# Innovative Big Data Analytics:
# A System for Document Management

Mariagrazia Fugini, Jacopo Finocchi
Dipartimento di Elettronica, Informazione e Bioingegneria
*Politecnico di Milano*
Milano, Italy
mariagrazia.fugini@polimi.it, jacopo.finocchi@i3lab.net

***Abstract*— This paper shows the solutions developed in the Project "Sistema Innovativo Big Data Analytics" named SIBDA. The needs of the three involved companies are described, which led to define an overall framework in the field of Big Data through application cases. The functional and technological requirements of an integrated Big Data Architecture are given. In particular, the criteria for selecting the solutions for Document Management for one company (Microdata Service) are described. The resulting Enterprise Content Management (ECM) system architecture and the overall system architecture are given. SIBDA stands for Sistema Innovativo Big Data Analytics, a project funded by Regione Lombardia within ``Accordi di Competitività", involving three ICT companies (Mail Up s.p.a, Microdata Service and LineaCom), belonging to the CRIT Consortium, Cremona, and Politecnico di Milano.**

***Keywords—Big Data architecture, document and content management, metadata extraction, machine learning for document classification***

## I. INTRODUCTION

In recent years, the term "Big Data" has spread more rapidly in the information technology market than in the academic field, while in the scientific literature the term was almost absent until a few years ago [1]. In general, quantitative thresholds are identified beyond which we deal with Big Data issues. These thresholds identify three dimensions: the volume of considered data, the speed of data acquisition and the variety of data [2]. The identification of precise quantitative thresholds for the definition of Big Data is influenced by the continuous progress made by technology, which extends the capabilities of the most general tools for managing and analyzing data. We can say we are talking about Big Data when these three combined dimensions make it impossible to manage a set of data with conventional techniques and technologies but require new tools, designed specifically to manage data at high volumes, high speeds and high variety. The tools include a set of techniques and technologies that, as a whole, allow extracting from the Big Data a value in terms of *information and knowledge* useful to companies and society [3]. Tools to manage large amounts of heterogeneous data in real time are increasingly important to address the new sources of data. Such data flow from devices and sensors of the Internet of Things – IoT (from smart devices to smart cities), from social media web platforms, from commercial and financial transactions, from digital media and multimedia systems and from data generated by business processes in the enterprise environment. The most significant techniques and technologies are discussed in [4].

This paper presents the main results of the Big Data Project concerning the study and development of an innovative solution in the field of Big Data, which can be shared among three companies, which created an "agreement for competitiveness". The perspective of this work is essentially experimental, giving an example of a Big Data Analytic solution to industrial problems in a real setting. In particular, the paper focuses on *Enterprise Content Management (ECM)* issues, which have been the focus or work of the authors. It also presents the overall system proposed architecture and discusses some experiments done as part of the project. It comments about some possible successful and unsuccessful solutions.

## II. THE SIBDA PROJECT

The most significant techniques and the technologies taken into consideration in the SIBDA Project belong to three areas described below, that are linked in the Project in a chronological succession in the overall *process* of Big Data manipulation. The process starts from the sources of origin of Big Data and reaches the delivery of the produced information. These areas are:

1. *Data acquisition*. Among the scenarios that characterize the stage of acquisition and initial processing of Big Data, one of the most *relevant* for its growing importance and for its technical specificities concerns data coming from the devices and sensors of the IoT, with the enabling middleware technologies and with the event processing techniques that allow an effective integration [5].
2. *Data storage*. For Big Data, we identify two technical problems: i) how to store large volumes of data while maintaining high performance; ii) how to archive unstructured or variable data. Two other issues related to storage techniques shared with more conventional storage solutions, such as relational databases (DBs), also play an important role: i) how to ensure the protection of archived data; ii) how to limit the complexity and costs of the hardware infrastructure when scaling up volumes. Different storage models have been proposed as an alternative to relational DBs, which

in the presence of large volumes of data cannot ensure adequate performance, particularly in response times. These new types of DBs include NoSQL DBs and NewSQL DBs [6]. An interesting model is that of a document-oriented DB, an associative DB that also manages complex internal data structures.

