

## BIG DATA INGESTION AND ANALYTICS FOR PHYSICAL THREAT INTELLIGENCE

Paolo Mignone<sup>1</sup>, Marco Anisetti<sup>2</sup>, Claudio Ardagna<sup>2</sup>, Costantino Mele<sup>1</sup>, Alessandro Balestrucci<sup>3</sup>, Chiara Braghin<sup>2</sup>, Ernesto Damiani<sup>2</sup>, Antongiaco Polimeno<sup>2</sup>, Donato Malerba<sup>1</sup>, Michelangelo Ceci<sup>1</sup>

### Abstract

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.

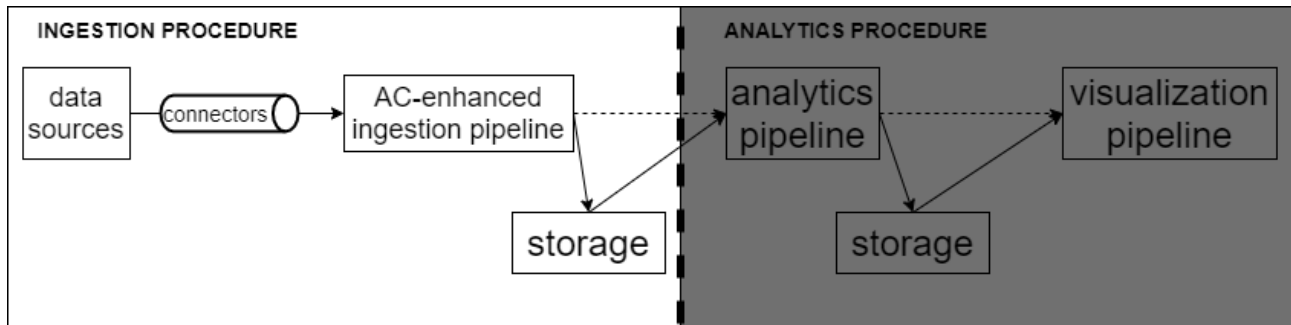


Figure 1. Ingestion procedure

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

<sup>1</sup>Università degli studi di Bari Aldo Moro - Piazza Umberto I, 1, 70121 Bari, Italy

<sup>2</sup>Università degli studi di Milano - Via Festa del Perdono, 7, 20122 Milano, Italy

<sup>3</sup>Consorzio Interuniversitario Nazionale per l'Informatica – CINI - Via Ariosto, 25, 00185 - Roma, Italy

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.

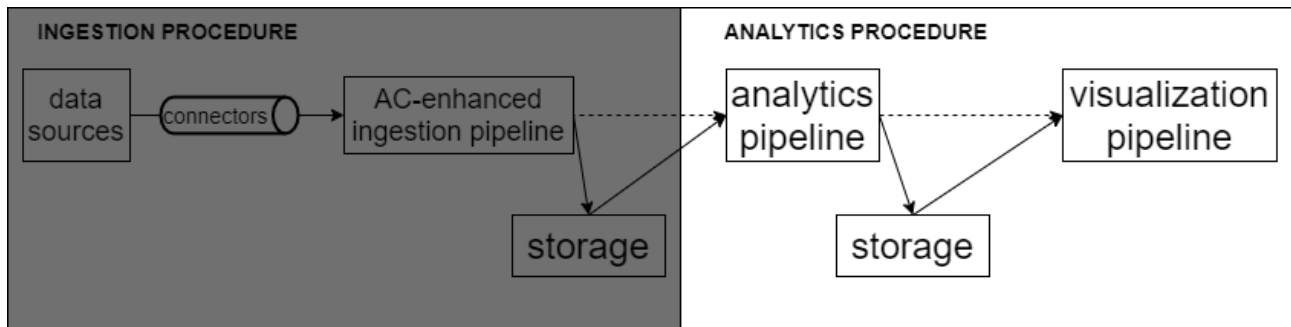


Figure 3. Analytics procedure

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

**Acknowledgments:** we acknowledge the project *IMPETUS (Intelligent Management of Processes, Ethics and Technology for Urban Safety)* that receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 883286.

<https://cordis.europa.eu/project/id/883286>