Multimedia Tools and Applications (2021) 80:23681–23700
https://doi.org/10.1007/s11042-020-10206-y

Multimedia analysis platform for crime prevention and investigation

Results of MAGNETO project

Francisco J. Pérez[1] · Victor J. Garrido[1] · Alberto García[1] · Marcelo Zambrano[2,3] · Rafał Kozik[4,5] · Michał Choraś[4,5] · Dirk Mühlenberg[6] · Dirk Pallmer[6] · Wilmuth Müller[6]

Received: 11 March 2020 / Revised: 29 July 2020 / Accepted: 25 November 2020 /
Published online: 6 February 2021

Abstract
Nowadays,the use of digital technologies is promoting three main characteristics of information, i.e. the volume, the modality and the frequency. Due to the amount of information generated by tools and individuals, it has been identified a critical need for the Law Enforcement Agencies to exploit this information and carry out criminal investigations in an effective way. To respond to the increasing challenges of managing huge amounts of heterogeneous data generated at high frequency, the paper outlines a modular approach adopted for the processing of information gathered from different information sources, and the extraction of knowledge to assist criminal investigation. The proposed platform provides novel technologies and efficient components for processing multimedia information in a scalable and distributed way, allowing Law Enforcement Agencies to make the analysis and a multidimensional visualization of criminal information in a single and secure point.

Keywords Law enforcement · Information extraction · Distributed infrastructure · Situational awareness

1 Introduction

In recent decades, human activities have progressively shifted from person to person to seamless interactions between the physical world and the world of information technology (IT); crime has naturally followed the same path, with imagination being the only limit. As a result, and in addition to the "traditional" criminal activities, new combinations of malicious routines appear every day in an exponential manner, ignoring political boundaries and legal jurisdictions [22]. Based on the possibilities offered by the evolution of technologies, especially Big Data analytics, representational models, semantic reasoning

✉ Francisco J. Pérez
frapecar@upvnet.upv.es

Extended author information available on the last page of the article.

and augmented intelligence, MAGNETO [23] will contribute to counteracting this threat to society [4, 14].

For this reason, researchers across the world are often faced with a variety of complex cases where large amounts of data have to be analysed. The analysis of these volumes of data often exceeds the capacity of police teams [13]. Moreover, occasionally, both the workload and time constraints of LEAs can be extreme, e.g. in the case of a sequence of terrorist attacks. They would involve pre-incident intelligence analysis, real-time incident processing and post-incident prediction potentially linked to other threats or subsequent attacks [39, 41].

Therefore, MAGNETO formalises a structural framework to improve the operational capacity of Law Enforcement Agencies (LEAs) in their fight against organised crime and terrorist organisations. Following recent terrorist attacks across Europe in London, Paris, Berlin and Brussels, the critical challenges the respective investigative bodies are facing include the processing of large volumes, heterogeneity and fragmentation of data that officials must analyse for the prevention, investigation and prosecution of criminal offences [20].

MAGNETO will achieve this goal by unifying different sources of evidence (video, audio, text/documents, social media and Web data, telecommunications data, surveillance system data, police databases, etc.), providing researchers with modular analytical tools that will generate comprehensive responses to their queries. To this end, MAGNETO will develop a new data model that will allow the joint exploitation of multiple and diverse multimedia data sources, together with modular and scalable analysis engines. This will help LEAs correlate data and find hidden relationships, select and classify pieces of evidence, evaluate and classify threats, reason out criminal cases, as well as visualize and understand the course of action, past or in real time (operational and situational knowledge).

To fulfil its purpose, MAGNETO will adopt a multidisciplinary scientific approach, conducting research, adapting and leveraging a number of advanced technologies, and focusing on:

- Representative data models, which allow to anticipate and predict future trends (e.g. threats), and establish the basis for reasoning and cognition (situational awareness). In particular, the value of the representative model of MAGNETO will be demonstrated through its application in decision-making, control and optimisation, and the extension of the usefulness and value of the data available for LEAs.
- Semantic information processing extracts knowledge from heterogeneous sources and transfers it to a knowledge base. For the Common Representational, an ontology has been developed using the OWL standards. In [7], the author describes a core ontology based on NATO standards to improve military intelligence analysis. The main concepts of this core ontology have been selected as a basis of the MAGNETO ontology: Organization, Equipment, Person, Place, and Event as depicted in Fig. 1.
- Semantic reasoning, allowing a computable framework for systems to treat knowledge in a formalized manner, allowing navigation through different pieces of data and the discovery of relationships and correlations between them, thus broadening the spectrum of knowledge capabilities for LEAs.
- Augmented intelligence, improving the thinking processes of human operators rather than replicating or replacing them. Thus, MAGNETO will ultimately function as a personalized partner in the thinking process of LEA officials It will present relevant suggestions guided by access to internal and external information, but more importantly, be user-guided, which will dynamically assess relevance through highly usable Human Machine Interfaces (HMIs).

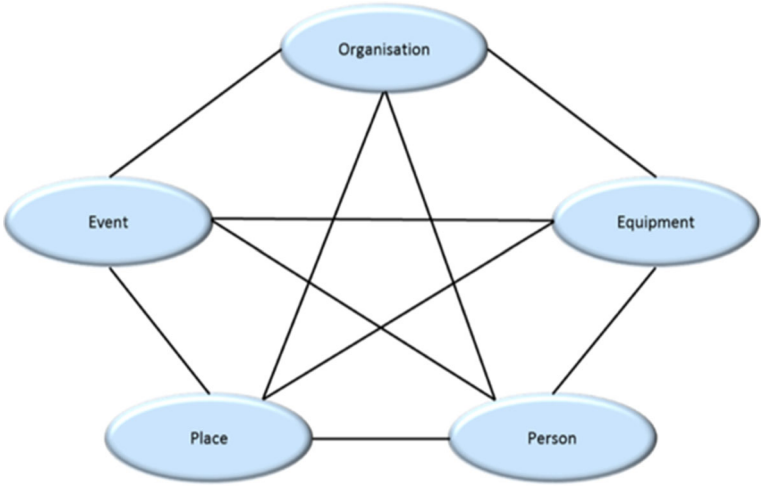


Fig. 1 Main concepts of MAGNETO ontology for criminal investigation

In Section 2, we will analyze the state of the art of the platform by looking at the different tools and analyzing the use case to be implemented; in Section 3 we will compare the platform with other solutions in the market that have a similar approach to MAGNETO; in Section 4 we describe the approach of the proposed architecture to be implemented in MAGNETO, while Section 5 is dedicated to analyzing and studying a specific use case to validate the results and our hypothesis.

