

A BUSINESS INTELLIGENCE PLATFORM IMPLEMENTED IN A BIG DATA SYSTEM EMBEDDING DATA MINING: A CASE OF STUDY

Alessandro Massaro, Valeria Vitti, Palo Lisco, Angelo Galiano and Nicola Savino

Dyrecta Lab, IT Research Laboratory, Via Vescovo Simplicio, 45, 70014 Conversano
(BA), Italy.

(in collaboration with ACI Global S.p.A., Viale Sarca, 336 - 20126 Milano, Via Stanislao
Cannizzaro, 83/a - 00156 Roma, Italy)

ABSTRACT

In this work is discussed a case study of a business intelligence –BI- platform developed within the framework of an industry project by following research and development –R&D- guidelines of ‘Frascati’. The proposed results are a part of the output of different jointed projects enabling the BI of the industry ACI Global working mainly in roadside assistance services. The main project goal is to upgrade the information system, the knowledge base –KB- and industry processes activating data mining algorithms and big data systems able to provide gain of knowledge. The proposed work concerns the development of the highly performing Cassandra big data system collecting data of two industry location. Data are processed by data mining algorithms in order to formulate a decision making system oriented on call center human resources optimization and on customer service improvement. Correlation Matrix, Decision Tree and Random Forest Decision Tree algorithms have been applied for the testing of the prototype system by finding a good accuracy of the output solutions. The Rapid Miner tool has been adopted for the data processing. The work describes all the system architectures adopted for the design and for the testing phases, providing information about Cassandra performance and showing some results of data mining processes matching with industry BI strategies.

KEYWORDS

Big Data Systems, Cassandra Big Data, Data Mining, Correlation Matrix, Decision Tree, Frascati Guideline.

1. INTRODUCTION: BASIC STATE OF THE ART DEFINING MAIN PROJECT SPECIFICATIONS OF THE BI PLATFORM

1.1 BASIC STATE OF THE ART

In this section is commented the state of the art concerning the specifications providing the basic architecture of Fig. 1. Basic form of business intelligence –BI- are performed by frameworks characterized by three operational levels as [1]:

1. Data access and analysis;
2. Data access;
3. Data capture and acquisition.

These levels indicate that digital data are fundamental for every BI operation thus confirming that an efficient information system should integrate facilities about data collecting and data

processing. BI strategies can be performed through big data systems [2]. These systems are of great importance in the case where there is a need of data retrieval speed, of a massive amount of information to process, and of a collecting of a wide range of data formats and types [3]. BI procedures can be applied in the field of transport and logistics [4], or can provide new service solutions starting from driver travel experience analysis [5]. Other authors in [6] studied workflows of Research Design Business Intelligence System –RDBIS- by enhancing the importance to embed into an unique data system BI, Data Mining –DM-, Decision Support System –DSS- and Strategic Management SM also by processing data from Customer Relationship Management –CRM- and Enterprise Resource Planning –ERP- tools. These studies enhance the importance to collect information coming from different software in order to structure an efficient BI platform.

Below is listed a series of tools and data systems suitable for BI integration[6]:

- Supply Chain Management –SCM-;
- CRM system;
- Data Mining –DM-;
- artificial intelligence- AI-;
- On-Line Analytical Processing –OLAP- techniques;
- Knowledge Management –KM-;
- Business Process Modeling –BPM-;
- Strategic Management –SM- tools;
- Analytic Network Process –ANP-;
- Quality Management System –QMS-;
- Decision Support System –DSS-;
- Performance Scorecard –PS-;
- Extract, Transform, Load –ETL- tools;

In particular customer data analysis is significant for BI strategies [7]. The concept of integrating CRM, DM and ERP into BI was also discussed extensively in [8], where has been proposed a reference framework model for research. Also the life cycle of a product or service can be considered into a BI plan [9]. Other studies have analyzed the phases of the BI life cycle by defining the following stages suggesting the operative steps [10]:

1. Feasibility study;
2. Project planning;
3. Business analysis;
4. Design;
5. Construction;
6. Implementation.

Another important aspect is the definition of the relationships involving the BI into an integrated information system [11]. However, an efficient integration of ERP processes into BI is of particular importance [12]. In [13] have been analyzed the processes associated with BI in terms of business value: these processes can be considered for the design of an efficient information system including:

- strategic alignment;
- process engineering;
- change management;
- BI technical development;
- BI project management.

However, the ERP can have a significant impact on the Business Decision Making –BDM- [14], and can constitute the first level of an efficiently structured operational workflow architecture [15].

Another important function for the BI analytics development is the data storage [16]. Concerning service innovation and engineering, below are listed some generic important processes discussed in [17] :

1. Concept Design.
 - Value Proposition;
 - Market Research;
 - Key Partners;
 - Distribution Channels;
 - Key Resources;
 - Integration Specifications;
2. Service Development.
 - Data Access & Processing;
 - Use of standards;
 - Software Development,
 - Data Security;
 - Certification Process;
3. Service Operation.
 - Distribution;
 - Customer Relationship;
 - Service Updates.

Also annual indicators [18], market prediction [19], and car accident analyses [20] can be useful for car services improvements.

1.2 MAIN PROJECT SPECIFICATIONS

Following the state of the art has been designed the BI architecture of Fig. 1 characterized by the following preliminary project specifications and requirements:

- The redefinition of company organization charts and the optimization of business processes;
- The feasibility study of a modern ERP replacing the current data flow system based on AS400 information system , adding the new BI system;
- The formulation of company dashboards able to generate customized operative/ financial reports and addressed to the management departments;
- The creation of a platform integrated with the ERP and CRM tools;
- The activation of digital facilities accelerating the service processes;
- The formulation of data extraction procedures able to keep data from different supporting BI and data mining processing;
- The implementation of a big data system [21] that can be interfaced with the database systems;
- The implementation of Graphical User Interfaces Interfaces -GUI- in data mining engines (such as Rapid Miner, or Weka, or KNIME, or Orange or others [22]-[25]) suitable for

the application and management field of the industry, and able to provide predictive outputs, data classification, data clustering, and association rules supporting the BI;

- The upgrade of the whole information system (email system, directory system, etc.);
- The determination of procedures for balancing the parameters of the data mining algorithms based on the comparative analysis of the data pertaining to the two different industry location;
- The testing of every project element and of the whole designed system.

Following these requirements has been developed a prototype BI platform where some facilities are discussed in this paper in order to provide information about methodologies and approaches suitable for the research. The paper is structured as follows:

- Description of the main Research and Development -R&D- scenario within has been developed the BI platform by citing the ‘Frascati’ manual;
- Discussion of some operating architectures supporting the knowledge gain and system testing involving big data;
- Presentation of an approach for the estimation of big data performance;
- Presentation of some examples of dashboards and data mining workflows interconnected to the big data system and supporting the decision making and improving the BI platform.