Values consisting of semi-structured data, are represented in a standard format such as XML, JSON (JavaScript Object Notation) or BSON (Binary JSON), and are organized as attributes or name-value pairs, where one "column" can contain hundreds of attributes, whose type and number can highly vary from one row to another. The most common examples are CouchDB (JSON) and MongoDB (BSON). This way of organizing information is particularly suitable for managing textual content. A further solution to manage data sets organized in complex and variable structures, as textual documents typically are, is based on text annotation, i.e., the incorporation of metadata directly into the data itself, always using XML or JSON syntax, even independently from the use of a document-oriented DB.

3. *Data analysis*. Analytical applications are the core of the Big Data phenomenon, as they are in charge of extracting a significant value, in terms of information and knowledge, from the *data* acquired and archived with the techniques described above. This result can be achieved either through Business Intelligence techniques, or through more exploratory techniques, which can be defined as Advanced Analytics. Therefore, the generated knowledge has to be made available to users and shared effectively with all the actors of the business processes that can benefit from it. In this area, we mention the techniques of Big Data Intelligence, Advanced Analytics, Content Analytics, Enterprise Search (or Information Discovery).

An overall view of the above themes is given in Figure 1.

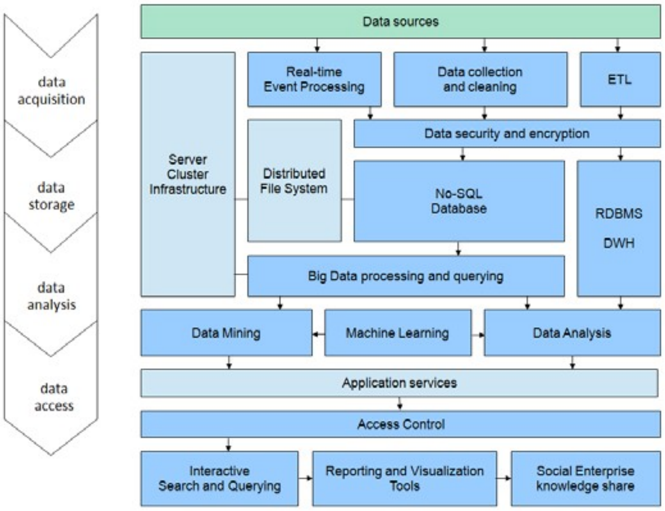
data acquisition
data storage
data analysis
data access
Data sources
Real-time Event Processing
Data collection and cleaning
ETL
Data security and encryption
Server Cluster Infrastructure
Distributed File System
No-SQL Database
RDBMS
DWH
Big Data processing and querying
Data Mining
Machine Learning
Data Analysis
Application services
Access Control
Interactive Search and Querying
Reporting and Visualization Tools
Social Enterprise knowledge share

Figure 1 Modules of interest for our Big Data System

*Areas of Interest in Big Data*

The interests of the three companies in Big Data are as follows.

*MailUp*, one of the leading Italian Email Service Providers, develops the main Big Data infrastructure (NoSQL Data Warehouse) and a series of Big Data analytical applications (Business Analytics).

*Microdata Service*, specialized in outsourcing services for management of document processes, develops an automatic document analysis environment, from which it will be possible to extract information of interest from the processed documents.

*Linea Com*, a member of the LGH Group operating in the IT and telecommunications services (which manages, among other things, the fiber-optic network of various municipalities in northern Italy), is in charge of creating a platform for the acquisition of sensor data.

The *logical model* of the envisioned Big Data system is the union of the assets characterizing each company, namely available hardware and software, knowledge and skills, data and information collected and managed. In the reminder of the paper, we will concentrate on Microdata and the ECM system portion of SIBDA.

## III. INNOVATION IN ENTERPRISE CONTENT MANAGEMENT SYSTEMS

Microdata is a company specialized in ECM process solutions. It has been on the market for over 20 years, as one of the leading Italian companies in the document outsourcing market. The links between the ECM activities and the use of a system for the management of Big Data are strong: on one hand, there are the volumes of documents managed and on the other the information contained in the documents themselves. The amount of data that populate the documents that Microdata must manage is part of the logic of Big Data, from which emerges the need to equip themselves technologically with solutions that allow exploiting the potential hidden in the data.

