BIG DATA INGESTION AND ANALYTICS FOR PHYSICAL THREAT INTELLIGENCE

Paolo Mignone, Costantino Mele, Donato Malerba and Michelangelo Ceci
Università degli studi di Bari[1]
paolo.mignone@uniba.it; c.mele22@studenti.uniba.it; donato.malerba@uniba.it; michelangelo.ceci@uniba.it

Marco Anisetti, Claudio Ardagna, Chiara Braghin, Ernesto Damiani and Antongiacomo Polimeno
Università degli studi di Milano[2]
marco.anisetti@unimi.it; claudio.ardagna@unimi.it; chiara.braghin@unimi.it; ernesto.damiani@unimi.it; antongiacomo.polimeno@unimi.it

Alessandro Balestrucci
Consorzio Interuniversitario Nazionale per l'Informatica – CINI [3]
alessandro.balestrucci@consorzio-cini.it

Abstract – Academic paper.

In this paper, we propose a big data engine supporting two main procedures to be performed on data coming from smart cities' sensor networks: data i) ingestion and ii) analytics. Data and resource management are fundamental functionalities for any kind of distributed system including a big data engine. We considered Hadoop HDFS and Hive to handle data storage, ingestion, and buffer storage, while Kafka is dedicated to messaging and Hadoop YARN to manage computational resources. We augment ingestion pipeline with access control (AC) capabilities through Atlas for data annotation, data governance and audit and Ranger to handle authorization/access policies to data and access audit. Our AC-enhanced ingestion pipeline allows us to enforce compliance to regulations (e.g., GDPR) at ingestion time prior to executing analytics.

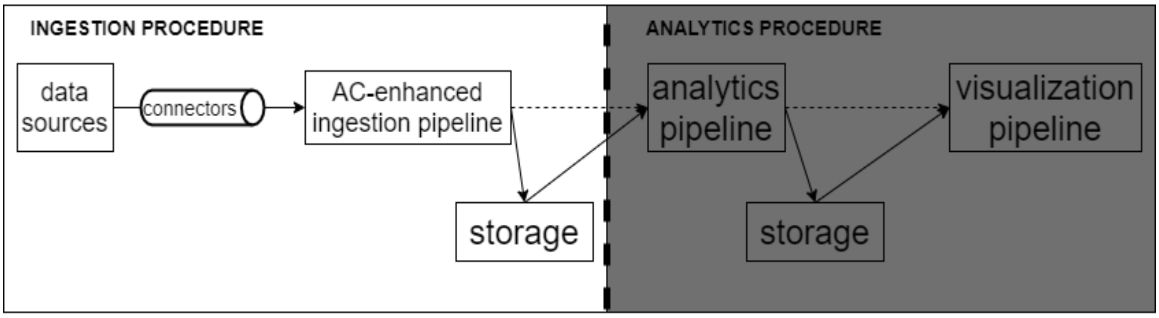

Figure 1. Ingestion procedure

[1] Piazza Umberto I, 1, 70121 Bari BA, Italy
[2] Via Festa del Perdono, 7, 20122 Milano MI, Italy
[3] Via Ariosto, 25, 00185 – Rome/ DIAG Sapienza, University of Rome - Via Ariosto, 25, 00185 Rome, Italy

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No. 883286, project IMPETUS (Intelligent Management of Processes, Ethics and Technology for Urban Safety).

As concerns the analytics procedure, after the data ingestion pipeline, we exploited Apache Spark framework in order to run machine learning algorithms in a distributed fashion. In particular, we aim to catch the spatio-temporal autocorrelations from geo-distributed sensor data. As a result, the proposed architecture is capable of handling batch, micro-batch, and streaming data to construct predictive models that effectively perform anomaly detection. Anomaly detection is a machine learning task that aims to identify rare items, events, or observations. Such observations are measured through the geo-localized sensors and when they are suspected by differing significantly from the majority of the previous data, they are considered anomalies.

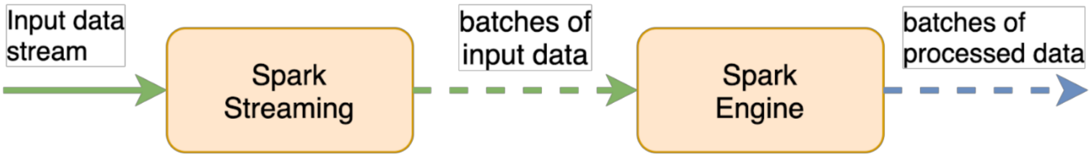

Figure 2. Data stream processing

Typical applications: bank fraud, medical problems, structural defects, malfunctioning equipment, etc. Our aim is to exploit machine learning algorithms for anomaly detection from data generated by multiple sensors for physical threat intelligence (PTI). PTI represents the analysis process that uses data-driven approaches to identify threats to physical systems. Therefore, we designed an anomaly detector that can promptly identify anomalies in data from geo-localized sensors to prevent any physical threats such as road accidents, attacks, and popular uprisings.

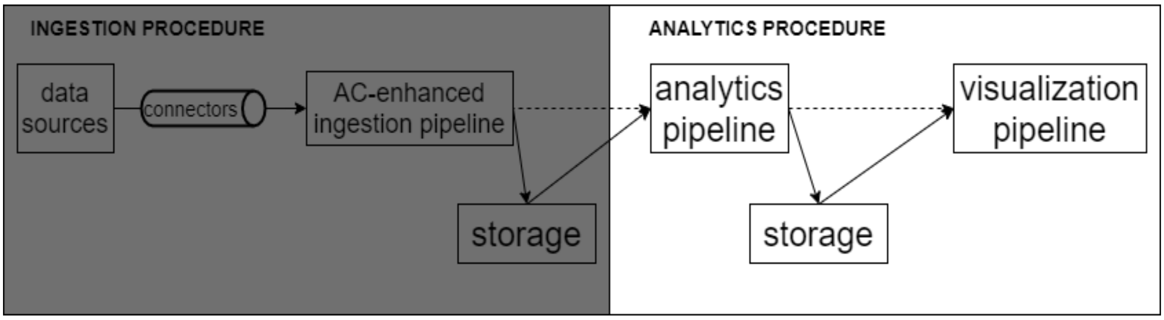

Figure 3. Analytics procedure

The proposed architecture helps us to ingest and analyze real-time data to support the local authorities for timely operations. The anomaly detectors in this context represent innovative models to deal with unexpected situations. Moreover, the proposed anomaly detector can provide an easy-to-explain output that could be interpretable also for non-confident operators, i.e., the end-users. Such users are typically front line operators that will inform the local authorities. An interactive user interface should support the operators to take actions promptly with respect to the anomalies automatically detected through our proposed architecture.

Keywords: big data analytics, data ingestion, anomaly detection, geo-distributed sensors, smart city, Apache Spark