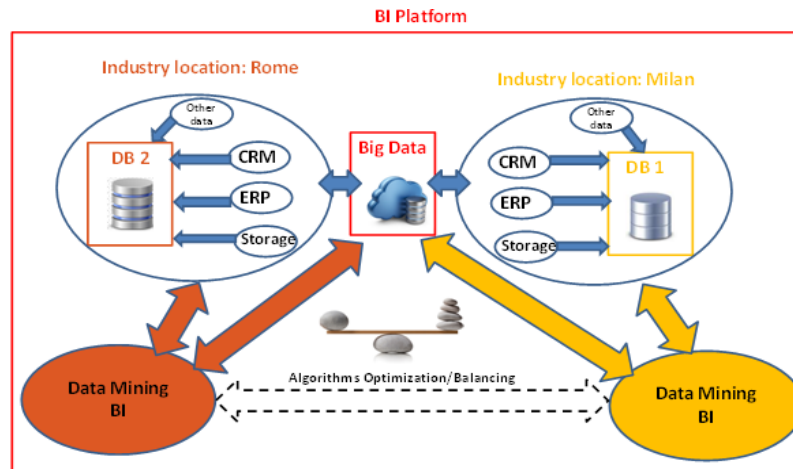


Figure 1. Basic architecture of the proposed BI platform.

2. FRASCATI R&D GUIDELINES

The research activities are aligned with R&D guidelines discussed in ‘Frascati’ manual [26]. According with the scheme of Fig. 2. The project specifications are able to increase the systemic knowledge stock of the car service company by providing a knowledge gain. Each software tool - SW_i- (level 2) facilitates the use of the Knowledge Base -KB- (level 1) of the industry by providing digitalized information and datasets. By means of information infrastructures such as Enterprise Service Bus -ESB- [27]-[29] and NoSQL big data technology (level 3), it is possible to improve the gain of knowledge and the BI by executing innovative algorithms (level 4). The level 3 of the architecture of Fig. 2 is also important to solve conflicts between hardware and software tools, and to integrate data having different formats and generated by different database technologies. The research project is mainly focused on the development of level 3 and level 4 of the architecture of Fig. 2.

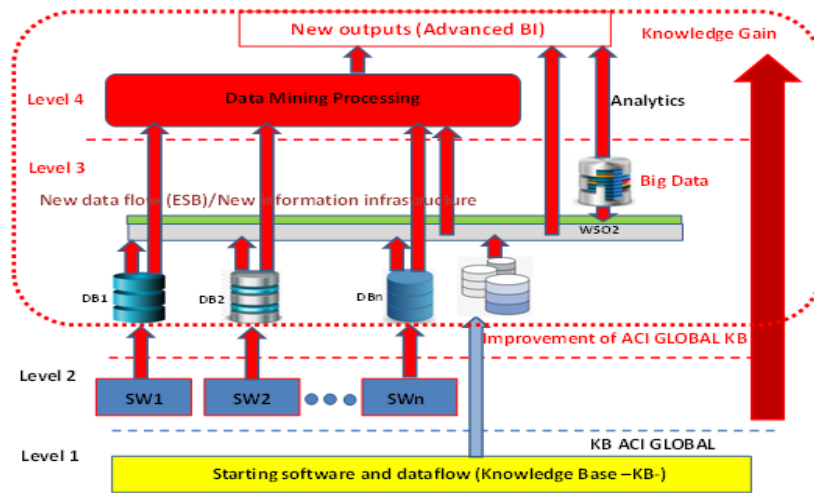


Figure 2. Knowledge Base Gain concept in ACI Global S.p.A. and four levels model upgrading KB.

3. OPERATING ARCHITECTURES

The general architecture of Fig. 2 is simplified by the architectures of Fig. 3 and Fig. 4. In particular the architecture of Fig. 3 is structured by the following elements: (i) industry operative systems (call center systems, customer services etc.) representing the basic software enabling the industry communication system; (ii) ERP platform representing the main part of KB of level 2; (iii) web tools (e-commerce, web system able to connect different industry locations, etc.); (iv) data warehouse (database system including interconnection between big data system and other databases) representing data integration of level 3; (v) advanced BI tools enabling data warehouse access and processing by executing innovative algorithms (level 4). The data flow related to level 2, level 3 and level 4 is sketched in Fig. 4: each software of the industry information system (SW_i) exchanges data by means of the WSO2 ESB [29] executing different scripts (java, php, python, etc.) able to transfer data from each database to Cassandra, to migrate some data to a local MySQL buffer DB (the buffer DB can be useful in order to test some data mining algorithms or for statistic analytics performed by standard tools), to lunch data mining algorithm, and to execute dashboards and data processing reports facilitating BI outputs reading. The choice of the open source Cassandra is due to its high performance if compared with other big data system performance for write, read and delete operations [21].

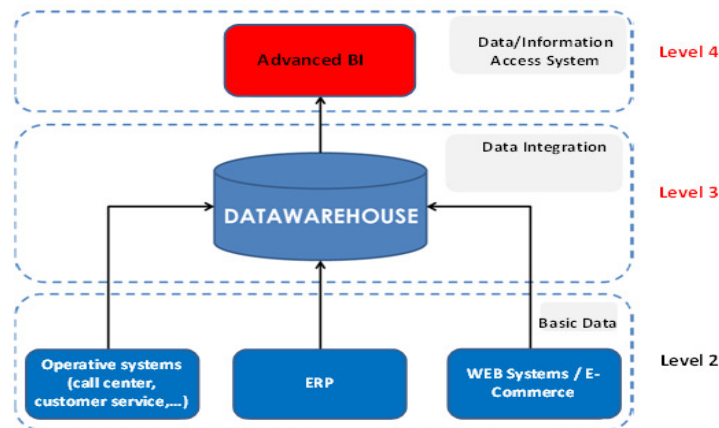


Figure 3. Testing architecture concerning level 2, level 3 and level 4 of Fig. 2.

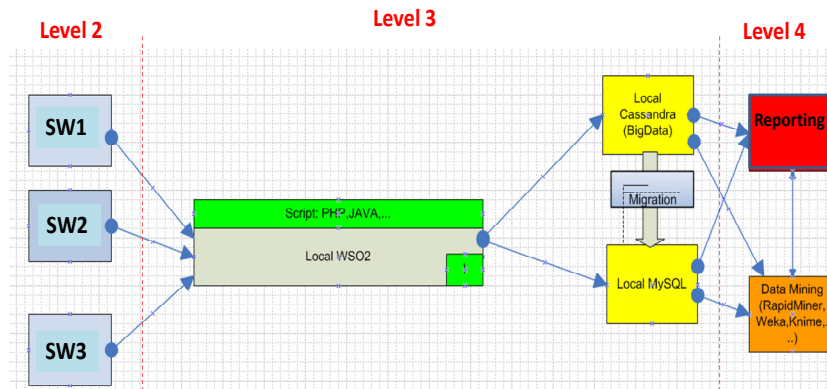


Figure 4. Architecture of data flow between level 2, level 3 and level 4 of Fig. 2.

