IBM Bluemix Reference Solution for IoT Data and Analytics

Travel & Transportation Industry - The Madrid Traffic Use-Case

General Purpose

This cloud-based reference architecture provides an IBM reference solution for handling the data component in an IoT application. It needs to address data collection and aggregation, online data analytics as well as analytics on historical data.

- **Data Collection and Aggregation:** collecting online data from end devices, sensors, etc. The data item from each end point is typically small in size and arrives at a relatively constant rate, but may aggregate to a large volume of data if accumulated from a very large number of devices (e.g. ten of thousands of devices).
- **Data Analytics:** executing on-line analytics on the data from a given time window as it arrives. Furthermore, execute analytics on the historical data, and produce output which may further enhance the on-line analytics.

The online data needs to be streamed across all components as it arrives from the data sources. Due to its large volume, historical data must be stored in a cost-effective and scalable storage, but at the same time should be easily and efficiently accessed for analytics purposes.

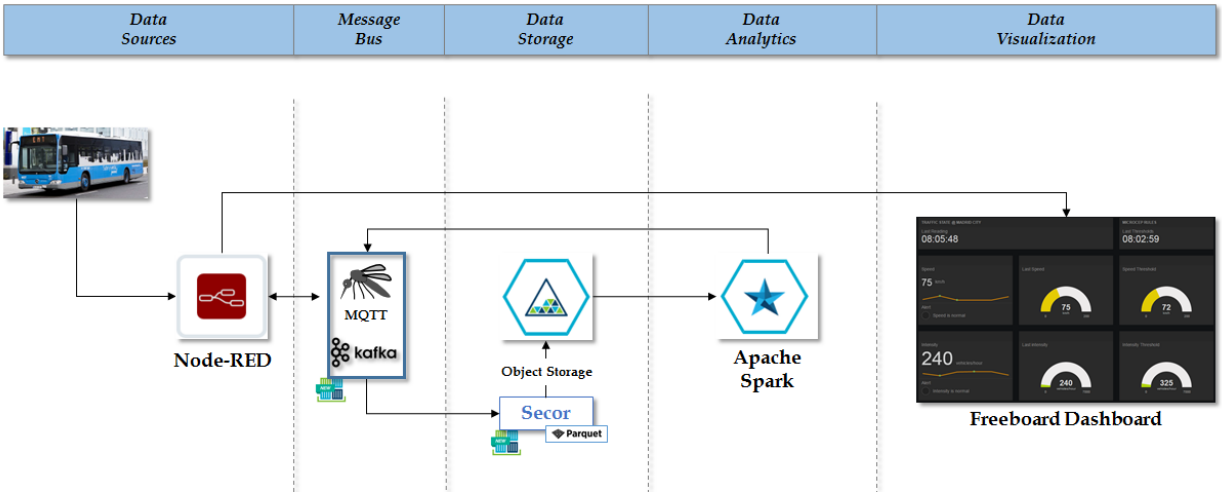
The Scenario

This scenario monitors and processes data from an online data source (http://informo.munimadrid.es/informo/tmadrid/pm.xml). The data corresponds to traffic data collected from 300 sensors spread across the City of Madrid, such as traffic speeds and intensities, and is updated every 5 minutes. Data is aggregated into objects and stored in a persistent data store for future use. The real-time data stream is presented on a dashboard, while indicating anomaly/exceptions in the traffic behavior. The anomaly is derived dynamically, based on thresholds that are computed using real-time analytics on the historical data. Thresholds and classifiers are also used to make the real-time decision over the traffic data, such as re-routing buses, modifying traffic lights, etc., based upon knowledge derived from historical data.

The Flow

- Node-RED retrieves the sensors' data from http://informo.munimadrid.es/informo/tmadrid/pm.xml via a web service in an XML format, converts the data into JSON format, and pushes and publishes the data on the Message Hub.
- In addition, Node-RED classifies the data using pre-computed thresholds and presents it visually on the Freeboard dashboard, highlighting "normal" or "exceptional" behavior.
- The Secor service, which is deployed in a container, is used to aggregate multiple messages into a single Swift object according to some policy (e.g. every 60 mins or according to dates), converts the data into Parquet format, and uploads it to the Swift Object Store.
- Data is streamed across all components through the Message Hub service.

- In parallel, a machine learning algorithm (which uses the MLib spark library) is executed in the Apache Spark service. It dynamically uploads historical data from the Object Storage into Spark and computes thresholds. Thresholds are published on the Message Hub through separate topic.



Glossary: Bluemix Components

- Node-RED – **Bluemix service**
  - Node-RED is a tool for wiring together hardware devices, APIs, and online services in new and interesting ways.
  - Bluemix service - using Bluemix SDK for Node.js - https://console.ng.bluemix.net/catalog/node-red-starter/
- Message Hub (Kafka) – **Bluemix service** (?)
  - Apache Kafka is publish-subscribe messaging rethought as a distributed commit log.
  - Right now waiting on a fix, if not there, will use container.
- Secor – **Running in Container**
  - Secor is a service persisting Kafka logs to the cloud.
  - https://github.com/pinterest/secor
- Object Storage - **Bluemix service**
  - IBM® Object Storage for Bluemix™ uses SoftLayer Object Storage, which is based on the OpenStack Swift project, to store data objects.
- Apache Spark - **Bluemix service**
  - Apache Spark is an open source cluster computing framework optimized for extremely fast and large scale data processing and analytics.
  - Bluemix service - https://console.ng.bluemix.net/catalog/apache-spark/
- Freeboard – **Bluemix service** - module in NodeRed
  - Open source tool for creating simple dashboards for IOT data visualization.
  - https://freeboard.io/
- CEP - Complex event processing developed by ATOS
  - Service that get streaming data, and take action under specific rules.
  - Right now not part of the solution.