Les bénéfices du streaming de données :

du stockage à l'analyse

Joulain, Cédric, cedric.joulain@kereon-intelligence.com, Kereon Intelligence, Poitiers

Pasquet, Jonas, jonas.pasquet@kereon-intelligence.com, Kereon Intelligence, Niort

Guillaume, Olivier, olivier.guillaume@kereon-intelligence.com, Kereon Intelligence, Niort

Thématique : Analyse des données (+ Structuration et Mise à disposition)

Résumé : *Grâce à une approche streaming pour l'ensemble de la chaîne de traitement des time series (séries temporelles), Kereon Intelligence est en mesure de stocker et d'analyser efficacement tout type de flux de valeurs. Notre démarche repose sur deux outils : un format d'archivage très compact et rapide et le moteur de calcul LispTick frugale et versatile. Cette approche permet de traiter indifféremment l'historique et le temps réel sur tout type d'architecture, de l'embarqué au centre de calcul.*

Mots clés : *timeseries, streaming, temps réel, edge, IoT, compression*

1. Introduction

Autrefois les time series, ou séries temporelles, étaient majoritairement cantonnées au monde de la finance. Aujourd'hui, avec l'explosion des objets connectés, les time series sont partout : dans l'IoT, la Météo, la santé… La majorité des données dites « Big Data » ont une notion de temporalité, ce sont donc des time series.

Analyser des évènements dont les temporalités sont différentes et avec de grandes volumétries constitue une vraie difficulté. Difficulté d'autant plus importante si nous voulons à la fois pouvoir travailler sur des données historiques et temps réel.

2. Méthodologie

Pour répondre à la problématique nous avons donc créé le moteur d'analyse LispTick qui permet de coder et d'appliquer des algorithmes de traitement en pur streaming. Tous les calculs se font à la volée sur un flux de données ou même sur plusieurs flux de données qui seront synchronisés de façon transparente et sans problématique de contention.

Notre moteur d'analyse LispTick reçoit une requête décrite dans un langage puissant et concis. Une fois la requête interprétée, il renvoie le résultat de manière continue jusqu'à épuisement des flux de données en entrée.

Les sources de données peuvent être multiples : fichiers plats, base SQL, Influx DB, MQTT, web socket, port série, I2C… Mais notre format de stockage propriétaire a l'avantage de mettre en exergue tout le potentiel du moteur d'analyse. Par sa compacité et son accès direct binaire, il permet de libérer des ressources mémoire qui vont être utilisées par l'OS comme cache fichier.

Nous avons donc ainsi gratuitement une « base in memory » avec une vitesse de lecture décuplée.

Pour arriver à un tel taux de compression sans perte, nous nous sommes basés sur notre expertise des flux financiers, les flux les plus complexes avec le plus de volumétrie. Ainsi nous arrivons à des taux de compression 2 à 3 fois supérieurs à la référence, le format Apache Parquet avec son meilleur niveau de compression.

Quelques chiffres : nous arrivons à stockons les données open source MétéoNet dans seulement 298MB. MétéoNet représente 2.6GB de fichiers tar gz. Ce sont 3 ans d'historique toutes les 6 minutes pour 862 stations météo soit 1,6 milliard de valeurs.

3. Originalité / perspective

L'originalité de l'approche réside dans sa mise en œuvre « pur streaming ». Tous les flux, historiques ou temps réels, sont lus en continu et analysés à la volée.

Le mécanisme de synchronisation interne permet de combiner tout type de flux, y compris des flux asynchrones avec des disparités de fréquences de mise à jour. Il permet aussi de s'affranchir des contentions classiques, « bouchon de données », en ne nécessitant qu'un point par série en mémoire.

Cette frugalité nous permet de faire tourner le moteur d'analyse sur des très petites configurations (Raspberry Pi, Onion Omega2) mais aussi d'utiliser toute la puissance des gros serveurs.

Nous avons ainsi pu appliquer des algorithmes de finances au suivi de la santé des abeilles dans des ruches connectées.

Nous travaillons activement sur la prochaine étape : l'IA. Nous disposons déjà de l'inférence pour n'importe quelle architecture de réseaux de neurones via ONNX et cherchons à intégrer l'apprentissage directement dans la chaîne streaming.

Références

https://meteonet.umr-cnrm.fr

https://kereon.lisptick.org