2 State of art

The increasing use of digital technologies is directly influencing the generation of information in three different ways [17]; the volume, type and frequency with which information is generated by both individuals and devices that law enforcement agencies need to use and take advantage of all available resources to conduct and effective criminal investigation [11]. This paper proposes a distributed approach that allows the use of different tools for processing and extracting knowledge, thus improving criminal investigation tasks. The novelty that this paper proposes is the use of new technologies based on Big Data to optimize the processing of heterogeneous sources, including images, audio files or text documents in different formats and provide analysis and visualization tools to improve the situational awareness [34]. In addition, the platform allows storing the processed information for its consultation in similar cases in the future, or for the training and optimization.

Within the large Big Data ecosystem, you can find numerous tools that allow the development of an open platform, expandable and upgradeable to future requirements and new generation tools. Currently, there are a large number of tools to encapsulate and automate the deployment of components, for which container-based solutions have been used to encapsulate and isolate components. Previously, this was done with virtualisation [36, 37] combined with solutions such as Puppet [28], Chef [5] or Ansible [1], which allowed us to simplify and automate the deployment of our components in isolated environments for both testing and production.

The ability to develop and perform container-based deployments makes the application modular, scalable and easily expandable. Without a doubt, communication between services and information exchange is fundamental part of any architecture. Taking into account the requirements, one type or another can be used, such as API Gateways (Kong [18], Express [10]), Service Meshes (Kuma [19], Linkerd [21]) or Message Queues (RabbitMQ [30], Kafka [3]).

A fundamental aspect when developing a Big Data platform is the storage of the information to be processed. Therefore, depending on the type of information, the solution that best suits your needs should be used. Solutions such as HADOOP HDFS [12] allow this type of storage to be deployed easily and allow the volume of data to be scaled by simply adding a new machine to the cluster. Due to the heterogeneity of the data currently handled, the frequency of insertion and the concurrence in the access to information, there are different types of databases for different purposes, such as MongoDB [24], Cassandra [2], Elasticsearch [8], among others.

As far as information extraction is concerned, different text mining algorithms must be analysed and applied. First of all the audio analysis is done through a Python library called pyAudioAnalysis [29], therefore the audio is transformed into text, for which the audio has to be segmented into sentences using the SpeechRecognition [32] library and then the transcription is done with the offline pocketsphinx [27] library. Finally, the text will be analyzed by running an unsupervised model and the Stanford CoreNLP [33] library which is capable of analyzing text through natural language processing and extracting patterns of behaviour and sentiment analysis.

In order to validate and verify the correct functioning of the components of the MAGNETO platform, the following use case is proposed:

Use case: prevention and investigation of terrorist attacks As illustrated by the recent terrorist events, anti-terrorist investigations call for preventive, real-time and post-event responses, always considering the possibility of a sequence of unprecedented and complex scenarios.

Problem statement Terrorist activity is a multifaceted type of crime (involving financing, purchase of weapons or prohibited goods, extortion, false documents, bombing or other attacks, etc.) that is strongly linked to organized criminal activity. Information about it is derived from various sources and in various formats – audio interception farms and the Internet, ancillary services (e.g. telephone billing details or bank documents from requisitions to banking organizations), interviews, surveillance, documents from international and European cooperation – but also from telephone interceptions, handwriting, audio, video, etc., as illustrated in the following use case. A terrorist apprentice typically tries repeatedly to join training camps in the Middle East (border crossings, visas, air passenger file, etc.), in contact with trainers, through a third party based in France that organizes his journey (detailed billing, telephone and geolocation, photo of contacts). Each traversed European country writes a warning card with his behaviour and contacts, and transmits this information (within the framework of European cooperation) to the country from which the suspect departed and to his destination country. A cross-check is then carried out with different contacts and a network mapping is made. Upon his return to France, he comes into contact with other disciples and proselytes. A geographical mapping and a relevant network are created. The interception of means of communication broadens the contact's so-called 'Tree', establishing his past, present and future movements. Before launching a suicide bombing attack with his associates, he is arrested by the police. After arrest, investigations continue,

looking for accomplices, and methodical exploitation of the information collected is carried out, with cross-checking via metadata, as well as full search looking for inconsistencies and weak signals. This same use case can also be extended or modified to address home-grown terrorists.

3 Relation to other products

Before the development of Magneto began, other solutions and tools that LEAs were using for gathering information, discovering, analysis and storage digital evidences were studied.

One of these solutions is i2 Analyst's Notebook [15] developed by IBM; i2 Analyst's Notebook is a visual analysis software that allows to convert data into information. Its main features are:

- Optimized analysis-ready data storage
- Information discovery via multiple search techniques
- Visual analysis tools help reduce the time to identify the "who, what, when, why and where"
- Multi-dimensional analysis, e.g. temporal, geospatial, histogram, social network views, to uncover hidden connection, patterns and trends
- Local analysis repository optimized for information management, discovery, and analysis

On the other hand, the EnCase Forensic tool [9] developed by OpenText is a powerful platform for acquisition, analysis and reporting data and evidences from computer and mobile devices. Its features include:

- Searching and analysis of physical and logical storage, both allocated and unallocated partition
- Mobile forensics
- Data recovery
- Analysis of "slack" space (space between physical and logical size of files)
- Analysis of files attributes and permissions
- Searching for alternate data streams
- File and metadata indexing
- Analysis of file signatures, RAM memory, backup files of mobile devices, EXIF files, entropy analysis, system register analysis
- Reporting capabilities

These solutions focus on the development of isolated tools for natural language processing, semantic media analysis, social network analysis, complex event processing and artificial intelligence. However, there is a large need among LEAs for tools that can process and correlate large amounts of data, and enrich the information across a broad range of heterogeneous sources, such as to speed up the investigation process, so that they may focus on proper investigative work, rather having to concentrate on technical aspects and manually correlating data.

The main MAGNETO vision is to augment the capabilities of LEAs in managing, investigating, correlating and reasoning upon huge volumes of heterogeneous and disjoint multimedia data. In short, MAGNETO offers an innovative open framework that:

- Interoperates with and processes massive heterogeneous data sources

- Represents the knowledge and relationships between different pieces of information in criminal cases in a homogeneous way
- Uses semantic fusion and reasoning technologies to discover hidden relations among data pieces
- Assesses the future development of certain incidents
- Increases the thinking process and efficiency of LEA investigators in their daily works and missions

4 System architecture

According to the first platform approach of MAGNETO platform shown on Fig. 2, the implementation should follow a modular architecture allowing the distribution and scalability of the provided services. In order to cover it, the proposed architecture is based on a services-oriented solution with a communication based on Publish/Subscribe paradigm for internal modules communication, and a REST API Gateway for external access. It should be noted that for a proper development process, and taking into account the heterogeneity of the different services developed, each module will be isolated in a Docker container [25]. With this approach, the development process is optimized, and each development group just needs to be focused on their services.