3.1 BIG DATA TESTING ARCHITECTURE

The first big data testing architecture is illustrated in Fig.5 and is described by the following open source facilities:

- Operative System configuring and testing Cassandra node: Ubuntu 14.04 LTS3 version, 64 bit [30];
- Cassandra [31] node (minimum RAM/CPU: 4 core , 8 Gbyte di RAM -INTEL XEON E 5-; Single node in Hadoop Platform/environment) ;
- Performance control: Datastax OpsCenter Enterprise [32];
- Windows platform: Datastax DevCenter 1.5.0 (platform useful to create table layouts and to manipulate by a GUI Cassandra records) [33];
- KNOWI web based platform [34]: plotting Cassandra table values.

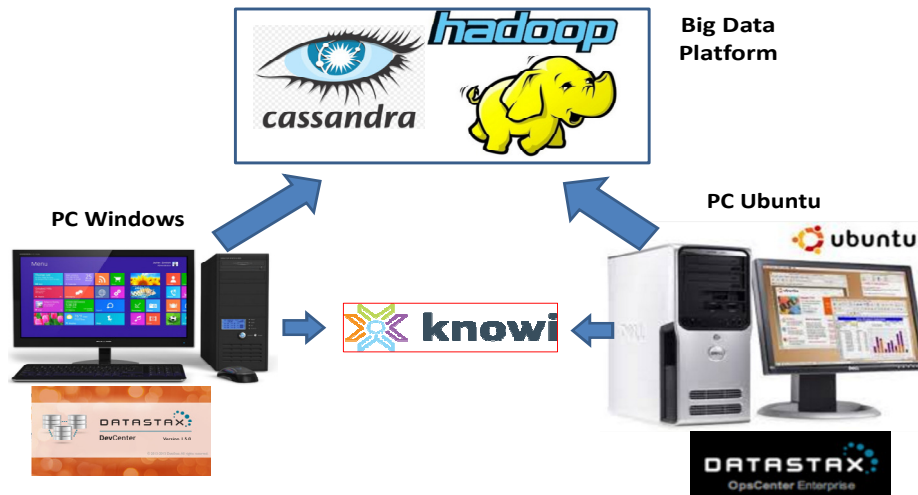


Figure 5. Big data testing architecture.

In Fig. 6 is illustrated the architecture related to the implementation of graphical dashboards for the plotting of Cassandra table values: the Eclipse platform is adopted to develop java code integrating the open source JFreeChart libraries [35] (free Java chart library suitable to display charts) and enabling the access to Cassandra data.

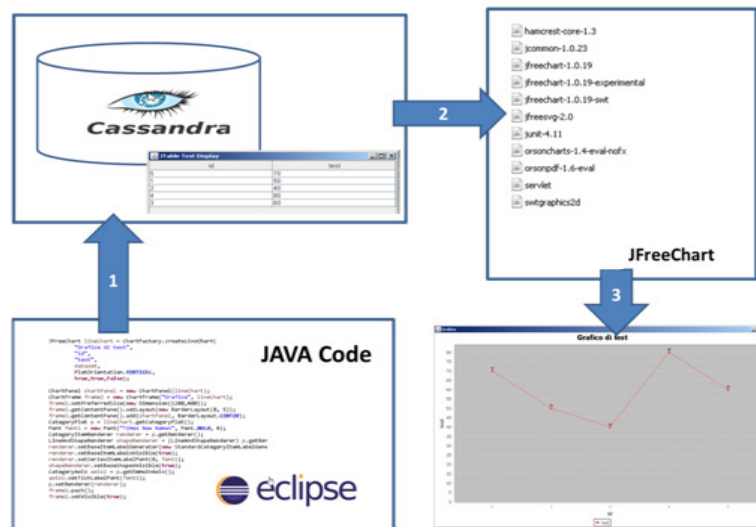


Figure 6. Testing architecture related to the dashboards and output plotting.

Below is commented a first check concerning plotting Cassandra data by the JFreeChart tool, by listing the executed steps:

1. Has been created the keyspace and the records by the following commands:

```
CREATE KEYSPACE test_aci
WITH replication = {'class':'SimpleStrategy', 'replication_factor': 3};

CREATE TABLE table_test (id int PRIMARY KEY, test int);
INSERT INTO table_test(id,test) VALUES (1,50);
INSERT INTO table_test(id,test) VALUES (2,40);
INSERT INTO table_test(id,test) VALUES (3,60);
INSERT INTO table_test(id,test) VALUES (4,80);
INSERT INTO table_test(id,test) VALUES (5,70);
```

2. Has been performed the Cassandra connection by the following java script:

```
CassandraConn.java Main.java
1 package connection;
2
3 import com.datastax.driver.core.Cluster;
4
5 public class CassandraConn
6 {
7     private Cluster cluster;
8     private Session session;
9
10    public void connect(final String node, final int port)
11    {
12        this.cluster = Cluster.builder().addContactPoint(node).withCredentials("cassandra", "cassandra").withPort(port).build();
13        final Metadata metadata = cluster.getMetadata();
14        out.printf("Connected to cluster: %s\n", metadata.getClusterName());
15        for (final Host host : metadata.getAllHosts())
16        {
17            out.printf("Datacenter: %s; Host: %s; Rack: %s\n",
18                host.getDatacenter(), host.getAddress(), host.getRack());
19        }
20        session = cluster.connect("test_aci");
21    }
22
23    public Session getSession()
24    {
25        return this.session;
26    }
27
28    public void close()
29    {
30        cluster.close();
31    }
32 }
33
34
35
36
37
38
39
```

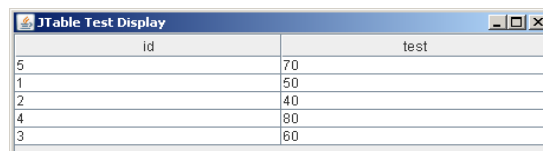
3. Has been checked the created basic testing table by the following java script:

```

73 //Tabella
74 JFrame frame = new JFrame("JTable Test Display");
75 JPanel panel = new JPanel();
76 panel.setLayout(new BorderLayout());
77 JTable tbl = new JTable();
78 DefaultTableModel dtm = new DefaultTableModel(0, 0);
79 String header[] = new String[] { "id", "test" };
80 dtm.setColumnIdentifiers(header);
81 for(int i=0; i<id.size(); i++) {
82     String[] stringArraySez = { id.get(i).toString(), test.get(i).toString() };
83     dtm.addRow(stringArraySez);
84 }
85 tbl.setModel(dtm);
86 JScrollPane tableContainer = new JScrollPane(tbl);
87 panel.add(tableContainer, BorderLayout.CENTER);
88 frame.getContentPane().add(panel);
89 frame.pack();
90 frame.setVisible(true);
91

```

thus providing the following output (see Fig. 7)



id	test
5	70
1	50
2	40
4	80
3	60

Figure 7. Check of Cassandra table used for the preliminary test.