Considering the three dimensions of Big Data [7], ECM here is concerned with the *variety* aspect of big data, due to the complexity of unstructured data, while *volume* and *velocity* have a marginal role. As shown for instance in [8], it will be more and more relevant for public and private organizations to use document analytics to significantly reduce the time needed to manage various types of documents, e.g., contracts, and their management process, e.g., checking the compliance against regulatory needs.

To be able to seize this opportunity, Microdata needs a system capable of *recognizing* the type of document to be processed, *extracting* the contained information and subsequently *processing* the information along with other parameters to free up the benefit. The *Innovative Big Data Analytics System studied in SIBDA* falls exactly in this area of research. In fact, the goal of Microdata is to develop an innovative ECM environment, from which a DB of data extracted from the processed documents can be created, to be integrated into the DB elaborated by the overall Big Data Analytics System of the three companies.

In more detail, the objectives are (in order of priority):

1. *extraction* and *completion* of the forms containing document information;
2. *tracking* of the document processing process;
3. *extraction* of additional information and *creation* of new services.

Currently, in fact, the typical activity of the ECM processes provided to customers involves off-line tasks and, often, physical paper management (depending on how the customer sends and manages documents and information). The company therefore intends to change the concept of ECM into the most advanced concept of *Content Management* thanks to the development of an *Enterprise Content Management (ECM)* system.

### *A. Complexity*

The complexity of automation of ECM depends on:

- The quality of off-line manually processing performed by operators is very high, since they are skilled to manage all the different layouts of a document. Moreover, the process steps can be adapted to each customer's specific requests.
- The amount of activity that today operators have to perform on documents already in digital format are:
  1. classification of the type of document;
  2. extraction of information contained in the document and storage of such information in the DB built for storage;
  3. update of the progress of work to allow the system to keep track of where a certain type of document is located.

To quantify the *variety* dimension, we mention that Microdata manages around 1000 different processing types for more than 100 customers. The most complex processing involves more than 50 types of different documents.

During phases 1 and 2 above, operators must take into account the variety of processing requests of each customer for each type of document. It is clear how challenging the search is for a system able to automate this part of the process. Besides the intrinsic difficulty in managing documents, it is necessary for Microdata to achieve improvements in terms of speed and quality of work compared to what the manual activity achieves.

### *B. Realization Targets*

The goal of Microdata is to optimize the process of cataloging a document in a number of document categories known a priori and, based on the obtained classification, of the identification and extraction of an information set within the document. The main requirement of the system is that it is completely integrated and improved in terms of speed and quality of work, compared to the current situation where documents are processed manually. Finally, the availability of these volumes and information will allow Microdata to aim at increasing the range of services offered to its customers. In fact, to date, only those necessary to meet the needs of the specific customer are used (e g: verification of a specific value in an invoice). Another target regards the possibility to provide processing services for the single bureaucratic practice in near-real time, which is unfeasible manually.

The approach used for the evaluation of the solutions has foreseen a series of *Proofs of Concept (PoCs)*, to be carried out on a representative sample of real working cases, whose results (processing speed and quality of the result) are compared with the performances resulting from the current modalities operational. Depending on the results obtained by the PoCs, with a view to continuous improvement, the approach provides for a subsequent expansion of the sample of documents to be processed, a possible production of the solution adopted or, in the event that the results are not satisfactory, dropping the solution and searching for an alternative that could better reflect Microdata's expectations.

To understand the complexity of this innovation project, it is useful to remember that in the initial situation, operators manually perform document classification and transcription of the document content in digital format (when it is in paper form) used to store information that compose the document. The quality of these activities is very high and valuable. For this reason, the *automation of the process* is expected never to provide an analogous quality. However, automation must try to get as close as possible to what the operators' activities provide in a way that is economically sustainable, i.e., that does not affect the current satisfaction of customers and that can reduce process costs. The complexity elements that the IT systems identified for the automation of the process will have to overcome not only lie in the diversity of documentation to be identified and in the different management required by each individual customer, but also lie in the recognition of the text. A document, which partly involves manual compilation, can be presented in very many different ways: the system will have to recognize the text better, almost close to how a human performs.