4.1 Architecture specifications

The use of microservices architectural style [31] in the context of MAGNETO project is beneficial due to a numerous capabilities this choice offers, e.g.:

- ease of integration of new services/functionalities into the framework,
- simplification of integration tests,

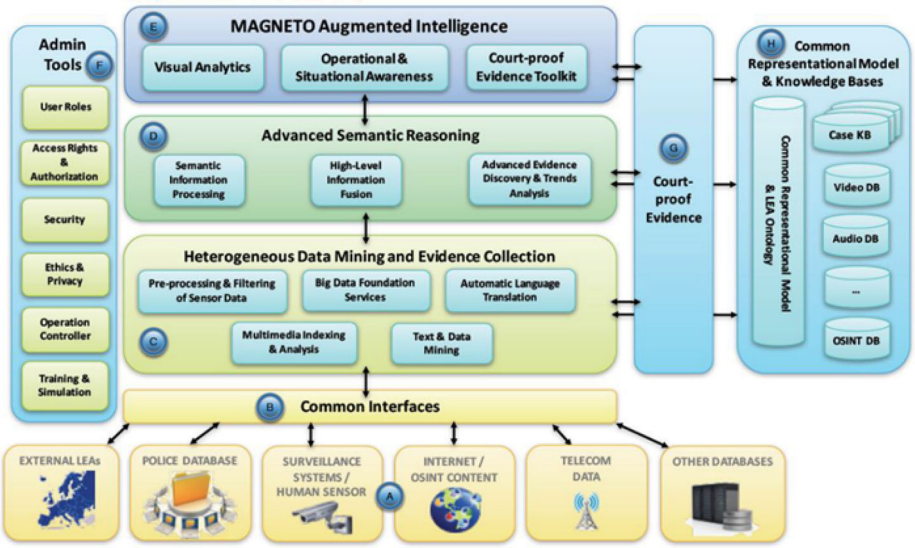


Fig. 2 MAGNETO platform architecture

– simplification of development updates (only the relevant parts of an application),
– better resiliency and stability of the application by eliminating a susceptibility to a single point of failure.

Microservices architecture style promotes the single responsibility paradigm and recommends loosely coupling. There are different approaches for the monolithic systems in order to be divided into several smaller autonomous components. The most obvious strategy is to use decomposition that is based on business capability [35]. In general, the process of decomposition produces smaller entities that can be developed individually by separate teams. This allows the teams to sustain autonomy in terms of architectural patterns and technologies selected to develop a specific service. This approach has been used in the MAGNETO project as shown in Fig. 3.

When the application is broken down into a set of separate services, these need to communicate in order to provide complex business capabilities. That capability usually needs to assemble the results obtained from multiple services. Depending on the specific application usage, this could impose several difficulties. In some cases, orchestration needs to be implemented in order to produce desired results. The orchestration is related with the chain of actions, which needs to happen to generate the final result. Moreover, these actions need to be sequenced in a time-based manner. Therefore, in this document we carefully investigated how various business processes need to be arranged.

Moreover, having a single consistent API for accessing the system is also of high importance. Therefore, in the architecture we have introduced an aggregator pattern called API Gateway. The pattern appears in many microservice frameworks and solutions, and it can be seen as a reverse proxy. Since the first version of the architecture, we have promoted Kong API Gateway due to the following reasons:

– the system is seen as a single monolithic application with a consistent API,
– thanks to plugin-based architecture it is easy to extend and adjust the portfolio of functionalities to user's needs,
– some specific function can be implemented at the gateway (e.g. authentication and authorisation).

On the other hand, the communication between the different services of the platform will be carried out asynchronously through the publish and subscribe bus.

4.2 Components deployed on scenario

In order to address the scenario requirements, a set of components have been deployed providing to the platform the capabilities to fulfil the case needs. The interaction of all of them is shown in Fig. 4.

- Speech to Text

  LEA researchers deal with huge and complex amounts of recorded call data or audio files; it is impossible to listen to them all to discover hidden or unsuspected connections within large datasets. This results in costly investigations with enormous delays in the resolution of crimes and difficulties in preventing them. Conversations about multilingualism are difficult to decipher and data sources are generally not contextualized. Because the MAGNETO platform provides new semantic technologies

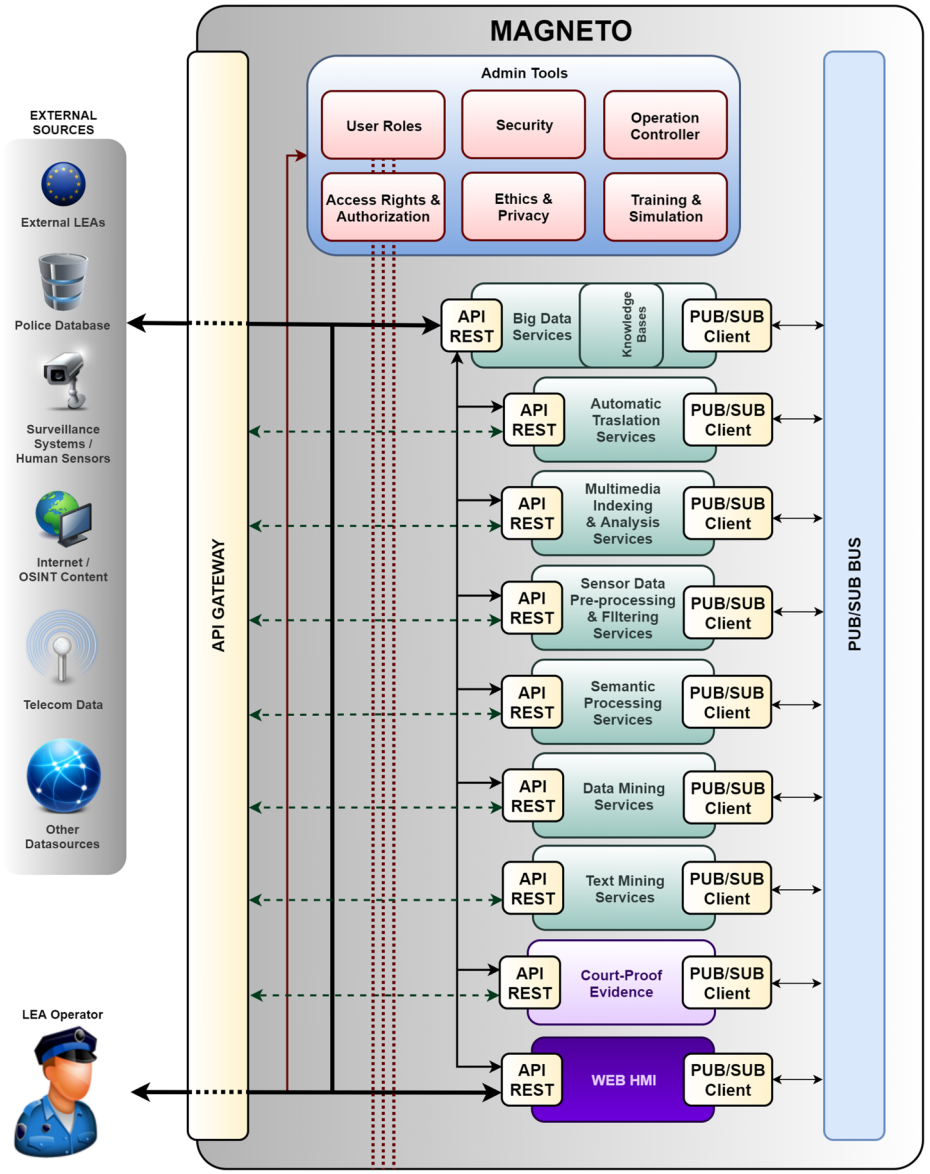


Fig. 3 MAGNETO platform architecture

and augmented intelligence tools for indexed textual content, a speech-to-text module provides MAGNETO with a solution that, combined with other modules, enhances investigative capabilities.

