A Case Study Of Innovation Of An Information Communication System And Upgrade Of The Knowledge Base In Industry By ESB, Artificial Intelligence, And Big Data System Integration

Alessandro Massaro[1,*], Angelo Calicchio[1], Vincenzo Maritati[1], Angelo Galiano[1], Vitangelo Birardi[1], Leonardo Pellicani[1], Maria Gutierrez Millan[2], Barbara Dalla Tezza[2], Mauro Bianchi[2], Guido Vertua[2], Antonello Puggioni[2]

[1]Dyrecta Lab, IT Research Laboratory, Via Vescovo Simplicio, 45, 70014 Conversano (BA), Italy.
[2]Performance in Lighting S.p.A., Viale del Lavoro 9/11 - 37030 Colognola ai Colli (VR), Italy.

Abstract

In this paper, a case study is analyzed. This case study is about an upgrade of an industry communication system developed by following Frascati research guidelines. The knowledge Base (KB) of the industry is gained by means of different tools that are able to provide data and information having different formats and structures into an unique bus system connected to a Big Data. The initial part of the research is focused on the implementation of strategic tools, which can able to upgrade the KB. The second part of the proposed study is related to the implementation of innovative algorithms based on a KNIME (Konstanz Information Miner) Gradient Boosted Trees workflow processing data of the communication system which travel into an Enterprise Service Bus (ESB) infrastructure. The goal of the paper is to prove that all the new KB collected into a Cassandra big data system could be processed through the ESB by predictive algorithms solving possible conflicts between hardware and software. The conflicts are due to the integration of different database technologies and data structures. In order to check the outputs of the Gradient Boosted Trees algorithm an experimental dataset suitable for machine learning testing has been tested. The test has been performed on a prototype network system modeling a part of the whole communication system. The paper shows how to validate industrial research by following a complete design and development of a whole communication system network improving business intelligence (BI).

Keywords

Frascati Guideline, ESB, Data Mining, KNIME, Gradient Boosted Tree Algorithm, Big Data.

1. Introduction

Frascati manual [1] represents a useful guideline for research projects. This manual is adopted in order to find real research topics in Research and Development R&D projects. Concerning to the software research, the manual affirms that "*the effort to resolve conflicts within hardware or software based on the process of re-engineering a system or a network*". In this direction an important issue to validate research is in the design and development of a communication system allowing the information and data transfer. It is known that data coming from technologies of different databases and tools are characterized by different structures and

DOI : 10.5121/ijaia.2018.9503

formats. For this reason the Enterprise Service Bus (ESB) [2]-[6] could solve many conflicts between server machine and software tools such as Customer Relationship Management – CRM-, Enterprise Resource Planning ERP-, Business Process Modeling (BPM), and Product Lifecycle Management PLM. These tools can be tailored by means of different methodologies such as [7]-[9]:

- The Linear model (Waterfall);
- Evolutionary development;
- Incremental Model (Waterfall in iteration);
- Spiral Model;
- Formal systems development;
- Agile Methods;
- Reuse-based development.

Starting from these approaches it is possible to formulate and to improve the new Knowledge Base (KB) by using different databases technologies related to different tailored tools, and by processing, by means of artificial intelligence algorithms [10]-[14], all data stored into a big data system [15]. In particular Cassandra big data system is suitable for execution of data mining workflows [6],[16]. Furthermore Frascati affirms that is scientific research [1] "*the creation of new or more efficient algorithms based on new techniques*", and that "*The software development component of such projects, however, may be classified as R&D if it leads to an advance in the area of computer software. Such advances are generally incremental rather than revolutionary. Therefore, an upgrade, addition or change to an existing program or system may be classified as R&D if it embodies scientific and/or technological advances that result in an increase in the stock of knowledge.*". Following in this direction the basic information industry system (basic KB) has been upgraded by developing software tools having the main specifications summarized in Fig. 1. The main functionalities of the software tools are also described by the block diagram of Fig. 2 able to provide the following advantages:

- KB gain;
- Optimization of data flow management;
- New inputs for business intelligence B.I.;
- Quality improvements of the production activity;
- Data security improvements;.

The choice of basic software requirements represent the tailoring of a new information system based on ESB/Bus communication suitable for Enterprise Application Integration (EAI) [2]-[4]. The bus systems together with big data system supports the overcoming of conflicts and problems: as mentioned in [1]. It is a research of "*the effort to resolve conflicts within hardware or software based on the process of re-engineering a system or a network*". According to this requirement, big data is considered as a massive information collector of different data (structured, semi-structured, and unstructured data [17]), which can be interfaced with the ESB in order to facilitate data migration between databases having different technologies and data transfer into a communication system. If compared with traditional relational database systems, big data is mostly indicated for the resolution of hardware inefficiencies linked to real time data processing [18]. In addition, big data systems, through cloud data computing, can overcome various problems related to memory and hardware calculation capacity [19]. According to Italian directive of the MISE (Ministero dello Sviluppo Economico), the Frascati Manual represents a reference for the evaluation of the industry research enabling tax incentives. For this reason all the stages concerning a study case of applied research are discussed in the paper thus validating the research activity developed within of the industry Pil Ligthing S.p.A.. In order to furthermore validate the internal research, a classifier predictive algorithm has been applied. A big family of classifiers consists of decision trees and their ensemble descendants. The basic principle of a decision tree is to split the original data set into two or more subsets at each algorithm step, so as to better isolate the desired classes (classes labeled for the prediction) [13]. Each step then produces a split on the data set and each split can be graphically represented as a node of a tree whose branches define a rule path classifying the labeled classes [13]. In the actual ensemble evolution of the decision tree, instead of working with only one decision tree, it is possible to process data by different trees. The joint data processing of all the decision trees support the decision process and makes the algorithm more robust in terms of possible misclassifications. Implementations of ensemble decision trees are:

- Decision Tree
- Random Forest
- Gradient Boosted Trees

In the proposed research, Gradient Boosted Trees approach has been adopted, which is a machine learning technique for regression and classification problems. It, generates a prediction model in the form of an ensemble of prediction models.

The paper is structured as follows:

- design and development of KB system of the industry by discussing the main software requirements necessary to improve the KB after a preliminary “As is” process mapping;
- design and development of the Big Data and artificial intelligence tools applied on the upgraded systems providing new predictive outputs oriented on Business Intelligence (B.I.).