To reach an approach based on continuous improvement, some scouting activities were necessary by a partner (in this case POLIMI-DEIB in the Project and third parties) who could offer Microdata the skills necessary to automate the process of classifying a document and extracting relevant information. This has led to the execution of many PoCs, aimed at continuing to refine the tools and therefore at improving the obtained results.

The *first experimentation* was performed in collaboration between Microdata and a Service Provider focused on analysis of images. This did not lead to any satisfactory results. The *second experimentation* was conducted through an appropriately parameterized and customized market solution. Two PoCs were created on processes related to insurance documents, both concluded at the end of 2016. In the first case, all the potential of the automation of a part of the process, although a series of problems that had to be resolved have also been brought into focus. The identification of the type of documents for testing was aimed at reducing the *time of occupation of the operators* and the *costs* when the process had been automated instead of being in processed via manual activities (from the perspective of economic sustainability mentioned above). Therefore, it was necessary that documents were numerically high and had a similar layout (to be able to train the algorithms to recognize where the text was placed). Subsequently, the design results were collected and analyzed. The result was positive and therefore the solution was put into production for the identified category of documents.

The next step is the expansion of the number of automated document types. The results of this are still under evaluation, although some aspects of improvement have convinced Microdata to continue the scouting activities with one of the project partners. The points that were considered as *improvable* are:

- Provide an unlicensed solution, with a view on expanding the automation of the document processing process and to "spread" the investment costs on the volume of processed documents;
- Streamline the system's education phase, while still guaranteeing high reliability;
- Provide a system able to adapt to multiple areas.

Another experimentation was conducted with a partner testing a tool developed ad hoc, with a Machine Learning (ML) approach. This approach would allow overcoming the limitations of the solutions based on the recognition of characteristic elements (features) of the documents. It allows instead the classification of documents not known a priori on the basis of the common characteristics that are identified in the set of documents processed (training set). Furthermore, the examined solution is not licensed, which allows Microdata to have savings when processing involves considerable amounts of documents.

To date, a PoC has been created on de-structured insurance documents. On the basis of the planned and under-construction functional architecture, we are identifying the pilot works on which the tool is to be applied, so as to be able to evaluate the effectiveness of the choices and the performances in relation to the current operating procedures (also making a comparison with the market tools).

This solution is in line with the Microdata implementation goal. Furthermore, the use of this new technology also makes it possible to detect process-tracking data in real time. The collection of these data in an automated manner makes it easier to carry out analyzes on load management in order to optimize the process. The new approach is based on a *finer granularity* of the process that hence can be highly optimized.

Besides, Microdata has the server infrastructure necessary to host the new solution. This action was necessary due to the type of data managed by Microdata and the strong regulatory constraints that customers impose on data management. Internal systems will be used for data processing, postponing any synergies with the data infrastructure of MailUp or Cloud solutions.

One of the most *important innovations* is the identification and implementation of a DB as a means to integrate all the vertical applications. Furthermore, another DB is set up for the storage of documents in the format in which they are transformed after the second phase.

The production choices adhere to the following directions:

- Choose components that can be integrated into an industrial process, to automate the processes while maintaining an acceptable level of quality.
- Favor solutions capable of minimizing the specific implementation work to be dedicated to each processing or to each type of document, exploiting tools with machine-learning capabilities, which progressively improve the quality of the result as the number of cases processed increases.
- Market solutions developed during the project respect the functional requirements of the project: in fact, the choice moved towards a NoSQL DB oriented to management of documents.

