al. (2015)) en tant que système de messagerie distribuée à haute performance, et Storm (Toshniwal et al. (2014)) comme moteur distribué de traitement de données massives à forte tolérance aux pannes.

Le workflow classique du système est le suivant :

— Les différentes mesures générées par les capteurs (e.g pression, débit, chlorine, etc) sont d'abord nettoyées par un algorithme qui élimine les valeurs aberrantes, remplace certaines données manquantes et réalise au besoin un échantillonage.

— Ces données sont ensuite compressées via un algorithme spécifique (García et al. 2014) pour réduire au maximum la taille des flux avant d'être soumises au broker Kafka pour être distribuées au composant principal de raisonnement et d'inférence.

— L'unité de transmission de données sur Storm, appelée Spout, lit ensuite les données compressées sous forme d'évènements atomiques qu'elle redistribue aux différentes unités de raisonnement, appelées Bolts, qui sont réparties sur un cluster de machines virtuelles. Les bolts appliquent durant cette étape des algorithmes d'apprentissage non supervisé dont les résultats seront stockés sur la base clé-valeur Redis puis visualisés ensuite par l'utilisateur final.

3 Scénario et Algorithme de détection d'anomalie

Le but de la démonstration est la détection d'anomalies dans le réseau d'eau potable grâce à l'analyse des données de consommation. Chaque jeu de données provient d'une zone géographique spécifique et de mesures horodatées. Ils sont divisées en secteurs de consommation constitués de quartiers ou de petites villes.La démonstration se concentrera sur la région de Versailles en île-de-France contenant 11 secteurs (environ 350.000 habitants) et comprenant 67 capteurs de débit étalés sur 900km de tuyauterie. Ces informations géographiques sont regroupées dans des fichiers de configuration, qui décrivent le réseau, les capteurs déployés (latitude/longitude), le type de mesure (pression, débit, température, etc.), et l'unité des mesures (m^3/h, $^\circ\mathrm{C}$, etc.).

L'apprentissage des modèles se fait sur des archives de données historiques sur près de deux ans. Les relèves sont rassemblées à raison d'un fichier par capteur, chaque fichier étant situé dans un répertoire en fonction du type de mesure effectuée. Les relevés se présentent sous la forme d'un couple date-mesure ; les dates seules figurent dans le cas où la mesure n'a pu être relevée.

On part du principe que les zones de consommations ou secteurs ont des périodicités dans leur consommation et que tout évènement qui sort de ces régularités est considéré comme anormal. Dans un premier temps, il s'agit de voir quels secteurs sont similaires, et sur quels jours. Ensuite, on fait une confrontation de la consommation des secteurs similaires sur 2 jours similaires (jour actuel et jour de référence). Théoriquement, on doit avoir un mélange homogène de nuages de points sur ces deux jours. Tout nuage de point ayant des évènements non balancés représente une anomalie.

Dans un premier temps, on extrait les différents profils en se basant sur des données historiques. Les méthodes d'extraction de profil sont assez simples. Pour les secteurs, il s'agit de moyenner leurs courbes de consommation journalières sur une période bien définie (6 à 8 semaines dans notre cas). On a pris soin de prendre une période pas très grande pour éviter l'impact de la variation en fonction des saisons. Ensuite, en utilise la corrélation de Pearson pour définir une mesure de similarité entre les différents profils. En ce qui concerne les jours, la courbe représentative du profil est extraite en faisant la moyenne des courbes de consommations des secteurs sur chaque jour.

Après avoir effectué les différentes étapes préliminaires, on a pu procéder à la détection d'anomalies par clustering. Pour ce faire, on a utilisé plusieurs modèles à savoir : KMeans, DBScan, Agglomerative Clustering, OPTICS et ROCK Clustering (Hari Krishna Kanagala (2016)). On a défini les paramètres de chaque modèle en s'ajustant préalablement sur des données test où on injecte nous-même des anomalies artificielles et en optimisant le taux de reconnaissance de ces anomalies. Il s'agit de trouver le modèle le plus correct possible ; leur évaluation est présentée dans le tableau 1.

Critère/Modèles	K-moyennes	Clustering Agglomératif	OPTICS	DBSCAN	ROCK
Précision	0.90	0.87	0.97	0.84	0.78
Recall	0.65	0.70	0.85	0.69	0.57
Seuil Alpha	0.60	0.80	0.75	0.80	0.78
Seuil Beta	0.45	0.54	0.32	0.44	0.78
Distance	Euclidienne	City-Block	Euclidienne	Manhattan	-

TAB. 1 – *Récapitulatif de l'évaluation des modèles*

Références

Redis, Dec 2015. `http://http://redis.io/`.

A. Rozyyev, H. et F.Subhan (2011). Indoor child tracking in wireless sensor network using fuzzy logic. *Research Journal of Information Technology. 3*, 81–92.

García, N. F., J. Arias-Fisteus, L. Sánchez, D. Fuentes-Lorenzo, et Ó. Corcho. RDSZ : an approach for lossless RDF stream compression. In *The Semantic Web : Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014*, pp. 52–67.

Hari Krishna Kanagala, J. R. K. (2016). A comparative study of k-means, dbscan and optics. In *2016 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6. IEEE Computer Society.

L.Lee et C.Chen (2008). Synchronizing sensor networks with pulse coupled and cluster based approach. *Information Technology Journal. 7*, 737–745.

S.Chauhdary, A.Bashir, S. (2009). Eoatr : Energy efficient object tracking by auto adjusting transmission range in wireless sensor network. *Journal of Applied Sciences. 9*, 4247–4252.

Toshniwal, A., S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel, S. Kulkarni, J. Jackson, K. Gade, M. Fu, J. Donham, N. Bhagat, S. Mittal, et D. Ryaboy (2014). Storm@Twitter. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, New York, NY, USA, pp. 147–156. ACM.

Wang, G., J. Koshy, S. Subramanian, K. Paramasivam, M. Zadeh, N. Narkhede, J. Rao, J. Kreps, et J. Stein (2015). Building a replicated logging system with apache kafka. *Proc. VLDB Endow. 8*, 1654–1665.

Summary

The analysis of massive amounts of data sent in real-time by sensors has experienced a major development in the last few years. Due to data heterogeneity, the application of machine learning models specifically calibrated for accurate use cases allowed to extract and infer valuable information. However, few systems propose a distributed implementation on a true industrial cluster permitting of taking advantage of increased computing capabilities. Here we present a demonstration of anomaly detection on an underground drinkable water network located in île-de-France, realized with an innovative platform: WAVES.