2. Upgrade Of The Information System And Of KB Following A Research Approach

The first step of the applied research concerns the gain of the KB. In order to achieve this gain the system has been tailored made into different modules as shown in of Fig. 1. Each one of the modules provides information and datasets to process them for the strategic marketing and B.I (Business Intelligence). All the implemented modules exchange and transfer data by the Local Area Network (LAN) interfaced with a unique communication bus. A preliminary management process design has been performed in order to plan the development activities. These modules allow upgradation of the initial KB by means of information digitalization concerning all the industry activities. We summarize below some of the important performed improvements related to each implemented module enhanced in Fig. 1 using different colored blocks (ARIS Express diagram).

2.1 Security Service Anti-malware

This research concerns the endpoint protection concept including virus identification engines, and more advanced secure systems oriented on network and Email security [20],[21]. Today there are different types of malware performing mass attacks with the aim of compromising the operation systems (DDOS or Distributed Denial Of Service). Furthermore modern ransomware are viruses able to “capture” through encryption, files on personal computers PCs and servers, thus requesting a “ransom” (Cryptolocker or Cryptowall function). The only anti-virus on the PC is not enough and today to ensure maximum protection, it is essential to use effective endpoint protection software, equipped with technologies of analysis of known patterns, being the automatic verification and real-time behavior of suspicious files insufficient. In addition, maximum protection is achieved when the endpoint protection integrates and collaborates actively with the devices included into the security perimeter of the network. The technologies adopted in the prototype system are able to provide maximum protection to clients, as they are equipped with different systems of analysis and inspection based on an Anti-malware system (called MaSS AM) suitable for remote management of endpoint security.

2.2 Firewall

Concerning firewall topic, the research has been focused on a new controller component integrated with the Information Technology (IT) architecture. The software component checks the data source using a firewall (named Barracuda) thus controlling the data flow and limiting actions in case anomalous situations. The integration of this new software component is part of the business integration concept, which allows to add new functionalities to the system, and to connect existing modules with the new ones. Firewall improves data traffic blocking for suspicious addresses. This according to the theory of Frascati [1], involves functionality in terms of security data properly designed for the company.

2.3 Web marketing

One of the frequent errors in the development of an internet site is to adopt a one-dimensional perspective. Some researchers in [22] suggest for this purpose to apply a particular model called "bar model". By this model a website is seen as:

- a set of contents, messages, interactions and possible transactions;
- a set of technical tools that make the contents accessible and feasible;
- an internet page which is implemented and updated by a group of users;
- an Iteration tools with visitors;
- a group of users who access and use it.

For a good design, the following logical phases of the production process can be executed:

- Phase 1: requirements definition;
- Phase 2: project start-up;
- Phase 3: web design;
- Phase 4: visual design;
- Phase 5: development;
- Phase 6: drafting of contents;
- Phase 7: publication.

where all the phases are analyzed in details in order to apply marketing strategies.
Concerning phase 1 the requirements are classified as follows:

- Content requirements: all the messages that the customers intend to communicate through the site;
- Structure requirements: these requirements explain how the contents are structured within the site;
- Content access requirements: these requirements show the way in which the user can access to web page contents;
- Navigation requirements: connections between the various contents that may be relevant to the visitor, in order to achieve the set objectives.
- Presentation requirements: presentation of contents, visual impact to be achieved, etc. (Web Design).

- Requirements for user operations (indications about operations, transactions and functionalities for the user);
- Requirements for system operations;
- Requirements related to the management of the site and its maintenance.

All the user feedbacks and sales results are used for the optimization of the web page by means of a data mining algorithms able to enhance clusters of products most sold and to predict demand marketing (time series forecasting prediction [23]-[24]).

2.4 Finance module

The Finance tool (called DocFinance) introduces an innovative business organization model that must be adopted to allocate efficiently resource and to optimize treasury operations thus supporting the relationships with the banks. This module is integrated into ERP SAP system thus allowing to use all the information in the ESB and Big Data informatics infrastructure.

2.5 Analytics and indoor/outdoor communication system

By means of Waterfall approach, a module useful to improve Knowledge Discovery, analytics, and indoor/outdoor communications has been developed. This module integrates Business Process Re-engineering (BPR-) and Business Process Management (BPM) facilities embedded in the ERP system. The Big Data analytics will provide important results mainly in the following topics:

- Logistics (reception, warehouse management, inventory control, transport planning, distribution management);
- Operations including maintenance and testing activities applied from the raw materials to the final products;
- Marketing and sales (marketing channel selection, promotion activities, price and space optimization);
- Services (customer support, installations, technology transfer).

2.6 ERP development

This module is related to a specific development of the ERP about management of accounting processes and management of commissions. These functions have been implemented after an accurate process analysis which enabled the reengineering process of the industry. The software has been developed by applying Waterfall and Agile methodologies. The implementation of this module allows the automation of different business processes thus optimizing the business intelligence (BI).

2.7 ERP basic platform

This module is related to the development of the following basic ERP SAP modules:

- Financial Accountig (FI);
- Controlling (CO);
- Sales and Distribution (SD);
- Material Management (MM);
- Production Planning (PP);
- Business Planning and Consolidation (BPC);
- Quality Management (QM);

- Asset Accounting (AA) ;
- Treasury (TR);
- Human Resource (HR);
- Warehouse Management (WM).

The ERP data represents the primary KB which will be transferred in the unique communication system.

2.7 Augmented Reality –AR- Building Information Model –BIM-

This module allows to adopt augmented reality-AR- for visual marketing. The AR is adopted for Building Information Modeling [25], specifically for virtual immersion in different environments involving different typologies of lights. The new AR objects are allocated in libraries stored in the a database of the information system, and are tailored by analyzing the behavior of the user during the AR visualization. AR is an enabling technology of Industry 4.0.

