THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

# DATABASE SYSTEMS JOURNAL

Vol. XI/2020

**LISTED IN**

RePEc, EBSCO, DOAJ, Open J-Gate,
Cabell's Directories of Publishing Opportunities,
Index Copernicus, Google Scholar,
Directory of Science, Cite Factor,
Electronic Journals Library

BIG DATA
NoSQL
DATA SCIENCE
BUSINESS INTELLIGENCE
CLOUD COMPUTING
DATA MINING
DATA WAREHOUSES
DATABASES

**ISSN: 2069 – 3230**
**dbjournal.ro**

# Database Systems Journal BOARD



**Contact**
CaleaDorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: 1
E-mail: editordbjournal@gmail.com

# CONTENTS

# Operational Research in the Emergency Medical System of Romania

Ionuț NICA
The Bucharest University of Economic Studies, Romania
ionut.nica@csie.ase.ro

*The explosive development of the human society in contrast to the limited character of resources determines the need for successful implementation of mathematic models in the decision-making process concerning the use of available resources. One of the critical areas where the need for rigorous criteria for resource allocation is strongly felt is the medical field. This issue appears to be currently affecting the great majority of nations in the world, being considered one of the most important challenges for modern states. The limited amount of resources allocated to the medical system brings forward the importance of optimizing the decision-making process concerning this field using models able to reflect the increasing complexity of the medical system, its interactions with the human society and its dynamics, therefore providing the perturbation control and adjustment instruments. From this point of view, the economical and mathematical modeling of the social phenomena provides strong, elegant and rigorous tools for the description of medical system that appears to be organized as a cybernetic system with a high level of complexity, focused on maximizing the social utility, and allowing the use of cybernetic methods designed for diagnosing, developing automatic medical archives, reducing time consumption and increasing overall efficiency.*
***Keywords:*** Markov network, Poisson distribution, cybernetics models, optimal decision, emergency medical system

## 1 Introduction

The resource allocation issue, especially for key sectors such as the medical one, seems to become more and more of a problem for the great majority of nations in the world. The limited character of resources, those allocated to the medical sector included, points out to the necessity of optimizing the decision-making processes involved in resource allocation based on rigorous methods capable of avoiding unnecessary expense and maximizing social utility.

The high complexity of the medical system and its multiple connections to human society determine the need for careful observation of its functioning and malfunctioning in order to produce the best suitable tools for perturbation control and adjustment.

The use of cybernetic and mathematical models for managing the functionalities of the medical system may lead at first to a mismatch between human expectations and the results achieved by following the path indicated through modeling. However, even these differences may provide useful information for further improvement, so that the medical system may eventually achieve the optimal resource allocation and maximize social utility.

In the pursuit of these goals (optimal utility/resource consumption ratio), cybernetic methods prove their utility for: designing diagnose systems for different clusters of diseases, developing automatic electronic medical archives (capable of minimizing searching times), as well as increasing overall efficiency by limiting waste, while keeping up with the evolution of biocybernetics.

## 2. The Romanian Medical System in the European Context

Due to Romania joining the European Union, the Romanian medical system started to use the similar structures in the EU as a reference point, which led to the need to increase efficiency by using mathematical methods applied in other European

countries. This would be the only way for the Romanian medical system to function as required by the new parameters implied by the efficiency standard implemented by the member states. According to data gathered in order to prove the need for implementing these changes, reforming the medical system using mathematical methods is more than critical, considering that the malfunctioning of the medical systems causes more than 60 000 deaths per year, which is the equivalent population of a small city.

Another challenge for the Romanian medical system is the shortage of medical staff (doctors, dentists, nurses, pharmacists), as compared to the other countries in the EU. Even though in Romania the financial efforts for sustaining almost all types of medical services have recently increased, there is still a general feel of system failure. By comparing the Romanian medical system to those of other European countries, and even by comparing the medical services provided in different regions in Romania, one can easily notice the massive differences in the access to medical services, as well as the gap between the values of most of the medical indicators, all picturing a worrying situation for the health of Romanians.

Romania has the highest mortality rate in the EU for both men and women, and this can be related to the difficult access to medical services of the general population, as well as to the fact that we have the smallest number of doctors, nurses and pharmacists reported to the size of the population. Moreover, in rural areas, where more than half of the population is located, there is even less medical personnel, and almost no functional hospital whatsoever.

According to a study by Ajay Tandon, Christopher JL Murray, Jeremy A. Lauer and David B. Evans in 2000 [0], Romania ranked 99 out of 191 countries in terms of medical system global performance. However, the percentage of the GDP allocated to the health system is not definitory for its efficiency, considering that the USA ranks 37, even if they have the highest percentage of the GDP allocated to health. Still, compared to the European average, Romania allocates to the health system only a third of the amount spent on health by other countries. For instance, in 2010, Romania spent 600 euros per capita, as compared to the 1800 euros per capita which is the European average, and the government only directs 4% of the GDP towards the health system, while in France the percentage is 11% and the European average is around 8%. The difference can be explained by considering the low number of tax payers (only approx. 30% of the total population) and corruption, and it is a direct indicator of the struggle Romania has to put up in order to improve the efficiency of its medical system.

Since 2007, international mobility became even more accessible for Romanians, especially medical staff: almost 10% of the doctors decided to emigrate to countries such as France, Germany or Sweden, allured by the latest medical technologies available there and the high wages, and this percentage is still growing.

Even when it comes to medical equipment, Romania is one of the lowest ranked countries in the EU regarding the use of modern medical technology, being severely underequipped, which brings up even more the necessity to optimize the use of the existing resources, considering the low capacity for budgetary investment in medical technology.

However, almost every nation in the world has yet some challenges to deal with when it comes to the health system: no country has enough resources, money or medical personnel to cover all medical needs. More and more people have to live in fear of getting ill and not being able to access medical care. Therefore, there is a real necessity to improve efficiency of the medical system in order to optimize the use of the existing resources and to meet as

much as possible the demand for medical services.

Beyond the theory of providing public health services and the specific legal frame, a mathematical approach of the issue is also recommended, since the medical system is a good example of a cybernetic system [1], [15], [17], which includes not only complex components, but also dynamic and sensitive interactions that need careful planning.

## The Legal Frame For The Functioning Of Emergency Units

***Law 95/2006*** states that qualified first aid should be provided within:

a) 8 minutes for urban areas, in at least 90% of the cases;

b) 12 minutes for non-urban areas, in at least 75% of the cases.

The emergency medical care service should be organized in such manner that the maximal time for an intervention must not be longer than:

a) **15 minutes**, for emergency and intensive care units in urban areas, in at least 90% of the cases;

b) **20 minutes**, for emergency and intensive care units in rural areas, in at least 75% of the cases.

In order to implement an integrated emergency services management at regional level, all hospitals within the region should be included in a network, each network consisting in one regional first degree emergency hospital and several 2nd and 3rd degree local emergency hospitals. Moreover, the emergency and paramedic services department has to function around the clock in 12 hours shifts.

The mobile intensive care and emergency services provided by **SMURD**[1] have to abide a series of restrictions as well: emergency teams should consist of at least 4 people, including a driver/firefighter and doctor trained in intensive care and traumatology, the rest of the team being supplied by other emergency structures, local authorities and local hospitals, or specially trained volunteers.

Also, according to law, the emergency and first aid structures in charge of the mobile units are responsible for providing functional medical equipment and drugs for the care of at least 20 critical patients.

***Order no. 1706/2007*** states that:

- County capital cities with less than 500 000 inhabitants have to provide at least one Emergency Unit or Emergency Department within the county hospital. In case there is a regional or county children hospital, it must include a paediatric emergency department.
- Paramedics transporting critical patients have to report the emergency with at least 10 minutes before arriving at the emergency unit and provide all necessary information regarding the medical condition and treatment received by the patient in question.

According to the ***National Triage Protocol***, the medical staff has to evaluate every patient presented to the emergency unit in order to determine the severity of the emergency and the urgency of accessing the medical services of each individual, and the average triage time should be 2 minutes or less. The triage procedure has to take into account two very important parameters:

- The time the patient was registered by the triage personnel;
- The time of the first medical consultation.

Since doctors are tempted to perform thorough examinations, and therefore become unavailable for patients who might need urgent interventions, the triage procedure is performed by other specialized personnel in order to optimize the use of doctors' time.

The National Triage Protocol states that patients registered to the emergency unit have to be sorted in order to be included into one of the following emergency levels:

Level 1 – CPR (code red): special room with life support equipment and defibrillator.

[1] *Mobile Emergency Service for Resuscitation and Extrication*

- The patient requires ***immediate*** life saving intervention.
- Time to be admitted in treatment area: 0 minutes.

Level 2 – Critical (code yellow): first degree emergency room.

- The patient is in severe pain or major disconfort, is of high risk or is in an altered mintal status.
- Time to be admitted in treatment area: 10 minutes.

Level 3 – Urgent (code green): 2nd degree emergency room.

- Stable patient requireing 2 or more of the resources defined in the Triage Protocol.
- Time to be admitted in treatment area: 30 minutes.

In case the time to be taken over by a doctor exceeds 15 minutes or there are changes in the patient's status the triage algorithm is repeated in order to update the procedures necessary for the patient in question.

Level 4 – Non-urgent (code blue)

- The patient is stable and requires the use of only one of the resources described in the Triage Protocol.
- Time to be admitted in treatment area: 60 minutes.

Level 5 – Consult (code white)

- The patient does not require emergency medical assistance and none of the resources described in the Triage Protocol.
- Includes people coming to the hospital for:
  - ✓ Getting vaccine shots;
  - ✓ Administrative resons such as medical permits, prescriptions etc.;
  - ✓ Social cases without medical complications;
  - ✓ Time to be admitted in treatment area: 120 minutes

In order to avoid overloading the Emergency Unit, the triage area can accommodate some of the medical procedures, so that the time to solve all cases is minimized.

Given all these restrictions and constraints, the need for a resource management algorithm becomes obvious, even more so considering the challenges the Romanian health system has still to deal with.

## 3. The mathematical model

Since emergency departments always focus on the quality of the medical services they provide, the 4 hour target set for a patient's waiting time is of critical importance, and this raises a problem that still has to be solved properly: how to allocate the human resources in order to meet this target.

The data collected from different emergency departments prove that most of them manage to meet the target and many other are close enough, but considering that lately the number of accidents increased, maintaining this target is still a priority. The so-called staffing algorithm appears to ease the managers' decision-making process regarding the efficient allocation of human resources in order to decrease the waiting intervals.

In all emergency units, the number of patients is a time variable, depending on the time of day, on the day of the week and even on the season (it is expected to have more patients presenting fall-related injuries and broken bones during winter, given the weather conditions). Therefore, the staff allocation differs on various intervals during a day.

The purpose of this research is determining the need for medical personnel (doctors, nurses, lab technicians, triage specialists etc.) for each interval during a day in order to reach the 4 hour target. [0] [0] [0] However, finding the optimal algorithm for such a complex system as the medical one is difficult, since the parameters taken into consideration (especially the patient's arrival time) are not constant. This characteristic was noticed by several specialists and therefore, there are multiple approaches to the matter. Unfortunately, all the approaches so far focused on single service systems, such as call-centers. The allocation of the human resource in an emergency department

is much more complex, because of the nature of the medical services: every patient arrived at the emergency unit is submitted to multiple tests, and therefore needs a number of different resources. Moreover, the severity of the case is also an important factor determining the access to certain resources, and resources can be used for more than one patient at a time, which means that, in order to be effective, the allocation algorithm has to take into account all these constraints. The suggested heuristic algorithm uses models based on waiting queues to estimate the amount of resources needed and the loading time for each resource in the system in order to optimize their allocation, while the quality of the medical services is measured through the probability for delays. The model includes a waiting queue M/M/1 with a single serving station [3] [4], with arrivals determined by a Poisson process and an exponential serving time. By using Wolfram Mathematica 9.0, the algorithm is implemented in order to optimize the allocation of doctors, nurses and triage specialists and maximize the number of patients treated.

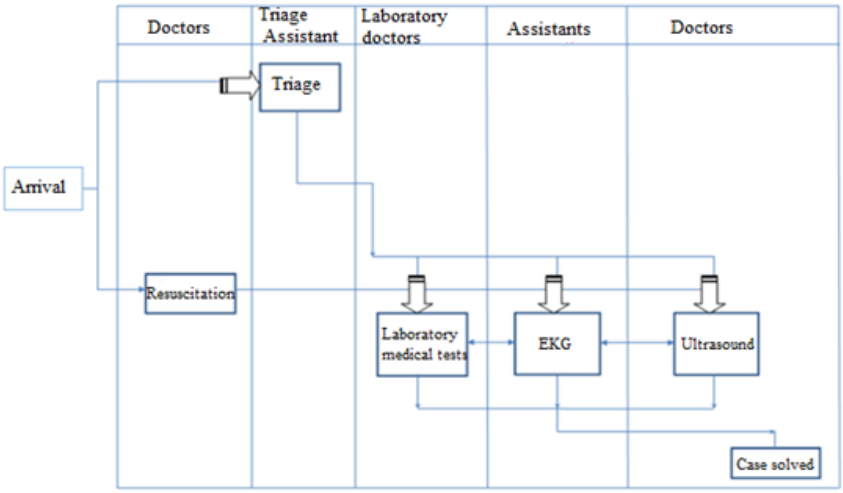



**Fig. 1**. The possible flow for a patient
(Source: Author prelucration)

According to the Romanian emergency procedures, the patients arrived at the ER are sent to the triage room, where they are sorted (using the National Triage Protocol) by the severity of the case and distributed to the medical disciplines required by their specific situations. This is where the decision regarding further examinations is made, and the patients can be submitted to additional testing, EKG or ultrasound examination. The results of these investigations determines whether the patient will be admitted or discharged. The possible flow for a patient who arrived at the emergency room is represented in Figure 1.

**Table 1**. *The flow of patients to the emergency room*

| | 00-04 | 04-08 | 08-12 | 12-16 | 16-20 | 20-24 | weekly | Percentage per ii | | Total / hour |
|---|---|---|---|---|---|---|---|---|---|---|
| CPR | 1 | 2 | 3 | 0 | 5 | 4 | 23 | 0.02 | 3 | **0.5** |
| Major emergency 1 | 64 | 86 | 90 | 60 | 40 | 26 | 374 | 0.33 | 53 | **8.83** |
| Major emergency 2 | 70 | 42 | 156 | 184 | 164 | 120 | 736 | 0.65 | 105 | **17.5** |
| **Total** | **135** | **130** | **257** | **252** | **209** | **150** | **1133** | **1** | **161** | **26.33** |

(Source: Author computation)

The table above (table 1) contains data regarding the patient flow at the emergency room within the Bucharest Emergency University Hospital. During the observed interval, the emergency unit registered a total of 1133 patients per

week, 161 patients per day and 27 patients per hour, included in three different emergency categories: CPR, Major Emergency type 1 and Major Emergency type 2, in order of decreasing priority. The highest number of patients is registered in the 8-12 interval, followed by the 12-16 interval and the 16-20 interval.

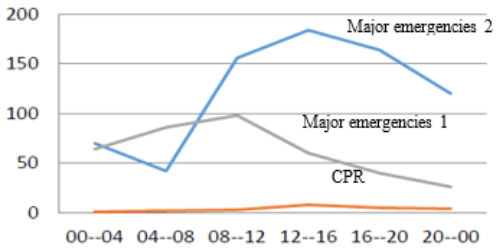



**Fig. 2.** Variation of the flow of patients by hourly intervals
(Source: Authors computation)

The variation of the patient flow during the 24-hour interval

| Basic Properties | |
|---|---|
| NetworkType | Open |
| NodeCount | 5 |
| SelectedNode | 1 |
| **Performance Measures** | |
| MeanSystemSize | 0.166667 |
| MeanSystemTime | 0.166667 |
| MeanQueueSize | 0.0238095 |
| MeanQueueTime | 0.0238095 |

| Basic Properties | |
|---|---|
| NetworkType | Open |
| NodeCount | 5 |
| SelectedNode | 2 |
| **Performance Measures** | |
| MeanSystemSize | 0.475117 |
| MeanSystemTime | 0.20103 |
| MeanQueueSize | 0.00243432 |
| MeanQueueTime | 0.00103 |

| Basic Properties | |
|---|---|
| NetworkType | Open |
| NodeCount | 5 |
| SelectedNode | 3 |
| **Performance Measures** | |
| MeanSystemSize | 0.214023 |
| MeanSystemTime | 0.0769622 |
| MeanQueueSize | 0.000108899 |
| MeanQueueTime | 0.00003916 |

| Basic Properties | |
|---|---|
| NetworkType | Open |
| NodeCount | 5 |
| SelectedNode | 4 |
| **Performance Measures** | |
| MeanSystemSize | 0.179108 |
| MeanSystemTime | 0.0909365 |
| MeanQueueSize | 0.0000539947 |
| MeanQueueTime | 0.0000274141 |

| Basic Properties | |
|---|---|
| NetworkType | Open |
| NodeCount | 5 |
| SelectedNode | 5 |
| **Performance Measures** | |
| MeanSystemSize | 105.806 |
| MeanSystemTime | 35.602 |
| MeanQueueSize | 104.815 |
| MeanQueueTime | 35.2686 |

**Fig. 3**. Output of Performance Measures using Poisson distribution.
(Source: Author computation)

In Figure 3 it is represented the variation of the patient flow during the 24-hour interval. We calculated the network load by determining MeanSystemSize. In the analysis we also took into account the fact that some of the patients are also admitted, thus diminishing the outputs. Thus, the number of medical interventions is 106/4, for one hour. Also, upon leaving the emergency reception network it is found that 105/4 (for one hour) hours of people / hours are waiting to be consulted at different offices. The average waiting time in the network is about 35 minutes and the number of people waiting is 105/4 hours.Based on the statistic data gathered using weekly reports, the average number of people arriving at the emergency room is 28/hour. The interval considered for simulation is 4 hours, since in Romania, reports are filed every 4 hours.

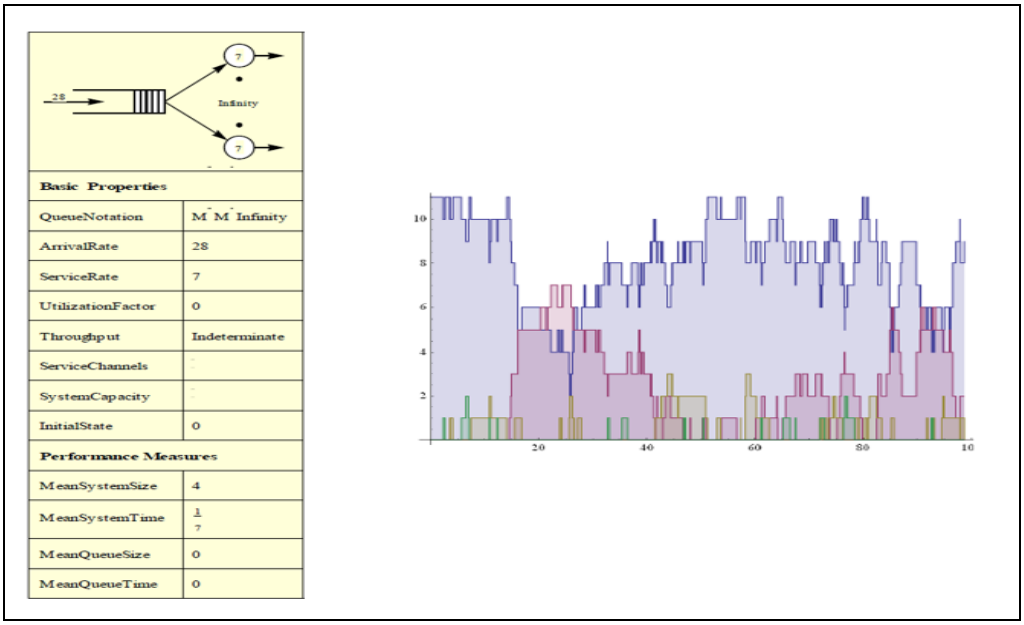
Considering that the events requiring medical response follow a Poisson distribution with an average of 28 and that serving times are exponentially distributed, a network structure can be identified, leading to the processing of approx. 106 people per simulation cycle. Moreover, there are 105 people waiting in the network, the average time for using a node in the network being of 35 minutes/patient.

Aiming to minimize the probability of delays ($\alpha = [1 + \beta \frac{\Phi(\beta)}{\phi(\beta)}]^{-1}$, where $\phi$ and $\Phi$ are the density function, respectively the normal standard distribution function) and considering that in fact it is approx. 50% (50% of the cases are solved within the system), the resulting loading coefficient is 56%.

In order to determine the number of personnel necessary in an ideal situation, a generic network is considered, characterised by Poisson [7] [9] random entries with exponential serving times and variable number of serving stations for the considered intervals. Once registered, a patient can be transferred to any of the medical specialties or can be discharged (either admitted or case solved and discharged), which means that the Romanian medical system appears to be organized as a Markov chain structure with probabilistic flows between the different medical specialties. [13] [14]
Based on the data gathered at the Bucharest Emergency University Hospital, the state transition matrix is:

$$\begin{pmatrix} 0 & 0,5 & 0,2 & 0,3 & 0 \\ 0 & 0 & 0,57 & 0,13 & 0,3 \\ 0 & 0,4 & 0 & 0,4 & 0,1 \\ 0 & 0,2 & 0,3 & 0 & 0,5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The medical emergency network includes three major states: a patient demanding medical care is called a potential patient and is identified when entering the system. Anybody entering the system has no possibility of going through it unless it transitions to one of the following states, which explains the transition probability associated to the first column of the matrix (0). The state corresponding to a person transitioning the stations associated to the specialized medical disciplines is the consultation state. Therefore, a person has a 0,5 probability to access the services of the medical lab, 0,2 to enter the EKG room and 0,3 for getting an ultrasound. Anybody entering the system cannot exit without going through at least one of the specialized consultations.
A patient entering the consultation state leaves the medical lab to enter the EKG room with a probability of 0,57, the ultrasound room with a probability of 0,13 or may get discharged with a probability of 0,3, if the case is considered closed by the emergency system. The patient getting an EKG may return to the lab with a probability of 0,4, be transferred to ultrasound with a probability of 0,4 or be discharged with a probability of 0,1.
Because the path the patient follows within the emergency department is rather complex, this is considered the cause for the queues that affect the efficiency of the Romanian medical system.
A patient sent directly to ultrasound may leave the system with a probability of 0,5, but can return to the lab or EKG with a probability of 0,2, respectively 0,3. A patient reaching the final transition state is considered cured (case solved), with no possibility of going back to any of the prior states, situation that is reflected on the bottom line of the matrix, which only contains 0 valued elements.




| Basic Properties | |
|---|---|
| QueueNotation | M M Infinity |
| ArrivalRate | 28 |
| ServiceRate | 7 |
| UtilizationFactor | 0 |
| Throughput | Indeterminate |
| ServiceChannels | |
| SystemCapacity | |
| InitialState | 0 |
| **Performance Measures** | |
| MeanSystemSize | 4 |
| MeanSystemTime | $\frac{1}{7}$ |
| MeanQueueSize | 0 |
| MeanQueueTime | 0 |

**Fig. 4**. The distribution of Markov network.
(Source: Author computation)

In Figure 4 s represented the system with infinite number of service stations, the average number of personnel needed is 4. In the graphic representation we simulated on a time horizon of 100 weeks, the allocation of the resources used by the emergency system. The blue line represent the doctors, the purple line are the allocation of the nurses, the yellow line represent the laboratory technicians and the green line is represented by the EKG specialists.As evidenced by a large number of scientific papers and studies, the Markov network structure is generally non-stationary; therefore it can be decomposed in serving networks with deterministic paths, equivalent to the situation of a single serving station with random entries but with unlimited order processing capacity (unlimited number of serving stations).[11] [12] [16]

This is justified by the sorting of patients into groups, by the type of consultation required, which makes the network sequential, with deterministic paths, so that the number of untreated patients is easily calculated by adding up the average unused factor of the station in question.

The ideal allocation of human resources for the emergency system is determined by taking into consideration a queue with an average of 28 entries (patients) per hour, which is equivalent to 112 patients for a 4 hour interval. This way, the system is used to its full capacity, which means that all patients arrived are treated and the average number of specialized consultation rooms in use in the network is 4. Theoretically, within such structure, the number of treated patients is undetermined, this being the ideal situation.

Considering the number of necessary specialists as 4, the model determines the over- and under-load of medical resource for each hour, as a difference between the actual number of doctors existing in the system and the ideal number of doctors necessary at a given time. Over-loading the system leads to an increase of the budgetary costs associated to the system, so this occurrence is penalized with the factor $p^o = 1$. Still, this situation is not that bad, since theoretically a larger number of doctors in the system would consequently mean a larger number of patients treated and an increased efficiency of the medical emergency system. The under-load of the system, on the other hand, means reducing the budgetary costs by

reducing the main resource in the system (doctors), which may lead to a diminished efficiency of the system. Therefore, this situation is penalized with the factor $p^u = 2$.