4. Has been plotted the data of the testing table by the following java script:

```

97
98 JFreeChart lineChart = ChartFactory.createLineChart(
99     "Grafico di test",
100     "id",
101     "test",
102     dataset,
103     PlotOrientation.VERTICAL,
104     true,true,false);
105
106 ChartPanel chartPanel = new ChartPanel(lineChart);
107 ChartFrame frame2 = new ChartFrame("Grafico", lineChart);
108 frame2.setPreferredSize(new Dimension(1200,800));
109 frame2.getContentPane().setLayout(new BorderLayout(0, 5));
110 frame2.getContentPane().add(chartPanel, BorderLayout.CENTER);
111 CategoryPlot p = lineChart.getCategoryPlot();
112 Font font1 = new Font("Times New Roman", Font.BOLD, 8);
113 CategoryItemRenderer renderer = p.getRenderer();
114 LineAndShapeRenderer shapeRenderer = (LineAndShapeRenderer) p.getRenderer();
115 renderer.setBaseItemLabelGenerator(new StandardCategoryItemLabelGenerator());
116 renderer.setBaseItemLabelsVisible(true);
117 renderer.setSeriesItemLabelFont(0, font1);
118 shapeRenderer.setBaseShapesVisible(true);
119 CategoryAxis axis2 = p.getDomainAxis();
120 axis2.setTickLabelFont(font1);
121 p.setRenderer(renderer);
122 frame2.pack();
123 frame2.setVisible(true);
124

```

thus providing the dashboard of Fig. 8 (the dashboards are adopted for the visualization of BI outputs and for reporting).

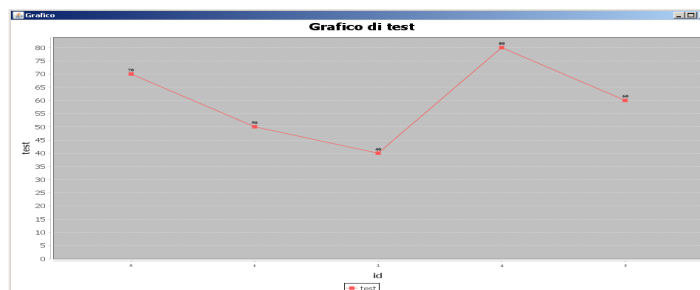


Figure 8. Test of graphical plot of testing Cassandra table.

Other test has been performed by opening a Virtual Private Network –VPN- channel allowing the connection with industry server (connection by means of the PuTTY tool [36]).

4. CASSANDRA BIG DATA PERFORMANCE ESTIMATION

In order to evaluate the Cassandra performance about the estimation of the insertion time of records (representing a critical parametr), have been created four testing tables having the following characteristics:

- table 1: 5 column with $1 \cdot 10^6$ records of randomic integer values;
- table 2: 10 column with $1 \cdot 10^6$ records of randomic integer values;
- table 3: 15 column with $1 \cdot 10^6$ records of randomic integer values;
- table 4: 30 column with $1 \cdot 10^6$ records of randomic integer values;

Observing results discussed in [21], the performance concerning the estimation of time of writing operation is more indicative, because the writing operation requires more time if compared to the reading or to cancellation times. In Fig. 9 is shown the performance time for the four testing table: a bigger slope angle is observed between 5 and 10 columns (the insertion time increases quickly), a plateau is checked between 10 and 15 columns (the insertion time does not increase in this range), and a lower slope angle is noted between 15 and 30 columns (the insertion time increase slowly for column number major than 15). The results enhance that increasing the column number the system will keep a good performance.

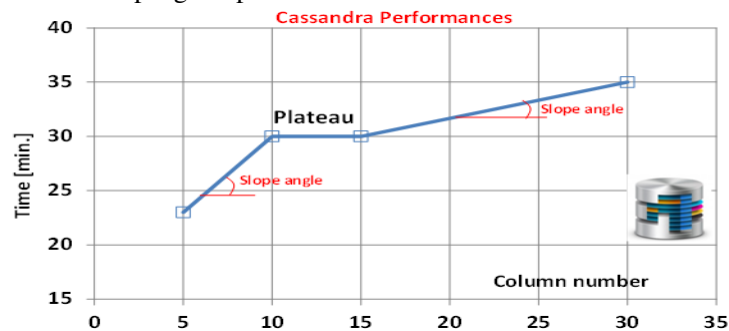


Figure 9. Insertion time in minutes versus column number (each column having 1.000.000 of records).

Below is listed as example the script used for the estimation of the time related to the writing of 1.000.000 of records:

-Creation of a table layout with 1.000.000 of records

```
CREATE TABLE table15_test_client1(
id int,
id_cliente int,
colonna1 varint,
colonna2 varint,
colonna3 varint,
colonna4 varint,
colonna5 varint,
colonna6 varint,
colonna7 varint,
colonna8 varint,
colonna9 varint,
colonna10 varint,
colonna11 varint,
```

colonna12 varint,
colonna13 varint,
colonna14 varint,
colonna15 varint,

PRIMARY KEY (id,id_cliente))WITH CLUSTERING ORDER BY (id_cliente ASC);

- Script for the random generation of data with 1 million rows on 15 columns:

```
#!/bin/bash
x=1
while [ $x -le 1000000 ]
do
numbers(){
    port=$((RANDOM%1000000))
    echo "$port"
}
p=$(numbers)
p1=$(numbers)
p2=$(numbers)
p3=$(numbers)
p4=$(numbers)
p5=$(numbers)
p6=$(numbers)
p7=$(numbers)
p8=$(numbers)
p9=$(numbers)
p10=$(numbers)
p11=$(numbers)
p12=$(numbers)
p13=$(numbers)
p14=$(numbers)
echo "INSERT INTO clienti.table15_test_client1
(id, id_cliente, colonna1, colonna2, colonna3, colonna4, colonna5, colonna6, colonna7,
colonna8, colonna9, colonna10, colonna11, colonna12, colonna13, colonna14, colonna15)
VALUES ($x,$x,$p,$p1,$p2,$p3,$p4,$p5,$p6,$p7,$p8,$p9,$p10,$p11,$p12,$p13,$p14);" >>
data15colonne.cql
x=$(( $x + 1 ))
done
```

-Script concerning the insertion of 1.000.000 of randomic data into the 15 column table:

```
cqlsh -u cassandra -p cassandra < data15colonne.cql
```

In Fig. 10 is illustrated a screenshot proving the correct insertion (by the following script: *SELECT * FROM clienti.table15_test_client1 WHERE id IN (102,104,300,2456,999999,1000000) ORDER BY id_cliente ASC;*)).

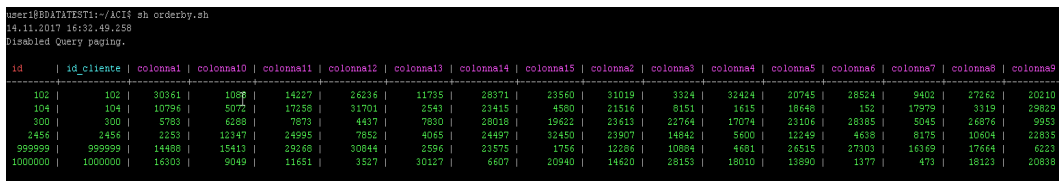


Figure 10. Screenshot proving the correct creation of a table having 15 columns each one with 1.000.000 of records.