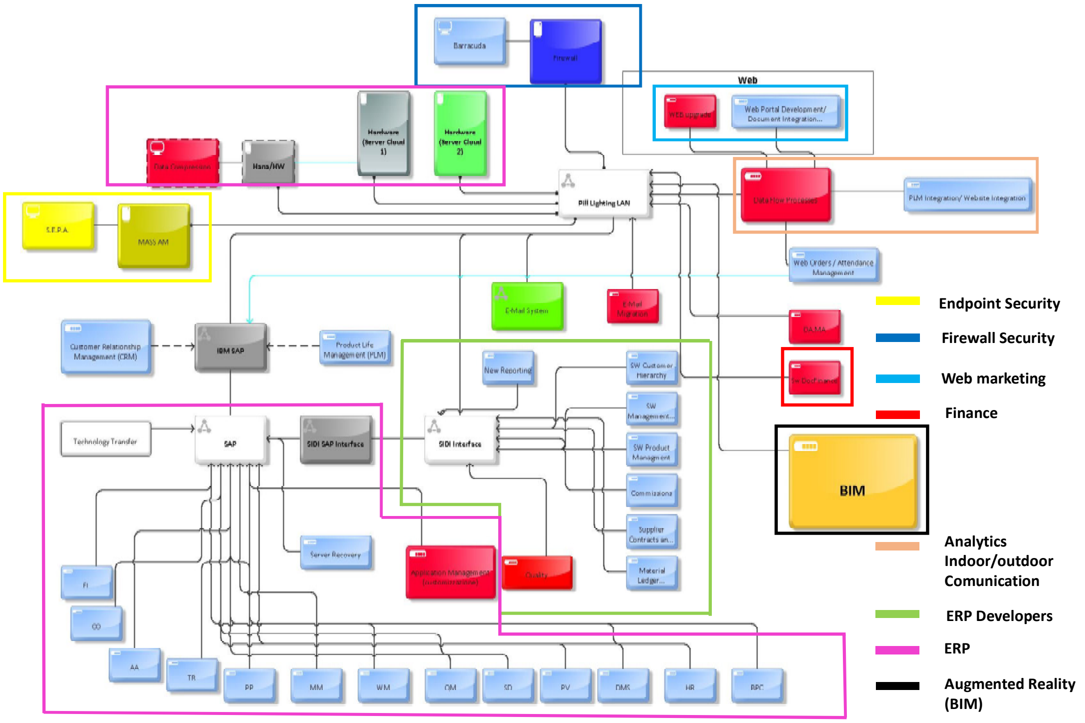


Figure 1. Aris Express Layout: Proposed upgraded Information System of Performance in Lighting S.p.A.

3. Gain of KB and Testing Prototype Basic Network

The gain of KB is achieved by integrating all the tools described in section 2 into an unique information system as sketched in Fig. 2. The diagram block of Fig. 2 summarizes the requirements of Fig. 1 and highlights project goals. The initial KB is upgraded by tools able to provide new structured and unstructured data and information useful for:

- Gain of knowledge related to different industry process in different area (administrative, sales, production, etc.);

- Optimization of the management of the new data flow by a communication bus;
- Improvement of data security;
- Improvement of product quality and product business.

These information can be collected into a big data and processed by an artificial intelligence engine able to deliver advanced outputs.

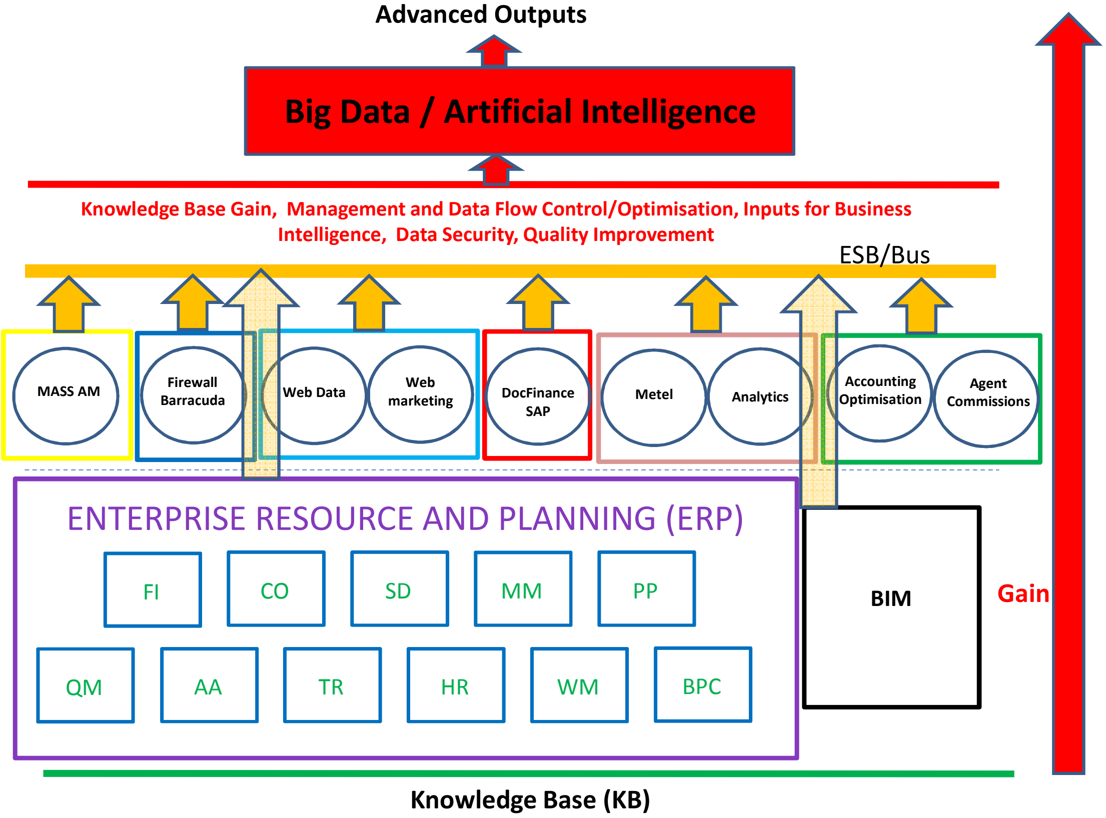

Advanced Outputs
Big Data / Artificial Intelligence
Knowledge Base Gain, Management and Data Flow Control/Optimisation, Inputs for Business Intelligence, Data Security, Quality Improvement
ESB/Bus
MASS AM
Firewall Barracuda
Web Data
Web marketing
DocFinance SAP
Metel
Analytics
Accounting Optimisation
Agent Commissions
ENTERPRISE RESOURCE AND PLANNING (ERP)
FI
CO
SD
MM
PP
QM
AA
TR
HR
WM
BPC
BIM
Gain
Knowledge Base (KB)

Figure 2. Proposed Knowledge Base Gain in Performance in Lighting S.p.A.

An example of a network prototype oriented on the implementation of the scheme of Fig. 2 is reported in Fig. 3, where the data flow is represented by the following three phases:

1) A Phyton script executed by a graphical user interface (GUI of the KNIME tool) activates a query for a data access (data stored into a MySQL database associated with a particular software tool);
2) The data are transferred into the big data system where it is combined with other ones;
3) The data collected in the big data system are processed by the KNIME workflow modeling artificial intelligence algorithm.