The available human resource is allocated so that the total number of people necessary per shift equals the total number of people allocated on that specific shift. Also, the difference between the variable measuring the overload and the variable measuring the underload should match the difference between the actual situation and the ideal situation resulted from the model. In the specific case of the Romanian medical emergency system, conventionally, the work hours are grouped in three 8-hour shifts. In this particular situation, the differences described above are considered positive (there are more doctors than in the ideal situation), while the maximum value for each resource in the system is considered 100 (maximal number of people available in the ER).

The model considers 49 constraints, the objective function being the minimizing of penalties paid for over- and under-loading the emergency medical system. Considering that the decision variables are whole numbers, the result is a whole number programming case, which can be solved using the Mathematica software. The solution offered by the model is that the first shift should be covered by 23 people, the second by 38 people, while for the third shift, 37 will suffice.

**Table 2***: The allocation of the medical staff*

| Time frame | 00-08 | 08-16 | 16-00 |
|---|---|---|---|
| Doctors | 6 | 9 | 8 |
| Nurses | 11 | 25 | 22 |
| Triage Specialists | 4 | 6 | 5 |
| Lab Doctors | 2 | 2 | 2 |
| Total | 23 | 38 | 37 |

(Source: Author computation)

Below, we evaluated the number of doctors needed in a system with an infinite number of service stations (equivalent to the ideal emergency system) is 4 doctors if it is considered that the number of patients is 26 per hour.

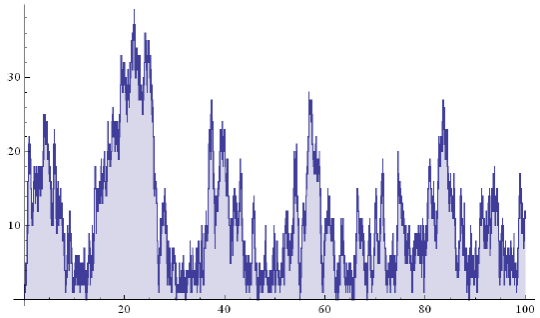



**Fig. 5.** Evolution of the flow of medical personnel. Simulation of the network in which the number of the patient in 26 and the number of doctors is 4 over a 100 – week horizon.
(Source: Author computation)

The algorithm for determining the necessary human resource consists of 5 steps:

- Step 1: Initialize a = 1;
- Step 2: Calculate parameter P, considering the constraint $= [1 + \beta \frac{\Phi(\beta)}{\phi(\beta)}]^{-1}$, where ϕ and Φ are the density function, respectively the repartition function of the standard normal distribution.
- Step 3: Determine the need for each resource for each specific interval, using the function:
  $s_k(t) = [x + \beta\sqrt{x}]$,
  where $x = m_\infty{}^k(t)$;
- Step 4: Estimate the percentage of discharged patients based on the resources estimated at step 3.
- Step 5: If the percentage calculated at step 4 is not 98%, a can be increased or decreased and we go back to step 2. Otherwise, STOP and save the solution.

$$\min[p^o \sum_{j=0}^{23} \Delta_j^+ + p^u \sum_{j=0}^{23} \Delta_j^-]$$

$$\sum_{i \in I} p_{ij} x_i = a_j, j = 0, \dots, 23$$
$$a_j - s_j = \Delta_j^+ - \Delta_j^-, j = 0, \dots, 23$$
$$\sum_{i=1} y_i \leq k$$
$$x_i \leq M y_i, i \in I$$
$$x_i \geq 0, i \in I$$
$$x_i \geq 0, x_i - integer \ and \ y_i = 0,1, i \in I$$
$$\Delta_j^+, \Delta_j^- \geq 0, j = 0, \dots, 23$$

$(s_0, s_1, s_2, \dots, s_{23})$ - *levels of resource allocation generated by the algorithm*;

$p^o$ - Penalties paid for each over allocation of the human resource per hour;

$p^u$ - Penalties paid for each underallocation of the human resource per hour;

$\Delta j+$ - overload at hour j, j=0...23;

$\Delta j-$ - underload at hour j, j=0...23;

I - the shift range allowed considering the legal constraints;

$xi$ - the decision variable expressing the number of employees scheduled to work on a shift, $i \varepsilon I$;

$$p_{ij} = \begin{cases} 1 & if \ the \ shift \ i \ includes \ the \ hour \ j \ as \ working \ time \\ 0 & if \ it \ doesn't \end{cases}$$

$$\sum_{i \in I} p_{ij} x_i - the \ total \ number \ of \ employees \ working \ at \ hour \ j, j = 0, \dots, 23;$$

M - *A very high number*;

k - The maximal number of shifts;

$y_i$ – Artificial variable - value 1 if at least one employee works on shift i, $i \varepsilon I$;

The suggested heuristic algorithm [8] [10] uses models based on waiting queues to estimate the amount of resources needed and the loading time for each resource in the system in order to optimize their allocation, while the quality of the medical services is measured through the probability for delays.[18]

## 4. Conclusions

To conclude, by comparing the Romanian emergency system to the British one, the first is found wanting, with a satisfaction level of only 56%, as opposed to 93% for the latter. According to the European standard, a medical emergency system is considered of high quality if it has a 98% quality indicator, which means that it can solve 98 cases out of 100. The Romanian emergency system is also affected by delays, the average time spent by a patient waiting for the different procedures necessary being 35 minutes, as opposed to only 7 minutes which is the European average. Since the main resource used in the emergency system is the human resource, the issue of staff allocation in the emergency departments is a major challenge for many medical systems. In the specific case of the Romanian medical system, it was found, using data gathered in a major Emergency Hospital, that a mathematical model is extremely helpful in determining the necessary human resource per shift, considering multiple factors such as the patient flow per hour and the budgetary constraint.

## References

[1] Ajay Tandon, Christopher JL Murray, Jeremy A. Lauer, David B. Evans - *The Comparative Efficiency Of National Health Systems In Producing Health: An Analysis Of 191 Countries*, GPE Discussion Paper Series: No. 29, EIP/GPE/EQC World Health Organization, 2000

[2] Coats, T.J., Michalis, S., 2001. Mathematical modelling of patient flow through an accident and emergency department. Emergency Medicine Journal 18 (3), 190-192;

[3] Eick, Stephen G., Massey, William A., Whitt, Ward, 1993. The physics of the Mt/G/1 queue. Operations Research 41 (4), 731-742;

[4] Feldman, Zohar., Mandelbaum, Avishai., Massey, William A., Whitt, Ward, 2008. Staffing of time-varying queues to achieve time-stable performance. Management Science 54 (2), 324-338;

[5] Fletcher, A., Halsall, D., Huxham, S., Worthington, D., 2006. The DH accident and emergency department model: A national generic model used locally. Journal of the Operational Research Society 58 (12), 1554-1562;
[6] Green, Linda V., Kolesar, Peter J., Whitt, Ward, 2007. Coping with time-varying demand when setting staffing requirements for a service system. Production and Operations Management 16 (1), 13-29;
[7] Green, Linda V., Soares, Jao., Giglio, James F., Green, Robert A., 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. Academic Emergency Medicine 13 (1), 61-68;
[8] Mayhew, L., Smith, D., 2008. Using queuing theory to analyze the Governments 4-h completion time target in accident and emergency departments;
[9] Gunal, M.M., Pidd, M., 2009. Understanding target-driven action in emergency department performance using simulation. Emergency Medicine Journal 26 (10), 724-727;
[10] Mortimore, Andy, Cooper, Simon, 2007. The ‘‘4-hour target: Emergency nurses’’ views. Emergency Medicine Journal 24 (6), 402-404;
[11] Munro, J., Mason, S., Nicholl, J., 2006. Effectiveness of measures to reduce emergency department waiting times: A natural experiment. Emergency Medicine Journal 23 (1), 35-39;
[12] Sinreich, David., Jabali, Ola., 2007. Staggered work shifts: A way to downsize and restructure an emergency department workforce yet maintain current operational performance. Health Care Management Science 10 (3), 293-308;
[13] Sinreich, David, Yariv, Marmor, 2005. Emergency department operations: The basis for developing a simulation tool. IIE Transactions 37 (3), 233-245;
[14] Whitt, Ward, 2007. What you should know about queueing models to set staffing requirements in service systems. Naval Research Logistics 54 (5), 476-484;
[15] Chiriță, N., Nica, I., 2019. Cibernetica Firmei. Aplicații și Studii de Caz. Ed. Economică;
[16] Ashour, O., Kremer G., 2013. A simulation analysis of the impact of FAHP-MAUT triage alghorithm on the emergency department performance measures. Expert Systems with Applications 40(1), 177-187;
[17] Nica, I., Chiriță, N, Fabian, C., 2018. Analysis of Financial Contagion in banking network, 32nd International Business Information Management Association Conference.
[18] Shih, C. L. and S. Su. 2003. Modeling an emergency medical services system using computer simulation. International Journal of Medical Informatics 72(3),57-72

**Ionuț NICA** (b. May 2, 1992) has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies in 2014. He followed a master’s degree in Cybernetics and Quantitative Economics, within the same faculty. Currently, he is a PhD student, teaching assistant in the department of the Faculty of Cybernetics, Statistics and Economic Informatics and work in bank as Basel II Expert in the department of Retail Credit Risk Methodology and Validation. He has high interest in areas such as Cybernetics, Operational Research, Economic Dynamics, Applied Mathematics and Big Data.

# Business Analytics Applications for Consumer Credits

Claudia ANTAL-VAIDA
The Bucharest University of Economic Studies, Romania

*The fast-paced and dynamic economical background determines all the industries to quickly adapt to change and adopt emerging technologies to remain competitive on the market. This tendency led to high volumes of data generated each second and to a decreasing ability of the manpower to analyze it and use if for beneficial purposes. This paper reviews the impact of Digital Transformation on the Banking area and how financial institutions leverage the advantage created by this trend, especially in the credit risk management field. Multiple papers on consumer credit scoring models written after the financial crisis from 2007 were reviewed and their results were summarized in this article, to increase the accuracy of future analysis by leveraging the already known results.*
**Keywords:** *Business Analytics, Machine Learning, Banking, Credit Risk Assessment, Scoring Models, Consumer Credits*

## 1 Introduction

Considering the current economic background, public and private companies are facing lots of challenges and difficulties when setting up their growth strategy due to the highly competitive environment where they operate. In order to overcome them, these entities require changes in their approach when making decisions, relying more on historical facts. Therefore, these companies should change focus – instead of intuition, they should aim to become data-driven companies, but for this approach to be successful, they need to invest in data-processing and methods of generating insights to get valuable information.

Change is all around us and if we analyze how fast technology, economy, medicine and other fields are evolving, we will notice that data and data-analysis are key players in it. As most of the industries are highly developed nowadays, any organizational decision should be taken based on results generated by business analytics to reach an ascending trend and a sustainable growth. Fortunately, companies understood the importance of data and the business value add they can bring when used correctly and consistent.

Gartner defines the Analytics area as a bunch of methods and techniques used for building analytical models and simulations to understand reality and predict the future state of a system. [1] It is a must-have area due to its impact on financial growth and profitability by improving company's competitive advantage and minimizing the error when making a decision. Examining the growing interest in this area and the resources invested in research by multiple companies, we can deduce that the popularity of this field has exponentially increased and companies acknowledged the impact of more sophisticated analytical decision-making tools for creating new opportunities, choosing the timing and enhancing the know-how.

The Banking sector is facing similar changes and needs to invest in business analytics as well. Its focus should be on understanding customer's behavior, how to improve interactions with customers and what motivates him to carry through their obligations. [2]

Even though this industry seems to be complicated, its activity is quite straightforward: the banking sector handles cash, credits and secures funds aiming to make more money out of them. Its products mainly consist of credit and debit accounts, loans and mortgages, helping people and companies to invest in their future, especially when they do not have enough liquidities to do that.

Players in each industry have a key focus on improving profitability, looking into cost reduction, revenue increase and process optimization and the banking field does not make an exception. Having said that, Business Analytics can play a key role in achieving those goals and there are already examples on the market which exceeded expectations. Some applications worth mentioning would be the algorithms of predicting risks, that analyze transactions and identify uncommon behavior of a customer or the credit-scoring algorithms that predict whether a customer will pay his debts, or he will miss them.

## 2. Algorithms used in the Banking Area

The following part of this paper explains why and how the concept of big data appeared in the banking field and how companies decided to leverage the advantages created. All the data created each second increased the influence of Business Analytics in business applications, with lots of solutions already implemented and many others being explored.

### 2.1. Digital Transformation of Banking

The Digital Era is characterized by new technologies which increase the speed of knowledge turnover [3]. The emerging development of analytics, cloud, social media and mobility technologies caused overall disruption across industries, Banking being one of the most impacted fields by this trend. [4]

Companies understood the importance and impact of technology in their activity, therefore they started to invest in research and development labs that use not only social media analytics, machine learning algorithms and big data, but also research on possible innovative scenarios that leverage artificial intelligence, robotics, automation, advanced data visualization and not only.

American Express, one of the giants in the Banking and Financial services from US, has set up a new tech lab to focus on big data, cloud computing, analytics and mobile technologies, as well as on futuristic ones. Thru this initiative, they aim to analyze customer's behavior on the market and quickly respond to it with customized products. [5]

Another example of company acknowledging the importance of new technologies and investing in its development would be the Fidelity Investments, one of the largest asset management companies in the world. In 2014, they announced the opening of Financial Labs, a research unit that will partner with The Massachusetts Institute of Technology (MIT) and Stanford University to get the "outside view" and develop innovative applications. [6]

These were only few examples of Financial companies investing into the tech area. The trend in the industry is to digitalize as much as possible and behave almost like tech companies. Banks must be quick in converting an idea into a service in order to survive on the market but also stay relevant in the years to come. Moreover, technology offers new opportunities to address untouched markets, by simplifying the communication and removing the geographical barriers. [7]

From previous examples, we draw the conclusion that companies have become aware of the importance of technology and the role they play in business performance management. Moreover, they started to massively invest in developing new technical capabilities to optimize their processes and improve relationships with customers.

What's the result of these investments? How digitalization translates in the real world? It results in high volumes of unstructured data that are harder and harder to analyze manually and that's the perfect scenario for Big Data and Analytics to come into the picture. As data is not meaningful enough, companies had to identify ways of converting data into information and insights to monetize digitalization.

### 2.2. Machine learning at a glance

Business Analytics represents the use of data, technologies, statistics, mathematics and computer-based models to help

management understand the business, solve issues and make fact-based decisions. [8] This area has four stages that have different business impacts, depending on their complexity and level of knowledge required [9]:

- **Descriptive Analytics** is answering the question "What happened?", by analyzing and displaying historical data in reports and dashboards to simplify the decision-making process;
- **Diagnostic Analytics** researches the causes and effects of a certain event in order to avoid it or increase its frequency in the future, depending on the impact. It usually answers the question "Why did it happen?";
- **Predictive Analytics** provides an answer to the question "When can it happen?" and implies statistical methods and Machine Learning techniques;
- **Prescriptive analytics** in the area that recommends decisions based on simulations and process optimizations, trying to answer the question "How can it happen again?".

Going forward, this paper will focus more On Machine Learning and its applicability. **Machine Learning** is defined as a tool or mechanism that uses statistical models to facilitate the solutioning of a problem, by studying the past behavior, identifying patterns and constantly improving itself based on the data analyzed. Its main purpose is to develop an adaptable algorithm to solve an issue, despite the external variables that might influence it or its complexity. [10] There are different types of algorithms used by machine learning listed below [11]:

- **Supervised algorithms** know the outcome from the beginning and its learning is guided by human observations and feedback thru tags and labels inputted from the beginning;
- **Unsupervised algorithms** rely exclusively on clustering separate data based on similar features and modifies the calculation process to respond to initial inputs; this type does not involve any external feedback, nor tags to be considered for data processing;
- **Reinforced algorithms** are about taking the most appropriate action to maximize the output, regardless the situation. If in supervised learning the expected outcome is known from the beginning, this type of algorithm has the liberty to decide what is best to do to perform a task and tends to learn from its own experience.

Even though Machine Learning sounds evolutionary and promises to revolutionize the way things work, having a positive impact on the areas where applied, it has both advantages and disadvantages that should be considered when deciding to use them for a real use case.

Thus, few of the advantages worth mentioning of Machine Learning are: [12]

- **Easily identified trends and patterns**. When given a large dataset to analyze, the Machine Learning algorithms can quickly identify specific trends that might not even be obvious to human beings.
- **Constant Improvement**. While exposed to computation of data, the ML algorithms gain experience and improve efficiency and accuracy, leading to more reliable decisions and results.
- **Various applications.** No matter the area you work on, you will find for sure an application that would involve Machine Learning algorithms and would be beneficial for your area.

On the other hand, Machine Learning has some limitations that should be known before deciding to use and invest in them:

- **Requires high volumes of data**. For excellent results, Machine Learning algorithms require high volumes of data to train on. Besides volumes, data quality of the train dataset plays an important role as well, outputs being highly dependent on inputs.
- **Time and Resources**. Depending on the complexity of analysis you want to perform, Machine Learning algorithms

may require time to learn and massive computer resources.

- **May produce biased results**. Machine Learning algorithms are highly susceptible to error, but this aspect mostly depends on the diversity of the dataset it trains on. If the train set is small and not inclusive enough, the results might be biased, leading to irrelevant interpretation.
- **Interpretability**. Unfortunately, it is hard to understand the reasoning behind a decision taken by an algorithm, reason why these are considered "black-boxes".
- **Scalability**. Once a model is proven to be efficient, companies implementing it need to overcome the challenge of scaling it. This can become expensive due to the resources required, the need of further optimizing it and integrating it with other systems.

All these challenges can be overcome if the company is willing to invest and few of them might not even appear, it always depends on the use case. Therefore, Machine Learning usage can surface the potential and value of unstructured data in each company. Its ability to form adaptive behavior in the process, without being programmed for this, makes them incredibly powerful when used for the right processes and analysis.

To sum up, the main benefits of the ML algorithms in this area is given by the complexity of the analysis performed (due to various parameters considered), reduction of approval time, less human resources involved, thus avoiding the human error, fraud of subjectivism

### 2.3. Algorithms for Risk Management

Risk management has become more important in the banking fields since the global financial crisis took place, moment since when banks started to research on how risks can be detected, measured and managed . [13]

There are different types of risks in the financial area that can be addressed by Machine Learning, but we will analyze the applicability only for three of them: Operational risk, market risk and credit risk.

**Operational risk management** assumes that a firm wants to foresee the direct or indirect risk of financial loss due to a host of potential operational breakdown. [14] The risk can be triggered by internal factors (people, system, deficient processes) or by external ones (global economic background, frauds, operational errors). Considering the increasing variety and complexity of risks, especially for financial institutions, machine learning and artificial intelligence applicability increased consistently and started to play a key role in predicting these events, assessing their impact and minimizing their effects. [15]

Banks pursuit the evaluation of the best ways to protect their data, systems and clients and machine learning can support that. Process automation can increase the execution of routine tasks, minimize human error, analyze data to outline the relevant content and increases the ability to evaluate risky clients and networks. Machine Learning can also generate and prioritize alerts for uncommon activities and asses the risk involved.

Another risks that is worth investigating is the **Market Risk** to which the banks (and not only) are exposed due to investing, trading and playing on the financial markets. Machine Learning is mainly proper for identifying inadvertent risk in trading behavior, for understanding the impact of the firms that trade on their market price, for establishing new patterns and connections between assets and how they influence each other or even for creating bots to constantly monitor the financial indicators and send alerts once a trade would be profitable.

Finally**, Credit risk** is one of the highest risks faced by banks and usually the one requiring the most capital, therefore its management is of high interest for the financial institutions.

The objective of credit risk management is to optimize the credit portfolio and reduce the risk of customers not meeting their obligations. The high and extensive complexity of credit risk assessment made

this area proper for machine learning applications.

## 3. Machine Learning for Consumer Credit

Utilizing Machine Learning techniques is not a new trend, but it is a growing one. Back in the '90s, a comparative analysis between traditional statistical models of distress and bankruptcy prediction and an alternative neural network algorithm proved to be an effective combinations, with a significant increase in accuracy. [16] And the research in this area just started at that point. Over years, there were multiple implementations of machine learning techniques supporting risk management, which proved to be very efficient, making the most optimal decision.

The following part of this paper mainly focuses on the applications that were developed and deployed for consumer credits in the risk management area after the financial crisis.

### 3.1. Scoring Models Overview

One of the tools most used in the credit risk management are the credit scoring models, defined as statistical methods that consider financial indicators to predict the default risk of individuals or companies. These indicators are given a relative importance and are considered when predicting the creditworthiness, pointing out the probability of default of the borrower. [13] In Table 3.1. are listed in a chronological order all the articles considered in this paper with the utilized algorithm(s). They are all tackling Credit Scoring models for Consumer Credit, reason why they were considered:

| Article | Author(s) | Year | Algorithm |
|---|---|---|---|
| *Credit scoring with a data mining approach based on support vector machines* | Huang, Chen, Wang | 2007 | Hybrid Support Vector Machines |
| *Support vector machines for credit scoring and discovery of significant features* | Bellotti, Crook | 2009 | Support Vector Machines |
| *Consumer Credit Risk Models via Machine-Learning Algorithms* | Khandani, Kim, Lo | 2010 | Classification and Regression Trees (CART), Linear Regression |
| *A Proposed Classification of Data Mining Techniques in Credit Scoring* | Keramati, Yousefi | 2011 | Artificial Neural Networks; Bayesian classifier; Discriminant Analysis; Logistic regression; K-Nearest Neighbor; Decision Tree; Survival Analysis; Fuzzy rule-based system; Support Vector Machine; Hybrid Models |
| *Loan Default Prediction on Large Imbalanced Data Using Random Forests* | Zhou, Wang | 2012 | Random forest |
| *Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions* | Harris | 2013 | Support Vector Machines |
| *Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble* | Wang, Xu, Zhou | 2015 | Lasso logic regression |
| *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research* | Lessmann, Baesens, Seow, Thomas | 2015 | Artificial Neural Networks, Support Vector Machine, Ensemble Classifier, Selective Ensemble Classifier, Threshold metric, Area under receiver operating characteristics curve, |

| | | | |
|---|---|---|---|
| | | | H-measure, Statistical Hypothesis Testing. |
| *redit scoring with a feature selection approach based deep learning* | Van-Sang, Nguyen | 2016 | Deep Learning |
| *A novel multistage deep belief network based extreme learning machine ensemble learning paradigm for credit risk assessment* | Yu, Yang, Tang | 2016 | Deep briefing Network, Extreme Machine Learning |
| *Ensemble Learning or Deep Learning? Application to Default Risk Analysis* | Hamori, Kawai, Kume, Murakami, Watanabe | 2018 | Bagging, Random Forest, Boosting |

Table 3.1. Articles list on Scoring Models for Consumer Credit

### 3.2. Machine Learning Algorithms for Consumer Credit Models

First reference found is the study presented by Huang, Chen and Wang in 2007 that proposed a hybrid SVM-based credit scoring model that analyzes the credit applications based on the information provided by applicants. [17] This application would help the banks decide whether to grant credit to consumers who require it. The authors compared the SVM classifiers with neural networks, genetic programming and decision tree classifiers, and even though the number of imputed features was low, the result was comparable with the other techniques' outputs. Moreover, the SVM algorithm was proven to be very effective in the classification area, successfully grouping credit applications into accepted or not, hence minimizing the money lose due to underperforming credits. Lastly, the researched revealed that if the SVM classifiers are combined with genetic algorithms, they can serve multiple purposes: performing feature selection tasks but also optimize the model parameters.

Even though most of the conclusions were positive, the authors outlined some negative sides of the hybrid Support Vector Machine – Genetic Algorithms models too. One of their downsides would be the long time required for training and the high computational complexity demanded for a good classification accuracy. Another inconvenience would be the "black-box" nature of SVMs, but this could be overcome with the use of SVM extraction techniques or the use in combination with other interpretable models.

In 2009, another article focusing on SVMs for credit scoring was published by Belloti and Crook. They compared three traditional techniques (logistic regression, discriminant analysis and K-Nearest neighbors) against SVMs using a dataset of around 25000 customers. The results reinforced the outcomes of previous researches (that SVMs are successful in classifying credit card customers), but also revealed that they can be well applied selecting the features that have the highest impact on likelihood of default. They also discovered that a very important indicator in this analysis is the type of credit card, as this could influence the other variables to be examined in the model.