5. DATA MINING ALGORITHMS ORIENTED TO BI

In this section are provided two main applications of data mining algorithms such as Correlation Matrix and Random Forest Decision Tree, adopted for the research project. These algorithms are performed by Rapid Miner tool [37] and are suitable for the BI integrated with the industry information system.

5.1 CORRELATION MATRIX ALGORITHM APPLIED ON CALL CENTER DATASET

Correlation Matrix is an algorithm determining determines correlation between all attributes of a table of the big data system generating a weights vector based on these correlations (weights of correlations). The attributes having higher weight are considered more relevant for the data processing (the operator calculates the weight of attributes with respect to the label attribute by constructing a single rule for each attribute and calculating the errors). The output of the algorithms are numbers estimated in a range between -1 and +1. The algorithm estimates the Pearson's correlation coefficient which is the covariance of the two variables divided by the product of their standard deviations [38]-[39]:

$$\rho_{x,y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

The correlation algorithm has been applied on an experimental dataset (simulation dataset) referred to call center operations. This dataset is stored into the big data table containing the following main attributes:

- Login: user login name;
- ID chiamata: identification –ID- number of the user call;
- Tempo ACD: time for the Automatic Call Distribution –ACD- process (the primary goal of the ACD system is to disperse automatically incoming calls to contact center operators with specific skills);
- Tempo Conv.: time of the telephone conversation;
- Tempo Coda: time of phone calls placed in the waiting queue;
- Tempo Hold: average time taken for an operator to answer a call or the time a customer waits in the queue before being answered;
- Anno: reference year;
- Mese: reference month;
- Giorno: reference day;
- Ora: reference hour;
- Durata: total duration time;
- Minuti: minutes;
- Sede: industry location.

In Fig. 11 and Fig. 12 are plotted some data of the experimental dataset (Scatter Multiple plot and Series plot).

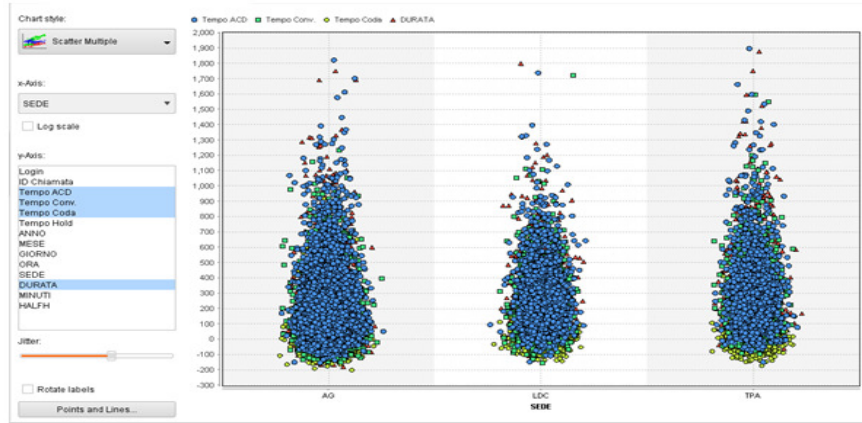


Figure 11. Rapid Miner dashboard: Scatter Multiple plot of the first experimental dataset (dashboard analysis).



Figure 12. Rapid Miner dashboard: Series Plot of the first experimental dataset (dashboard analysis).

In Fig. 13 is illustrated the Rapid Miner workflow enabling the Correlation Matrix algorithm by processing data of the Cassandra testing database. By properly set the rule of the analysed attributes has been estimated the output matrix of Fig. 14. This simulation output highlights a strong correlation between ACD time and call duration thus suggesting to revise the automatic processing enabling operator phone linking.

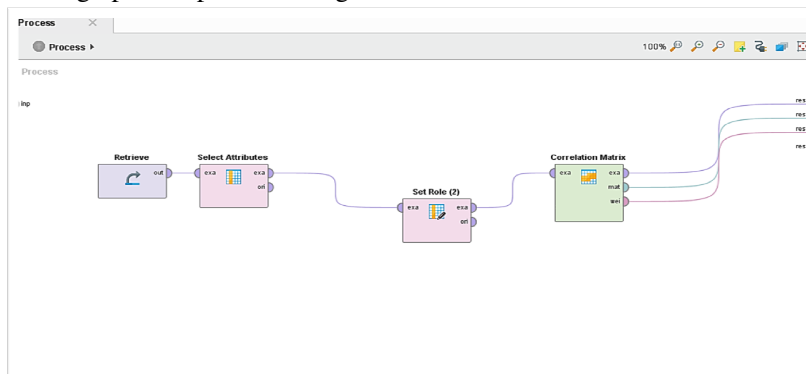


Figure 13. Rapid Miner: workflow of the Correlation Matrix algorithm.

Attributes	Login	Tempo ACD	Tempo Conv.	Tempo Coda	Tempo Hold	MESE	GIORNO	ORA	DURATA
Login	1	0.000	0.001	0.004	0.000	0.236	0.225	0.063	0.001
Tempo ACD	0.000	1	0.799	0.127	0.256	0.002	0.005	0.000	0.917
Tempo Conv.	0.001	0.799	1	0.005	0.036	0.001	0.011	0.000	0.871
Tempo Coda	0.004	0.127	0.005	1	0.001	0.036	0.006	0.000	0.005
Tempo Hold	0.000	0.256	0.036	0.001	1	0.001	0.000	0.002	0.279
MESE	0.236	0.002	0.001	0.036	0.001	1	0.304	0.006	0.000
GIORNO	0.225	0.005	0.011	0.006	0.000	0.304	1	0.017	0.009
ORA	0.063	0.000	0.000	0.000	0.002	0.006	0.017	1	0.000
DURATA	0.001	0.917	0.871	0.005	0.279	0.000	0.009	0.000	1

Figure 14. Correlation Matrix simulation output.

The proposed tool is suitable also to formulate Key Performance Indicators –KPI- of the call center operators and to optimize the human resources allocation.

5.2 DECISION TREE ALGORITHM

The same attributes of the experimental dataset has been processed by Decision Tree algorithm [40],[41] used as a data mining classifier. The decision tree is a tree structure collecting attributes as nodes, where each node represents a splitting rule for one specific attribute. The tree separates values belonging to different classes. A workflow implementing this algorithm, and connecting to the big data system is illustrated in Fig. 15 (Rapid Miner workflow).

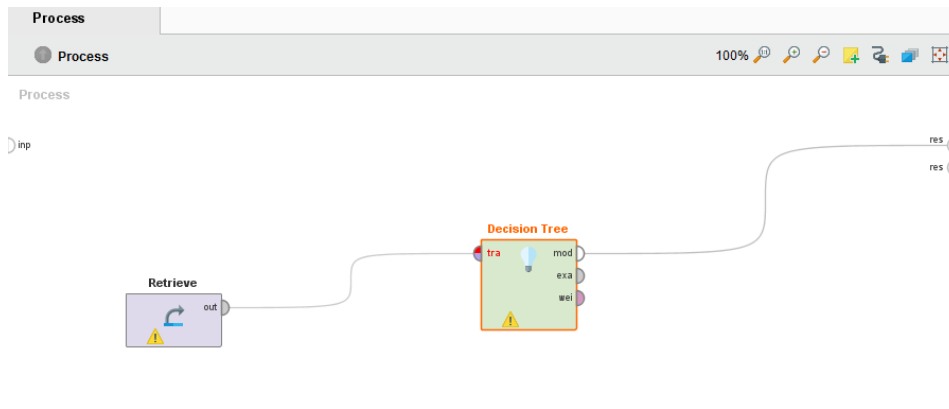


Figure 15. RapidMiner: workflow implementing Decision tree algorithm.