The speech-to-text module provides MAGNETO with a powerful tool capable of combining online conversions through APIs exposed by the main providers (GOOGLE, MICROSOFT, IBM, WIT, ...) in case Internet access is allowed in LEA premises, or

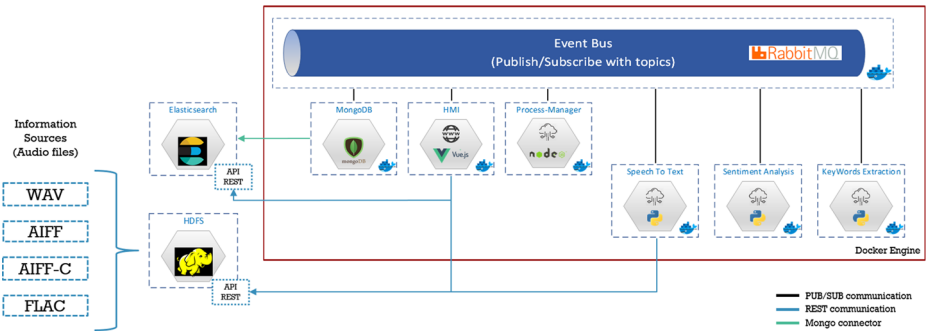


Fig. 4 Scenario deployed. Modules interaction

offline conversions based on local dictionaries in case of security policies or privacy restrictions, as shown in Fig. 5.

The result obtained from MAGNETO's Speech to Text module is the transcription of the audio file. The file can be analyzed by other MAGNETO modules responsible for pattern identification, sentimental analysis or text classification. This automated approach is clearly more feasible than extracting information by listening directly to audio files.

To improve the performance of the speech-to-text module, it has been designed with integration into MAGNETO's Big Data Services in mind. The solution proposed in Fig. 4 uses the Hadoop Distributed File System (HDFS) as the repository for all the audio files to be transcribed. Once a file is uploaded to the repository, a Docker container starts the conversion, and the result of this transformation is associated as

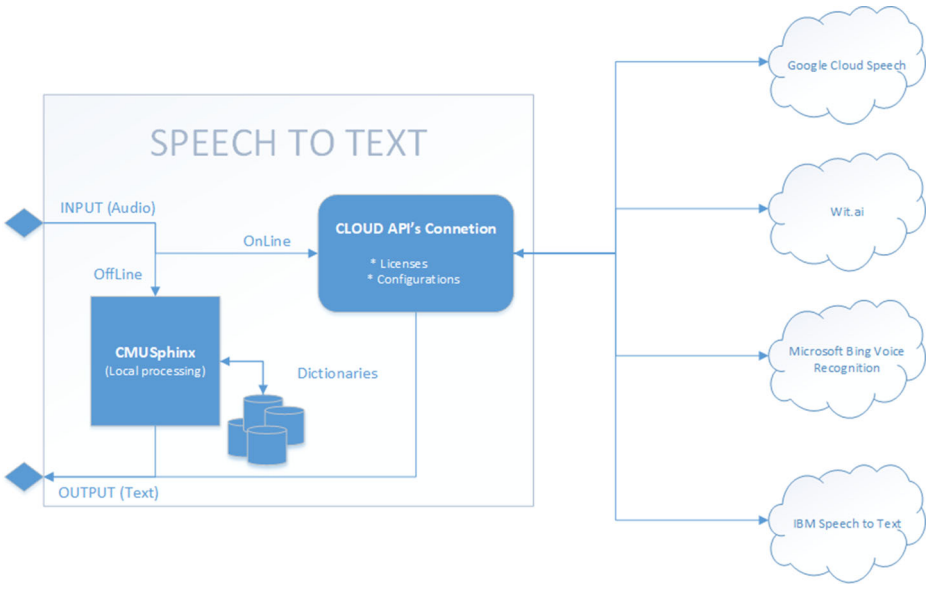


Fig. 5 Speech to text transformation flow diagram using CMUSphinx

metadata of the audio file and stored in a NoSQL database for future analysis by other modules.

- Text Mining

  An ongoing investigation produces vast amounts of textual data, some of which comprises natural language with all its vagueness and ambiguity, such as email and documents confiscated during house-searches, transcribed telecom/audio data, and other documents used in an investigation. The purpose of the text mining module is to provide the methods and algorithms for retrieving high-quality information from these textual data, which can be incorporated into the MAGNETO knowledge database. In order to extract these fragments from the data, linguistic analysis of the textual data is necessary, and then clarifies the meaning of a given word in the context of the textual data. These analyses rely on theoretical models used either in rule-based natural language processing (NLP) or in statistical NLP, which is based on algorithms from machine learning or on a new approach using deep neural networks. The following are some of the analyses that have been deployed on the platform.

  – **Sentiment Analysis** is the automated process of analyzing text data and classifying opinions or emotions as negative, neutral or positive. Thus, it involves the identification of:

    Polarity: if the speaker expresses a positive or negative opinion.
    Subject: the thing that is being talked about.
    Opinion holder: the person or entity that expresses the opinion.

    The process to determine the sentiment score of a text is as follows:

    1. Break each text into its components parts (sentences, phrases, parts of speech)
    2. Identify each sentiment-bearing phrase and component.
    3. Assign a sentiment score to each phrase and component (-1 to +1, 0 to 4) For example, in the case of use a terrorist contacts one of his trainers to prepare an attack and names the expression “attack”, which denotes a negative feeling, so this phrase gets a score of “-1”, indicating that it is very negative.

    For the deployment, the Stanford CoreNLP library has been used. This open source library contains tools to convert a string contain human languages text into lists of sentences and words, to generate base forms of those words, their parts of speech and morphological features, and to give a syntactic structure dependency parse.