All the three phases are characterized by the data transfer occurring in ESB. The KNIME workflow is divided into the following main blocks, which will be described in details in section 3.2:

a) data preparation in local repository: the phase 1), 2) and 3) decribed above allows to prepare data into the local repository;
b) data pre-processing: some attributes are selected and filter to be processed;
c) data processing; data are processed by a training model used to learn the algorithm and by a testing data processing used to obtain predicted values.

We observe that the prototype network has been implemented by connecting a standard MySQL database. Authors checked the same connectivity of ESB with Sybase ASE database technology (ERP SAP database).

We observe that the data migration from a DB to a big data system can be also managed by the ESB by Application Programming Interfaces (API) executed as web services (see scheme of Fig. 4).

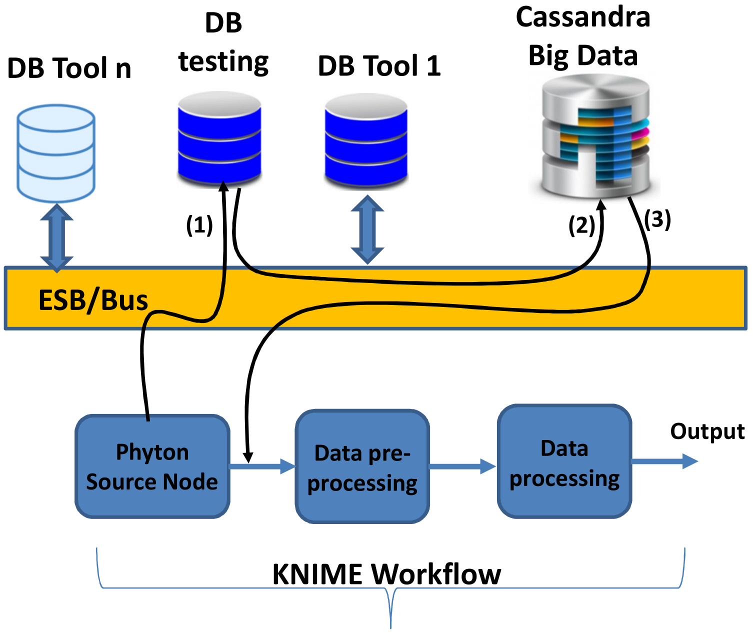


Figure 3. Testing prototype network enabling gain of KB by KNIME data mining workflow connected to a Cassandra Big Data system. The database DB testing refers to a database having the same technology (MySQL) of a developed tool of Fig. 1 and Fig. 2.

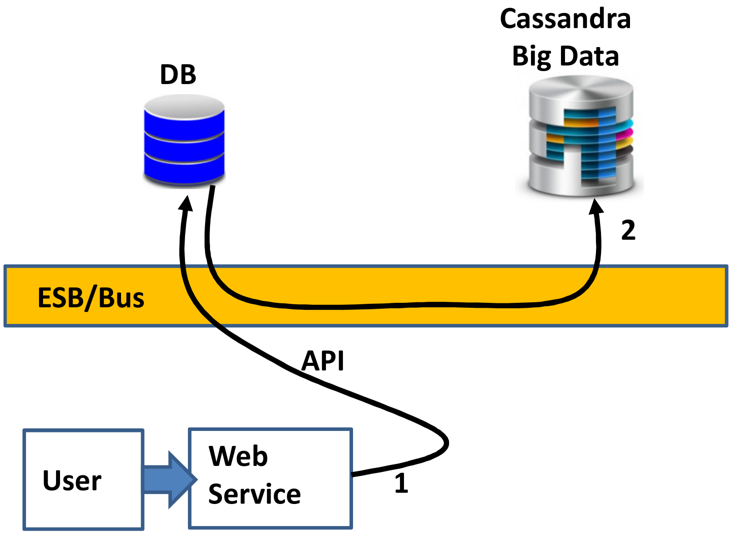
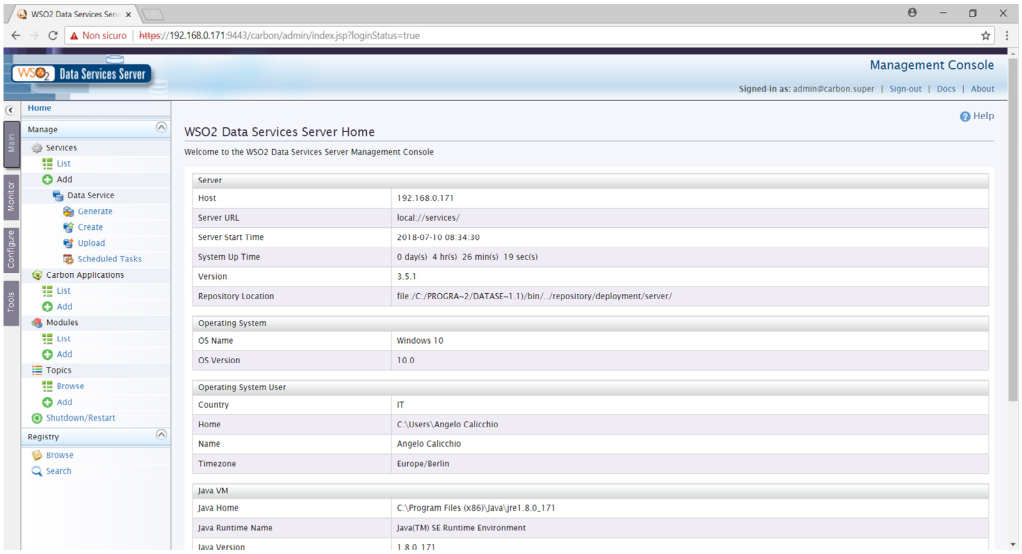


Figure 4. Web service enabling data migration from a DB to a Big Data.

3.1 ESB implementation

In order to test the prototype network of Fig. 3 the WSO2 Data Service Server module (WSO2 DSS) has been implemented [26]. The main function of the prototype ESB is to connect different databases having different technologies with the Cassandra big data systems, to manage data transfer into the local information system, and to generate data sources which will be imported in the local repository to be processed [6]. This module is also suitable for Service Oriented Architecture (SOA) for the creation of multisource data systems, and for hosting data services. It supports secure and managed data, data service transactions, and data transformation using an agile development approach according with tools specifications and data security. The WSO2 graphical interface facilitates the multiple sources management and the formulation of structured queries useful to select different data from different databases. In Fig. 5 is illustrated a screenshot of the implemented WSO2 DSS ESB. The ESB has been tested for database management involving all the tools indicated in Fig. 1 and Fig. 2 and Cassandra big data system (WSO2 includes a Cassandra database connector). In Fig. 6 is illustrated a screenshot of WSO2 indicating Sybase ASE connectors thus confirming the compatibility of the ESB with the ERP.