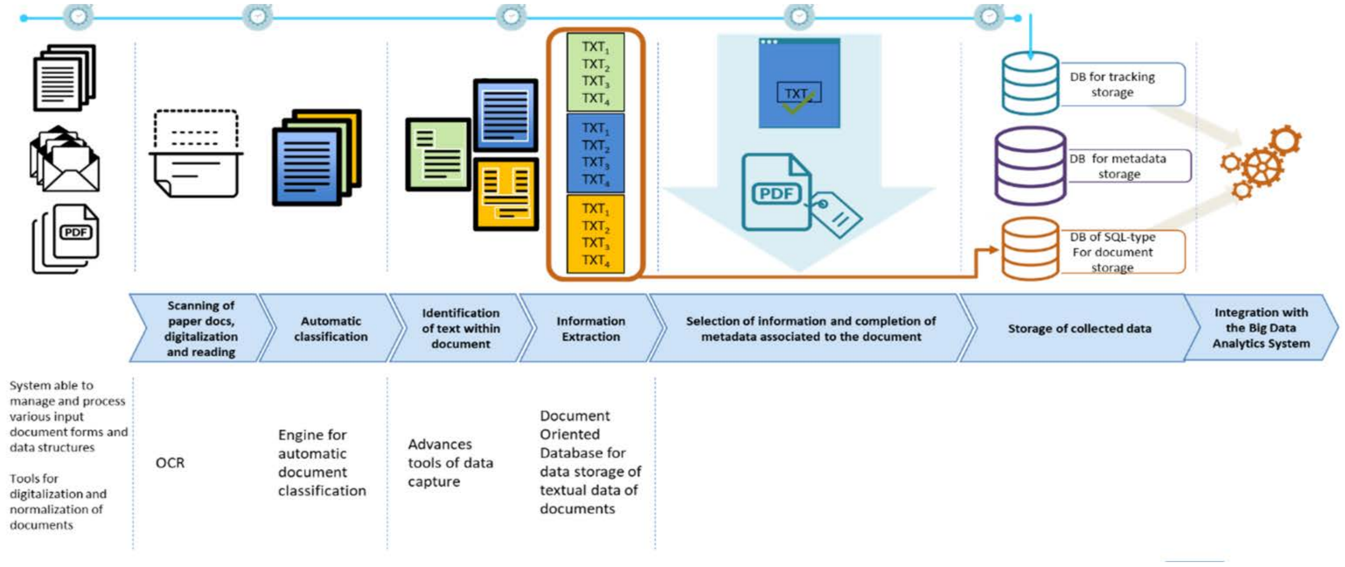
TXT1
TXT2
TXT3
TXT4
TXT
PDF
DB for tracking storage
DB for metadata storage
DB of SQL-type For document storage
Scanning of paper docs, digitalization and reading
Automatic classification
Identification of text within document
Information Extraction
Selection of information and completion of metadata associated to the document
Storage of collected data
Integration with the Big Data Analytics System
System able to manage and process various input document forms and data structures
Tools for digitalization and normalization of documents
OCR
Engine for automatic document classification
Advances tools of data capture
Document Oriented Database for data storage of textual data of documents

Figure 2. Logical model of the ECM system

## IV. SYSTEM LOGICAL MODEL

Figure 2 represents a schematization of the system designed to fit the realization objective of Microdata, and configured by taking into account the *project horizon* and the ECM *processes.*

The SIBDA project consists of seven phases, each foreseeing technological requirements able to satisfy the emerging requests for the most efficient and at the same time efficient automation of the process. The seven phases are:

*1. Multi-channel input*: The needs of Microdata customers require them to be able to manage multiple types of incoming documents and heterogeneous information transmitted, for example, via e-mail. The complexity lies in not being able to define a standard model that can be compatible with all managed formats and sometimes the information coming from mail is not structured. The different types of documents that characterize the offered ECM service are: paper mail and documentation, fax, microfilm, electronic documents, structured data flows, e-mails and PECs, electronic invoices, graphometric signature, etc. Microdata must therefore be ready to manage these elements of complexity thanks to appropriate technological solutions.

2. *Digitization and Optical Character Recognition* (OCR), for paper documents.

3. *Automatic document classification* (for all types of documents), identifies the type of document among the possible categories.

These phases 2 and 3 aim at rendering all documents in the same format. Considering the areas of the document where the text appears, the system is trained to recognize the type of document and classify it.

4. *Extracting text from the document*: after having correctly classified the incoming document, the computer system extracts the text contained in it, thanks to the knowledge of where it is located.