Though, one major disadvantage acknowledged is the high number of support vectors required for the best performance, mainly due to the broad indicative nature of credit data. [18]

In 2010, Khandani, Kim and Lo [19] published a study in which they were using machine-learning techniques for forecasting models of consumer credit risks. The framework used consisted of generalized classification and regression trees (CART)

By combining credit bureau data with customer transactions and applying linear regression $R^2$ with a delinquency rate of 85%, they reached a better accuracy of classification rates for the default of an

application. They also analyzed the patterns of the time-series of delinquency rates and concluded that aggregated consumer credit-risk analytics may have a high influence in forecasting systemic risks. Moreover, by applying machine learning forecast models on the decision to cut credit lines, they estimated a cost saving between 6% to 25%.

On the same topic, Keramati and Yousefi presented a study in 2011 in which they acknowledge the importance of analyzing the huge amount of data generated by credit scoring in a fast-growing credit industry, but also the human impossibility of manually reviewing and interpreting it, hence the need of data mining techniques to support this effort. In their paper, they analyzed ten different data mining approaches and their results outlined the following [20]:

- K-Nearest Neighbor (K-NN) proved to obtain the best results for the credit scoring purposes;
- The Employ Multi-Group Hierarchical Discrimination (M.H.DIS) resulted to have better classification abilities than the traditional models;
- Support Vector Machine – MARS (SVM MARS), logistic regressions and neural network are very good for classification, but LDA and CART are easy to use in building such a model,
- Integration of Self Organization Maps (SOMs) with supervised classification methods proved to bring more advantages.
- Kernel Based RBF neural network was the best choice in identifying the true positive.
- The comparison between discriminant analysis, logistic regression, neural network and regression trees for predictions and classification tasks outlined that CART and neural network are the best to apply for best results.
- When evaulating the accuracy of K-NN, SVMs and neural networks, it resulted that the integration of all these methods with effective feature selection improved the accuracy of the classification.

One of the main conclusions they draw was that calculating the probability of default for an applicant if more meaningful that classifying them into the binary classes.

One year later, in 2012, Zhou and Wang proposed a study in which they applied improved random forest algorithms in the binary classification field, by attributing weights to the decision trees in the forest. These weights were calculated based on the previous performance, namely errors in training. Their approach outperformed expectations, beating the result of benchmark algorithms like traditional random forest, SVMs, KNNs in terms of accuracy and proved that parallel random forests can considerably reduce the learning time [21].

In 2013, Terry Harris published an article on Support Vector Machines (SVMs) applied for credit-scoring models from two perspectives: a broad one, considering the credits that are less than 90 days past due, and a narrow perspective, analyzing the credits that are more than 90 days past due, reaching the conclusion that the last produced more accurate, mainly for severe cases of default. The main explanation for this conclusion could be the greater number of cases fed to the model, leading to a better learning of the pattern for un-creditworthiness. [22]

Wang, Xu and Zhou published a new article in 2015, in which they outlined a new mix of algorithms for credit scoring that exceeded expectations and previously known results. Their approach consisted of applying clustering and bagging algorithms to generate balanced training data and diversify data, applying Lasso-logistic regression ensemble to evaluate credit risks. [23]

During the same year, Lessman, Baesens, Seow and Thomas updated the study started by Baesens et al., including new classification algorithms used in the credit scoring area. They considered 41 classifiers for 8 credit scoring data sets and their results proved that there are more performant classifiers than the standard logistic regression. More than that, they outlined the

business value add of improving the prediction models, variable selection and data quality and suggested that focus should change into these areas. [24]

One year later, in 2016, Van-Sang and Nguyen published a study on Deep Learning, a powerful classification tool that provides training stability, generalization and scalability with big data. This method surpassed results previously obtained with baseline methods and showed competitive performance with other feature selection models extensively used in credit scoring area. The study also outlined that fewer features considerer for the evaluation procedure allow for collecting essential variables, hence reducing the resources allocated on performing the research. Moreover, parallel processing proved to decrease processing time, whilst obtaining the same results. [25]

During the same year, Yu, Yang and Tang proposed a novel multistage deep belief network based extreme learning machine (DBN – based ELM) ensemble learning methodology as a promising mix for credit scoring problems. These 2 techniques were already known for the time-saving characteristics and for the high-learning capacity thru hidden layers. The structure of multistage ensemble learning model, working in three stages, conducted to better results than typical single classifiers. The steps followed for this analysis are the following: in the first stage, the bagging sampling algorithms are applied to generate multiple and diverse training subsets of data; during the second stage, the ELM is utilized as classifier and different ensemble components can be properly defined with right subsets and different initial conditions. The last stage merges the individual results to form the final classification output thru the DBN model, which can effectively outline the relevant information hidden in ensemble members. [26]

In 2018, Hamori, Kawai, Kume, Murakami and Watanabe published an article in which they assessed payment data and compared the prediction accuracy and the classification ability of three ensemble-learning methods with neural-network methods. The three methods assessed were bagging, random forest and boosting. The study outlined that the boosting method has a superior classifying ability, even when compared to neural networks. The performance of the lastly mentioned proven to be highly dependent on the choice of activation function, dropout inclusion and number of hidden layers. [27]

### 3.3. Discussions

All these articles are approaching the same topic from different angles and thru different methods. The observations and conclusions obtained by the authors can be further leveraged in analysis, hence improving efficiency and accuracy by using the already known results.

Overall, the main idea that was outlined by each article is that Business Analytics plays a key role in the evolution of the financial institutions and in the optimization of their processes. When applied, it minimizes costs by reducing the number of human resources involved and increases the accuracy of the decisions.

Credit scoring algorithms assigns numerical values to the client outlining whether the entity is likely to default or not. Most of the studies were focused on this area, treating it as a classification problem in order to facilitate the credit decision, but also minimize the credit risk exposure.

Machine Learning algorithms performed better than the traditional techniques in classification steps and obtained increased prediction accuracy. The SVMs were widely tested and proved to be very effective in the classification area and in the feature selection process, especially when combined with genetic algorithms. Another algorithm that exceeded expectations was the Random Forest applied in the binary classification area, which showed outstanding results and a reduced learning time. Deep learning was outlined as one of the tools that provides training stability, generalization and scalability.

Beside all these, another valid point that should be considered is the dataset used for researches: it should be varied, integer, divers, to cover as many scenarios as possible and reduce biased results. Having said that, if the availability of real data would increase, if would encourage more researches on evaluating all the problems encountered in credit risk management, and not only.

## 4. Conclusions

Companies acknowledged the importance of adopting new technologies and the business value it creates and it is expected that financial institutions will increase the machine learning applications in the risk management fields to enhance their capabilities.

Even though these applications have some known limitations, a major one being the inability of understanding the mechanism used to reach a decision (mentioned in the literature as “operating like a black box”), the business value it creates it significantly higher, main benefit worth outlining being: high complexity of analysis performed, limited number of human resources involved, minimal error and reduction of performing time.

Machine Learning proved to be evolutionary and promises to revolutionize the way things work, hence it has the potential to transform the risk management area and enables the discovery of complex, nonlinear patterns in broad datasets.

This paper introduced an assessment of the researches around credit scoring algorithms for consumer credit within the banking industry, mostly because credit risks is considered the highest risk for a financial institution. However, the advantages and disadvantages of various machine learning tools for credit scoring can be further studied to refine them, improve results and maximize their values.

In conclusion, even though there have been different studies performed in this area, there’s still room for research and improvement to extend the beneficial applicability and impact of machine learning in the financial field and not only.

## References

[1] "Gartner Glossary," Gartner, Available: https://www.gartner.com/en/information-technology/glossary/business-analytics.

[2] M. Dwight, A framework for Applying Analytics in Healthcare – What can be Learned from the Best Practices in Retail, Banking, Politics and Sports, Pearson Education Inc, 2013.

[3] G. Doukidis, N. Mylonopoulos and N. Pouloudi, Social and Economic Transformation in the Digital Era, IGI Global, 2004.

[4] B. Raghynathan and R. V. Maiya, SMACing the Bank - How to Use Social Media, Mobility, Analytics, and Cloud Technologies to Transform the Business Processes of Banks and the banking Experience, CRC Press, 2018.

[5] T. Groenfeldt, "Forbes," 24 Dec 2014. [Online]. Available: https://www.forbes.com/sites/tomgroenfeldt/2014/12/24/american-express-opens-tech-lab-in-palo-alto/.

[6] I. Schmerken, "WallStreet and Technology," 10 Feb 2014. [Online]. Available: http://wallstreetandtech.com/asset-management/fidelity-labs-takes-innovation-to-the-next-level/d/d-id/1316288d41d.html?.

[7] R. Browne, "CNBC," 18 Nov 2019. Available: https://www.cnbc.com/2019/11/18/banks-must-behave-like-tech-companies-to-survive-amid-fintech-threat.html.

[8] J. R. Evans, Business Analytics - Methods, Models and Decisions, Second Edition, Pearson Education, Inc., 2016.

[9] H. Chahal, J. Jyoti and J. Wirtz, Understanding the Role of Business Analytics, Spinger, 2019.

[10] G. Sunila, Practical Machine Learning, Packt, 2016.

[11] O. Theobald, Machine Learning from Absolute Beginners, Scatterplot Press, 2017.

[12] D. TEAM, "Data Flair," 1 Jan 2019. [Online]. Available: https://data-

flair.training/blogs/advantages-and-disadvantages-of-machine-learning/.
[13] L. Martin, S. Suneel and K. Maddulety, "Machine Learning in Banking Risk Management: A Literature Review," Risks, 2019.
[14] I. A. Moosa, Operational Risk Management, Palgrave MACMILLAN, 2007.
[15] T.-M. Choi, H. K. Chan and X. Yue, "Recent Development in Big Data Analytics for Business Operations and Risk Management," IEEE Transactions on Cybernetics, 2007.
[16] E. I. Altman, M. Giancarlo and F. Varetto, "Corporate distress diagnosis: Comparison using linear discriminant analysis and neural networks (the Italian experience)," Journal of Banking & Finance, 1994.
[17] C. L. Huang, M. C. Chen and C. J. Wang, "Credit scoring with a data mining approach beased on support vector machines," ScienceDirect - Expert Systems with Applications, vol. 33, pp. 847-856, 2007.
[18] T. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," ScienceDirect - Expert Systems with Applications, vol. 36, pp. 3302-3308, 2009.
[19] A. E. Khandani, A. J. Kim and A. W. Lo, "Consumer credit-risk models via machine leraning algorithms," Journal of Banking & Finance, vol. 34, pp. 2767-2787, 2010.
[20] A. Keramati and N. Yousefi, "A Proposed Classification of Data Mining Techniques in Credit Scoring," in International Conference on Industrial Engineering and Operations Management, Kuala Lumbur, Malaysia, 2011.
[21] L. Zhou and H. Wang, "Loan Default Prediction on Large Imbalanced Data Using Random Forests," TELKOMNIKA Indonesian Journal of Electrical Engineering, vol. 10, pp. 1519-1525, 2012.
[22] T. Harris, "Quantitative credit risk assessment using support vector nachines: Broad versus Narrow deafult definitions," Elsevier - Expert Systems with Applications, vol. 40, 2013.
[23] H. Wang, Q. Xu and L. Zhou, "Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble," PloS Ones, 2015.
[24] S. Lessman, B. Baesens, H.-V. Seow and L. C. Thomas, "Benchmarking state-of-the-art classiÞcation algorithms for credit scoring: An update of research," European Journal of Operational Research, 2015.
[25] V.-S. Ha and H.-N. Nguyen, "Credit scoring with a feature selection approach based deep learning," MATEC Web of Conference, vol. 54, 2016.
[26] L. Yu, Z. Yang and L. Tang, "A Novel Multistage Deep Belief Network Based Extreme Learning Machine Ensemble Learning Paradigms for Credit Risk Assessment," Flexible Services and Manufacturing Journal, vol. 28, 2016.
[27] S. Hamori, M. Kawai, T. Kume, Y. Murakami and C. Watanabe, "Ensemble Learning or Deep Learning? Application to Default Risk Analysis," Journal of Risk and Financial Management, 2019.

**Claudia ANTAL-VAIDA** is a graduate of the Faculty of Economic Cybernetics, Statistics and Information at the Bucharest Academy of Economic Studies, bachelor's degree in Cybernetics and master's degree in Business Analysis and Enterprise Performance Control. She is currently a PhD Student at the same University, mostly interested in business analytics, machine learning and data structures.

# Business Intelligence and Machine Learning. Integrated cloud solutions providing business insights for decision makers

Laura - Gabriela TĂNĂSESCU
The Bucharest University of Economic Studies, Romania
lauratanasescu@gmail.com

*The aim of this paper is to present the latest trends in business intelligence and ways in which nowadays organizations can implement cloud technologies. This work is going to present challenges of the market, providers of integrated cloud business intelligences tools, advantages and disadvantages of moving to the cloud. A real life use case will argue the importance of taking advantage of data, as well as the necessity and the obvious benefits of having the right tools of transforming data into correct business decisions.*
***Keywords**: business intelligence, cloud computing, artificial intelligence, analytics, machine learning, innovation, data*

## 1 Introduction

Technology has evolved a lot in the last decade and nowadays we can even talk about a new paradigm related to cloud and how cloud technologies are going to influence organizations and their development.

In the same time, there are plenty of talks about a revolution related to data. How big data has been developing in recent years, how is going to challenge artificial intelligence our everyday work and what is the way in which organizations adopt business intelligence in order to gain insights from this data.

Therefore, the aim of this paper is to talk about the recent trends in technology, offering clear but relevant information about the most important concepts. In addition, the paper is going to provide an example about business intelligence applications in real life use cases, using cloud technologies from one of the top providers of cloud.

In the following chapters, this work is going to talk about all the theoretical concepts mentioned before in order to provide a clear image of the domain. Afterwards, it is presented an analysis of Oracle Corporation and its analytics solution in cloud, with advantages and disadvantages, competitors and benefits. Finally, a use case is going to be realized with this technology.

## 2 Cloud technology

This chapter is going to present some theoretical aspects related to cloud, as well as types of it and what are the advantages and disadvantages of using it.

### 2.1 Cloud computing

We can refer to cloud computing as the possibility to provision computing services with the help of the internet, services where we can include networks, software, servers, analytics and databases. All these cloud capabilities are used to offer a faster innovation and a more flexible way to use resources. [1]

A concept (**Fig. 1**) that comes with cloud technology is that the locations of the service used, the hardware, all the operating systems and also many more other details remain irrelevant to the final user. [2]

Practically, cloud providers offer services that enable the users to access, store or transmit file or applications on different remote servers as well as the power to access all the data using the internet. This being said, it is not required for any user to be in a specific place in order to gain access to it. [3]


**Fig. 1**. Cloud computing concept [4]

### 2.2 Cloud classification

Cloud computing is known as public or private cloud. The first one refers to those services that are offered to users for free over the internet. The second one provides access just for a number of people, offering services that are a system of networks. Here, we can also mention a third category, known as a hybrid cloud, that is a combination of the two described above.

Cloud computing cannot be seen as one piece of technology, but it is divided in three different services: software-as-a-service (SaaS), infrastructure-as-a-service (IaaS) and platform-as-a-service (PaaS). The first one is related to the part of the license for software applications that is offered to customers, license that is provided with a pay as you go model.

The second one provides customers the opportunity to practically rent infrastructure that includes servers, storage, operating systems and networks from any of the cloud providers of infrastructure.

The last one is especially designed in order to make it easier for developers to create web and mobile apps, without having the need to manage or set an environment and infrastructure for the development process. [5]

In **Fig. 2** can be seen a complete and detailed architecture of cloud.

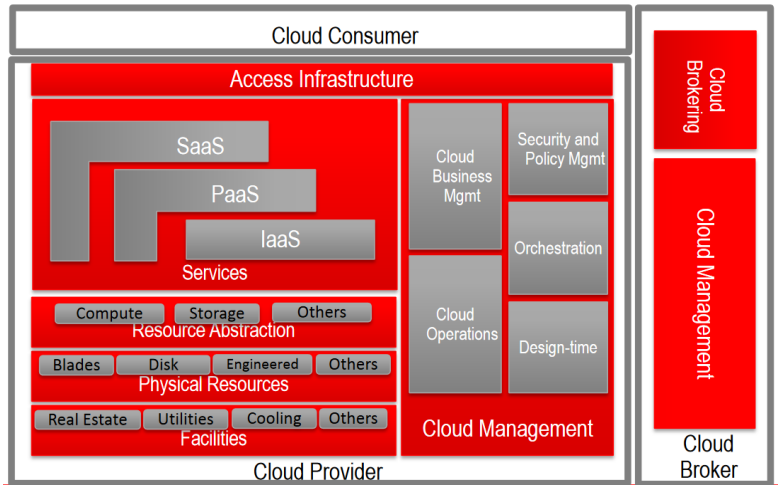



**Fig. 2.** Cloud architecture

### 2.3 Advantages and disadvantages

Looking at this new technology that impacts our life nowadays, it is important to talk about the benefits that come with it. So, the first one to mention is related to costs. The fact that cloud computing eliminates all the expenses that were coming with the hardware and the software, as well as with setting up and running all the data centers, it is needless to say that the overall costs are being reduced. The second one is effective for all the customers and also for the technical users and it is about speed. All the services are provided as self-service and on demand, as well as the fact that every service can be provisioned within minutes and without a great amount of knowledge.

The third benefit that deserves being mentioned is related to performance that comes from the fact that cloud services run

on a worldwide network of secure data centers. In addition, the performance comes also from the upgrades that are regularly being made to the systems, making them always faster and more efficient.

Last, but not least, we must talk about security. There are plenty of systems that cannot assure a good security due to the lack of knowledge or the missing budgets for improvements, so all the technologies and policies that are offered from the cloud providers offer a very important and needed secured system for customers.

Having mentioned all the benefits from moving to cloud and accept the innovations, it is equally fair to also talk about the downsides. The most important one that can be identified is again related to security. Moving and working with sensible data to a cloud that runs on a different country, for example, can cause concerns.

There are also several regulations that are unclear when talking about whether or not some critical national data can be stored in another country where the data center physically is. So, this is a risk that an organization should consider when trying to adopt cloud technologies.

Furthermore, the fact that just one portal is used by multiple employers at the same time, manipulating data and making changes too, can cause damage to the overall course of work.

## 3 Business Intelligence

This chapter is going to present some aspects of business intelligence as known today, as well as how this technology is used in cloud.

### 3.1 BI concept in nowadays technology

Business intelligence can be defined as a software application that is realized in order to analyze, report and offer visualizations of data. The entire procedure that includes reporting data, analyzing it and also accessing all the sources are achieved by a business intelligence software. This concept covers multiples directions like applications, technologies, processes and tools, as well as practices of translating relevant conclusions. [1]

As mentioned above, business intelligence is, in fact, a process driven by and with data, where data storage and knowledge management make a combination that helps in the business decision process. [1]

BI technologies (**Fig. 3**) are used in order to help organizations achieve better decisions about the existent processes, requiring skills, relevant data and innovative technology. BI can be extended as a concept where it can include not only applications and tools, but also infrastructure and practices that enables those organizations to analyze information faster and better, optimize processes and formulate relevant conclusions and taking decisions. [1]

We can admit that a successful BI implementation should be focused on software development or hardware, but on the value that comes from information.

Taking this into considerations, it is important to understand the way data is created and used, what is the quality of that data, how it is constructed that system and the service levels.

So, coming to a point where a conclusion of business intelligence concept is needed for nowadays technology, we can best define it as a set of business data that is taken from multiple sources which are translated into information using different applications in order to support decision making and help organizations to achieve their needs. [6]


**Fig. 3.** Business Intelligence concept

### 3.2 BI in cloud

When using business intelligence solution in cloud computing environment, we should underline the great opportunities this combination can offer. Even though both of these technologies are at a starting point of their development, they are in trends for most of the organizations, having some difficulties to solve though. One of the main problems with these two is related to integration. Of course it is relevant to add here the costs that come from reorganization of processes and work, as well as from people trainings. Not only will these costs appear, but it is also possible that many employees will have a bad attitude towards change. Here it can be added the lack of resources to support these changes, the possibility of downsizing the targeted departments and also the uncertainty that comes with adopting new technology. Nevertheless, every risks and disadvantage that were mentioned about any cloud technologies can be translated to this combination of applications too.

### 3.3 Methods used to create a business intelligence system in cloud

In order to adopt such a system, there are several steps to follow so that the final result should be the one expected.

The first thing to consider is about data collections, where it is important to have the means of accessing and integrating all the places from where the data can be taken. In addition, an architectural model has to be proposed in order to have the best way to collect data.

Another step in this process is the validation part that also comes with reliability. So it is proposed to solve the reliability issues to ensure correct measurement accuracy and also the right measuring instrument used in the measuring process. Reliability is in fact used to phrase the measurement to which a metric provides correct results and no random errors.

The third part is about data preparation. In order to prepare data, we have to first collect all the data, combine it from all the sources where the data was found, structure it in order to be clear and organize it so that it can be easily analyzed.

Analysis of the data comes with a process where statistical or logical techniques can be applied in order to illustrate, recap and evaluate data.

The last part and the one that also brings business value and insights is about data

analytics. The most common way used was descriptive analysis, where reports were added. After big data has started entering in the biggest companies on the market, the traditional business intelligence has changed due to speed and ways of storing. Therefore, predictive and normative analysis has emerged lately, the firs ones being in the spotlight as well.
The evolution of big data and analytics has affected the overall way of business intelligence delivery. Information needs to be quickly extracted from data, organizations being more and more concerned about normative and predictive analytics that include machine learning capabilities and rapid ways of building relevant visualizations.

## 4 Artificial intelligence

In this chapter, the paper presents some theoretical information, as well as a brief introduction in artificial intelligence. Moreover, it is going to be made the obvious relationship between data and artificial intelligence in nowadays businesses, as well as the important part named machine learning.

### 4.1 AI definition and structure

While it is mentioned before that business intelligence works with the aim of collecting, reporting and analyzing data, artificial intelligence comes with another approach that impacts data.
In fact, artificial intelligence enables computers to make their own decisions. Thus, we can define artificial intelligence as the ability of a machine or computer to learn and think like human's brain.
Artificial intelligence contains subfields like machine learning, neural network, deep learning, compute vision and natural language processing (**Fig.4**). Explicitly, machine learning is working to automate analytical model building. This field uses different methods like neural network, operations research and statistics so that to find hidden insights from data. [7] [8]

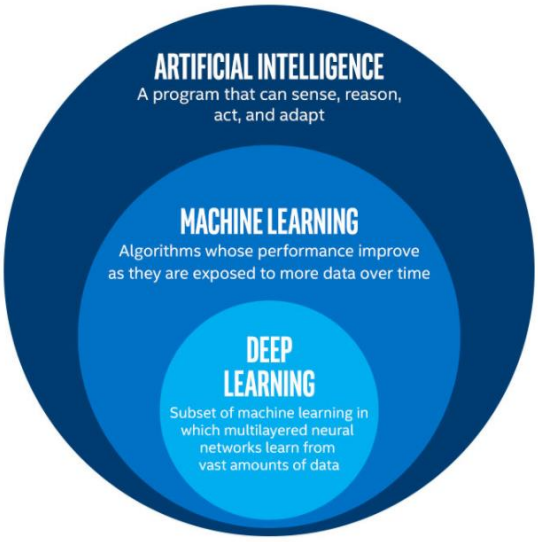



**Fig. 4.** Artificial intelligence and its categories

### 4.2 Data and AI

There are multiple sectors of economy that deal with huge amounts of data which are available in different formats and sources. This enormous amount known as big data is becoming available and easily accessible due to the progress of technology. Multiples data applications of machine learning are formed through complex algorithms build into a machine or computer. The code used creates a model that identifies the data and, after data, it is building predictions around it. The model is going to use parameters built in the algorithm in order to form patterns that are going to help the decision making process. When new data is added to the process, the algorithm used will adjust those parameters mentioned before in order to check if the patter has changed. However, the entire model should remain the same.
AI along with Machine Learning and Deep Learning present multiple technologies that are utilizing Tensor Process Unit (TPU) and Graphics Processing Unit (GPU).

### 4.3 AI applications in cloud

Apart from the visualizations of data that are done using a business intelligence software, we can talk separately about

what is the value that cloud brings to the machine learning component.
Therefore, there are many reasons to talk about regarding using machine learning in cloud, along with business intelligence. First of all, it is about the leverage and speed provided by the power of the GPUs that are needed to train different algorithms, without investing a lot in hardware. Moreover, the scale up and down capability make it efficient and easier for users to improve the power depending on the needs and measure a project have.
In addition, the new picture offered by cloud providers in terms of business intelligence and machine learning does not require advanced skills and lots of knowledge in data science and programming.

## 5 Oracle as a cloud provider of analytics platform

In the following lines, the paper introduces Oracle as a cloud provider, as well as an interesting player in data and analytics market for cloud users. Finally, this chapter also propose a demo of Oracle analytics platform that is going to demonstrate the benefits of using cloud for analyzing data.