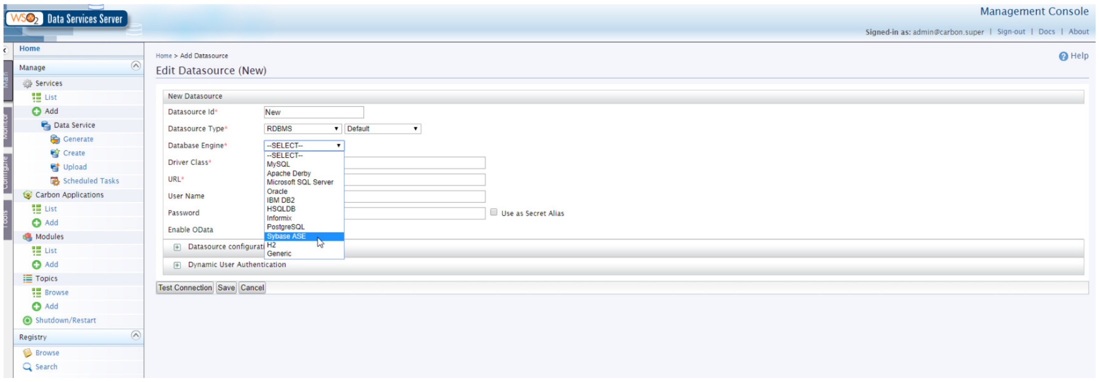


WSO2 Data Services Server Home

Welcome to the WSO2 Data Services Server Management Console

Server	
Host	192.168.0.171
Server URL	local://services/
Server Start Time	2018-07-10 08:34:30
System Up Time	0 day(s) 4 hr(s) 26 min(s) 19 sec(s)
Version	3.5.1
Repository Location	file:/C:/PROGRA~2/DATASE~1.1)/bin/../repository/deployment/server/
Operating System	
OS Name	Windows 10
OS Version	10.0
Operating System User	
Country	IT
Home	C:\Users\Angelo Calicchio
Name	Angelo Calicchio
Timezone	Europe/Berlin
Java VM	
Java Home	C:\Program Files (x86)\Java\jre1.8.0_171
Java Runtime Name	Java(TM) SE Runtime Environment
Java Version	1.8.0_171

Figure 5. Screenshot of WSO2 DSS ESB implementation.



Home > Add Datasource

Edit Datasource (New)

New Datasource

Datasource Id* New

Datasource Type* RDBMS ▾ Default ▾

Database Engine* --SELECT-- ▾ (Sybase ASE highlighted)

Driver Class*

URL*

User Name

Password □ Use as Secret Alias

Enable OData

⊞ Datasource configurati

⊞ Dynamic User Authentication

Figure 6. Screenshot of WSO2 DSS ESB implementation: Sybase ASE connectors (SAP ERP connectors).

3.2 Cassandra Big Data and KNIME testing

Apache Cassandra on a Windows machine has been installed in order to execute testing. The only prerequisites are [27]:

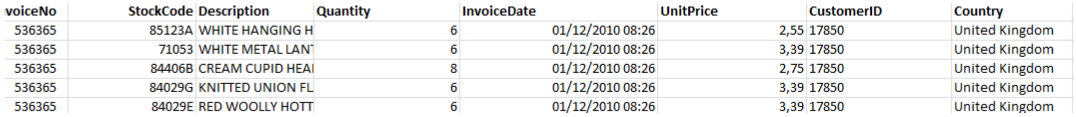
- Windows 7 or Windows 2008 server;
- Java;
- Either the Firefox or Chrome Web browser for DataStax OpsCenter;
- Visual C++ 2008 runtime (Windows 7 and Windows 2008 Server R2 has it already installed).

In Cassandra database, data available in the testing dataset of [28] has been loaded, which is a dataset suitable for machine learning testing .

In Fig. 7 is shown an example indicating the following attributes:

- VoiceNo: Invoice number (Nominal type): 6-digit integral number uniquely assigned to each transaction;
- StockCode: Product (item) code (Nominal type): 5-digit integral number uniquely assigned to each distinct product;
- Description: Product (item) name (Nominal type).
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time (Numeric type) indicating the day and time when each transaction has been executed;
- UnitPrice: Unit price (Numeric type) indicating product price per unit in sterling units.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

The used experimental dataset is described in [28],[29]: this is a dataset which contains all the transactions of a store occurring between 01/12/2010 and 09/12/2011 A dataset of about 50000 records (more precisely 45909 records) is extracted from the whole original dataset.



voiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING H	6	01/12/2010 08:26	2,55	17850	United Kingdom
536365	71053	WHITE METAL LAN	6	01/12/2010 08:26	3,39	17850	United Kingdom
536365	84406B	CREAM CUPID HEA	8	01/12/2010 08:26	2,75	17850	United Kingdom
536365	84029G	KNITTED UNION FL	6	01/12/2010 08:26	3,39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTT	6	01/12/2010 08:26	3,39	17850	United Kingdom

Figure 7. First records of the experimental dataset [28].

In Fig. 8 is illustrated the designed KNIME workflow able to predict sales validating the

prototype of Fig. 3. This algorithm is composed by 12 blocks named nodes. These node are:

- Node 1 (Puthon Source). This block allows to load in local repository the experimental dataset by means of the following Puthon script:

  ```
  from cassandra.cluster import Cluster
  from pandas import DataFrame
  cluster= Cluster()
  session=cluster.connect()
  query="""SELECT * FROM SAP.getVendite;"""
  rows=session.execute(query)
  data=[]
  for row in rows:
      data.append(row)
  output_table=DataFrame(data);
  ```

- Node 2 (Column Filter able to filter code, date, unit price and quantity);
- Node 3 (Rule-based Row Filter used to clean not congruous fields);
- Node 4 (String Manipulation block extracting from string Date only values in day / month / year format);
- Node 5 (String To Number-PMML- block enabling conversion from string to an integer number of quantity and price);
- Node 6 (Math Formula –Multi Column- block allowing multiplication of quantity by price);
- Node 7 (Column Filter block allowing elimination columns duplicated);

- Node 8 (Partitioning block splitting the experimental dataset into 80% of training dataset and 20% of testing dataset by taking data from top of the list);
- Node 9 (Domain Calculator block applied to define the attributes for the training model);
- Node 10 (Gradient Boosted Trees Learner –Regression-): this block represents the learning of the model and embeds the algorithm implementing very shallow regression trees and a special form of boosting to build an ensemble of trees; the implementation follows the algorithm described in [30] (Gradient Boosted Trees algorithm produces competitive, highly robust, interpretable procedures for both regression and classification. This algorithm is suitable to solve "predicting learning"[30]);
- Node 11 (Gradient Boosted Trees Predictor -Regression-): this block represents the testing of the model and applies regression from the Gradient Boosted Trees model; the output is the sales prediction;
- Node 12 (Numeric Score): this node computes certain statistics between the a numeric column's values and predicted values (an error of about ± 5 receipts is estimated);
- Node 13 (Excel Writer –XLS- block which saves outputs into an excel file).