    When applying the code of Sentiment Analysis over a text, it is obtained a JSON with sentiment distribution, probability from 0 to 1 of each sentence is:

    0: very negative
    1: negative
    2: neutral
    3: positive
    4: very positive

    This module allows LEAs to save time because it is not necessary for them to extensively inspect each document Instead, they will be able to detect any anomalous pattern that triggers alarms and detect possible terrorist attacks,

as well as, like in this use case, intercept communications between terrorist networks and be able to disrupt them.

– **Keywords extraction** is an automated process of text data analysis. The main function is to extract important words in order to group the most relevant ideas and get an overview of the project.

  For module development, the YAKE [40] library has been used. It is a featured-based system for multi-lingual keyword extraction, which supports texts of different sizes, domain or languages. Unlike other extraction patterns, Yake is not based on the training of a corpus composed by dictionaries or word bags. Rather, it follows an unsupervised approach based on the features extracted from the text, which makes it applicable to documents written in different languages without the need for further knowledge. This can be beneficial for a large number of tasks and situations where access to training corpus is limited or restricted.

  The main features are:

  Unsupervised approach
  Independent training corpus
  Domain and language independent
  Single-Document

  Regarding the output, when applying the module on a text, an array is obtained with the keywords and their corresponding values, which oscillate between 0 and 1. The lower the point score, the more relevant the keyword will be.

  For example, in the case of use mentioned above, the keyword extraction analysis could be executed when we already have the transcriptions generated. In this way, it will be possible to analyse the degree of importance and store it to be able to show a complete analysis.

- Semantic Reasoning Tool

  It provides probabilistic reasoning which aims at the enrichment of existing information, as well as the discovery of new knowledge and relations between different objects and items of data. The technique employed, Markov Logic Networks (MLN), allows probabilistic reasoning by combining a probabilistic graphical model with first-order logic and allows to deal with uncertainties in the rules and the evidence. It is based on the open-source implementation of MLN reasoning implementation Tuffy [6], which expects the input as text in First-Order-Logic. An adapter has been developed to integrate the MLN reasoner into the MAGNETO framework, transforming the knowledge that is fetched from Apache Jene Fuseki (RDF Triple Store) to the formats required by the MLN reasoning tool. In cooperation with LEAs, a set of rules for different use cases has been developed to detect persons suspected in homicide investigations, terrorist threats or financial frauds. In Fig. 6 we see an example of the application of a rule in a tax fraud case, where low taxed heating oil is decolourized and sold at fuel stations without paying the fuel duty; the reasoning results are the relations marked in orange.
- Advanced Search

  It is a component based on ElasticSearch opensource solution. This component provides LEAs the capability to perform specific text queries to the stored information. This functionality is accessible through the MAGNETO HMI in two different ways. The first one - if a LEA needs to find a specific topic or text in the NOSQL DB, the

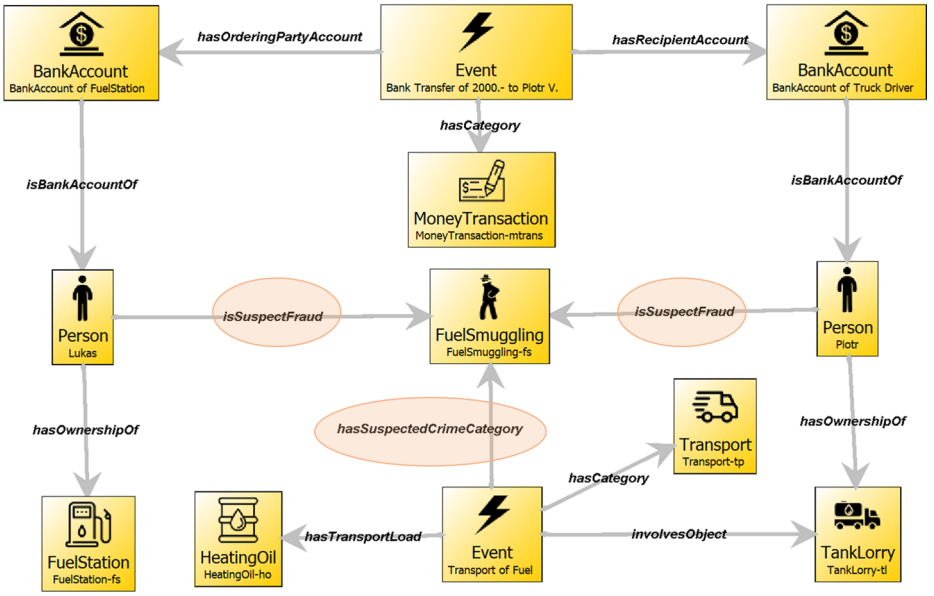


Fig. 6 Example of a reasoning result in a tax fraud case

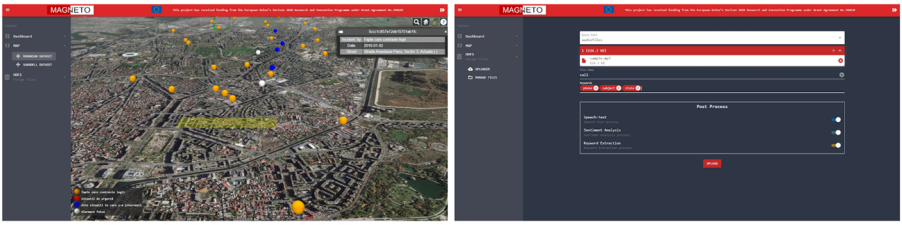
HMI provides an easy interface to find the elements that fit with the search patterns. The second one is via the MAGNETO dashboard. From this HMI component, the LEA is able to see the information stored in the platform in a graphical way. Both components take the advantages of ElasticSearch engine, which is able to index huge volume of textual data and perform fast queries.

- WEB HMI

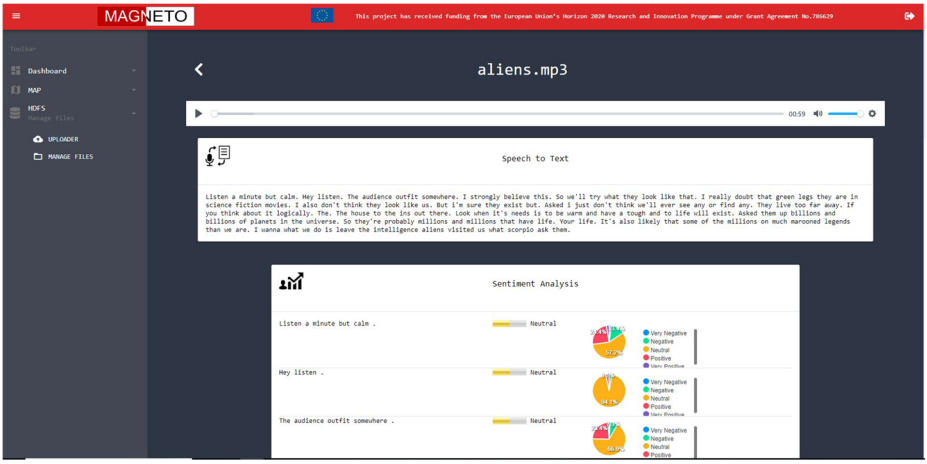
  It is a WEB solution based on NodeJS [26] and VueJS [38]. The development provides a common user interface where LEAs can access the stored data and interact with the different services provided by the MAGNETO platform. This approach tries to unify all the possible interactions with the platform in a single entry point where the LEAs will visualize the result of all the analyses performed (including fusion data, multimedia processing, semantic extraction, etc.). Due to the heterogeneity and fragmentation of the data, the information will be accessed through MAGNETO Web-HMI by default. However, in some cases, access to raw data may be necessary for external analytical tools or for internal processes that need the stored data to be accessible in some way.