### 5.1 Oracle Analytics Cloud

One of the top cloud providers that also comes with an analytics platform is Oracle, which is proposing a comprehensive tool in a unified platform, including data preparation for enterprise reporting, self-service visualizations, advanced analytics, self-learning analytics and machine learning integration on top. (**Fig.5**)

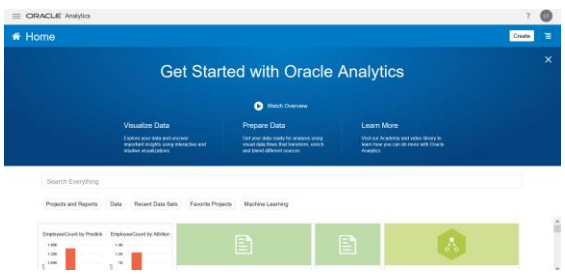



**Fig. 5.** Oracle Analytics Cloud interface

In the capabilities of this cloud platform we can enumerate data discovery, which helps users to easily collaborate with others, building intelligent analysis, machine learning models and statistical modeling.
Another thing to mention is related to the fact that developers can utilize interfaces that help them extend and customize all the analytics experiences in the flow.
It is very interesting the fact that in Oracle Analytics Cloud users can take data from any source, collaborate on project with others and explore real time data. Furthermore, unlike other providers that require the user to compromise between self-service, governed and centralized analytics, Oracle Analytics Cloud (OAC) solves this problems by offering a single solution that also incorporates Machine Learning and Artificial Intelligence.
Through the capabilities of OAC we find the data preparation enrichment that is built into the analytics platform. Another one is the business scenario modeling, a self-service engine for industry that helps in multidimensional and visual analyses. Moreover, we see here that proactive mobile that always learns from your work and offers contextual insights in daily activities. Last, but not least, is the enterprise reporting capability, the power of security and governance having a semantic layer which maps complex data into familiar business terms.

### 5.2 Augmented analytics – features of OAC

Keeping in mind the concepts mentioned above, we can converge business intelligence, artificial intelligence and more specific machine learning, into a term named augmented analytics.
We can see this concept as an evolution for the foundation build from analytics and business intelligence as well as big data, combining different and emerging technologies.

While business intelligence is about creating and finding data insights, AI and ML are about learning from different datasets in order to offer machine-driven decisions.

As it is known at the moment, a BI platform actually ingest a lot of data from multiples sources before anyone can prepare and reorder data.

An augmented analytics system is taking these latter steps and automates them using machine learning and artificial intelligence technologies. As an explanation, machine learning handles data preparation and artificial intelligence handles initial analysis.

Looking at the benefits of such a concept, we can tell that, in spite of those that are offered by multiples providers, there are some that offer a level of efficiency and accuracy that is possible due to computer processing. Thus, one of the most important aspect of augmented analytics includes accuracy. If the analysis is made by data scientists, there is likely that a mistake is going to occur. When using machine learning for that, these situations are eliminated from the beginning. [9]

Another thing to consider is speed. There are gaps that can appear when we first initiate a project using a BI platform like when we manually prepared data and also wait some time in order to receive an answer from different parties. Using augmented analytics, this process begins immediately, launching AI to cull the specific and needed data and also to begin the drilling down for the specific output needed for the project.

One more aspect to consider is the reduction of bias. Bias does not have to come as a personal shortcoming, but as a habit or a routine. Humans tend to revert to patterns so there can be a blind spot for data scientist that can lead to overlooked insights. In this case, computers and machines are going to work more efficient without inherent bias.

Last thing on this list is about the resources used. Augmented analytics can increase the resources by having them do more important things than some manual labor. So, for data scientist, it is going to mean more time to create different business problems and extract deeper insights form data. [9]

### 5.3 Oracle versus competitors

One of the advantages that Oracle Analytics Cloud has, as CEO of Red Pill Analytics said [8], is the fact that Oracle offers all in a single solutions. In fact, there is known that other providers offer multiple products in order to satisfy the same need and the problems is this process takes more time, resources and configurations before getting value from the investment.

Another thing to consider is the ability to scale up or down in order to adjust the resources, depending of the nature of workloads.

Also, its ability to offer not only visualizations, but a comprehensive view that helps the enterprise is considered important by another group director of Qubix International [8].

## 6 HR Attrition case study using OAC

Having all these concepts about business intelligence, artificial intelligence and cloud technology in mind, a small demo can be easily provided. An HR data set added in Oracle analytics cloud is going to be used so that to present the advantages and extended possibilities for data analysis.

### 6.3 Data loading and hypothesis formulation

We are going to use a public data set about employees and some details about them, as well as staff attrition.

These data set contains details about age, department, hours worked, over time, distance from home, daily rate, education, employee satisfaction, gender, job level, job role, marital status, relationship satisfaction and years since promotion.

These variables are considered suitable in order to make an analysis for the

organization's employees (in order to see their satisfaction and problems based on work life balance, benefits and capabilities) as well as realizing an algorithm in order to predict whether or not an employee that we do not know anything about is going to leave the company or not.

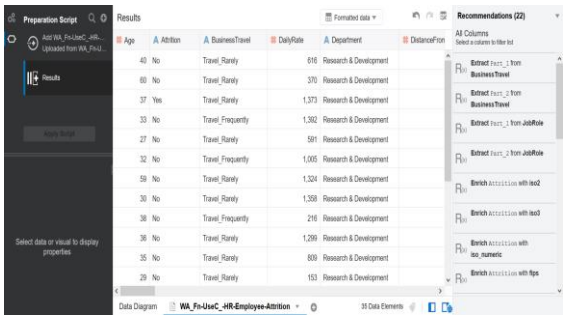


**Fig. 6.** Data loading menu for OAC

### 6.4 Data preparation

As mentioned before, Oracle Analytics Cloud, the tool used for the analysis, is offering intelligent recommendations and possibilities to arrange and filter data, as well as change a measure into an attribute and vice versa.

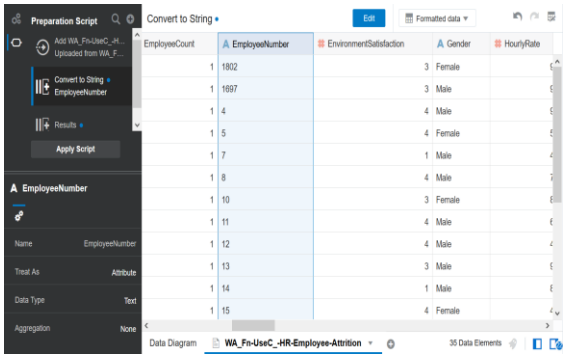


**Fig. 7.** Data preparation menu

There is also a part provided by some machine learning capabilities of the tool that is helping the users to enrich the actual data. Depending on this, we can add this recommendations or not.

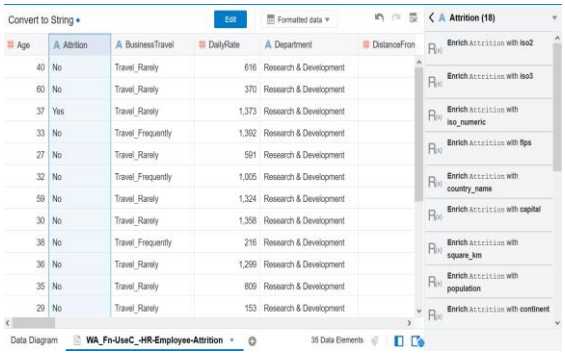


**Fig. 8.** Data enrichment capability

In the same time, the explanation mode that can be seen below is part of the augmented analytics. Thus, for a numeric variable like monthly income, we see that the tool offers us different graphics that are relevant for the analysis.

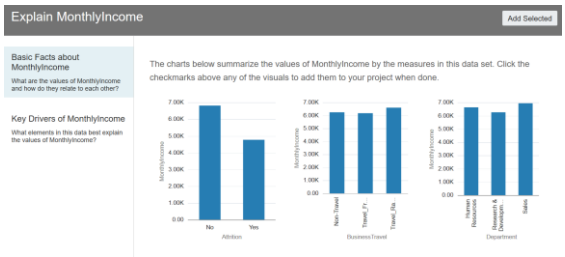


**Fig. 9.** Explanation mode for monthly income – basic menu

Finally, the same thing can be seen in key driver tab that shows us which are the variables that best explain monthly income (**Fig. 10**).

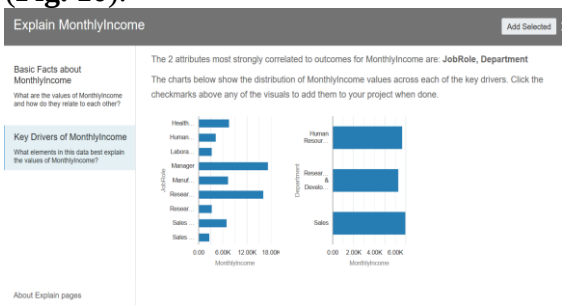


**Fig. 10.** Explanation mode for monthly income – key driver menu

### 6.5 Data visualization

One type of dashboard we can make is a general one that provides us with general information about that data we use.

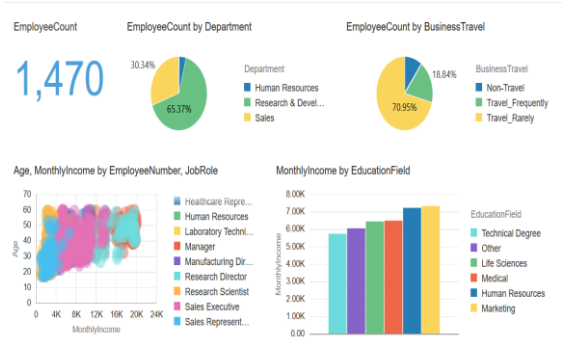



**Fig. 11.** Data visualization for general overview

Therefore, we can see in **Fig. 11** that the total number of people analyzed are 1470. The majority are working in research and development and after that sales. In addition, we see that 70% of the people travel rarely, while 19% travel frequently.

In the same time, if we look at the monthly income of the people, as well as the role they have and the age, we see that the greatest income is for research scientists and research directors, with age between 35 and 60. On the other hand, we have the minimal income values for sales representatives and laboratory technicians, where the start age begins with 20.

Last, but no least, we see that marketing provides the greatest monthly income, while a technical degree provides the smallest value when analyzing this organization.

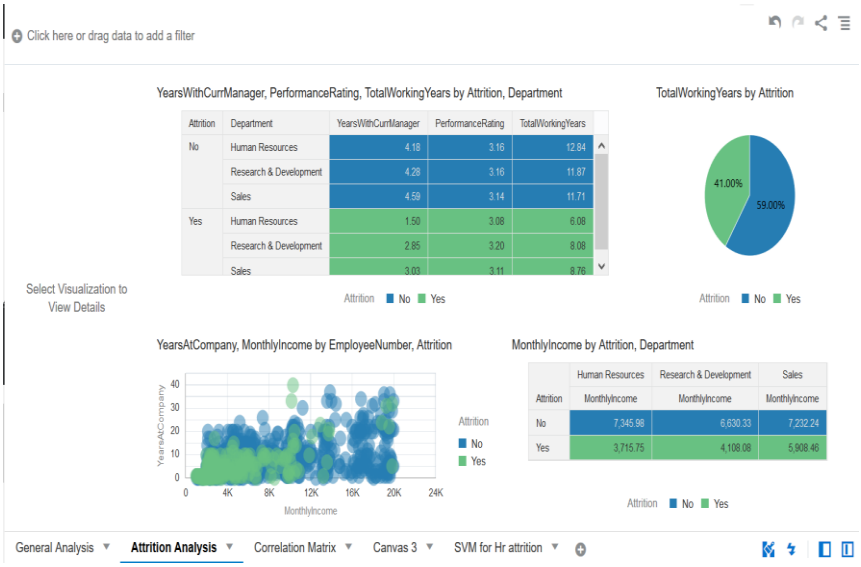



**Fig. 12.** Data visualization for attrition

The second dashboard (**Fig. 12**) provides information about attrition. We can see from the charts that most of the people that are going to leave the organization are not for many years with the current manager, that they have lower performance ratings and lower total working hours. This shows us that people who will stay within the organization have a history in it, they dedicate a lot of time to it and they perform.

In the same time, most of the people that are going to leave have a smaller income and they have been with the organization for little time than the others.

Nevertheless, we see that the younger employees are those that will leave the organization.

Another interesting thing to consider is the fact that the tool provides an instant visualization of correlation between variables like in the picture below.

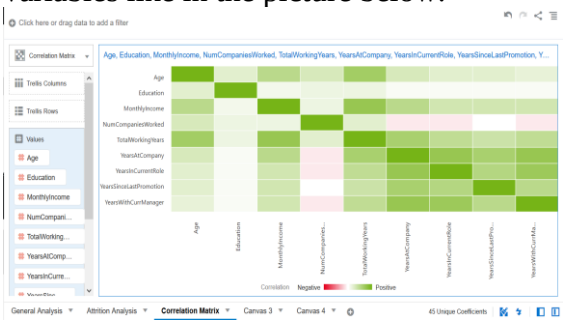



**Fig. 13.** Correlation matrix

Not only (**Fig. 13**) is this correlation easy to build because it does not request any technical knowledge, but it instantly provides useful information about our data like the fact that there are powerful and positive connections between age and total working years, monthly income and total working years, age and total working years.
On the other hand, we see some negative connections between years in current role and number of companies worked and number of companies worked and years with current manager.

### 6.4. Machine learning for HR use case

For the machine learning part of this project, we are going to build different machine learning algorithms of classification. After building them, an analysis is useful in order to decide which model is the best one for the use case and that model will be used for prediction.
For the HR attrition use case we are going to use support vector machine algorithm and Naïve Bayes algorithm, both useful for the binary type of classification.
For building this models, we are going to use the data flow that is available in Oracle Analytics Cloud. (**Fig. 14**)
As mentioned before, this is step does not require programming or very advanced technical knowledge.

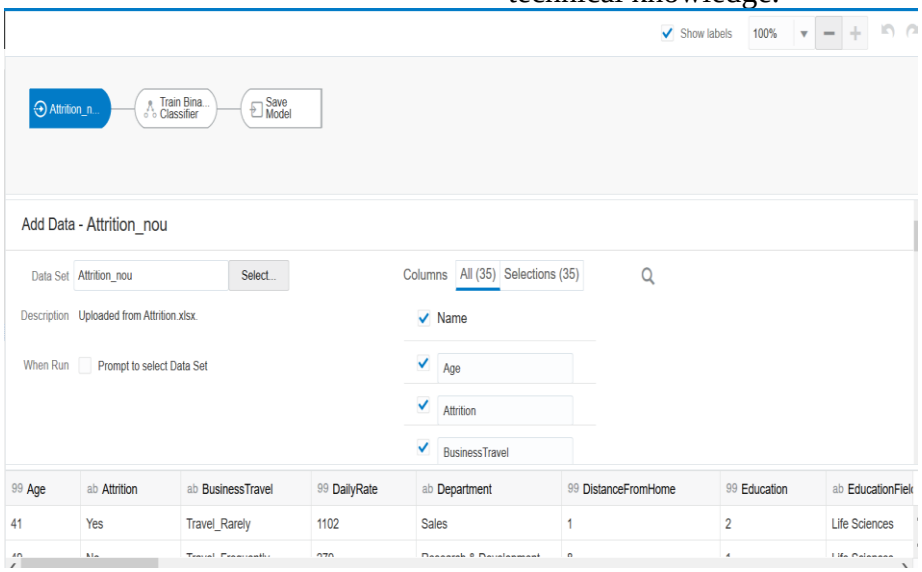



**Fig. 14.** Data flow menu

After applying the chosen algorithm to the data set, we obtain a model that is going to be inspected in order to see how good or bad that model is.

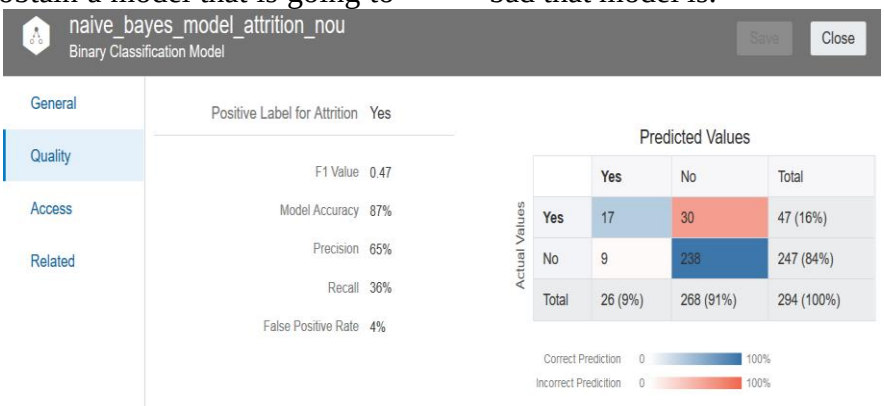



**Fig. 15**. Model analysis for Naïve Bayes

The first one is built for Naïve Bayes and we have an accuracy of 87%, with a precision of 65%, a recall 38% of and a false positive rate of 4%. (**Fig. 15**)

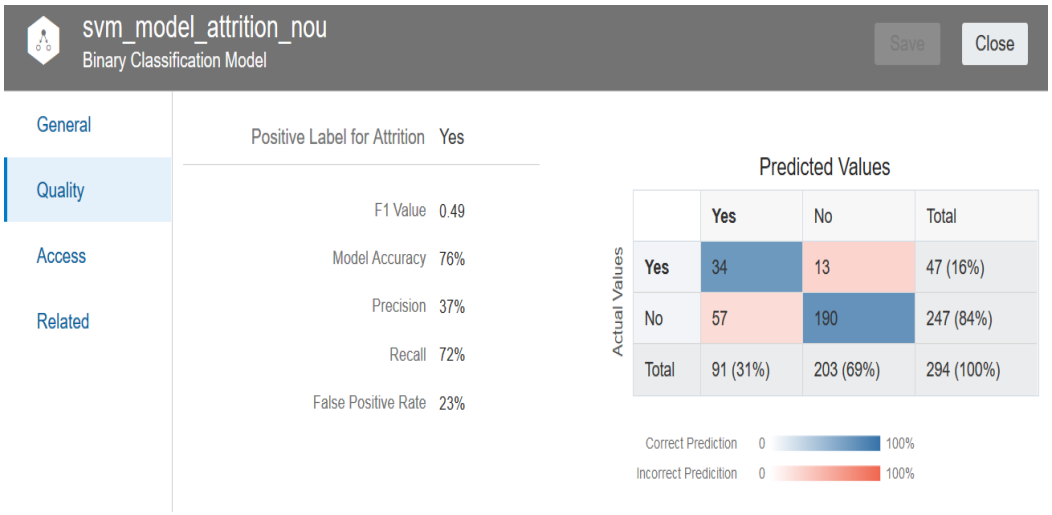



**Fig. 16.** Model analysis for SVM

For the second one we have an accuracy of 76%, a precision of 37%, a recall 72% and a false positive rate of 23%. (**Fig. 16**)

Therefore, in order to choose the best model, we are looking at the fact that the model is predicting attrition. In other words, we want to maximize the recall because we want to predict correctly all truly positive cases.

Lastly, a false negative for these models points out to the idea that we are wrongly going to conclude that a person is going to leave the company, fact that might decrease the chances to prevent a person from leaving.

That being said, we can use support vector machine in the next steps so that to predict new values for possible leavers.

Using the same data flow where, to the initial set of data, the support vector machine model is applied to the data, we are going to obtain a prediction that is visualized in the pictures below.

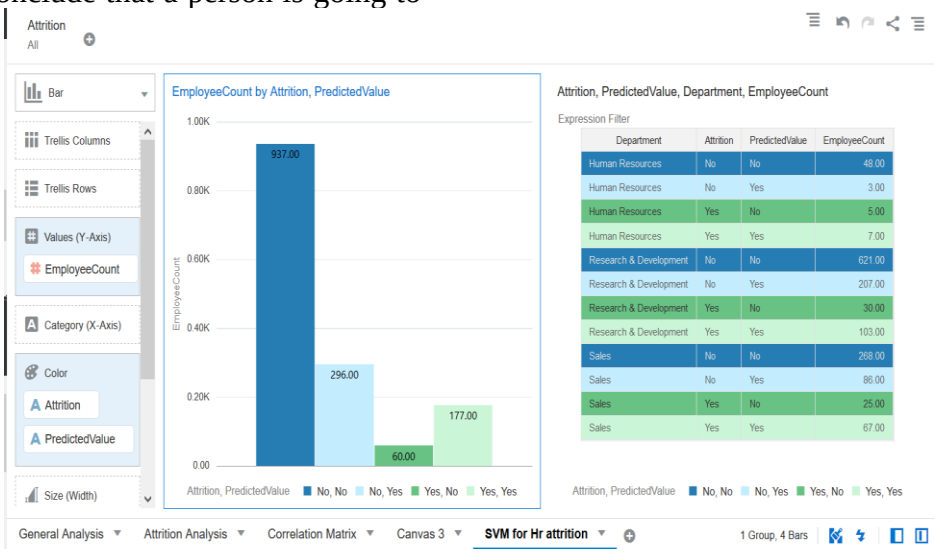



**Fig. 17.** Visualization for SVM model prediction

So, from the total number of people, we can see that most of them were predicted for attrition correctly. 937 have no for attrition and they were also predicted with no, 177 have yes and their prediction was yes too. The interesting part is that, like

remembered in the previous paragraph, 296 were mistaken, but for our analysis is better to think that 296 are going to leave, even though they are not. Finally, the smallest number is 60 for those that are going to leave and they were actually predicted as non-leavers. (**Fig. 17**)

In the next chart, we see the attrition and prediction for attrition grouped by department. (**Fig. 17**)

We see that most of the leavers are from research and development, as well as from sales. These two departments have also the biggest numbers for the false negatives, but the predictions are overall very good.

### 6.5 Use case results and proposal for improvements

We have seen through this use case the population that we analyzed. So, we have different people working in sales, research and human resources, that have experimented different levels of income, working years, type of managers, number of trainings, levels of job satisfaction, work life balance and more.

For these people, we have seen that those with lower levels of income, people that work for little time in the organization, that are little experience or that have been working for less time with a manager are going to be exposed to attrition.

So, in order to prevent this event, a machine learning algorithm of classification is being used so that to predict the possible employees that can leave. We have chosen the best one, more exactly the support vector machine that offered the minimal false negative rate.

With this algorithm we have predicted those employees that might leave and, the targeted organization can now address to them in order to find solution to the existing problems.

As seen from the data used, some of the solutions to propose might include a solid plan for development that includes levels of income, ways of promotion and trainings. Moreover, organizations should adapt to younger people that tend to leave early when something is wrong, on contradiction to those that are older and that have spent many years in just one place.

## 7 Conclusions

First thing to mention in this final part of the paper is the fact that, using all the technologies presented before, a business problem was solved within days and with little technical knowledge.

All things being considered, a BI tool like Oracle Analytics provides us an integrated platform that is going to support the work from preparation till predicting future behavior, facts that are going to help business decision makers to act faster and better in their daily work.

Artificial intelligence and machine learning was useful not only for the suggestion area, the explanation mode or enrichment part, but also as providers of useful algorithms that can be applied right away.

Last, but not least, the cloud has offered multiple benefits in the entire work process. First of all, the permissions and roles part that helps more users to work on the same project or use it at the same time. One administrator can be responsible to create all these users and give them the right privileges. [4]

Second of all, the power offered by the cloud in order to run machine learning algorithms is very important. Considering the fact that a support vector machine classification can take a lot of resources, it is clearly an advantage when we can run in just one minute an entire algorithm in order to build a model and a prediction.

Lastly, the platform can be integrated with other solutions, it can take data from other applications, database or personal computer and everything is going to be in one place.

## References

[1] A. Pyae, "Cloud Computing in Business Intelligence," Asia Pacific

University of Technology and Innovation, November 2018.
[2] https://www.zdnet.com/article/what-is-cloud-computing-everything-you-need-to-know-from-public-and-private-cloud-to-software-as-a/, "What is cloud computing?," Steve Ranger, 2018.
[3] https://www.investopedia.com/terms/c/cloud-computing.asp, "Investopedia Cloud Computing".
[4] A. Banafa, "Ten Myths about Cloud Computing," https://www.experfy.com/blog/ten-myths-about-cloud-computing, 2019.
[5] https://azure.microsoft.com/en-gb/overview/what-is-cloud-computing/, "Azure Microsoft - Cloud computing".
[6] https://www.saksoft.com/information-management-services/business-intelligence/, "Information management services".
[7] https://www.sas.com/en_us/insights/articles/big-data/artificial-intelligence-machine-learning-deep-learning-and-beyond.html, "SAS Big Data and Artificial Intelligence".
[8] R. Clayton, "Oracle Analytics Cloud Succeeds Where Others Fall Short," Oracle Blogs, 2019.
[9] https://blogs.oracle.com/analyticscloud/what-is-augmented-analytics-v2, "What is Augmented Analytics?," Oracle Blogs, 2019.