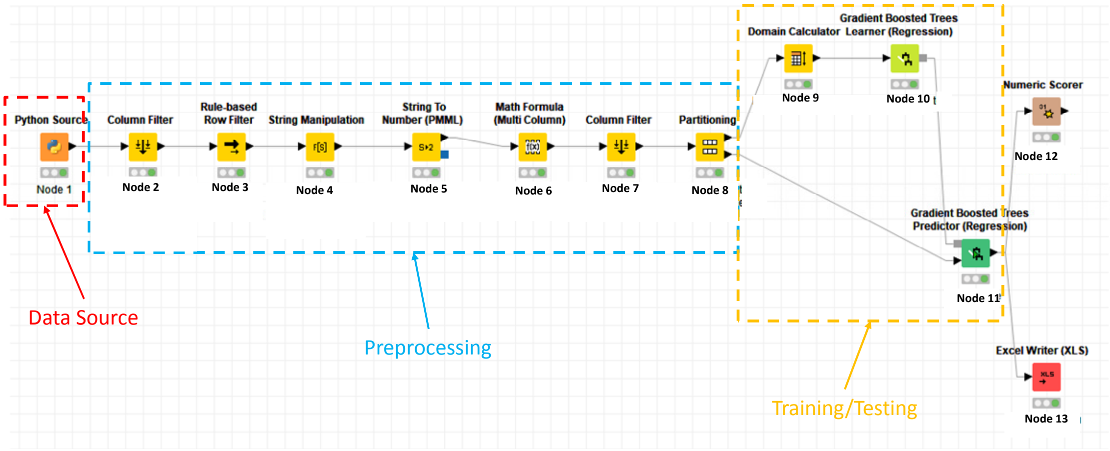


Figure 8. KNIME prediction algorithm data flow. In the figure are divided the main functions of the predictive workflow

In Fig. 9 is shown the comparison between real values and predicted ones of daily number of receipts by taking into account a specified day for each month of the year. The prediction is referred to the same period of the next year. A good convergence is found thus proving a truthful prediction. Figure 7 is obtained by means of plot facilities of the exported excel file (node 13). A part of the whole output in table format is illustrated in Fig. 10.

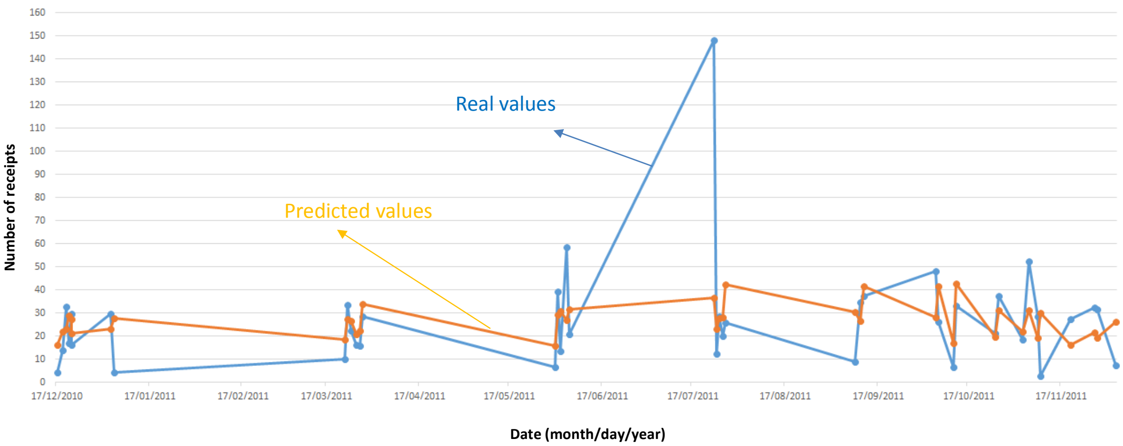
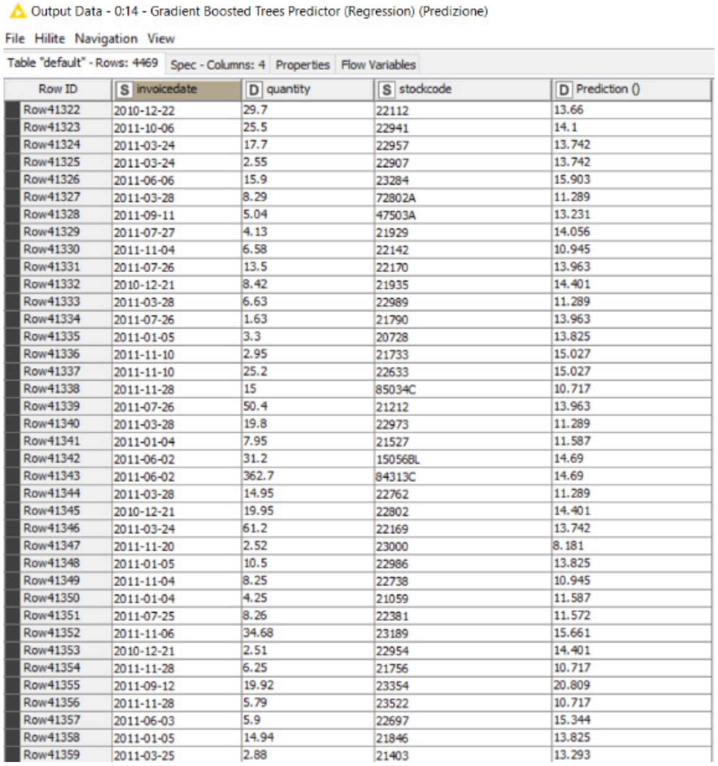


Figure 9. KNIME: Gradient Boosted Trees KNIME outputs indicating the comparison between real values and predicted ones of number of daily receipts. The predicted values are to consider for the 2012 year.