5. *Data selection and processing completion*: data contained in the documents are completed with metadata for indexing and for search activities within the storage systems. Moreover, part of the extracted data is used to complete the information requested by the customer, integrating it directly on the interface of the client company that is integrated into the processing process.

6. *Storage of collected data*: data collected during the whole process are stored. Those deriving from process tracking, which is also automated in the system, are collected in a document- oriented DB together with the data contained in the documents. The DB is oriented to management of documents and is aimed at managing extremely high data rates (Big Data logic), used for archiving and managing the metadata linked to documents within the manufacturing processes. Each record stored in the DB is a single document, progressively updated as processing progresses, with all the process-specific information (processing times, file references, etc.) stored internally. The extraction of information becomes very rapid, even for extremely high data volumes.

*7. Integration into the analysis system:* this last phase is in line with Microdata's objective of integrating the data collected in the infrastructure created by MailUp for the subsequent analysis.

Process tracking, which is managed in a completely automated way, is not a standalone phase but rather covers all the previous ones, up to the "Storage of the collected data".

The module of the Big Data system dedicated to the acquisition of data from multi-channel document sources addresses one of the characteristic problems faced by Big Data solutions, that of managing data of a complex and variable structure, not completely known a priori, as typically occurs in the case of documental data. This module, considered independently of the Big Data management system, represents for Microdata an important support to the analysis of the processes, given that the partner directly involved in this part of the system has important needs for optimization of the processing of documents, which will undergo a significant transformation from the introduction of new solutions.

The approach was prototyped, providing the basis for the experimentation of different specialized software tools, applied in a defined context, consisting of the definition of some PoCs. The purpose was to evaluate the *performances*, in terms of quality and speed of execution, and to estimate the *costs* of adoption, both as an implementation commitment and as preparation of the necessary infrastructure, before the subsequent passage into production on a set of pilots. The design of the modules started from the need to extract information from documents *even in the presence of a structure not described a priori* formally, and to carry out a *complete extraction* of all the textual information contained in the document, without being limited only to the information necessary for the required processing.

A first requirement of the solution to be implemented was to integrate in the solution effective functions of Machine Learning, in order to progressively develop the ability to recognize the texts and their structure starting from a set of training documents. In order to store the new contents extracted, it was also necessary to have a storage solution of high capacity and with the flexibility necessary to manage the complexity of document data.

Secondly, a high degree of modularity and flexibility of the processes was required, in order to compose the processes by assembling the operations on the single document types and to enrich the process over time with new types of documents and processing, based on customer indications.

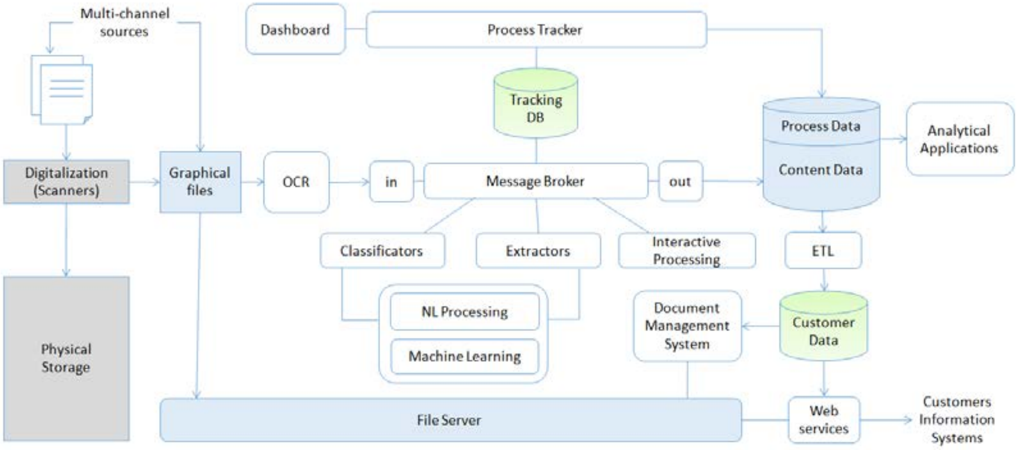

Figure 3 Phases and tools of the Enterprise Content Management system

Technologically, this need has led us to move towards a service architecture, which ensures the separation between the individual tasks and the overall process, with a message-broker that plays the role of orchestration.