# A Big Data Modeling Methodology for NoSQL Document Databases

Gerardo ROSSEL, Andrea MANNA
Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación. Buenos Aires, Argentina
grossel@dc.uba.ar, amanna@dc.uba.ar

*In recent years, there has been an increasing interest in the field of non-relational databases. However, far too little attention has been paid to design methodology. Key-value datastores are an important component of a class of non-relational technologies that are grouped under the name of NoSQL databases. The aim of this paper is to propose a design methodology for this type of database that allows overcoming the limitations of the traditional techniques. The proposed methodology leads to a clean design that also allows for better data management and consistency*
***Keywords:*** *NoSQL, Document Databases, Conceptual Modeling, Data Modeling, NoSQL Database developing.*

## 1 Introduction

The need for analysis, processing, and storage of large amounts of data has led to what is now called Big Data. The rise of Big Data has had strong impact on data storage technology. The challenges in this regard include: the need to scale horizontally, have access to different data sources, data with no scheme or structure, etc. These demands, coupled with the need for global reach and permanent availability, gave ground to a family of databases, with no reference in the relational model, known as NoSQL or "Not Only SQL".

The NoSQL databases can be classified by the way they store and retrieve the information [1][2]:

- Key-Value databases.
- Document databases.
- Column Families databases.
- Graph Databases.

The development of conceptual modeling and general design methodology associated with the construction of NoSQL databases is at an early stage [SS17]. of data modeling is to highlight in [3]: "*Data modelling has an impact on querying performance, consistency, usability, software debugging and maintainability, and many other aspects"*

There are previous works on development methodologies we can cite, like the BigData Apache Cassandra methodology, proposed by Artem Chebotko [4][13]. It uses the Entity Relationship Diagram as a conceptual model, but it is oriented to a specific engine, Apache Cassandra. Thus, it is not generic and does not adapt to a design of other NoSQL Databases. Another proposal using a conceptual model for the design of NoSQL is described in [5]. It suggests the use of the various NoSQL databases common features to obtain a general methodology, in which an abstract data model called NOAM is used for conceptual data modeling. Such data model is intended to serve all types of NoSQL databases using a general notation.

Recently, an attempt to generate a universal modeling methodology adapted to both relational and non-relational database management systems was also presented, on the grounds of overcoming the constraints that the entity relationship model has, according to the author [6].

The use of conceptual modeling is also proposed in [7], although the background is not sufficiently studied, such as our work on interrelation of documents and the relationship between them and the conceptual model [8]. They use UML as a tool for the realization of the conceptual model and simple rules to transform it into a

logical model using UML stereotypes.
These efforts show that traditional methodologies and techniques of data modeling are insufficient for new generations of non-relational databases.
It is therefore necessary to develop modeling techniques that adapt to these new ways of storing information. In this sense, this paper will provide the tools to solve these limitations for document database design. As indicated in [14] the methodology should allow: "*describe the data-model precisely*"
The rest of the paper is organized as follows: Section 2 outlines the definition of document database; Section 3 describes the main elements of the methodology and phases of document database develop; Section 4 presents the logical design using the document interaction diagram or DID by extending our previous work: moving from logical to physical model using *JsonSCHEMA* is presented in section 5 and finally Section 6 presents conclusions and future work.

## 2 Document Databases

The proposed methodology is oriented to the design of databases based on documents. A document is a collection of field name and value pairs. The values can be a simple atomic value or a complex structure such as lists of values, another document or lists of child documents.
NoSQL documents are generally referred to as schema-less, which seems to suggest that it is not necessary to make a model before the development starts. The fact that the structure of the data does not need to be defined in advance has many advantages for prototyping or exploratory development, but as data expands and the applications make use of them, the necessity to have them organized in some way arises. In that sense it is more appropriate to say that they are agnostic with respect to the internal structure of the data. It is, therefore, necessary to make a design of the data organization.
to as *schema-less*, which seems to suggest that it is not necessary to make a model before the development starts. The fact that the structure of the data does not need to be defined has a priori many advantages for prototyping or exploratory development, but as data expands and the applications make use of them, the necessity to have them organized in some way arises. In that sense it is more appropriate to say that they are *agnostic* with respect to the internal structure of the data. It is, therefore, necessary to make a design of the data organization.

## 3 Methodology

The proposed design methodology has as its starting point the conceptual model, that can be considered as a high-level description of data requirements. Conceptual modeling is usually performed using some form of entity-relationship diagram ([9]) for conceptual class diagram in UML. Conceptual modeling is intended to describe the semantics of software applications.
In traditional relational database design methodologies, conceptual modeling gave way to a logical design that was later transformed into a physical design. It operates by transforming models from higher levels of abstraction to a model that maps directly into the structures of the database.
Phases of proposed NoSQL document database develop consists of high or conceptual level (conceptual model and access patterns), logical level (types of documents, interrelations and specifications), and physical design in steps like phases of traditional relational database.
In the high-level phase, a conceptual data model is developed in a similar way to the design of relational databases. In the current era, with the emergence of Big Data, the need for conceptual modelling is even more important than before.
As a tool of specification and communication with the other phases, the entity relationship diagram is used (ERD) [9]. In this phase, it is also necessary to

specify the query patterns that have been obtained in the analysis requirements. Query patterns can be specified in natural language or in a more formal language like ERQL [10].

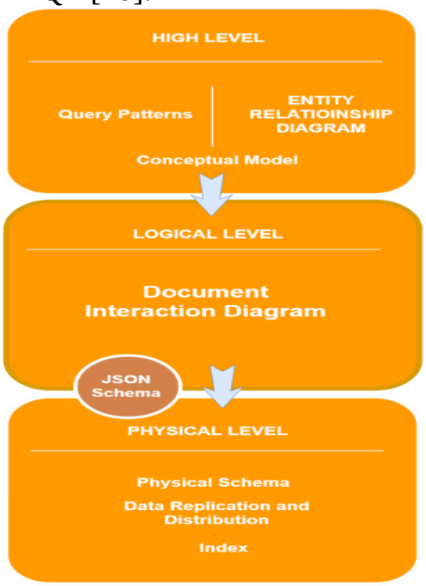



**Fig. 1.** Phases

The Logical Level is the heart of the proposed methodology, in which the types of documents and their interrelationships are established. To represent the logical design, we use a new type of diagram that extends the ERD and that we call document interrelation diagram (DID)[8]. Each type of document is later specified using *JSONSchema*.

There are two ways of relating documents: referencing or embedding. The ability to embed documents allows the designer to store related data as a simple document.

In this way, what is called impedance mismatch can be solved (that is, the difference between the structures of data in memory and the way in which they are stored) [2]. The decision whether to embed or reference is a design decision that is guided by query patterns.

The last phase of our methodology is the analysis and optimization of a logical model to produce a physical data model. In this phase, topics such as index creation, sharding, data distribution, and adapting the data types to the software of the database are considered. The utilization of JSONSchemes is essential in this regard.

## 4 Logical Design

The more important task in this phase is the development of the document interrelation diagram. The DID represents the logical model for a document-based database that captures the classes or types of documents, their structure and interrelation. The documents can be grouped into different classes. Each database uses its own terminology as collections in MongoDB or tables in RethinkDB. we use classes or document types as terminology to indicate a group of documents with similar characteristics.

In the DID each entity of the ERD corresponds to a class or document type, unless it is specifically indicated that this entity will have an independent existence as document type.

In order to exemplify, the entity relationship diagram of Fig.2 will be used. This ERD represents, in a simplified way, the conceptual model of a database that stores orders, products and customers.

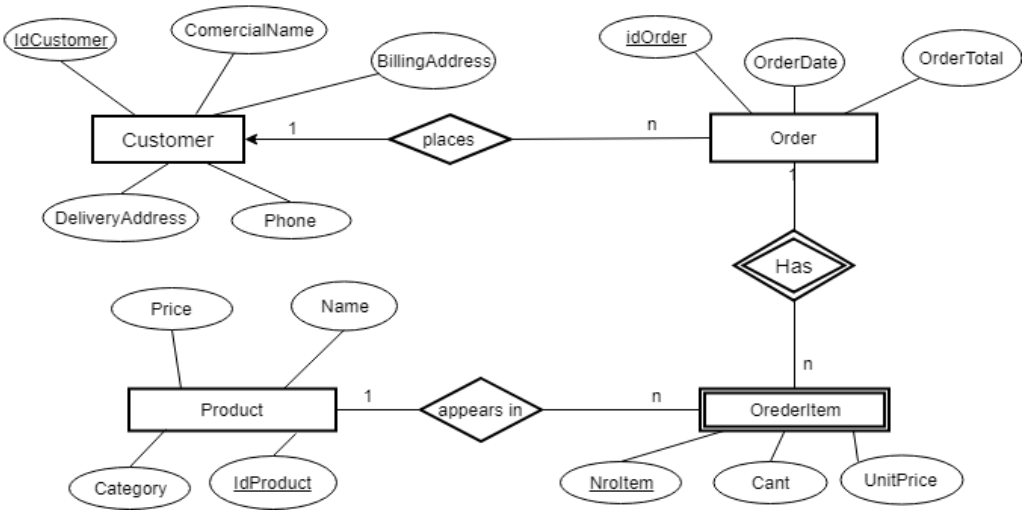



**Fig. 2.** ERD

The entities *Customer*, *Order* and *Product* becomes three document types. *OrderItem* is a weak entity so it is a special case.

To complete the document interaction diagram, it is necessary to decide how the interrelationships will be solved. For this it is necessary to consider the query patterns.

Let's start with the relationship "places". Many design decisions are possible:

- Reference from both sides
- Embed on both sides
- Reference from Order and embed from Customer (or vice versa)
- Embed partially from one side and reference from the other.
- Embed partially from both sides
- Embed total / partial or reference from one side and do nothing from the other

Fig. 3 shows how the reference of both sides is specified while Fig. 4 does the same with embedding of both sides. The arrow indicates reference and curly brackets indicates embedding [8].

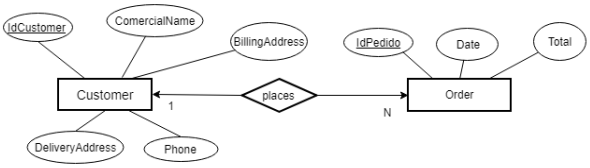



**Fig. 3.** DID: reference

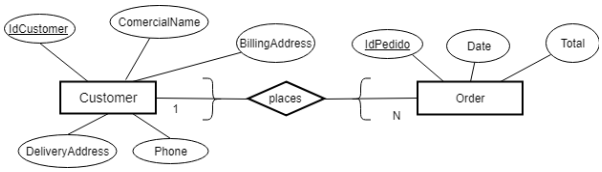



**Fig. 4.** DID: embedding

Embedding simplifies access by minimizing the number of times it should be read from persistent storage. The goal is to keep data that is frequently used together in one document. Although it might be better for a document not to incorporate all the information of the document with which it is interrelated, but only the necessary information that arises from the query patterns.

Suppose that the query patterns indicate that a common way of access to the data is the printing of the order for which the customer's commercial name and shipping address are needed, in addition to all the associated order items. Also suppose that you want to get the dates of the orders made and the total amounts of the same. If the interrelation is solved using only references, the applications are being forced to make several roundtrips to server for to obtain the necessary data. In these cases, a partial embedding can be a better solution.

Fig.5 shows how partially embed is represented. It is necessary to indicate which
fields of the other entity that will be embedded.

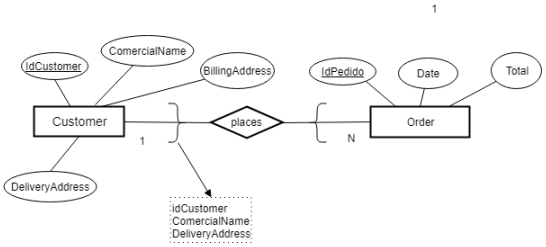



**Fig 5** Embed Partially

Weak entities generally form an aggregate with the strong entity that determines them. It is the case of *ItemOrder* and *Order* in which *Order* can be considered as an aggregate or "a collection of related objects that we wish to treat as a unit" [1].
The simplest way to deal with this is to embed the weak entity in the type of document generated by the strong entity. It is also necessary to indicate that the weak entity will only have an embedded existence, which is done by placing a cross on it as shown in Fig. 6
The cross over any entity indicates that it is not generating a type of document that will be stored independently.

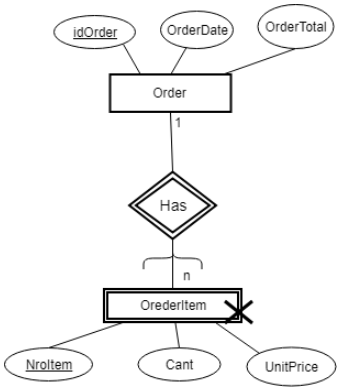



**Fig. 6.** DID: week entity

Although *ItemOrder* entity does not generate a document type, it has an interrelation with the *Product* entity that must be resolved in the logical model. The product information needed in the *ItemOrder* will depend on the domain over which the model was made and what are the access patterns. In this case, it can be assumed that only the name of the product is needed, for which we partially embed the name of the product in the item. When embedding the *ItemOrder* in Order it is embedded with everything it contains including references and embedded fields of other types of documents, in this case the name of the product. The final diagram is as in Fig. 7.

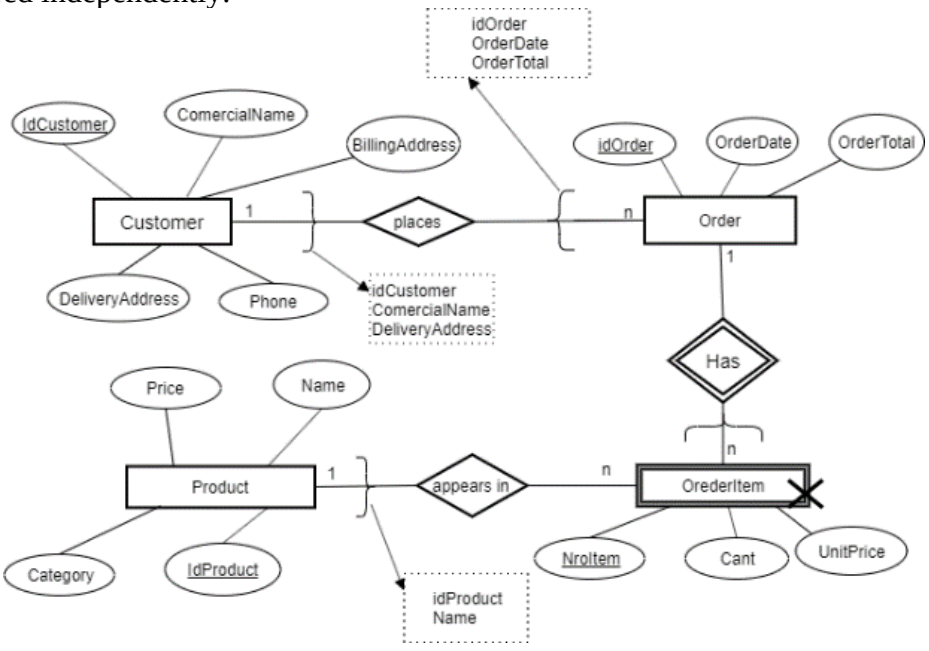



**Fig. 7.** DID

In some cases, it is not enough with the types of documents generated from the ERD to resolve all interrelationships. Assume the case of a database that must save user access to different modules and that a large number of daily accesses are made by each user. The most important query is to know on a given date which modules a user accessed. The ERD in Fig.8 is the conceptual data model.

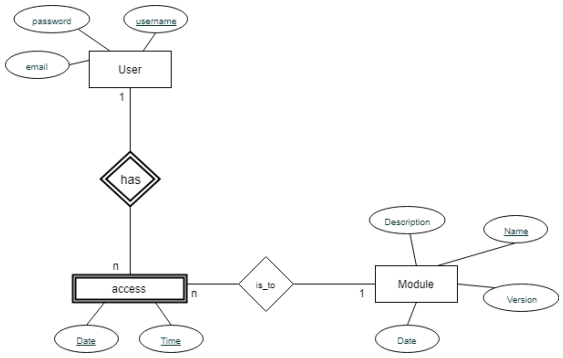



**Fig. 8**. ERD Users and Modules

How to resolve the interrelation between *User* and *Access*? At first glance it seems to be a case like that of the previous order and item. But there are two important differences that change the design decision:

1. The immutability or not of the data: In the previous case, once the order has been sent to the client, the items can no longer be modified. However, in this new domain accesses are added frequently.
2. The volume of data: The items in an order have a limited amount of data. On the other hand, user accesses grow permanently and frequently.

In a document-based database the document is the unit of access, changes in their sizes may generate the need to reorganize the physical space where they are stored, if this is done very often there may be a degradation of performance.

The query patterns in the example indicate that, in general, accesses for a given date are consulted, so it would be a good design decision to divide the accesses by date. Also, once the date is finished, the accesses of the same are immutable. To have a document by date it is necessary to create an auxiliary document type. Fig. 9 shows how that document is specified.

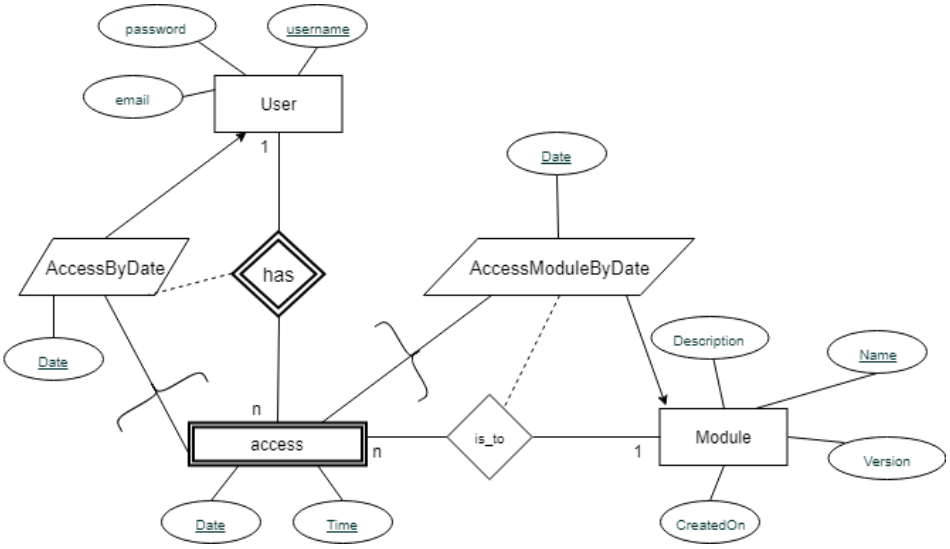



**Fig. 9.** DID: Partition

The new document that does not correspond to any entity of the ERD is drawn as a parallelogram with two inclined sides. It is also necessary to indicate which interrelation that document is representing that is achieved with a dotted line from the interrelation to the symbol of the intermediate document.

The auxiliary document has on one hand a reference to the user and on the other it embeds the accesses. The key will be the date and user id. We must explicitly mark as

a key the *Date* taken from the accesses to indicate that it is the partition key and therefore there is a single date per document, the user identifier does not need to indicate it since the arrow indicates reference to the key of the user and also the cardinality of the user-measurement relationship indicates that the measurements are of a single user. It is not necessary to keep the measurements as an independent document, so the cross is placed on that entity.

The extended entity relationship diagram also supports hierarchies between entities.

The hierarchies in the ERD can be with full or partial coverage, with overlapping or without overlapping. The possibility that documents of the same type have different schemes facilitates the design. We can generate a single type of document corresponding to the super-entity that also has the attributes of the sub-entities. For this, it is enough to indicate that the sub-entities do not generate a type of document as seen in the Fig. 10.

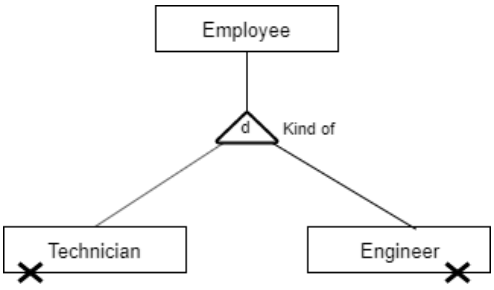



**Fig. 10.** DID Herarchies

Depending on the pattern of consultations, other decisions may be made:

- Mark the super-entity as not generating a document type and then generate one for each sub-entity. This is possible if the hierarchy has no overlap.
- Specify that both super-entity and sub-entities generate one document type each. Indicating which attributes would be placed in super-entity.

Another type of relationship that is necessary to model is ternary relationship.

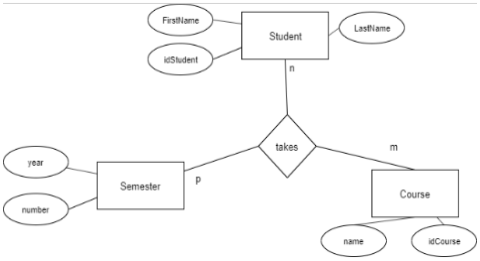



**Fig. 11.** ERD: ternary relationship

Suppose a ternary relationship between *Student*, *Semester* and *Course* entities. The cardinality in this case is n:m:p, for a student and a semester there are many courses he takes, a semester and a course has many students enrolled, for a course and a student can be many semesters where he takes it. The DER of Fig. 11 shows this relationship.

The most complex part is deciding how to model the relationship takes. The decision on how to model will, as always, depend on the query patterns. The basic case is to generate a type of document that simply contains the information of the relationship with the identifiers of each of the entities involved. To do this, an auxiliary document is drawn with the name of the new document type and a dotted line that binds it to the entity as seen in Fig.12.

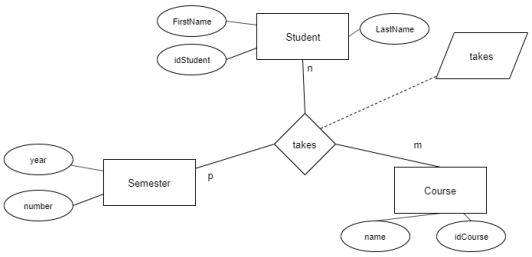



**Fig. 12.** DID: ternary relationship

That is the simplest model, but suppose that a very common query is to know which students are enrolled in a course in a semester, in fact you want to know first name, last name of them for a given course and semester. While the previous model

allows you to answer this query, you might decide to have a document type that stores the complete information to optimize access to it.

The semantics of this diagram (Fig. 13) are that only key attributes are added or that allow you to group data from another entity from those participating entities in the interrelationship that are not related by any link to the new document type.

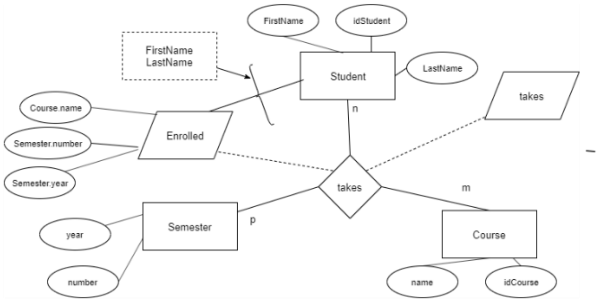



**Fig. 13.** DID: complex ternary relationship

One case to consider is when it becomes necessary to group multiple instances of an entity, by one or more attributes, into a single document. To exemplify let's assume a part of a DER where users and their searches are modeled.

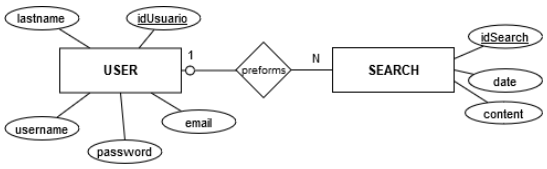



**Fig. 14.** DER: User/ search

The relationship between user and search can be modeled in various ways, either by embedding searches in the user or by referencing. The relationship could also be resolved by partitioning by user and date in the same way as shown in Figure 9 for user and access. Let's say that a very frequent query is to know the searches performed on a given date. The solution of partitioning by user and date is not efficient for this because access should be made for each user who has a search on that date. In this case, the ideal is to have a single document with all the searches for a date. This would involve grouping by the date attribute, i.e. generating a document for each date that has all searches. An auxiliary document should be created to save all searches with the date as key. The notation is similar to that seen before, although in this case the auxiliary document refers only to the entity on which it is being grouped.