An example of Decision Tree output is illustrated in Fig. 16 where are classified the user operators for location and for call duration. Also in this case the data mining algorithm is useful for KPI evaluation and for the human resources management.

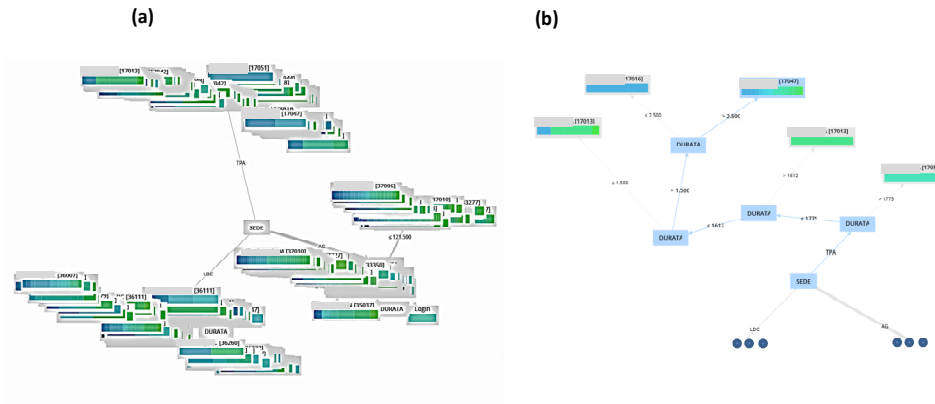


Figure 16. Decision Tree Model outputs.

The adopted Decision Tree calculation parameters are: gain ratio as criterion, 10 as maximal tree depth, 0.1 as confidence, 0.01 as minimal gain, 2 as minimal leaf size, 4 as minimal size of split, and 3 as number of prepruning alternatives. In table 1 are listed some BI functionalities which can be activated by the correlation and by the decision tree classification applied on the first experimental dataset.

Table 1. BI potential functionalities due to the Correlation Matrix and Decision Tree algorithms (first experimental dataset).

Function	BI Facilities
Graphical plot of the experimental dataset (dashboard analysis)	Overview of the general trend of the call centers of different industry location: this analysis allows to understand the efficiency of the whole human resources of each industry location, potentially allowing the balancing of performances of call centers.
Data processing by Correlation Matrix algorithm	The correlation between the attributes: - facilitates to understand the causes of possible inefficiencies by finding quickly solutions; - supports the definition of KPI and scorecards mainly related of human resources pertaining of a call center location; - provides correlations between the period the call service actuation and inefficiencies due to the delay of the service; - allows to manage the scheduling activities of the operators.
Data processing by Decision Tree algorithm	The Decision Tree algorithms: - Facilitates the operator classification supporting the formulation of his skill scoring based on an efficiency indicator (for example based on the output of the call and on the response time); - Classifies the industry location efficiency; - Classifies the typology of service thus facilitating the future operator assignation (each operator can be associated to particular call service thus optimizing the ACD system).
Data set of a large dataset (big data dataset)	The historic data stored into Cassandra database system (massive training dataset) will optimize the evaluation of the skills and of the efficiency indicators thus reducing the evaluation error.

5.3 RANDOM FOREST DECISION TREE: BILLING CLASSIFICATION

The Random Forest algorithm [40],[42] has been applied on a second experimental dataset in order to provide knowledge gain and to enable BI outputs. This algorithm has been implemented by a proper Rapid Miner operator ('Random Forest' operator) generating a random forest model, suitable for classification and regression. The random forest is based on the concept to consider for the data processing an ensemble of a definite number of random trees, which are created and trained on bootstrapped sub-sets of an input dataset. Each node of a tree defines a splitting rule of a specific attribute. The data processing mechanism is able to separate values following a defined criterion. The building of new nodes is repeated until the stopping criteria are satisfied. The pruning function allows to reduce the model complexity by replacing sub-trees and actuating the pruning. The algorithm is randomized in two levels, working with the following steps [42] (n training records with m attributes, and let k be the number of trees) applied for each tree:

1. An n random sample is selected with replacement.
2. A random number of attributes to be considered for node splitting ($D \ll m$).
3. A decision tree is started; for each node of the processed tree, the number D of attributes are considered for the split (this step is repeated for each node).
4. As in any ensemble data process, the greater the diversity of the base trees, lower will be the error of the ensemble.

If compared with Decision Tree algorithm, Random Forest achieves increased classification performance and yields results that are accurate and precise in the cases of large number of instances (large number of rows as for the second simulation case analysed in this paper). This algorithm also overcomes the missing values problem typical for large datasets (as for big data systems), providing a good in accuracy and overcoming the over-fitting problem [43]. The dataset of the second experimental dataset simulating the BI output is constituted by the following attributes:

LINEA: typology of vehicle (truck, car, etc.);
 MERCATO: reference market (automotive industry, other industry sectors and customer typologies);
 ACCOUNT: account information;
 CLIENTE: customer information (automotive industries);
 CONTRATTO: contract associated to a billing typology (the contracts are classified by means of a calculus engine implementing association rules);
 AZIENDA: identification number of the customer industry;
 PRODOTTO: product/service identification number;
 DESC. PROD.: description of the offered product/service;
 FATTURAZIONE: billing classification (FxV is billing for vehicle, FxP is billing for services, FxD is billing for report/dossier);
 PRESTAZIONE: identification number of the performed service;
 DESC.PRE.: description of the performed service.

In Fig. 17 is illustrated the Rapid Miner workflow implementing the Random Forest model applied to the second experimental dataset. In the inset of the same figure it is also represented the sketch indicating the processing logic of the algorithm. As examples, in Fig. 18 are illustrated some representative outputs of the algorithm simulating the billing classification. This algorithm could provide BI facilities which are summarized in table 2:

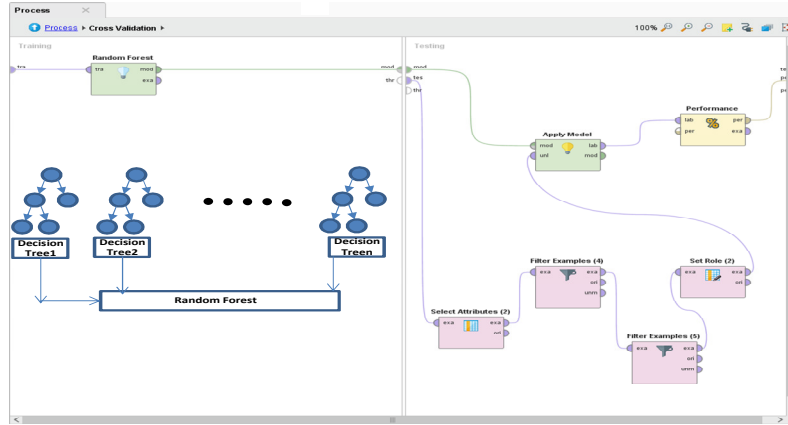


Figure 17. Random Forest Tree Model.