The need for a complete integration of solutions in an industrial-type IT context, created the requirement of a server-based solution, avoiding the need to install software components on client stations and minimizing the operators' interventions. The solution had to start and complete the different phases of processing preferably in an automated way. The implemented software tool is able to interface directly with DBs and file servers and expose the operations performed to be able to interface with the other components of the SEBDA information system and in particular with the work supervision and monitoring system. The solution had to ensure the traceability of the various operations carried out, in order to allow subsequent process analysis.

To realize the document acquisition module and satisfy the described requirements, we built an integrated solution that includes all the phases: from the application of OCR functions, to the classification of the document and the extraction of the textual data, up to its registration in the pertinent destinations. An interactive intervention of the operators is limited to the processing of the fraction of documents that the automated tools are not able to process.

The *experimentation of two alternatives* has been performed. The first consisted of the adoption of a specialized product for the classification and extraction of texts, capable of autonomously managing the individual processing phases and based on the prior definition of the document models (through the development of a "template"). The second alternative has instead envisaged the development of a series of custom modules (in the ".Net" framework) each devoted to the automatic acquisition of data from a specific type of document, using a ML component able to acquire and improve the quality of the recognition and extraction of textual data. This second solution allows restructuring the document process by bringing the granularity to the single document level (rather than to a batch of several documents), with potentially significant consequences on the optimization and control of the process

For both scenarios, the activation of the different components is orchestrated by a message-broker (see Figure 3), coordinating the sequence of the various automated phases. The whole process will instead be monitored and controlled by a tracking system able to take into account the different types and granularity of processing, which will be realized as a custom component as an evolution of the current tracking tool already used by the partner (MD-Tracker). When operational, the automated solution will cover most of the partner's document acquisition activities, although some non-automated processes will remain, where operators extract and store contents interactively.

Even within the automated operations, still the software cannot classify a certain percentage of documents or cannot identify the information sought. This fraction of documents therefore constitutes a *residue* to be assigned to operators for interactive processing. The target is to minimize this residue.

To decide when an operator has to be invoked for an intervention, *three indexes of confidence* are computed regarding the automated phases (OCR, classification, data extraction). Each manually elaborated document becomes a part of the training set of the ML engine, so that next time the automatic systems has greater probability of processing it. The tools for control of the acquisition process and for process analysis concern both the automated and the manual mode. This is necessary both to allow the overall monitoring of the processes and to carry out the comparative analyses and evaluate the reliability and efficiency of the automated processes.

As for *storage*, the adoption of a NoSQL DB has been selected, in particular a document-oriented type, in order to store both the data extracted from the documents and the process data. Document-oriented DBs provide structured data types, typically in XML or JSON format (JavaScript Object Notation), which allow storing a multi-level text structure in a single point, thus facilitating the insertion and recovery operations. For this purpose, the MongoDB DB server was chosen, a document- oriented NoSQL DB, which natively uses the JSON format, in an extended variant with binary encoding, called BSON.

The possible extraction of a data flow towards the Big Data storage offered by the central core of the overall system is expected to be limited only to process data, especially due to strong regulatory constraints that limit the possibility of transferring content data. For *process data*, this option will be evaluated as a solution to generate more advanced process analysis. The analysis of process data occurs on a platform shared with MailUp, which provides integrated storage and analytic tools, with the aim of process improvement.

Thanks to the possibilities of the document module to extract a greater amount of information than those strictly requested by customers, it was also assumed the possible development of solutions dedicated to generating *new added value services*. Starting from the analysis of new content data that will be available, examples of services are envisioned for updating/removing obsolete information or for automated generation of suggestions and proposals to end customers.