  As can be seen in Fig. 7., specifically in the sub-image a), the aim is to represent the data provided in a georeferenced form in order to obtain a global view of the incidents. Each colour represents a different type of incident, in order to be able to make groupings and make the treatment simpler. On the other hand, in the sub-image b) you can see how a multimedia file loading service is used, in order to be able to store them in a distributed file system (HDFS). In post-process, it will run all the services marked at the time of loading, for instance Speech to text, Sentiment Analysis or keyword extraction.

  Finally, in sub-image c) you can see the results extracted from the post-processing execution of the selected services. It is possible to extract the transcription of the audio to the text, as well as the sentimental analysis carried out in an exhaustive way, since it is executed for each sentence, and also the extraction of key words is analyzed.



a) Georeferenced Data Viewer b) Platfom manual data ingestion



c) Performed analysis summary

Fig. 7 Web HMI platform's components

4.3 External communication and security

Due to several aspects such as security, but also accountability and ethical requirements, the MAGNETO platform needs an efficient AAA mechanism. Commonly, the **authentication** is carried out starting from the user that logs in. Typically, the approach is described as follows:

- Security module verifies user credentials and creates a session containing information such as their name, roles and permissions/constraints.
- The session is then returned to the client and recorded at the user side as a cookie.

As for the **authorization**, in various scenarios, it is carried out by the security module. It happens each time the backend service is interacted, and it commonly includes the following aspects:

- Requests of particular actions are sent to a security module alongside security information,
- Security information (carried out in the request) is verified,
- In the next steps, the user's rights are checked, i.e. whether the request can be executed or not.

In the final architecture specification of the Magneto platform, we have decided to adapt Role-based Access Control mode (RBAC); implemented by the Keycloak Identity Management Service (a key building block of Authorisation and Accountability service).

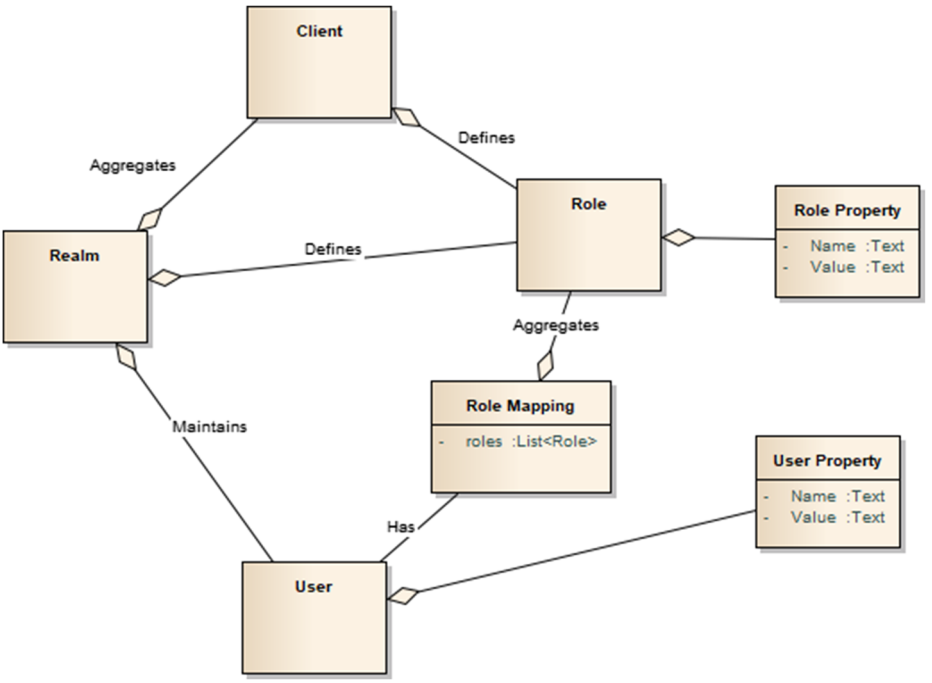


Fig. 8 General view on the proposed RBAC model

The model has a graph-like structure with the key node called a Realm (See Fig. 8). It maintains the set of users, credentials, roles, and clients. In the RBAC model, Roles play an important part in access authorisation, because application often applies permissions and access rights based on roles rather than individual user's name. Roles can be assigned at a level of Real or Client, which is an entity that could be related with a specific access channel (e.g. mobile device, web-based application, desktop application, etc.).

Authentication and authorisation (AA) microservice is one of the elements composing the entire MAGNETO platform. Its general concept and core functionalities interrelated with other MAGNETO platform components have been shown in Fig. 9. In this diagram,

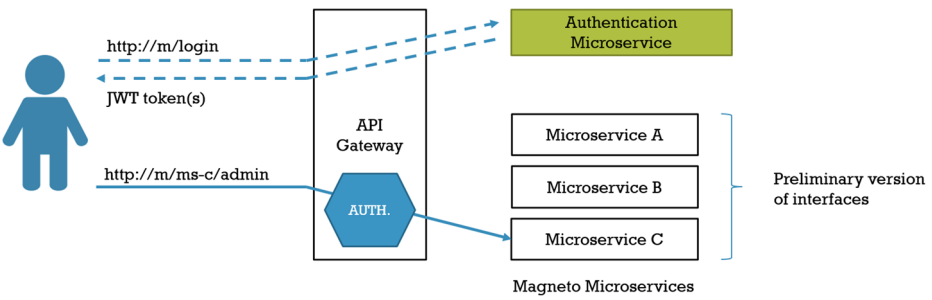


Fig. 9 General overview of authorization and authentication framework

we deliberately indicated the MAGNETO services as a black box (MAGNETO Microservices) because the authentication and authorization capabilities are presented as a general framework adapted within the project.

Here we emphasize that the communication between the "external world" and the internal services of the Magneto platform is realized via the API Gateway. In the proposed approach, the AA capabilities are implemented around the Keycloak server [16]. It is an open source Identity and Access Management system. It gives the user all kinds of mechanism to secure a broad spectrum of backend services and applications.