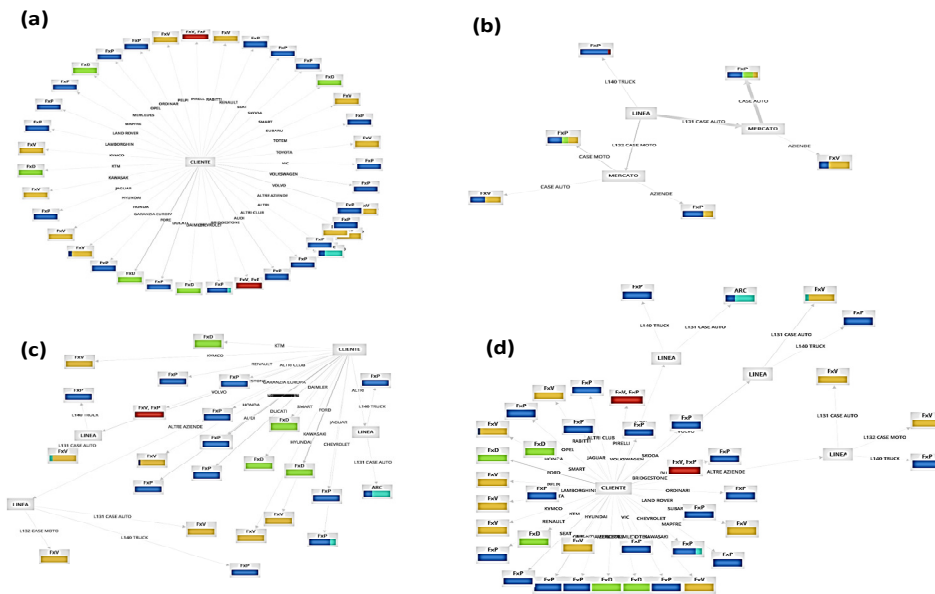


Figure 18. Random Forest Tree Model outputs.

Table 3. BI potential functionalities due to the Random Forest model (second experimental dataset).

Function	BI Facilities
Graphical plot of the bill classification	Overview of the billing classification combined with customer information (BI analytics).
Customer classification	Customer segmentation.
BI prediction	Prediction of typologies of contracts which will be applied in the future.
Contract optimization	Optimization of service contracts and service offers.
Economic and financial impact prediction	Estimation of the future impact for typologies of services and for industry customers.
Service optimization	Balancing of the offered services

The adopted Random Forest calculation parameter are: gain ratio as criterion, 100 as number of trees, and 10 as maximal depth. According with the accuracy matrix of Fig. 19, a good performance is checked during the output simulation thus confirming the correct choice of the data mining algorithm for the analyzed simulation dataset (accuracy of 97.75%).

accuracy: 97.75% +/- 0.49% (mikro: 97.75%)

	true FxP	true ARC	true OR	true FxD	true FxV	true FxV, FxP	class precision
pred. FxP	1870	22	0	0	0	0	98.84%
pred. ARC	8	19	0	0	0	0	70.37%
pred. OR	0	0	0	0	0	0	0.00%
pred. FxD	0	0	0	975	0	0	100.00%
pred. FxV	11	40	0	0	631	0	92.52%
pred. FxV, FxP	0	0	0	0	0	22	100.00%
class recall	98.99%	23.46%	0.00%	100.00%	100.00%	100.00%	

Figure 19. Matrix accuracy.

6. CONCLUSION

The paper propose a methodological approach for the construction of an industry information system gaining the knowledge base by following R&D ‘Frascati’ guideline. The upgrade of the information systems has been designed and executed by implementing different knowledge levels. The upgraded information system can be adopted to compare data of different industry locations, balancing the resources and improving a weighted business intelligence by processing operator and customer data. The proposed architectures proves how it is possible to implement a Cassandra big data system interconnected with data mining algorithm. In order to validate the choice of the Cassandra system has been estimated the big data performance showing how increasing the number of attributes, the system the system shows good performance. The data mining outputs of the Correlation Matrix, and of Decision Tree and Random Forest algorithms proved the possibility to manage services and resources by applying BI strategies with a good solution accuracy. The paper proposes some data mining algorithms suitable for BI outputs which can be improved by other algorithms implementing marketing prediction such as artificial neural networks.

ACKNOWLEDGEMENTS

The work has been developed in the frameworks of the Italian projects: “Modello Big Data integrato di business intelligence dinamico basato sull’ottimizzazione di algoritmi di data mining operanti su due piattaforme ACI Global distinte: ‘ACI B.I. Balancing’” [Integrated Big Data model of dynamic business intelligence based on the optimization of data mining algorithms operating on two distinct ACI Global platforms ‘ACI B.I. Balancing’]. The authors would like to thank the following researchers and collaborators: D. Barbuzzi, B. Boussahel, V. Calati, A. Colonna, R. Cosmo, V. Custodero, G. Fanelli, R. Guglielmi, M. Legrottaglie, A. Leogrande, A. Lombardi, G. Lonigro, A. Lorusso, S. Maggio, N. Malfettone, V. Maritati, S. F. Massari, L. Pellicani, R. Porfido, D. D. Romagno, G. Sicolo, M. Solazzo, M. M. Sorbo, D. Suma, and L. Patruno.