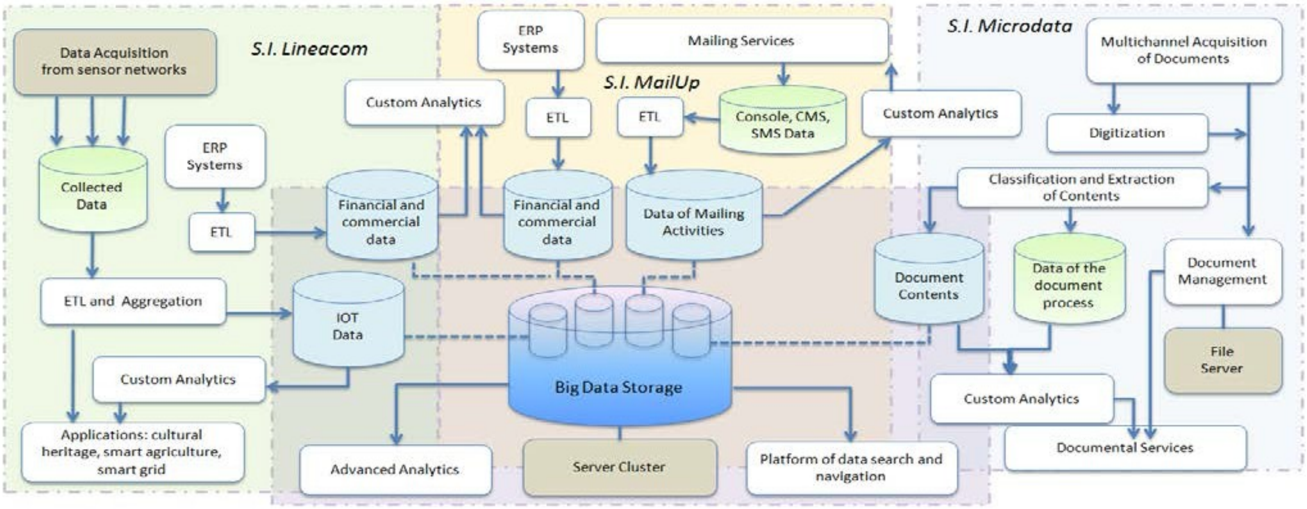

Fig. 4 Overall Architecture of the Big Data System

## V. CONCLUDING REMARKS

The transition into production of the pilot of the SIBDA system ad in particular of the described ECM systems will allow us to test the competitiveness of the solutions performances, precision and effectiveness of the new system. This will allow the companies to select the best tools to implement and develop the architecture of Figure 4, extending its use to an ever-increasing number of documents and processes and integrating IoT data and other types of big data characterizing the three companies' operations. A set of indicators is being tested for applicability. An analysis of information contained in texts is also aimed at discovering new services as a future work.

## ACKNOWLEDGMENTS

We are thankful to the partners in the Project “Sistema Innovativo Big Data Analytics”, particularly to the staff of Microdata and to the Project Manager R.Cerella - Fondazione Politecnico di Milano.

## REFERENCES

[1] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35.2 (2015): 137-144.

[2] Kaisler, Stephen H., et al. "Advanced Analytics--Issues and Challenges in a Global Environment." System Sciences (HICSS), 2014 47th Hawaii International Conference on. IEEE, 2014.

[3] Wang Y, Kung L, Byrd TA. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change. 2018 Jan 1;126:3-13.

[4] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information Sciences 275 (2014): 314-347.

[5] Cai H, Xu B, Jiang L, Vasilakos AV. IoT-based big data storage systems in cloud computing: perspectives and challenges. IEEE Internet of Things Journal. 2017 Feb;4(1):75-87.

[6] Moniruzzaman, A. B. M., and Syed Akhter Hossain. "Nosql DB: New era of DBs for big data analytics-classification, characteristics and comparison." arXiv preprint arXiv:1307.0191 (2013).

[7] Storey VC, Song IY. Big data technologies and Management: What conceptual modeling can do. Data & Knowledge Engineering. 2017 Mar 1;108:50-67.

[8] Joshi KP, Gupta A, Mittal S, Pearce C, Joshi A, Finin T. Alda: Cognitive assistant for legal document analytics. In AAAI Fall Symposium 2016 2016 Sep 28.