There are several scenarios of adapting the AA service:

- When an unauthenticated human client (who wants to access a particular backend application) will be forced to first access the Keycloak login page. After successful login, the client is brought back to the original application with a security token. In that scenario, the backend application just needs to confirm the token is valid.
- When an internal Magneto service (without interaction with human operator) will call an API request on another service that requires the caller to be authenticated. In such situations, the service will securely store the credentials that will be associated with that service (not the human user) and perform the authentication via HTTP(S) channel.
- When an internal Magneto service requires user credentials to call an API request on another services. In such a scenario, the authorisation is delegated to Keycloak and (when successful) a security token is returned. The service uses this token to call another services on behalf of the user.

The detailed interaction between a human user, the requested resource server and the microservice-based AA module is as follows:

1. User provides login and password requesting access to the service,
2. The request is forwarded through the API gateway to the service,
3. Authentication server generates JSON Web Token (JWT) with the information about user and forwards it to the API Gateway,
4. The Gateway grands the user the token,
5. User uses the token to request access to the service,
6. The request is forwarded via the Gateway to service,
7. Service validates the token in terms of its legitimacy/validity,
8. Service extracts user's privileges (roles) from the token.

Access to the resource is granted whenever the role matches the required privileges. Moreover, signatures and Message Encryption techniques will be employed by the security system for the protection of confidentiality.

The Authorisation and Accountability module provides (via Keycloak Identity Management Service) a dedicated web interface for authorisation and authentication configuration. This interface allows for centralised management of realms, clients, users, roles, sessions or identify broker.

5 Validation and results

Solutions and added value offered by MAGNETO: The main expectation is a full source-merging capability allowing cross searches, despite possible obstacles such as data volumes, heterogeneous formats, multilingualism, and alphabets (Latin, Cyrillic, Arabic, Korean,

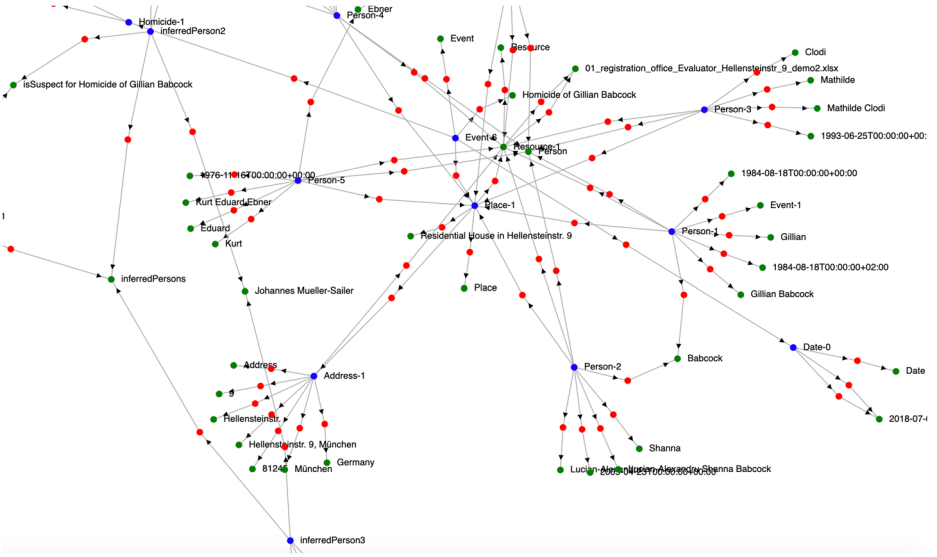


Fig. 10 General scenario overview

Chinese ideograms, etc.), the non-operability of existing tools on some project spectra, irrelevant data indexing, the frequently poor-quality data storage and intangibility, the limited geolocation accuracy and possibilities, etc. For its corresponding validation, a real use case was proposed in which data from attestations have been used in order to collect case information and to be able to cross-check data so that the search for a possible suspect can be much more efficient Fig. 10.

Inferred depiction of the residential addresses of persons

Reasoning Results
Person-48 "Christiane Anna Grossegger" Residential House in Leopoldstr. 8
Person-89 "Adalbert Melke-Lepges" Residential House in Leopoldstr. 8
Person-131 "Andre Weber-Dorsch" Residential House in Leopoldstr. 8
Person-103 "Thien Rameil" Residential House in Leopoldstr. 8
Person-70 "Pavel Emilia Kiss" Residential House in Leopoldstr. 8
Person-116 "Marie Katharina Schoeneck" Residential House in Leopoldstr. 8
Person-42 "Sully Eygermann" Residential House in Leopoldstr. 8
Person-83 "Norshima Lehmann" Residential House in Leopoldstr. 8
Person-129 "Ilse Wakas" Residential House in Leopoldstr. 8
Person-55 "Mathilde Hesselbach" Residential House in Leopoldstr. 8
Person-27 "Elise Belik" Residential House in Leopoldstr. 8
Person-5 "Kurt Eduard Ebner" Residential House in Hellensteinstr. 9
Person-96 "Tobias Oemler" Residential House in Leopoldstr. 8
Person-68 "Nadja Kendl" Residential House in Leopoldstr. 8
Person-110 "Maria Jakovlevic Schmid" Residential House in Leopoldstr. 8
Person-123 "Vladimirovic Vatersname Thunig" Residential House in Leopoldstr. 8
Person-21 "Radal Tundan" Residential House in Tegernseer Landstraße 210

Fig. 11 Reasoning result: persons in neighbourhood

Fig. 12 Inferred car models and owners

The different formats should not prohibit multi-criteria searching and relevant indexing, while search for weak signals should be made possible across this wealth of multi-origin information, giving minimum tangible guidance and decision support to the ongoing investigation.

In order to implement the envisaged solution, full interoperability of the different components is expected, to avoid multiple manipulations (e.g. passage of a text on the translator, the indexing, etc.), as well as easy recognition of sources and quotes together with consistency checking of data (e.g. location information) and statements of suspects or witnesses (in a neighbourhood survey) Fig. 11.

All this must take into account the EU and national legal frameworks (Penal Code, Privacy and Data Protection, etc.), the time constraints (duration of custody or detention, pending responsiveness of police), the human resources constraints (lack of personnel, training, equipment support) and, last but not least, the huge data volumes to be addressed, while resources may be scarce.

Thus, through the networks and connections generated on the MAGNETO platform thanks to the Semantic information processing service, it is possible to establish the connection that the suspect has with the rest of the people and thus quickly reach a potentially much larger group Fig. 12.

6 Conclusions

In this paper, a summary of the proposed architecture and the different components used for information extraction and analytics are presented. The prototype has been validated through a real scenario as shown in the results in the section above, with a set of modules deployed to show the interaction of a distributed file system, a NoSQL database, a search engine, a parallel processor based on containers for analytic, and a graphic user interface for control and visualization. The selection of the deployed components has been made, taking into account the improvement of LEAs. MAGNETO allows them to automate processes for extracting relevant data from police reports and documents, merging information and carrying out analyses that provide valid evidence for the prevention and investigation of crimes. This translates into greater ease and efficiency in the daily tasks of LEAs, providing them with the necessary resources for better decision-making and allowing them to focus on activities that require greater human intervention.