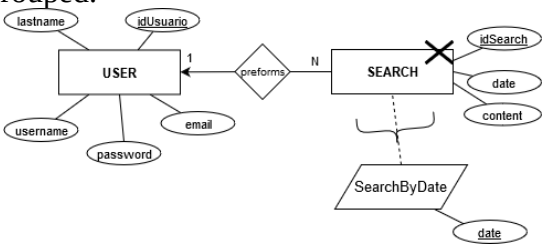



**Fig. 15**. DID: User/Search

Figure 15 shows the corresponding DID. Note that the reference from *Search* to *User* is important, because marking the entity as not generating a type of document would lose the relationship.

It is also possible to generate an intermediate document to resolve the relationship between *User* and *Search.* There would be data redundancy in favor of access speed. The complete DID is shown in Figure 16.

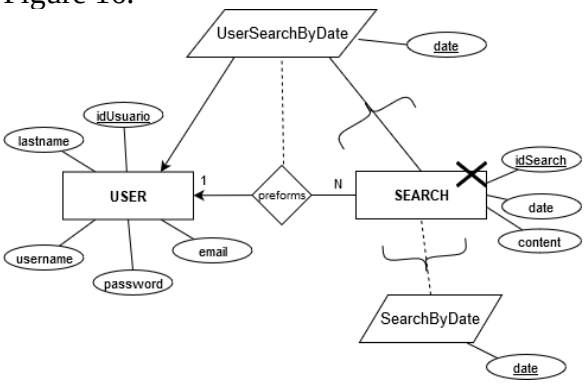



**Fig. 16.** DID: Complete User-Search

## 5 From logical to physical level

Upon completion of the development model interrelationship of documents, which is equivalent to logic design relational database, it continues with the physical design.

The physical design implies making decisions about specific aspects of implementation such as: data distribution, index generation, use of engine facilities of the selected database, etc.

Many document databases support indexes.

Index creation must be based on query patterns. It's about doing a trade-off so you don't have a few indexes that could lead to poor read performance, but not so many that affect the write performance.

The use of *JSONSchema* for a more detailed specification of each type of document facilitates decision making process and implementation. A *JSONSchema* is a JSON document which describes the structure of another document.

The steps to follow are as follows.

1. For each document type in the DID:
    a. Define the appropriate data types for each attribute
    b. Write the specification using *JSONSchema*.
2. For each query analyze the ease of documents to respond to it. Ideally a single access should be enough for the most used queries.

From the DID each type of document is mapped to a *JSONSchema* which allows to specify in detail the structure of each document. For example, the document type *AccessByDate* in Fig. 9 is mapped to the the following scheme:

```
{"title":"AccessByDate",
 "type":"object",
 "properties":{
            "userId":{"type":"integer"},
            "date":{"type":"string","format":"date"},
            "accesses":{ "type":"array",
            "items":{ "type":"object",
                        "properties":{
                                    "modulename":{"type":"string"},
                                    "timestamp":{"type":"string",
                                    "format":"date-time"}
                                    }
                        }
                }
    }
```

From the DID in Figure 16 JSONSchema will be generated for each of the following document types:

*User*: With the attributes in the diagram, specifying the type of each.

*UserSearchByDate*: having the *userid* and *date* as keys and a vector with that user's searches on that date.

*SearchByDate:* The key is the *date* and has a vector with the searches and in each the corresponding *userid*.

No other document types are generated. By indicating that an attribute is key we are claiming that it is unique and that it identifies each document, even though the database always generates an identifier attribute.

The flexibility of the *JSONSechema* to establish optional properties makes it an ideal tool for specifying document types of variable structure. In the case of hierarchies this facility is extremely useful because you can specify conditions for which an attribute exists or not. Looking at *JSONSchemas* it is possible to realize that in some case it is convenient to reserve space the same in such a way that the document does not resize it during its lifetime. If the document grows larger than the size allocated for it, the document may be moved to another location with the consequent input/output cost [12].

Some document-based databases have tools to validate if a document complies with a *JSONSchema.*

## 6 Conclusions

A methodology that allows obtaining a detailed design from a conceptual model has been presented. This work extends and completes previous work on document modeling in the design process.

The proposal presented allows flexibility to establish detailed design decisions. There is not currently, to the best of our knowledge, complete methodology such as that presented for document-based databases that have the same level of flexibility and specification capability.

The presented methodology was used successfully in several developments using different database engines. In future work we plan to report in detail the cases of success in the use of this methodology.

## References

[1] Adam Flowler, “The State of NoSQL”, 1st edition, 2016

[2] Pramod J. Sadalage, Martin Fowler, “NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence”, *Addison-Wesley*

[3] Gómez, P., Casallas, R., & Roncancio, C. (2016). “Data schema does matter,

even in NoSQL systems!" 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), 1-6.

[4] Artem Chebotko, Andrey Kashlev, Shiyong Lu, "A Big Data Modeling Methodology for Apache Cassandra", IEEE International Congress on Big Data (BigData'15), pp. 238-245, New York, USA, 2015.

[5] Francesca Bugiotti, Luca Cabibbo, Paolo Atzeni, Riccardo Torlone. "Database Design for NoSQL Systems". *International Conference on Conceptual Modeling*, pp. 223 - 231 Atlanta, USA, Oct 2014.

[6] Ted Hills, "NoSQL and SQL Data Modeling", *Basking Ridge, NJ: Technics Publications*, 2016

[7] Shin, K & Hwang, C & Jung, H. (2017). "NoSQL database design using UML conceptual data model based on peter chen's framework". International Journal of Applied Engineering Research. 12. 632-636

[8] Gerardo Rossel, Andrea Manna, "Diseño de Bases de Datos Basadas en Documento: Modelo de Interrelación de Documentos" *XIII Workshop Bases de Datos y Minería de Datos. Congreso Argentino de Ciencias de la Computación* CACIC 2016 San Luis Argentina..

[9] Peter P. S. Chen, "The entity-relationship model: toward a unified view of data", *Proceedings of the 1st International Conference on Very Large Data Bases*, ACM, New York, NY, USA, 1975.

[10] M. Lawley and R. W. Topor, "A query language for EER schemas," in *Proceedings of the 5th Australasian Database Conference*, 1994, pp.292–304.

[11] Storey, Veda & Song, Il-Yeol. (2017). Big data technologies and management: What conceptual modeling can do. Data & Knowledge Engineering. 10.1016/j.datak.2017.01.001.

[12] Dan Sullivan. 2015. NoSQL for Mere Mortals (1st. ed.). Addison-Wesley Professional

[13] Jeff Carpenter & Eben Hewitt. (2020). Cassandra: The Definitive Guide (3st. ed.). O'Reilly Media, Inc.

[14] Pivert, Olivier. NoSQL Data Models: Trends and Challenges. 2018. Wiley-ISTE

**Gerardo ROSSEL** graduated as Ms. Sc. in Computer Science from the Faculty of Exact and Natural Sciences of the University of Buenos Aires. He has a Doctor's degree from the National University of Tres de Febrero. At present, he is an assistant lecturer at Computer Department of FCEyN UBA. He has more than 20 years of experience in software industry and is Chief Scientist of UpperSoft software company. Her specific area of competences is in Databases, NoSQL, Data Science, Machine Learning, Patterns and Software Architectures, Epistemology and Philosophy of Computer Science. He is co-author of book "*Algoritmos, Objetos y Estructuras de Datos".* He has published several papers in national and international conferences and journals. He was a member of the International Program Committee of several international conferences.

**Andrea MANNA**, graduated from the Faculty of Exact and Natural Sciences of the University of Buenos Aires in 2000. She got the title of Ms. Sc. of Computer Science. At present, she is assistant lecturer in the Faculty of Exact and Natural Sciences of the University of Buenos Aires. She has been working in the software industry since 1995. She is Chief Software Architect of UpperSoft Sofware Company and work in software development for more than twenty years. She is co-author of book "*Algoritmos, Objetos y Estructuras de Datos".*

# Learning view over the implementation of business process optimizations

Radu SAMOILA
Bucharest University of Economic Studies, Romania
radusamoila2@gmail.com

*Current context of great development and changes in the technological matters, national and global economies have to keep pace and undergone major changes. The ultimate aim of the companies and organization is to improve or optimize its business processes to cope with increased competitiveness in order to deliver more efficiently better products or services. The success of current businesses is linked to the efficiency but also the effectiveness of their core processes. Most part of the latest researches have recognized this importance leading to efforts concentrated on analyzing and optimizing the business processes. There are many techniques which are considered but also others which are not causing potentially significant opportunities of improvement not being addressed. However, currently, there is a scarce of an universal technique or methodology that can be used by the organizations to address business optimizations. This paper addresses different topics on how companies are analyzing the decision to initiate a business process optimization and how optimizations landed within organizations.*
***Keywords:*** *business process, optimization, efficiency and effectiveness of internal processes, business automation, optimization methodology.*

## 1 Introduction

Business processes have received ample attention during the last years. Many approaches and techniques have been discussed and proposed, there were many promises made, but the spectacular results that the reengineering and optimization revolution vowed were never fully realized. This made more and more people hesitant about the whole concepts.

As defined by the main relevant literature publications, Business Process Optimization ("BPO") is the concept of redesigning the internal processes to promote efficiency and effectiveness in order to strengthen the alignment of individual processes with the overall strategy of the company. While the optimization of a singular process or of the processes in a particular business function could trigger real business improvement, organizations that gather their efforts across the entire organization can see significant competitive advantage, better customer service (internal and external), and much more efficient operation.

## 2 Business processes' optimization approaches

Zhou and Chen [1] suggest that business process optimization should aim at reducing lead time and cost, improving quality of product and services, and enhancing the satisfaction of customer and personnel so that the competitive advantage of an organization can be maintained. Reijers [2] suggests that the goals of business process optimization are often the reduction of cost and flow time. However, Hofacker and Vetschera [3] underline that the concept of "optimality" of process designs is not trivial, and the quality of processes is defined by many, often conflicting criteria.

Zhou and Chen [4] remark that there is still no structured optimization methodology or technique for business processes. Optimization is not an option for diagrammatic process models. This is because optimization requires quantitative measures of process performance that cannot be offered by diagrammatic models. However, there are many qualitative improvement approaches applied to diagrammatic process models such as that by Zakarian [5] and Phalp and Shepperd [6]

to name a few. **Table 1** summarizes the main business process optimization approaches identified in literature, mostly related to Petri nets and mathematical process models. Taking into consideration the emphasis that has been put on Petri nets for their analysis capabilities, one would expect that they would also fit for optimization purposes. But, according to Lee [40], Petri nets are not appropriate to solve optimization problems except using graph reduction techniques. Although they can capture system dynamics and physical constraints, they are not suitable for optimization problems with combinatorial characteristics and complex precedence relations.

*Table 1:*

| MODEL of business process | modelling SET(S) | TYPES of business process optimisation | APPROACHES to business process optimisation |
|---|---|---|---|
| –Petri–nets (and workflows) | –Diagrammatic models<br>–Mathematical/formal models | –Graph reduction techniques | – (Sadiq and Orlowska, 2000)<br>– (van der Aalst *et al.*, 2002)<br>– (Lin *et al.*, 2002) |
| –Mathematical models | –Mathematical/formal models | –Algorithmic approaches | – (Han, 2003)<br>– (Gutjahr *et al.*, 2000)<br>– (Jaeger *et al.*, 1995)<br>– (Hofacker and Vetschera, 2001)<br>– (Soliman, 1998)<br>– (Tiwari *et al.*, 2006)<br>– (Vergidis *et al.*, 2006)<br>– (Volkner and Werners, 2000)<br>– (Zhou and Chen, 2003a)<br>– (Zhou and Chen, 2002)<br>– (Zhou and Chen, 2003b) |
| | | –Activity/Task consolidation | – (Dewan *et al.*, 1998)<br>– (Rummel *et al.*, 2005) |

Zhou and Chen [1] developed a structured design methodology for business process optimization from strategic, tactical, and operational perspectives using quantitative methods that support the design. This optimization optimally assigns resource capabilities, organizational responsibilities and authorities, and organizational decision structure. Another approach to optimization is the consolidation of the activities (or tasks) of a business process. Rummel [9] proposes a model that focuses on decreasing the cycle time of an internal process by consolidating activities—assigning multiple activities to one actor—thereby eliminating the coordination and handoff delay between different activities that occurred when assigned to different actors. As this approach is activity focused, it ignores interactivity delay that may contribute significantly to overall process cycle time. Dewan [10] claims that there is no structured methodology to determine the optimal re-bundling of information-intensive tasks. They present an approach to optimally consolidate tasks in order to reduce the overall process cycle time. The authors present a mathematical model to optimally redesign complex process networks but a limitation of the paper is that it refers to business processes with information flows only. Its main contribution is the effective business process restructuring and the reduction of the overall task time using handoff delay reduction or elimination as a result of a unified methodology applicable to multiple task-based business processes. Although formal languages have associated analysis techniques that can be used for investigating properties of processes, an optimization approach based on executable process languages was not observed in the literature. Since most of the optimization approaches—as discussed before—are based on algorithmic methods, these could be easily translated to executable software

programs. Analysis and optimization of business processes can be done best using an approach based on explicit and executable process models. Such models would allow evaluating performance in terms of flows, calculating costs against objectives, recognizing constraints, and evaluating the impact of internal and external events. Therefore, by being able to assess the process execution quality and costs, it is possible to take actions to improve and optimize process execution.

## 3 Key factors in financial and operational optimizations

There are various warning factors that signal the degradation of business processes. These factors are triggered either from the internal or external environment of the business. The signals are usually presented before the company enters into a crisis. There are cases where managers do not observe these signals or consider them to be some one-off or periodic difficulties. This approach does not only aggravate the outcome of the business, but, on a long term, threaten the existence of the organization itself [7]. Most companies agree to implement a change management plan. The overall review of the process can be split into five stages, as follows [8]:

- Stage 1: Analysis, usually takes from one week to one month.
- Stage 2: Planning – takes from one to three months.
- Stage 3: Implementation - six months to one or more than one year.
- Stage 4: Monitoring - six months to one year.
- Step 5: Return to business growth - from one year to two years.

### 3.1 Corporate approvals for business optimizations

The preparation of an effective restructuring plan (Slatter, Lovett, Barlow, 2006) is based on the following elements: crisis ending and business stabilization, appointment of a new director, stakeholders' management, strategic orientation, critical improvement process, implementation of organizational changes and financial optimization.

Six major milestones must be attended by an organization: (according to Downey, 2009):

**Step 1:** Changes in the management structure. It refers to bringing a new CEO or an external specialist. Involves the board of directors or senior management to recognize that a change is required and initiate a corporate review program.

**Stage 2:** Business Review. Rapid identification of the problems faced by society and assessment of business survival chances: strategy, operations, finance, infrastructure, people, commitment and ability to change.

**Stage 3:** Business Restructuring Plan. Establishing appropriate strategies and a well-structured recovery plan to deliver lasting results.

**Stage 4:** Implementation. Organizations can use sharp actions to save the company's performance: layoffs, department dismantling, and drastic cuts in all nonessential costs. Positive cash flow is critical and needs to be set quickly. In addition, the cash will be needed to implement the review strategy and must come in a timely fashion.

**Stage 5:** Stabilization. In this stage, the main aim is to increase the efficiency and the effectiveness of the business operations the focus is on. It is necessary to improve the profitability, but also to ensure the good functioning of the existing technologies.

**Stage 6:** Implement the change(s). The final stage is to implement the planned change, the organization / corporation restoring its financial loneliness. At this stage motivates staff and employees to achieve profitability and return on investment.

### 3.2 General corporate model for business restructuring through optimizations

Strategic analysis, monitoring, and strategic planning of the organization's/ corporations' activities can be monitored and controlled on the basis of several methods, models and

process diagrams. One such method is the **Critical Success Factors** method provided by Rockart (1979), which is based on the 80/20 Pareto rule for strategic management needs. The method involves identifying the critical success factors of the company as the results in important areas to ensure corporate success. This method identifies the most important business processes and the performance indicators (or KPIs) of these processes. It allows the plan to be compared with the results obtained, as well. The effectiveness of strategy implementation should be measured continuously in order to ensure continuous improvements of the operations. Key areas of corporate restructuring include: sales, finance, production and supply chain, management activities, services, business development, organization and human resources. Supply and logistics are usually embedded under the production processes or activities.

## 4. Managing Processes versus Projects

One important aspect is to make the difference between processes and the projects. Business processes consist of providing value to a customer through value-added activities, moving work across functional area boundaries, and controlling process performance indicators and standards and measuring process execution. Business processes are usually driven by facts or events, such as the maintenance of a factory, printing a product catalogue, the close of a billing cycle, or solving customers' issues in reconciling a checking account. These are activities that are typically replicated and repeated with specific resources allocated to an individual steering group such as factory line workers or customer service employees, to give some examples. Business processes are looking to the following core features: efficiency, agility and meeting customers' demands. While efficiency seeks to cut operational costs and cost of capital, agility strives to cut the time required to develop products and services, and to respond to customer and market demands (thus through improving the effectiveness). Customer demands focuses on retaining the customers and their overall level of satisfaction.

A project, as defined by the Guide to the Project Management Body of Knowledge (PMI, 2004) is represented by a series of activities, and related tasks with a dedicated objective, bounded under a starting and an end date, and resources. A project consumes cash, people, time and equipment for the specified time period and defines what is planning to be done, when to do it, and ensures that the planned results are reached. The tasks of a project are unique and usually not repeated, and once the project is planned, changes to the plan are avoided to ensure meeting the schedule. While business processes look for efficiency, agility (effectiveness), and meeting customer demands, projects look to deliver the related objectives within established budget and time boundaries.

Thus, a processes optimization project is a short to medium term effort an organization puts upon to identify all necessary inefficiencies/ redundancies, decide the action plans to resolve them and ensure the optimization enhancements will meet the desired results without any negative outcome for the current operations.

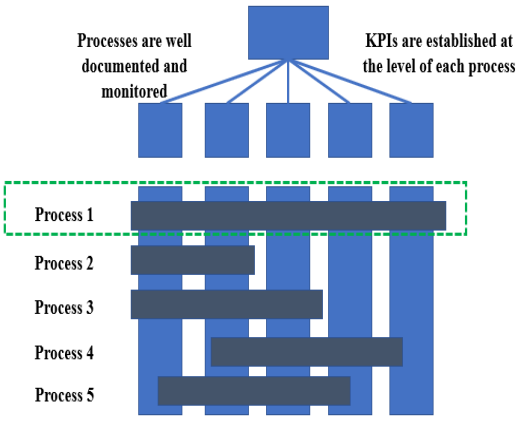
### 4.1 Keys to Optimizing Processes

The executive management have to be equipped with the necessary tools to make the right decisions to realize the organizational course corrections with agility. The keys to optimizing process performance and execution capability are tied to the organizations' commitment to define and continuously assess and update the documentation of internal business processes. These documentations, including process maps, inputs and outputs, resource allocations, cycle times, etc., formally define the scope of the process from initiation to delivery and serve as the "process

roadmap".

Once the decisions are made, the Business Process Optimization ("BPO") projects concentrate the organization to identify and scrutinize opportunities to reduce costs, eliminate waste and reduce cycle times, while increasing products and services quality. There are cases when the vast majority of organizations continue to operate in silos. This is where the core business processes' activities must cross the traditional functional view and have neither an owner assigned nor measures of process execution success. Many successful organizations have mature business process management structures where the core and support business processes are well-defined, documented and assigned to an owner, measured for execution success, and scrutinized for efficiency, agility, and quality. Process management focuses on the management of the cross functional processes. This involves continuous monitoring, evaluation and measurement in terms of costs, quality, cycle time, etc. (Figure 1 below).

Fig. 1




1. The management of processes is clearly and structurally defined.
2. The performance of each process is monitored and looked after continuous improvements (response time, gaps, delivery time, etc.)

Formally defined processes and related documentation bring an organization with visualization of high-level and detailed core, supporting and management processes. Processes maps and supporting detail documents define how the processes look like, their interactions across business functions, what each the process step delivers and how it produces its intended deliverables. Process owners own the core or supporting business process, not the individuals assigned to work the tasks within the process. The process owner is responsible to ensure that the execution of the process is successful and that it works to identify process issues, root causes and training needs. In order to manage a process that will bring continuous successful execution, process communications must ensure that process the information flow is established vertically and horizontally across the organization. A process owner must be able to monitor the external and internal flows of information. The purpose of process communication is to make sure that all employees are informed of process performance information and to control the company's progress toward its objectives and goals.

### 4.2 The Efficiency should never compromise the Quality

For many consumers, the quality of the products and services is as important, if not more important than the cost of the product or service. While the focus of many business process optimization projects is centered on optimizing efficiencies and reducing the cycle times, businesses must continue to ensure that the business process optimizations does not compromise the quality of the product or services that are delivered by the internal processes. Six-Sigma methodologies are usually used to embed quality into process optimization projects. Six Sigma originated at Motorola in the early 1980's and is a methodology for disciplined issues solving and quality improvement. Six Sigma's goal is the near elimination of defects from any process, product, or service, limiting defects to just 3,4 defects per million opportunities. To ensure organizational alignment, Six Sigma methodology requires all improvement projects must be integrated with the goals of an organization. The DMAIC methodology

Six-Sigma (The Black Belt Memory Jogger, 2002) employs the following activities:

- Define: the phase whereby the customer needs are established and the processes and products to be improved are identified.
- Measure – determine the baseline and target performance of the process, defines input and output variables of process steps, and validates the measurement methods.
- Analyze – analysis of data to identify critical factors required for process execution.
- Improve – identification of necessary improvements/ optimizations (process, procedural, systemic, etc.) to optimize the outputs and eliminate and or reduce defects and variation. Statistically validates the new process operating conditions.
- Control – establishes the development of documents, monitors, and assigns overall responsibility for sustaining gains made by the implementation of process improvements.

## 5 BPO Project Management Methodology

Business Process Optimization projects follow the same method as defined by the Project Management Institute. Project management is obtained through the use of the process group such as: initiating, planning, executing, controlling, and closing. As stated in the Project Management Institute's Guide to the Project Management Body of Knowledge 2000 edition, "these process groups are linked by the results they produce- the result of one often becomes an input to another." Specifically:

- Initiating- approving the project or phase is part of the scope management;
- Planning- defining and refining objectives and selecting the best of the alternative courses of actions to reach the objectives that the project was undertaken to address.
- Executing- coordinating people and other resources to carry out the plan.
- Controlling- ensuring that the project objectives are met by monitoring and measuring the KPIs periodically to identify variances from plan to identify corrective actions that can be taken, when necessary.
- Closing- formalizing acceptance of the project and bringing it to an orderly end.

### 5.1 Combining Six Sigma and Project Management Best Practices in the Initiating and Planning Process Groups

One of the single, most critical activities to ensure the success of a project, whether it be in the development of a software application, drug compound or optimizing a key business process, is the clear and concise definition of project objectives, goals and milestones in the projects planning phase. The purpose of the project should support the vision and mission statements of the organization and it requires the support and commitment of the management. Business process optimization projects should contain a section in its charter that defines the specific business process to improve. This formalized definition of the process optimization scope eliminates any confusion and formally defines the subject boundaries. Additionally, it assists in the identification of the established deliverables. For example, a fulfillment organization receives customer complaints on low order fill level. The customer places an order for a quantity of 100 for a particular item and receives only a quantity of 90. The project objective could be "to optimize the warehouse picking process to ensure an increase in the fill rate on customer orders from 90% to 99% by 4th Quarter 202X". The process scope has been narrowed specifically to the picking process and provides the basis for the process goal.

Six Sigma is a data driven problem solving methodology that requires the formal definition of performance key indicators. When planning for a process optimization

project, specific Six-Sigma tools and activities are used to characterize customer needs, and processes to be improved. These tools include the mapping of the high-level process in its current state, identification of the processes existing performance measures (i.e., pick time, product staging time) and a process financial analysis (i.e., resource cost, overhead). Specifically, Six-Sigma seeks to identify the Costs of Poor Quality (COPQ). COPQ includes costs of repairs, rework, rejections, inspection, testing and in the case of our fulfillment example, the cost of customer complaints. While a process optimization project's benefit can be measured financially (hard) or non-financially (soft) most business cases are based on the hard benefits. In the above example, the soft benefits of "improved customer satisfaction" should be considered as well.

While discussion of the "customers", Six-Sigma projects take the time to understand the needs of the customers. The project team must understand how the process issues link to the eventual customers. Six-Sigma mentions about the "*voice of the clients"* research to gain this important insight. There are many different methods to researching the customer's voice. These include, but are not limited to the following:

- Customer Complaint database- this is an acceptable place to start if the organization formally tracks issues;
- Direct Contact- if allowed, considers phone call surveys, focus groups, interviews at the point of provision.
- In-Direct Contact- includes mail surveys, feedback cards, market research and competitor analysis.
- Becoming the customer- order from your own distribution center, buy your own brand products, set up a new account with your own financial institution.