Output Data - 0:14 - Gradient Boosted Trees Predictor (Regression) (Predizione)
File
Hilite
Navigation
View
Table "default" - Rows: 4469
Spec - Columns: 4
Properties
Flow Variables

Row ID	S invoicedate	D quantity	S stockcode	D Prediction ()
Row41322	2010-12-22	29.7	22112	13.66
Row41323	2011-10-06	25.5	22941	14.1
Row41324	2011-03-24	17.7	22957	13.742
Row41325	2011-03-24	2.55	22907	13.742
Row41326	2011-06-06	15.9	23284	15.903
Row41327	2011-03-28	8.29	72802A	11.289
Row41328	2011-09-11	5.04	47503A	13.231
Row41329	2011-07-27	4.13	21929	14.056
Row41330	2011-11-04	6.58	22142	10.945
Row41331	2011-07-26	13.5	22170	13.963
Row41332	2010-12-21	8.42	21935	14.401
Row41333	2011-03-28	6.63	22989	11.289
Row41334	2011-07-26	1.63	21790	13.963
Row41335	2011-01-05	3.3	20728	13.825
Row41336	2011-11-10	2.95	21733	15.027
Row41337	2011-11-10	25.2	22633	15.027
Row41338	2011-11-28	15	85034C	10.717
Row41339	2011-07-26	50.4	21212	13.963
Row41340	2011-03-28	19.8	22973	11.289
Row41341	2011-01-04	7.95	21527	11.587
Row41342	2011-06-02	31.2	15056BL	14.69
Row41343	2011-06-02	362.7	84313C	14.69
Row41344	2011-03-28	14.95	22762	11.289
Row41345	2010-12-21	19.95	22802	14.401
Row41346	2011-03-24	61.2	22169	13.742
Row41347	2011-11-20	2.52	23000	8.181
Row41348	2011-01-05	10.5	22986	13.825
Row41349	2011-11-04	8.25	22738	10.945
Row41350	2011-01-04	4.25	21059	11.587
Row41351	2011-07-25	8.26	22381	11.572
Row41352	2011-11-06	34.68	23189	15.661
Row41353	2010-12-21	2.51	22954	14.401
Row41354	2011-11-28	6.25	21756	10.717
Row41355	2011-09-12	19.92	23354	20.809
Row41356	2011-11-28	5.79	23522	10.717
Row41357	2011-06-03	5.9	22697	15.344
Row41358	2011-01-05	14.94	21846	13.825
Row41359	2011-03-25	2.88	21403	13.293

Figure 11. KNIME table: output of the Gradient Boosted Trees Predictor node (node 11) indicating quantity prediction.

About 11 seconds For the execution of the workflow of Fig. 6 by using a personal computer having the following characteristics: Asus Intel Core i5 -7200U, 2.5GHz- 2,71 GHz, 8 GB Ram, OS Windows 10, 64 bit. For the data prediction have been adopted 100 as numnber of levels, a learning rate of 0.5, and 10 as three depth. We observe that the number of models is the number of decision trees to learn, that the limit number of levels (three depth) is the number of tree levels to be learned, and that the learning rate influences how much influence a single model has on the ensemble result (usually a value of 0.1 is a good starting point but the best learning rate also depends on the number of models).

The same algorithm of Fig. 8 can be applied to other dataset of the upgraded communication systems by predicting sales, financial trends, costs, AR contents impacts, revenues, visitors number of the industry web page, social trend and other factors introduced in the new KB. However databases of software could have some missing values due to data storage failures or to conflicts. This problem can be overcomed by the "missing value handling" procedure related the data- pre processing where the missing value handling can be:

- XGBoost (default): the learner will calculate which direction is best suited for missing values, by sending the missing values in each direction of a split; the direction that provides the best result (largest gain) is then used as default direction for missing values; this method works with both, binary and multiway splits.
- Surrogate: this method estimates for each split alternative splits that best approximate the best split [31]; this method can only be used with binary nominal splits.

4. CONCLUSION

The goal of the paper is to show results of a case of a study regarding research and development applied in industry. Following "Frascati" guidelines the research has been focused on the design and development of an ESB managing data contained into different databases and information sources. The data source is structured by tailoring different software tools enabling the gain of the knowledge base. The software tools have been structured by reengineering data flow process of Pil Ligthting S.P.A. information infrastructure, and by adding new functions suitable for B.I.. These tools have been implemented by following research methodologies such as Waterfall and Agile approaches. After the implementation and the testing of the properly designed tools has been developed a WSO2 ESB network able to solve different conflicts between hardware and software generated by the transfer of unstructured data and by the use of different database technologies. The ESB allows to connect different data source technologies by managing data transfer and collecting them into a Cassandra big data system. Big data allows to collect during the time massive data performing analytics and advanced statistical analyses about production trend, sales, costs etc.. Finally in the work has been developed a test a prototype communication system proving the connectivity of WSO2 ESB with Cassandra and with data mining KNIME algorithms. The test proved that all data stored into a data system of an application tool can be transferred and managed by the ESB, and, consecutively, processed by a data mining predictive algorithm. The prototype is tested by applying Gradient Boosted Trees predictive algorithm generating sales prediction outputs with good performances. Gradient boosted algorithms are defined by scientific community as innovative algorithms thus proving that the knowledge base is "gained by new algorithms". The gain of knowledge base is improved by all the communication system of the case of study composed by software tools, ESB, Big Data and Gradient Boosted Trees algorithms. Other innovations of the research can be found in the flowchart able to manage data transfer and data pre- processing/data processing of the whole workflow of the KNIME predictive algorithm. The correct execution of the KNIME algorithm proves that the research is well matched according to the industry requirements. This work can be considered as a first example case study demonstrated as a specified case study of Frascati research and development (R&D) theories. The proposed work can be a reference for the researchers working in industry projects.

ACKNOWLEDGEMENTS

The work has been developed in the frameworks of the Italian projects: “Process, Design, Management and Marketing in Lighting Production ‘De.Li.Pro’ ”. The authors would like to thank the following researchers and collaborators: B. Boussahel, V. Custodero, M. Le Grottaglie, A. Leogrande, F. Tarulli, and L. Patruno.