In addition, due to the importance of data protection, the platform has been made more secure with authentication and authorization mechanisms by a security module. The future work will focus on the usage of the platform in different use cases and scenarios, and the development and integration of various components required to perform additional tasks.

Acknowledgements This work has been performed under the H2020 786629 project MAGNETO, which has received funding from the European Union's Horizon 2020 Programme. This paper reflects only the authors' view, and the European Commission is not liable to any use that may be made of the information contained therein.

References

1. Ansible. https://www.ansible.com
2. Apache Cassandra. https://cassandra.apache.org
3. Apache Kafka. https://kafka.apache.org
4. Chauhan C, Sehgal S, et al. (2017) A review: crime analysis using data mining techniques and algorithms. In: International conference on computing, communication and automation (ICCCA). IEEE, Greater Noida, India. https://doi.org/10.1109/CCAA.2017.8229823
5. Chef. https://www.chef.io
6. Computer Science Department Stanford (2019). Tuffy: a scalable Markov logic inference engine. http://i.stanford.edu/hazy/tuffy/doc/
7. Dragos V (2013) Developing a core ontology to improve military intelligence analysis. International Journal of Knowledge-Based and Intelligent Engineering Systems, IOS Press
8. Elasticsearch (2019). https://www.elastic.co/
9. EnCase Forensic (2020). https://www.guidancesoftware.com/encase-forensic
10. Express JS. https://expressjs.com
11. Feng M, et al. (2019) Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access 7:106111–106123. https://ieeexplore.ieee.org/document/8768367
12. Hadoop Distributed File System (2019). https://hadoop.apache.org/
13. Hossein H, Xu H, Emmanuel S, Mansi G (2016) A review of data mining applications in crime. Statistical Analysis and Data Mining 9:139–154. https://doi.org/10.1002/sam.11312
14. Hussain D, et al. (2012) Criminal behaviour analysis by using data mining techniques. In: International conference on advances in engineering, science and management (ICAESM). IEEE, Nagapattinam, Tamil Nadu, India. https://ieeexplore.ieee.org/document/6215921
15. IBM I2 Analyst's Notebook (2019). https://www.ibm.com/us-en/marketplace/analysts-notebook
16. Keycloak (2019). https://www.keycloak.org/
17. Komalavalli C, Laroiya C (2019) Challenges in big data analytics techniques: a survey. In: 9th international conference on cloud computing, data science and engineering (Confluence), Noida, India, pp 223–228. https://ieeexplore.ieee.org/document/8776932
18. KongHQ (2019). https://konghq.com/kong/
19. Kuma. https://kuma.io
20. Ku C, Nguyen JH, Leroy G (2012) TASC - crime report visualization for investigative analysis: a case study. In: IEEE 13th international conference on information reuse & integration (IRI), Las Vegas, NV, pp 466–473. https://doi.org/10.1109/IRI.2012.6303045
21. Linkerd. https://linkerd.io

22. Luay A, Mahmoud S, Al-Sharif Z (2017) Towards hierarchical cooperative analytics architecture in law enforcement agencies. In: 8th international conference on information, intelligence, systems & applications (IISA), Larnaca, Cyprus. https://doi.org/10.1109/IISA.2017.8316400
23. Multimedia analysis and correlation engine for organised crime prevention and investigation. http://www.magneto-h2020.eu/
24. MongoDB. https://www.mongodb.com
25. Naik N, et al. (2017) Docker container-based big data processing system in multiple clouds for everyone. https://doi.org/10.1109/SysEng.2017.8088294
26. NodeJS Foundation (2019). https://nodejs.org/en/
27. Pocket Sphinx. https://pypi.org/project/pocketsphinx/
28. Puppet. https://puppet.com
29. py Audio Analysis. https://pypi.org/project/pyAudioAnalysis/
30. RabbitMQ. https://www.rabbitmq.com
31. Sill A, et al. (2016) The design and architecture of microservices. IEEE Cloud Computing 3(5):76–80. https://doi.org/10.1109/MCC.2016.111
32. Speech Recognition. https://pypi.org/project/SpeechRecognition/
33. Stanford CoreNLP (2019). https://stanfordnlp.github.io/CoreNLP/
34. Trunzer E, Kirchen I, Folmer J, Koltun G, Vogel-Heuser B (2017) A flexible architecture for data mining from heterogeneous data sources in automated production systems. In: IEEE international conference on industrial technology (ICIT), Toronto, ON, pp 1106–1111. https://ieeexplore.ieee.org/abstract/document/7915517
35. Truyen E, Bruzek M, Van Landuyt D, Lagaisse B, Joosen W (2018) Evaluation of container orchestration systems for deploying and managing NoSQL database clusters. In: IEEE 11th international conference on cloud computing (CLOUD), San Francisco, CA, pp 468–475. https://ieeexplore.ieee.org/document/8457834
36. VirtualBox. https://www.virtualbox.org
37. VMWare. https://www.vmware.com
38. VueJS Organization (2019). https://vuejs.org/
39. Yang M, Chow K-P (2015) An information extraction framework for digital forensic investigations. In: 11th IFIP international conference on digital forensics (DF). Orlando, FL, pp 61–76. https://hal.inria.fr/hal-01449071/document
40. Yet Another Keyword Extractor (2019). https://github.com/LIAAD/yake
41. Yu H, Hu C (2016) A police big data analytics platform: framework and implications. In: IEEE first international conference on data science in cyberspace (DSC), Changsha, China, pp 323–328. https://doi.org/10.1109/DSC.2016.84

Affiliations

Francisco J. Pérez[1] · Victor J. Garrido[1] · Alberto García[1] · Marcelo Zambrano[2,3] · Rafał Kozik[4,5] · Michał Choraś[4,5] · Dirk Mühlenberg[6] · Dirk Pallmer[6] · Wilmuth Müller[6]

Marcelo Zambrano
omzambrano@utn.edu.ec

Michał Choraś
mchoras@itti.com.pl; chorasm@utp.edu.pl

Wilmuth Müller
wilmuth.mueller@iosb.fraunhofer.de

1 Universitat Politècnica de Valencia, Valencia, Spain

2 Universidad Técnica del Norte, Ibarra, Ecuador

3 Instituto Superior Tecnológico Rumiñahui, Rumiñahui, Ecuador

4 ITTI Sp. z o.o., Poznań, Poland

5 UTP University of Science and Technology in Bydgoszcz, Bydgoszcz, Poland

6 Fraunhofer IOSB, Karlsruhe, Germany