Another effective tool to use in a process optimization project is the SIPOC High Level Process Mapping tool. The acronym SIPOC stands for Suppliers, Inputs, Process, Outputs, Customer. It is a simple, but effective tool to align the project team and all stakeholders as to the core process within the scope of the project. It is important to mention that it is too early in the project to mention the existing process (that comes later in the Measure Phase).

The general approach to the SIPOC process identification includes the following steps:

- Begin with a simple definition of the in-scope process;
- Identify key steps of the process (expand these at the bottom of the SIPOC diagram);
- Have the project team identify the major inputs and outputs of the process;
- Have the team identify key suppliers of the inputs, and customer for each output.

Accordingly to any other project, a business process optimization project requires the formal identification of a project team with clear structure, roles and responsibilities. It can be used the SIPOC High Level Process Map to ensure all process stakeholders are represented on the main project team.

The Initiating and Planning Phase of a business process optimization project starts by formally identifying the process problem, not with the identification of the solutions to the problems/ issues. Six-Sigma tools such as the SIPOC, COPQ, and VOC help the project team identify the potential issues, process scope and essential process representatives, before the organization invests substantial time and money in the initiatives.

### 5.2 Measuring and Analyzing Current Process Performance

During the execution phase of a BPO project, the project manager is concentrated on executing the process optimization plan. These integral activities include the development of individual and team skills through the use of various team building exercises, reward and recognition systems and locating team member in the same physical area. The project manager is also focusing efforts to ensure the process

optimization plan is being carried out through regularly scheduled status meetings to exchange information about the specific project. During the execution phase, team efforts are focused in the identification of measurements to determine the effectiveness and efficiency of the process. There is necessary to develop the process measures which are critical for the process optimization project. It must identify and capture data on key performance indicators to determine process effectiveness and efficiency. Process effectiveness measures a customer's quantifiable service or product specifications. In addition, a process optimization project must track key performance indicators that reflect the internal efficiency of the process. In general, the following main steps are completed to measure the performance of a business process:

- Develop a data collection plan for the process;
- Identify process efficiency data collection sources;
- Identify process effectiveness data collection sources;
- Collect efficiency and effectiveness data to determine process performance baseline measurements.

### 5.3 Controlling Key Business Process

Following the development and testing of systemic, procedural, or responsibility enhancements, the BPO project team efforts should focus on ensuring the solutions are implemented and measured for their effectiveness. The project team must identify measures to be monitored after the desired state process is landed. This activity includes the identification of the persons responsible for collecting and analyzing the process data and reporting process efficiencies and effectiveness to the entire organization in the form of dashboards or status reports. Six-Sigma projects typically employ Statistical Process Control charts that monitor the stability and variation of a particular process. A typical Statistical Process Control chart tracks the performance of a process over time and shows control boundaries which the results will lie between if the process is “in-control”. Use of any Statistical Process Control charts require regular updating and review to ensure their feasibility. This ensures that process performance doesn't decline again.

Process change control is another key that ensures ongoing alignment with an organization's strategic goals. Processes are enabled by technological change, not hindered and that the appropriate organizational structure is in place to provide resources to support the business process. As documented in the Six Sigma Black Belt Guide (2001), a classical model for managing the change process has three phases: (1) unfreezing, (2) movement and (3) refreezing. Once a process change is identified and ready for deployment, the “unfreezing” of existing behavior patterns must be addressed. Typically, most work groups are resistant to changes and this must be solved. People or practices must then be moved (movement) to the process change by training or through technology adoption. Once, process resources have acquired the necessary skills and technology is in place, the process is then *refrozen* to ensure the process or function is aligned for organizational effectiveness. One effective technique used to facilitate the transition from existing processes to the new process is the use of a formal “White Paper Fair” where all functional areas impacted by the process changes have an opportunity to visualize the process enhancements.

## 6. Implementation of optimization processes- market study results.

Through a questionnaire developed according to the basic conceptual model of the project optimization implementation, process includes the following major specific modules, namely:

- Changes in business processes;

- Optimizations targets (planned and achieved);
- Optimizations implementation issues;
- Obtained benefits;
- Impact of the optimizations on corporate performance;
- Success factors of the optimizations.

The market study run over Romanian market (mostly energy sector) shows which are the main items of the above modules impacted by the optimizations.

### 6.1 Changes in business processes

The study shows that on average product design/ development, costs reductions aims, inventory management and production planning have led to a change in business processes to the largest extent. The sales and ordering, product design/ development, distribution, inventory management and production planning have determined the change in business processes measure.

On average, at the level of the study, only advertising/ promotion, billing/ payments and business planning received less attention.

### 6.2 Optimizations targets (planned and achieved)

The study shows that, as far as the objectives were included in the optimization plans, on average, at Romania level, the improvements based on automation, reduction of the costs and production costs, increase of competitiveness through costs reductions and the utilization of the novel technologies represented the main goals for which the objectives and targets were included in project plans. On the other hand, the increase in competitiveness through increased quality, concentration on the main results and the establishment of aggressive objectives received the lowest scores with regards to the optimization targets.

### 6.3 Optimizations implementation issues

The study shows that, having in mind the main implementation issues list at the planning level, on average, at the level of Romania, the available IT infrastructure does not support the planned optimizations, management reticence to allocate the funds and the business mistakes under the pressure of delivering the expected results were have been the major issues for the implementation of the optimizations.

### 6.4 Obtained benefits

The list of possible issues considered to be under the list of benefits of the optimizations' implementation found at the company level, as resulted from the market study, at the level of Romania, shows that the customer satisfaction level (improved response to clients' requests), concentrating the resources towards the selling aspects, increased flexibility through the adoption of new IT technologies and more efficient marketing and selling processes were the major benefits of the optimizations' implementation.

### 6.5 Impact of the optimizations on corporate performance

Based on the market study performed, the main items that were considered to be of impact of optimizations' implementation found at the company level, in Romania, shows that in terms of impact on corporate performance, on average, improving the development of new products, improving the costs' reductions and improving the investments return level, had a greater impact on corporate performance. On the other hand, the improvement of the sales rate, increasing the market rate and the improvement of the operational profits had the lowest impacts on corporate performance.

### 6.6 Success factors of the optimizations.

Based on the market study, the list of possible matters considered to be success factors at the optimizations level, found at the company level, show that the utilization of experts or external support, optimizations

driven by customer requirements and competitive pressures and involving all important employees represented to a greater extent impact on success in optimizations' implementation. On the other hand, the process mapping approach, the development of a well-defined project structure and using the internal surveys scored the lowest in terms of the main success factors for the optimizations done.

### 6.7 Summary of the study

The approach to optimizations is a broad one and aims at a sharp change in the quality of services offered, costs and production, including the analyses of the current state of scientific research in this field, based on the most recent and representative references in the relevant literature and interpretations and own contributions. This study aimed to provide certain contributions in the BPO environment, namely:

- to highlight factors and conditions necessary for the optimizations of financial and business companies;
- establish the optimization methodologies, but also how to apply them;
- Identify the common traits of business optimizations and/ or related methodologies, along with the actions and measures used;
- Developing a general optimizations model for companies, through which managers will be able to determine the reasons for the changes they want, as well as changes in the environment, all of which stimulate the need for improvements.

The results conclude that organizations do not focus on some of the most important tasks and actions recommended in the literature as a basis for optimizations, such as the use of time as a competitive advantage, changes in customer/ market business/ processes, the value-added item of each business activity and the application of the right innovative technology. Therefore, one can assume that there is a major reason why many of the optimizations project objectives were only modestly achieved.

On average, the most common problems encountered in optimizations' implementation appear to be basic and difficult to solve in practice: implementation difficulties due to communication barriers between the organization/ functional sub-units, unexpected amount of optimization efforts required, interruption of operations, failure to achieve the expected benefits, pressure business mistakes to produce quick and overestimated results, and reluctance of top managers to commit the funds needed for the project.

Given that most optimizations benefit from innovative uses of information technology, an organizational problem that could condemn optimization projects to the failure of a particular company is the lack of communication between CEOs / top executives and CIO / IS managers.

## 7. Conclusions

Organizations can achieve sustainable and effective process improvement by combining project management best practices with certain Six Sigma methodologies and automation solutions. The ability to combine these proven methodologies provides the structure and discipline required to identify process improvement and optimization opportunities, develop sustainable solutions and lead the organization through the strategic change process. Use of these integrated techniques allows business processes to be efficient, agile, and meet the organization's customer demands. In today's challenging, global economy it is essential for organizations to combine the disciplines of Project Management, Six-Sigma and business process optimization to realize process gains that ensure "faster", "better", "cheaper" for their products or services, while maintaining a high level of quality in the marketplace.

**References**

[1] Y. Zhou and Y. Chen, "Project-oriented business process performance optimization,", 2003, vol. 5.
[2] H. A. Reijers, "Product-based design of business processes applied within the financial services," vol. 34, 2002.
[3] I. Hofacker and R. Vetschera, "Algorithmical approaches to business process design," 2001.
[4] Y. Zhou and Y. Chen, "The methodology for business process optimized design," 2003
[5] A. Zakarian, "Analysis of process models: A fuzzy logic approach," vol. 17, 2001.
[6] K. Phalp and M. Shepperd, "Quantitative analysis of static models of processes,", 2000.
[7] Soininen, J., Puumalainen, K., Sjögrén, H., Syrjä, P., The impact of global economic crisis on SMEs, Management Research Review, 2012.
[8] Scherrer, P. S., Management turnarounds: diagnosing business ailments, Corporate Governance: The international journal of business in society, Vol. 3, 2003
[9] J. L. Rummel, Z.Walter, R.Dewan, and A. Seidmann, "Activity consolidation to improve responsiveness," vol. 161, 2005.
[10] R. Dewan, A. Seidmann, and Z. Walter, "Workflow optimization through task redesign in business information processes," 1998, vol. 1.
[11] Tristan Boutros și Jennifer Cardella, "the Basics of process improvement", Eng., 2016.

**Radu SAMOILA** has graduated the Master of Economy and Information Technology in 2011 at Bucharest University of Economy. Currently he is a PhD Student at this university, since 2019. Main fields of interest are business process optimization, automation of business processes and the continuous improvements concept.

# Natural Learning Processing based on Machine Learning Model for automatic analysis of Online Reviews related to Hotels and Resorts

Bogdan-Ştefan POSEDARU, Tiberiu-Marian GEORGESCU, Florin-Valeriu PANTELIMON
The Bucharest University of Economic Studies
posedarubogdan10@stud.ase.ro, tiberiugeorgescu@ase.ro, pantelimonflorin15@stud.ase.ro

*This article describes the development and implementation of a natural language processing (NLP) model based on machine learning (ML) for automatic analysis of customers' reviews on hotels and resorts written in English. The model performs named entity recognition (NER), relation extraction (RE) as well as sentiment analysis (SA). The performance indicators validate the model, as we obtained an F1 score of 0.79 for ER and 0.61 for RE. Our results prove to be remarkable compared to other models that use similar techniques and technologies. Furthermore, we developed a web application which allows users to benefit from our model to automatically analyze customers' reviews about hotels and resorts.*
***Keywords:*** *Natural Language Processing, Machine Learning, Entity Recognition, Relation Extraction, Sentiment Analysis*

## 1 Introduction

The growth of internet in the last decade has had a massive impact in the hospitality industry, as in most industries. Nowadays, both customers and service providers invest plenty of time in reading and analyzing online information. When choosing to book a product or a service in the hospitality domain, a critical role is played by consumers' online reviews [1], [2]. Study [3] considers that more than 75% of people are taking online reviews into account when booking a hotel. Therefore, the companies from hospitality industry need to carefully monitor customers' reviews, as they are an important decision factor for their clients [4].

In this context, the volume of reviews increased drastically and it became more and more time consuming to analyze them manually. This study aims to automate the hotel review analysis, by using natural language processing (NLP) based on machine learning (ML). We developed a model which is able to perform named entity recognition (NER), relation extraction (RE) and sentiment analysis (SA) for online reviews regarding hotels and resorts written in English. Based on a list of reviews, the model identifies the main entities and relations, the sentiment for each entity and the general sentiment for the whole document analyzed. Furthermore, we created a web application available at https://hotelinsights.live/ which serves as a portal to the model. By this manner, users can benefit from the model's capabilities when choosing a hotel for their holiday.

### Related work

The tremendous growth of data related to tourism led to the necessity of new tools to manage it. Many web mining techniques are used for automatically extracting useful insights from web content related to hospitality industry. In the 2000s, the main techniques used were business intelligence for structured and semi-structured content and web analytics for unstructured data [5]. The Web 2.0 brought to the table big volumes of data, especially user-generated content. The existing techniques weren't efficient in extracting insights from large volumes of data, therefore, new ones were developed, in particular NLP [6].

In the early ages of the NLP, rule-based techniques were used. Article [7] describes a rule-based NLP model for the analysis of online reviews, while paper [8] explores ways to implement rules-based NLP for documents about the hospitality industry.

However, such techniques are limited to predefined sets of rules. Since the

statistical revolution, numerous algorithms were developed which are capable of discovering various patterns outside the predefined sets of rules. Subsequently, the ML paradigm has become increasingly popular, and in the last decade, ML-based NLP is considered by far the main approach. Many articles discuss NLP based on ML models for the hospitality industry, such as [9], [10].

Several studies that aim to automatically extract relevant information from tourism reviews have been identified. Article [11] uses NER on travel texts and the paper [12] studies the application of NER and RE to achieve the proposed goal. The article [13] discusses the SA techniques applied on hotel reviews.

**Our work**

The main purpose of our work consists in developing a NLP model which automatically analyses reviews related to hotels and resorts. In this regard, we studied related papers and identified the main methods, techniques and instruments used.

We started by developing a domain ontology specific to the hotel industry field. Section 2 describes in detail the process of developing the ontology. Subsequently, the ontology was used as the structure for the NLP based on ML model. Section 2 also presents the process of designing and training the ML model. We developed custom-made solutions to automatically download training data, consisting in online reviews from popular hotel booking platforms, such as Tripadvisor (https://www.tripadvisor.com) or Booking (https://www.booking.com/).

Section 3 presents the design and implementation of an application for automatic text analysis of hotel reviews. The application architecture is described in detail, the model described in the previous section being the main component. The application is available online at https://hotelinsights.live/, so that any interested individual can use it. Performance evaluation was conducted regularly, based on which we adjusted the model and the training process.

Section 4 describes a detailed performance analysis of the model. We calculated the values of the F1 score, precision and recall indicators and examined the correlation matrix. The remarkable results validate our model and prove its robustness. A comparison of our model's performances to the ones obtained by other projects was conducted in order to highlight our results in relation to other studies. Section 5 concludes our work and discusses the main contributions, future work as well as the limitations.

## 2 The development of the NLP based on ML model

This section describes the process of developing a specialized NLP based on ML model for the hotel industry. The model can perform the following three tasks: NER, RE and SA. Based on the documents received, the model identifies the relevant entities and the relations between them, the sentiment associated with each entity, as well as the general sentiment associated with each class of entities.

In order to build the model, the first step was to create a domain ontology. Once completed, it served as the structure of the model.

### 2.1. The ontology development process

An extensive study was done in order to understand the particularities of the documents of interest. In order to develop a robust ontology we applied two different approaches: (1) we studied the state-of-the-art related to similar ontologies and (2) we selected over 500 reviews available on platforms such as Tripadvisor or, from which we identified the tokens relevant for our purpose. The tokens, consisting in sequences of characters such as one word or more words combined, were considered potential entities for our ontology.

**The study of relevant papers**

Various articles discuss the process of building ontologies for the tourism field. Article [14] uses semantic web technologies to develop an ontology whose aim is to enable users to acquire useful information regarding their trip. Paper [15] describes a framework for tourists knowledge representation based on online reviews. Other studies focus more specifically and describe the development of ontologies for tourism in particular regions [16], [17]. Our study is concentrated particularly to the hotels and resorts, therefore we focused on the articles which describe ontologies designed for hotels services, such as [18] or [19]. Although paper [19] was partly written in a language unknown to us (Korean), we managed to comprehend its content by using translating applications. Since our work is designed to perform not only NER and RE, but also SA from online reviews, we studied various articles which discuss models suitable in this regard, such as precum [20] or [21].

**Our ontology**

A robust domain-ontology was essential in order to develop the NLP model. After we studied the related work, we explored online reviews and extracted the relevant tokens, which lately were implemented as entities in the model's dictionary. The ontology's structure was designed in such a manner to be suitable for the NLP based on ML model.

Our ontology consists of 6 classes and 14 subclasses. The 6 classes are connected to each other by 14 relationships, as shown in Figure 1. An extended representation of the ontology, which contains the subclasses as well, can be observed in Annex 1. In order to design the ontology, we used WebProtégé (https://webprotege.stanford.edu/). For the graphical representation, we used WebVOWL, a web application for visualization of ontologies available at (http://vowl.visualdataweb.org/webvowl.html).

In order to ensure the development of an optimal ontology, various strategies were used. In the initial stage, we decided that each author will develop its own version of the ontology and then compare them to each other. Out of the three ontologies, two were selected as candidates. We tested several versions of the ontology, in order to identify the optimal one for the NLP model described later in this paper. During this process, we eliminated the redundant classes or the ones with low relevance for our work. As an example, initially we considered the class *Hotel's Specifications*. However, we decided to include most of its entities into the class *Amenities* and to eliminate the others, as they were increasing the model's complexity. Also, several classes were merged. Instead of having two classes named *Incident* and *Vacation*, we decided to create a new one – *Experience* and included those two as subclasses of the class *Experience*.

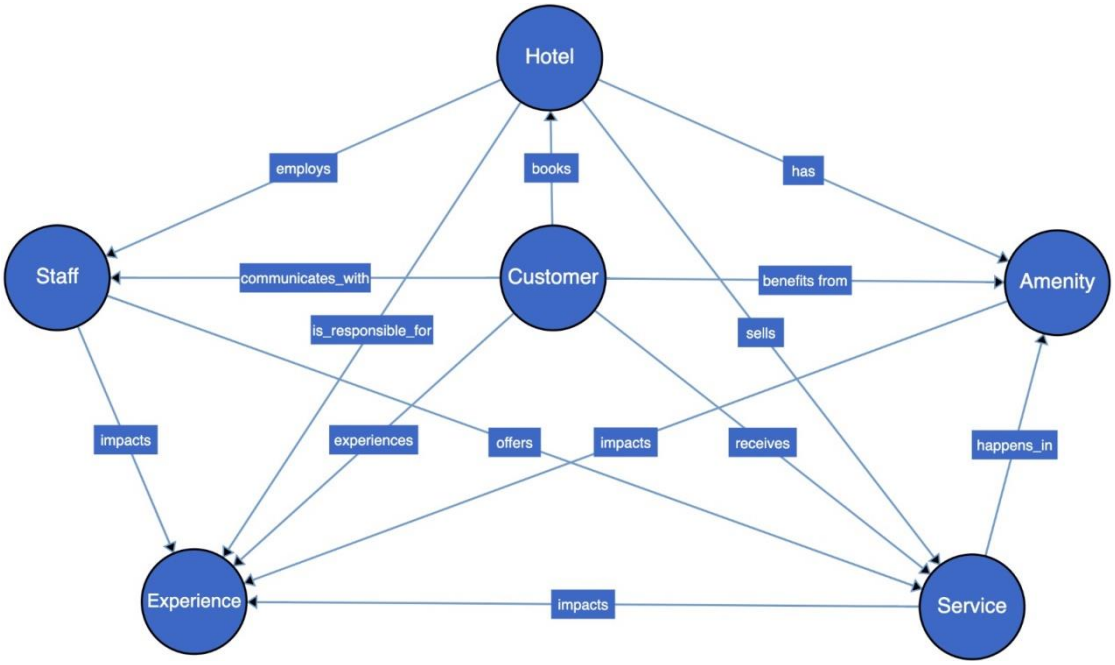



**Fig. 1**. The ontology specialized in the hotel field

Table 1 describes both the classes and their subclasses and table 2 illustrates the relationships between classes. Four of the classes are divided into subclasses. We considered that the complexity of the model based only on the main six classes was relatively low; therefore, it could bring results with better accuracy. However, by including the subclasses the model can provide more detailed information. After conducting various performance tests, we chose to keep the subclasses as part of our model's structure.

**Table 1.** The list of classes and subclasses

| No. | Class | Subclass |
|---|---|---|
| **1** | Amenity | FoodDrinks_Amenity |
| | | Leisure Amenity |
| | | General_Amenity |
| | | Room_Amenity |
| **2** | Experience | Incident |
| | | Vacation |
| **3** | Staff | FoodDrinks_Staff |
| | | General_Staff |
| | | Leisure_Staff |
| | | Room_Staff |
| **4** | Service | FoodDrinks_Service |
| | | General_Service |
| | | Leisure_Service |
| | | Room_Service |
| **5** | Hotel | - |
| **6** | Customer | - |

Table 2 illustrates the relations between classes. Once we implemented the NLP based on ML model, we noticed that its performance was very low for some of the relations; therefore we adjusted them to the current variation.

**Table 2** The relations between classes

| No. | Relation | Parent | Child |
|---|---|---|---|
| 1 | benefits_from | Customer | Amenity |
| 2 | books | Customer | Hotel |
| 3 | communicates_with | Customer | Staff |
| 4 | employs | Hotel | Staff |
| 5 | experiences | Customer | Experience |
| 6 | happens_in | Service | Amenity |
| 7 | has | Hotel | Amenity |
| 8 | impacts | Amenity | Experience |
| 9 | impacts | Staff | Experience |
| 10 | impacts | Service | Experience |
| 11 | is_responsible_for | Hotel | Experience |
| 12 | offers | Staff | Service |
| 13 | receives | Customer | Service |
| 14 | sells | Hotel | Service |

### 2.2. Choosing the proper technologies and techniques

Paper [22] presents a detailed comparison of the main NLP solutions based on ML, and paper [23] describes the current state-of-the-art on NER based on ML. Considering the studies presented, as well as the research objectives of this paper, we chose to use the services Knowledge Studio(https://www.ibm.com/cloud/watson-knowledge-studio) and Natural Language Understanding (https://www.ibm.com/cloud/watson-natural-language-understanding) from IBM Watson (https://www.ibm.com/watson). Regarding the ML approach, supervised learning was preferred, as in recent years studies show that it provides the best performance for NLP [24].

Knowledge Studio was used to define the model's structure, to configure it according to the domain particularities and to train the model, while Natural Language Understanding was used to deploy the model and served as a gateway to the analysis results provided by our model.

We implemented the ontology in Knowledge Studio as follows: we defined the classes, their subtypes and the relations between classes. Each class contained a dictionary which consisted in a list of entities. Annex 2 illustrates the dictionary afferent for the class Amenity.

The dictionary was used as a rule-based model only in the training process, accelerating the document annotation process. It contains of approximately 1000 entities, plus their respective surface forms. For example, in annex 2, the entity "a la carte restaurant" has four possible surface forms.

### 2.3. Training the model

Supervised ML models require the provision of correct result sets on which the machine builds the *ground truth*. In the training process, we annotated the documents for two tasks NER and RE. For NER, the annotation process consisted in identifying and labeling the relevant tokens and assigning each one to the proper class. In order to automate the process, the rule-based model was used as it automatically identified the tokens defined in the dictionary. However, much attention from the annotator was required, as many of the tokens weren't correctly labeled by the dictionary. Figure 2 shows a screenshot taken during the training process for NER.

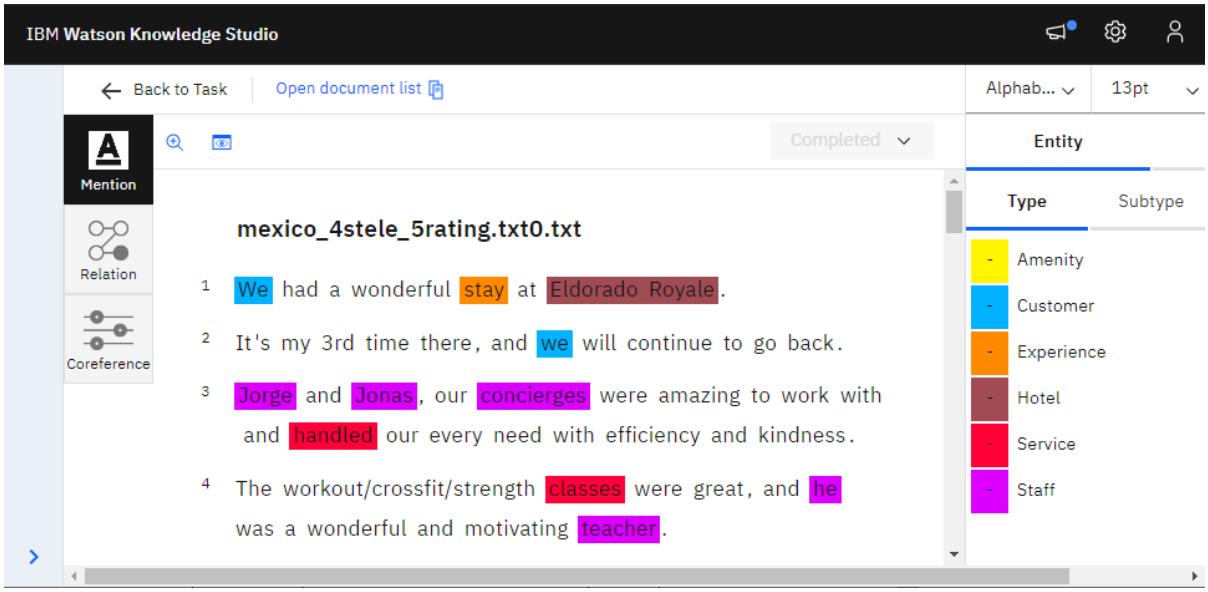



**Fig. 2.** Training the model for the NER task

As can be observed in figure above, in the first sentence the token *We* was annotated as part of the class *Customer*, the token *stay* as entity of the class *Experience* and the token *Eldorado Royale* as type *Hotel*. In addition to classes, subclasses were also annotated. Figure 3 illustrates a screenshot taken during the annotation process; as can be observed, the token *stay* was labeled as entity of the subclass *Vacation*.