REFERENCES

[1] Frascati Manual 2015: The Measurement of Scientific, Technological and Innovation Activities-Guidelines for Collecting and Reporting Data on Research and Experimental Development. OECD (2015), ISBN 978-926423901-2 (PDF).

[2] Hohpe, G., & Woolf, B. (2004) “Enterprise Integration Patterns Designing, Building, and Deploying Messaging Solutions”, Addison-Wesley.

[3] Polgar, J. (2009) “Open Source ESB in Action”, IGI Publishing.

[4] Górski, T., & Pietrasik, K. (2016) “Performance analysis of Enterprise Service Buses”, Journal of Theoretical and Applied Computer Science, Vol. 10, No. 2, pp 16-32.

[5] Yenlo (2016) “ESB Comparison How to choose a reliable and fast ESB that fits your business needs”, white paper.

[6] Massaro, A., Maritati, V., Galiano, A., Birardi, V., Pellicani, L. (2018) “ESB Platform Integrating KNIME Data Mining Tool Oriented on Industry 4.0 based on Artificial Neural Network Predictive Maintenance”, International Journal of Artificial Intelligence and Applications (IJAIA), Vol. 9, No. 3, pp1-17.

[7] Bassil, Y. (2012) “A Simulation Model for the Waterfall Software Development Life Cycle”, International Journal of Engineering & Technology (IJET), Vol. 2, No. 5, pp1-7.

[8] Ragunath, P. K., Velmourougan, S., Davachelvan, P., Kayalvizhi, S., Ravimohan, R. (2010) “Evolving A New Model (SDLC Model-2010) For Software Development Life Cycle (SDLC),” IJCSNS International Journal of Computer Science and Network Security, Vol.10 No.1, January 2010.

[9] Rather, M. A., Bhatnagar, V. (2015) “A Comparative Study of Software Development Life Cycle Models”, Vol. 4, No. 10, pp23-29.

[10] Dilek, S., Çakır, H., Aydın, M. (2015) “Applications of Artificial Intelligence Techniques to Combating Cyber Crimes: a Review”, International Journal of Artificial Intelligence and Applications (IJAIA), Vol. 6, No. 1, pp21-39.

[11] Linoff, G. S., Berry, M. J. (2011) “Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management”, 3rdEdition, John Wiley & Sons Inc, 2011.

[12] Maimon, O., Rokach, L. (2006) “Data Mining and Knowledge Discovery Handbook”, 2nd edition, Springer US, 2010.

[13] Kotu, V., Deshpande, B. (2015) “Predictive Analytics and Data Mining”, Elsevier book.

[14] Adhikari, N. C. D. (2018) “Prevention of Heart Problem Using Artificial Intelligence”, International Journal of Artificial Intelligence and Applications (IJAIA), Vo., 9, No. 2, pp21-35.

[15] Khan, N., Yaqoob, I., Ibrahim, Hashem, A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., Shiraz, M., Gani, A. (2014) "Big Data: Survey, Technologies, Opportunities, and Challenges", Hindawi Publishing Corporation The Scientific World Journal, Vol. 2014, No. 712826, pp1-18.

[16] Massaro, A., Maritati, V., Savino, N., Galiano, A., Convertini, D., De Fonte, E., Di Muro, M. (2018) "A Study of a Health Resources Management Platform Integrating Neural Networks and DSS Telemedicine for Homecare Assistance," Information, Vol. 9, No. 176, pp1-20.

[17] Prasad, B. R., Agarwal, S. (2016) "Comparative Study of Big Data Computing and Storage Tools: A Review", International Journal of Database Theory and Application, Vol. 9, No. 1, pp45-66.

[18] Zheng, Z., Wang, P., Liu, J., Sun, S. (2015) "Real-Time Big Data Processing Framework: Challenges and Solutions," Applied Mathematics & Information Sciences An International Journal, Vol. 9, No. 6, pp3169-3190.

[19] Hashem, I. A. T.,Yaqoo, I., Anuar, N. B., Mokhtar, S., Gani, A., Khan, A. U. (2015) "The Rise of 'Big Data' on Cloud Computing: Review and Open Research Issues," Information Systems, Vol. 47, pp98–115.

[20 Pandove, K, Jindal, A., Kumar, R. (2010) "Email Security", International Journal of Computer Applications, Vol. 5, No. 1, pp23-26.

[21] Ruotsalainen, P. (2013) "Endpoint Protection Security System for an Enterprise", Master's Thesis, Jamk University of Applied Sciences.

[22] Cantoni L., Di Blas, N., Bolchini, D. (2010) "Comunicazione, Qualità, Usabilità, una Nuova Prospettiva per la Valutazione di Siti Web", Maggiolini Editore, ISBN: 9788838788888.

[23] Massaro, A., Maritati, V., Galiano, A. (2018) "Data Mining Model Performance of Sales Predictive Algorithms Based on RapiMiner Workflow", International Journal of Computer Science & Information Technology (IJCSIT),Vol. 10, No. 3, pp39-56.

[24] Massaro, A., Barbuzzi, D., Vitti, V., Galiano, A., Aruci, M., Pirlo, G. (2016) "Predictive Sales Analysis According to the Effect of Weather", Proceeding of the 2nd International Conference on Recent Trends and Applications in Computer Science and Information Technology, Tirana, Albania, November 18 - 19, pp53-55.

[25] Johansson, M., Roupé, M., Tallgren M. V. (2014) "From BIM to VR", Proceedings of the 32nd eCAADe Conference, Vol. 2 (eCAADe 2014), pp1-9.

[26] "WSO2 Data Service Server" 2018. [Online]. Available: https://wso2.com/products/data-services-server/

[27] "Getting Started with Apache Cassandra on Windows the Easy Way" [Online]. Available: https://www.datastax.com/2012/01/getting-started-with-apache-cassandra-on-windows-the-easy-way

[28] "Machine Learning Repository" [Online]. Available: https://archive.ics.uci.edu/ml/datasets/online+retail#

[29] Chen, D., Sain, S. L., Guo, K. (2012) "Data Mining for the Online Retail Industry: A Case Study of RFM Model-Based Customer Segmentation Using Data Mining", Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp197-208.

[30] Friedman, J. H. (2001) "1999 REITZ LECTURE Greedy Function Approximation: A Gradient Boosting Machine", The Annals of Statistics, Vol. 29, No. 5, pp1189-1232.

[31] Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A. (1984) "Classification and Regression Trees", Taylor & Francis.

Corresponding Author

Alessandro Massaro: Research & Development Chief of Dyrecta Lab s.r.l.