REFERENCES

- [1] Khan, R. A. & Quadri, S. M. K. (2012) "Business Intelligence: an Integrated Approach", Business Intelligence Journal, Vol.5 No.1, pp 64-70.
- [2] Chen, H., Chiang, R. H. L. & Storey V. C. (2012) "Business Intelligence and Analytics: from Big Data to Big Impact", MIS Quarterly, Vol. 36, No. 4, pp 1165-1188.
- [3] Andronie, M. (2015) "Airline Applications of Business Intelligence Systems", Incas Bulletin, Vol. 7, No. 3, pp 153 – 160.
- [4] Iankoulova, I. (2012) "Business Intelligence for Horizontal Cooperation", Master Thesis, Univesitiet Twente. [Online]. Available: https://www.utwente.nl/en/mbit/final-project/example_excellent_master_thesi/master_thesis_bit/IankoulovaID.pdf
- [5] Nunes, A. A., Galvão, T. & Cunha, J. F. (2014) "Urban Public Transport Service Co-creation: Leveraging Passenger's Knowledge to Enhance Travel Experience", Procedia Social and Behavioral Sciences, Vol. 111, pp 577 – 585.
- [6] Fitriana, R., Eriyatno, Djatna, T. (2011) "Progress in Business Intelligence System research: A literature Review", International Journal of Basic & Applied Sciences IJBAS-IJENS, Vol. 11, No. 03, pp 96-105.
- [7] Lia, M. (2015) "Customer Data Analysis Model using Business Intelligence Tools in Telecommunication Companies", Database Systems Journal, Vol. 6, No. 2, pp 39-43.
- [8] Habul, A., Pilav-Velić, A. & Kremić, E. (2012) "Customer Relationship Management and Business Intelligence", Intech book 2012: Advances in Customer Relationship Management, chapter 2.
- [9] Kemper,H.-G., Baars, H. & Lasi, H. (2013) "An Integrated Business Intelligence Framework Closing the Gap Between IT Support for Management and for Production", Springer: Business Intelligence and Performance Management Part of the series Advanced Information and Knowledge Processing, pp 13-26, Chapter 2.
- [10] Bara, A., Botha, I., Diaconița, V., Lungu, I., Velicanu, A., Velicanu, M. (2009) "A Model for Business Intelligence Systems' Development", Informatica Economică, Vol. 13, No. 4, pp 99-108.
- [11] Negash, S. (2004) "Business Intelligence", Communications of the Association for Information Systems, Vol. 13, pp 177-195.
- [12] Nofal, M. I. & Yusof, Z. M. (2013) "Integration of Business Intelligence and Enterprise Resource Planning within Organizations", Procedia Technology, Vol. 11 (2013), pp. 658 – 665.
- [13] Williams, S. & Williams, N. (2003) "The Business Value of Business Intelligence", Business Intelligence Journal, FALL 2003, pp 1-11.
- [14] Lečić, D. & Kupusinac, A. (2013) "The Impact of ERP Systems on Business Decision-Making", TEM Journal, Vol. 2, No. 4, pp 323-326.
- [15] Ong, L., Siew, P. H. & Wong, S. F. (2011) "A Five-Layered Business Intelligence Architecture", IBIMA Publishing, Communications of the IBIMA, Vol. 2011, Article ID 695619, pp 1-11.
- [16] Raymond T. Ng, Arocena, P. C., Barbosa, D., Carenini, G., Gomes, L., Jou, S., Leung, R. A., Milios, E., Miller, R. J., Mylopoulos, J., Pottinger, R. A., Tompa, F. & Yu, E. (2013) "Perspectives on Business Intelligence", A Publication in the Morgan & Claypool Publishers series Synthesis Lectures on Data Management.
- [17] "NTT DATA Connected Car Report: A brief insight on the connected car market, showing possibilities and challenges for third-party service providers by means of an application case study" [Online]. Available: https://emea.nttdata.com/fileadmin/web_data/country/de/documents/Manufacturing/Studien/2015_Connected_Car_Report_NTT_DATA_ENG.pdf
- [18] "Cognizant report: Exploring the Connected Car Cognizant 20-20" [Online]. Available: <https://www.cognizant.com/InsightsWhitepapers/Exploring-the-Connected-Car.pdf>

- [19] Sarangi, P. K., Bano, S., Pant, M. (2014) "Future Trend in Indian Automobile Industry: A Statistical Approach", *Journal of Management Sciences And Technology*, Vol. 2, No. 1, pp. 28-32.
- [20] Bates, H. & Holweg, M. (2007) "Motor Vehicle Recalls: Trends, Patterns and Emerging Issues", *Omega*, Vol. 35, No. 2, pp 202–210.
- [21] D'Aloia, M., Russo, M. R., Cice G., Montingelli, A., Frulli, G., Frulli, E., Mancini, F., Rizzi, M., Longo, A. (2017) "Big Data Performance and Comparison with Different DB Systems", *International Journal of Computer Science and Information Technologies*, Vol. 8, No. 1, pp 59-63.
- [22] Wimmer, H. & Powell, L. M. (2015) "A Comparison of Open Source Tools for Data Science", *Proceedings of the Conference on Information Systems Applied Research*, Wilmington, North Carolina USA.
- [23] Al-Khoder, A. & Harmouch, H. (2014) "Evaluating four of the most popular Open Source and Free Data Mining Tools," *IJASR International Journal of Academic Scientific Research*, Vol. 3, No. 1, pp 13-23.
- [24] Antonio Gulli, Sujit Pal, "Deep Learning with Keras- Implement neural networks with Keras on Theano and TensorFlow," BIRMINGHAM – MUMBAI Packt book, ISBN 978-1-78712-842-2, 2017.
- [25] Kovalev V., Kalinovsky A., and Kovalev S. Deep Learning with Theano, Torch, Caffe, TensorFlow, and deeplearning4j: which one is the best in speed and accuracy? In: XIII Int. Conf. on Pattern Recognition and Information Processing, 3-5 October, Minsk, Belarus State University, 2016, pp. 99-103.
- [26] 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).
- [27] 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, pp 1-17.
- [28] Massaro, A., Calicchio, A., Maritati, V., Galiano, A., Birardi, V., Pellicani, L., Gutierrez Millan, M., Dalla Tezza, B., Bianchi, M., Vertua, G., Puggioni, A. (2018) "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", *International Journal of Artificial Intelligence and Applications (IJAIA)*, Vol. 9, No.5, pp. 27-43.
- [29] "WSO2" [Online]. Available: <https://wso2.com/products/enterprise-service-bus/>
- [30] "Ubuntu" [Online]. Available: <https://www.ubuntu.com/>
- [31] "Apache Cassandra" [Online]. Available: <http://cassandra.apache.org/>
- [32] "DataStax Enterprise OpsCenter" [Online]. Available: <https://www.datastax.com/products/datastax-opscenter>
- [33] "About DataStax DevCenter" [Online]. Available: <https://docs.datastax.com/en/developer/devcenter/doc/devcenter/dcAbout.html>
- [34] "Knowi" [Online]. Available: <https://www.knowi.com/>
- [35] "JFreeChart" [Online]. Available: <http://www.jfree.org/jfreechart/samples.html>
- [36] "PuTTY" [Online]. Available: <https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html>
- [37] "Lightning Fast Data Science for Teams" [Online]. Available: <https://rapidminer.com/>
- [38] Massaro, A., Meuli, G. & Galiano, A. (2018) "Intelligent Electrical Multi Outlets Controlled and Activated by a Data Mining Engine Oriented to Building Electrical Management", *International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI)*, Vol.7, No.4, pp 1-20.
- [39] Myers, J. L. & Well, A. D. (2003) "Research Design and Statistical Analysis", (2nd ed.) Lawrence Erlbaum.

- [40] Kotu, V., Deshpande, B. (2015) “Predictive Analytics and Data Mining”, Elsevier book, Steven Elliot editor.
- [41] Quinlan, J. (1986) “Induction of Decision Trees”, Machine Learning, pp 81–106.
- [42] Breiman, L. (2001) “Random Forests”, Machine Learning, Vol. 45, pp 5–32.
- [43] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood (2012) “Random Forests and Decision Trees”, International Journal of Computer Science Issues, Vol. 9, No. 3, pp 272-278.

AUTHOR

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