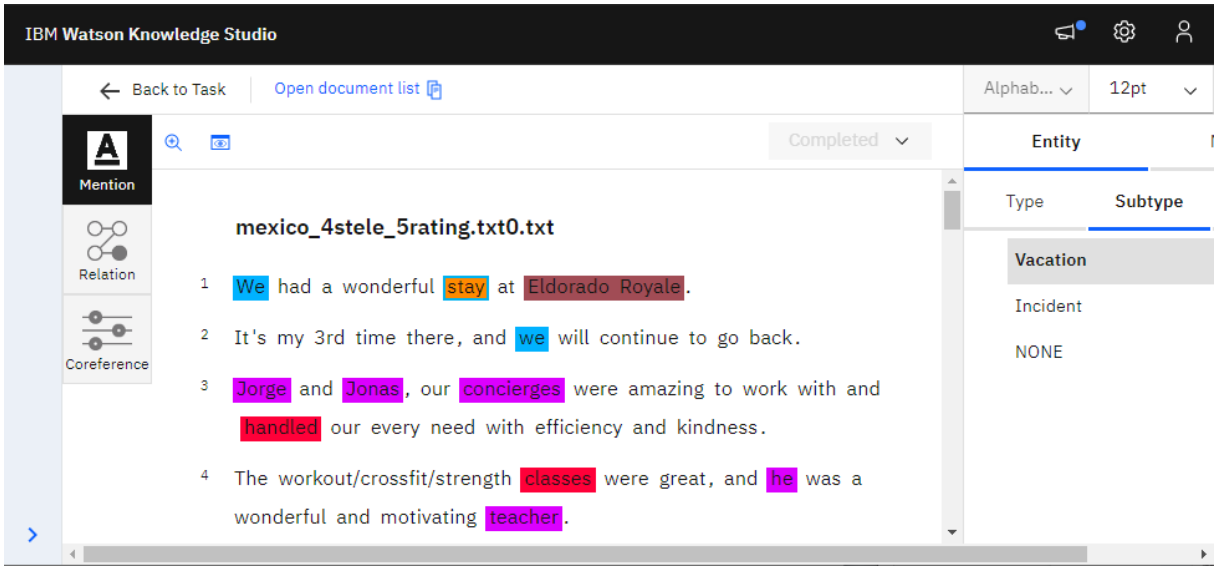



**Fig. 3**. The annotation of a subclass

The training of the RE functionality is performed by marking the relations between the identified tokens. Figure 4 is a screenshot taken during the annotation process. Within it, the annotated relations are: (1) *Customers experiences stay*, (2) *Customer books Eldorado Royale*, (3) *Eldorado Royale is responsible for stay*.

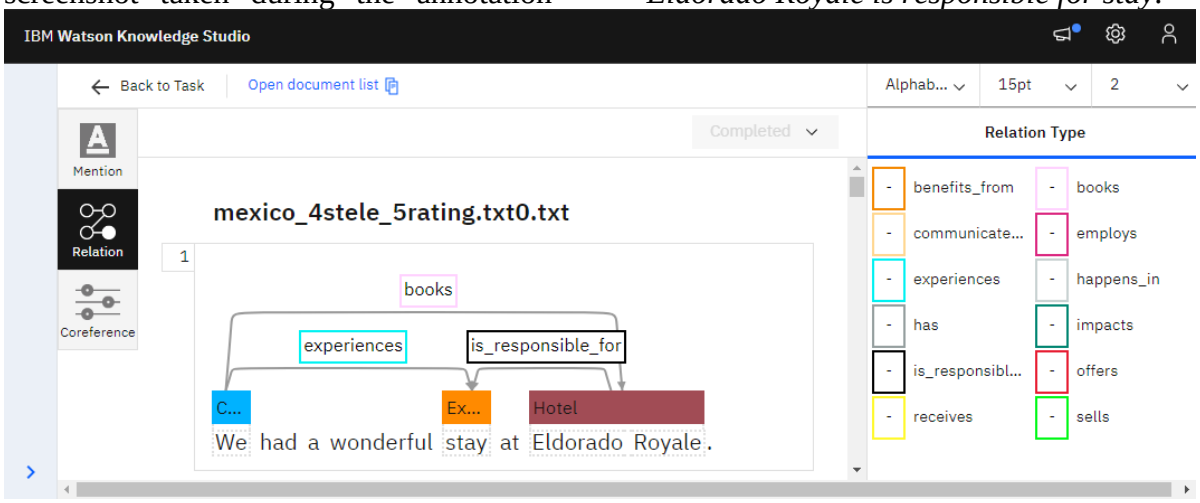



**Fig. 4.** Training the model for the RE task

For training, we downloaded reviews available online. We used documents totalizing approximately 100,000 words. In order to speed up the process, we developed custom-made scraping solutions to automatically download and structure the relevant data.

According to Watson's methodology (https://cloud.ibm.com/docs/watson-knowledge-studio-data?topic=watson-knowledge-studio-data-documents-for-annotation), for best performances it is recommended to use training documents of maximum 1000 words each. Therefore, we built a script that splits the training data in documents of at most 1000 words each.

The human annotation defined the model's *ground truth*. It was essential to make sure that the annotations were consistent. For our work, we decided that only one author (B.-Ş.P.) to perform annotation. Furthermore, the annotations were checked and amended by all three authors together before validating. During the process, periodic performance tests were conducted based on which we adjusted the annotations. A detailed description of the performance tests is presented in section 4.

### 2.4. Deploying the model

In order to deploy the model, we used IBM Watson Natural Language Understanding (NLU). Watson's Natural Language Understanding service is a collection of content analysis features that can extract semantic information from the desired input. The data can be either text, public URLs or HTML content, the service benefiting from sophisticated NLP techniques to quickly obtain features extracted from the content, such as: NER, keywords identification, RE, SA, emotion extraction, semantic roles extraction, etc.

In order to use the developed Model for Natural Language Processing analysis, it must be deployed from Knowledge Studio (where the model was trained) to Natural Language Understanding Service, communicating through a RESTful API.

Following various configurations, a private access key (API key) is offered, which requires secure storage and which must be attached to every API call to the service. In order to ensure the security of such sensitive data, a robust solution was designed, its implementation being presented in greater detail in subsection 3.1 of the present paper.

The Natural Language Understanding API receives all requests and sends all responses using the JSON (JavaScript Object Notation) standard, which facilitates data manipulation at all the levels of our web application, described in section 3 of the present paper.

Following the analysis of the content, the NER task of the NLU service sends a response consisting of a list of identified entities and their respective properties (figure 5).

```
entities: [{
 type: "Amenity",
 text: "Spa",
 mentions: [
   {
     text: "Spa",
     location: [332, 335],
     confidence: 0.998892,
   },
 ],
 disambiguation: {
   subtype:
["Leisure_Amenity"],
 },
 confidence: 0.998892,
}]
```

**Fig. 5.** Watson NLU service response for the NER functionality

We can notice from the above code snippet many properties of the identified entity, such as:

- the class and subclass to which it belongs;
- the entity text;
- location of the entity in the text (both starting and ending indices are provided);
- the associated confidence score: the closer the score is to 1, the higher the confidence label.

The RE functionality, which is closely related to the above described process of NER, provides valuable insights in relation to the present connections at a sentence level.

The IBM Watson Natural Language Understanding service sends a response consisting of a complete list of identified relationships, with their associated properties, such as linked entities, relations types, confidence scores, etc.

Regarding the SA functionality, the NLU service assesses the sentiment for the document in its entirety and also performs an individual sentiment evaluation for every entity identified. An example of the

response provided by the relationship extraction functionality can be consulted in figure 6.

```
{
 type: "Amenity",
 text: "Spa",
 sentiment: {
   score: 0.972212,
   label: "positive",
 },
 disambiguation: {
   subtype:
["Leisure_Amenity"],
 },
 count: 1,
}
```

**Fig. 6.** Watson NLU service response for the SA functionality

From the above code snippet, it is noticeable that the *Amenity* identified in the given text as *Spa*, belonging to *Leisure Amenity* subclass, had attributed a positive sentiment, with an almost perfect score of 0.97.

The sentiment score provided by the service can take any value between -1 and 1, where a score of 0 represents a completely neutral sentiment towards the subject. The closer a value gets to -1, the more negative the sentiment it denotes, and the closer the value is to 1, the stronger the positive sentiment related to the entity.

## 3. Implementation of an application for automatic text analysis of hotel reviews

In order to demonstrate the capabilities of the NLP model based on ML, a web application was designed for the cognitive analysis of documents specific to the hospitality industry.

### 3.1. The Application's Architecture

During the documentation and development phases, the question arose as to whether the web application should also include a server component, or whether a direct Client-NLP Model communication should suffice. It was finally decided to implement the stand-alone server component, acting as a bridge between the interface and the IBM Watson API Service, for the following reasons:

- The stand-alone NodeJS Server allows to securely store sensitive data, such as API keys, by using environment variables stored in special *.env* files that are inaccessible to the external environment;
- The stand-alone NodeJS Server allows to execute preprocessing functions on input data, such as extracting text content from compatible files. This is an important aspect in terms of IBM Cloud usage savings;

Taking the above into account, the complete architecture of the web application can be consulted in the figure below:

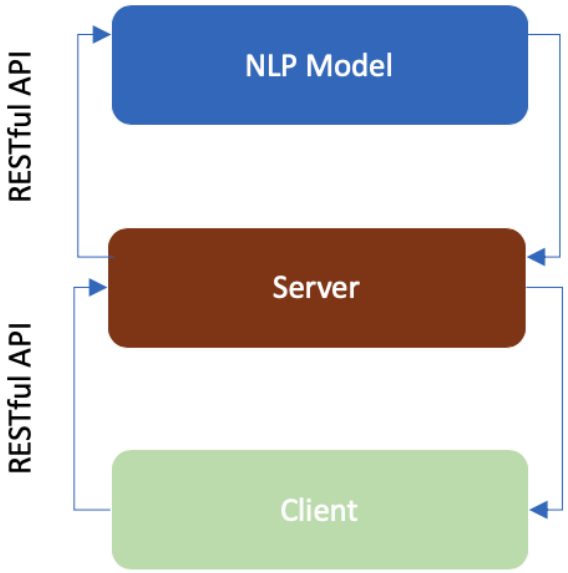



**Fig. 7.** Web Application Architecture

### 3.2. Technologies used

Considering the fact that one of the main requirements of the application was for it to be readily available for any potential user, we decided to develop it as a web application, published and accessible from a public domain (http://hotelinsights.live). In addition to satisfying the basic functionalities of the app, many other factors were taken into account when

deciding on the technology stack, such as general speed and performance, Client-Server communication stability or session persistence.

In order to develop the web application, a solution was chosen that consists entirely of Javascript libraries and frameworks, both for the client-side (ReactJS) and for the server-side (NodeJS). Developing both components using the same programming language ensures the possibility of fast services integration or code reusability. The chosen technology stack presents many advantages, such as full compliance with all the functional requirements of the application or the high availability and reduced costs related to deploying solutions.

The complete list of technologies used includes: Javascript, NodeJS, ExpressJs, ReactJS, Create-React-App CLI, Jest, Axios, GoJS, Textract, Dotenv, Git, Heroku and Postman.

Analyzing the stack of technologies presented above, we observe a complete synergy between all the application's components, which is possible only due to the use of a single programming language for all architectural layers.

### 3.3. An overview of an application scenario

This subsection describes a typical use case scenario, with all the associated steps from loading the document to seeing the analysis results, structured in an intuitive way.

Accessing and using the web application is very simple both from a laptop and a mobile device, the only requirement being a modern web browser, such as Google Chrome or Mozilla Firefox.

When the application loads, it presents the user with a welcome screen and displays a series of three buttons representing options of entering data, by uploading a document from the user's personal computer, using sample hotel reviews documents or entering text directly in the web app. In the following paragraphs, we describe the file upload method of entering data, component that also allows drag-and-drop technique, improving the general user experience. After the desired document is selected, the user confirms the operation, the data is sent to the stand-alone server and a *loading* component is presented, acting as a buffer screen between the input and the output components.

In case the stand-alone server sends back an error-free response, the user is redirected to the results page, from where he can visualize the analysis done by the NLP based on ML model. The results are structured into three tabs, one for NER task, one for RE and the final for SA.

For visualizing the NER task results, the application highlights the recognized entities in the text, using custom colors according to its corresponding class.

Results from the RE task are presented as a Force Directed Graph consisting of nodes – representing entities, and links – representing identified relations, where the user can rearrange these components, for a better overview of complex sentences, with a high number of elements identified (figure 8).

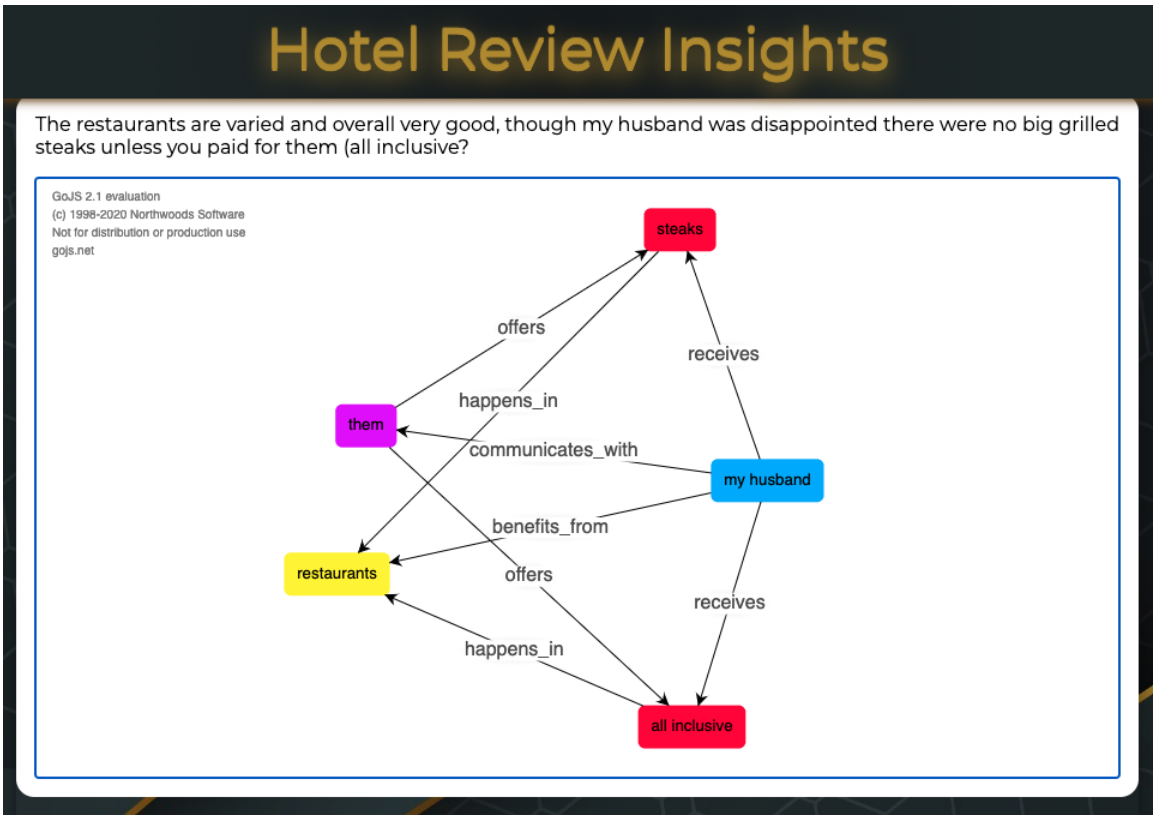



**Fig. 8.** The graphical representation of the RE results

Regarding the SA results, the web application presents an entity-level analysis, aggregated by class, as follows: *Experience*, *Amenity*, *Customer*, *Hotel*, *Service* and *Staff*. Each class' sentiment score is graphically represented by the help of a red circle that is partially or completely covered with a green trail: the longer the green trail, the better the sentiment score, as can be seen in figure 9. The user can also get a more detailed look on one class' sentiment score by clicking on the corresponding button and consulting the table of identified entities and associated sentiment score.


**Fig. 9.** The graphical representation of the SA results

A special role in the analysis is played by the *Incident* subclass, belonging to the *Experience* class. The web application builds an *Incident Report* based on recognized entities belonging to this subclass and assigns a score representing the probability of an incident being

described in the provided document.
Taking into account the previously described functionalities, the user can benefit from the NLP model analysis capabilities in an intuitive web application, and can extract invaluable insights from documents related to the hospitality industry.

## 4 Results

This section describes the performances evaluation process and the model's results. F1 score was calculated for the NER and RE functionalities and the results validate the model. We describe the values of the indicators, for each entity type and relation type as well as overall. Besides these, the confusion matrices are presented and analyzed, in order to better understand the model's performances and identify new strategies for improvement. Also, we compare our results with other similar projects and discuss our model's advantages as well as its limitations.

### 4.1. Methodology

In order to evaluate the performance of the model, the methodology developed by IBM, specific for NLP based on ML models, was used (https://cloud.ibm.com/docs/watson-knowledge-studio?topic=watson-knowledge-studio-evaluate-ml).
The documents were split into three categories: training sets, test sets and blind sets. The training sets which were used to define the *ground truth* represented 70% of the total documents. The test sets were documents annotated by human, which weren't included in the *ground truth*, but were kept aside to be used in the performance evaluation process. The test sets represented 23%. For the blind sets, we kept 7% of the documents aside not only from the *ground truth*, but also from the annotator (B.-Ş.P). They were managed by the other two authors of this paper. Knowing all the documents, the annotator could have influenced the model so that it would have adapted particularly for the test sets and not in general, as desired. By this manner, we reduced the bias tendency of the annotator as much as possible.
We took into account the F1 score, precision and recall, which are established indicators used for the supervised ML evaluation process [25] The value of each indicator is between [0,1], the bigger the value, the better the result.
The precision indicator illustrates how accurate are the model's annotations in comparison to the human's annotations, considered as *ground truth*. The precision's formula is:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}},$$

where *true positive* consists in tokens annotated as relevant in a correct manner, contrary to *false positive*, which represent tokens incorrectly labeled as relevant. The worst value for the precision indicator is 0 and the best is 1. If the value of precision is 1, it means that the model marked correctly all the selected tokens.
Recall measures how many mentions were actually annotated with the proper labels.

$$(2)\ \text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}},$$

where *false negative* consists in the mentions which the model fails to recognize. The highest value (of 1) indicates that the model identifies all the relevant mentions, on the contrary, a recall of 0 means that the model wasn't able to identify any entity.
The F1 score is calculated based on precision and recall, as a harmonic mean of the two indicators, showed in the formula below. In order to validate the model, the F1 score value has to be higher than 0.5.

$$(3)\qquad \text{F1 score} = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}.$$

Usually, a low precision indicates that the machine generates incorrect annotations (https://cloud.ibm.com/docs/watson-

knowledge-studio?topic=watson-knowledge-studio-evaluate-ml#evaluate-mllowp). A low recall illustrates that the machine is not able to identify and annotate the right mentions (https://cloud.ibm.com/docs/watson-knowledge-studio?topic=watson-knowledge-studio-evaluate-ml#evaluate-mllowr).

### 4.2. The performances of the model

The F1 score of the current version of the model is 0.79 for NER and 0.61 for RE. Figure 10 illustrate the evolution of the model, as well as the values for the Precision and Recall for both NER and RE.

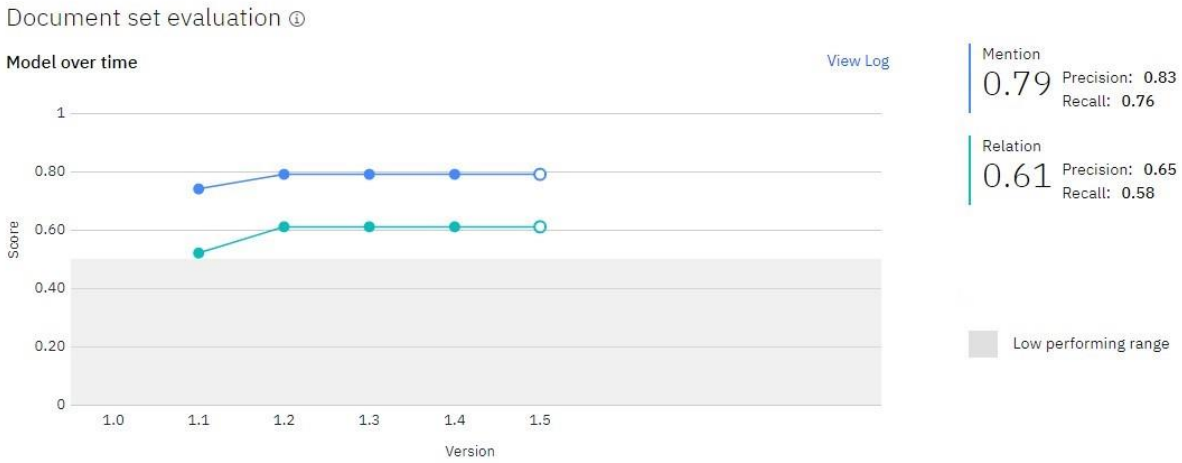



**Fig. 10.** The evolution of the model's performances for NER and RE

As can be observed, the F1 score has known a significant improvement from version 1.1 to version 1.2 both for NER and RE. For version 1.1 documents totaling around 50,000 words were annotated, compared to version 1.2 where their number was approximately 100,000. The performance tests corresponding to 1.3, 1.4 and 1.5 are performed with the same corpus as in version 1.2, but with different test sets. Since all the last 4 versions have very similar F1 score, we consider that the results weren't casual and that the tests present sufficient consistency.

**The model's performances for NER**

The F1 score for NER of 0.79 indicates good performances. The precision indicator is higher than the recall, therefore we plan to perform further training session which may improve the model. Moreover, as can be observed in figure 11, the recall is lower than the precision for each entity class, which accentuates our assumption that further training may raise the performances.

| Entity Types | F1 | Precision | Recall | % of Total Annotations |
|---|---|---|---|---|
| Amenity | 0.74 | 0.8 | 0.7 | 30% (633/2114) |
| Customer | 0.89 | 0.9 | 0.87 | 18% (378/2114) |
| Experience | 0.69 | 0.71 | 0.67 | 7% (139/2114) |
| Hotel | 0.75 | 0.89 | 0.66 | 7% (156/2114) |
| Service | 0.68 | 0.72 | 0.64 | 28% (619/2114) |
| Staff | 0.7 | 0.75 | 0.66 | 9% (189/2114) |

**Fig. 11.** The performance indicators for each class

The F1 score is over 0.68 for each class type. The best performances are obtained for the class Customer (0.89) and the biggest percent of annotation for the classes Amenity (30%) and Service (28%). An essential tool for improving the model is the confusion matrix. Based on it, we can get a better understanding about the model's flaws.

Based on it, we can get a better understanding of the classes for which the model has trouble annotating properly, as it offers details about the reasons for the wrong labels. The analysis of the matrix can be done from two perspectives: (1) the confusion between each type of class and the others and (2) the confusion between each type of class and tokens which are not relevant for our model marked in table 3 with the label *Other tokens*. The table below illustrates the confusion matrix for NER in absolute values.

**Table 3.** The confusion matrix for NER

| **Entity Types** | **Amenity** | **Customer** | **Experience** | **Hotel** | **Other tokens** | **Service** | **Staff** | **Total** |
|---|---|---|---|---|---|---|---|---|
| **Amenity** | 462 | 0 | 1 | 3 | 137 | 21 | 9 | 633 |
| **Customer** | 0 | 329 | 0 | 0 | 47 | 2 | 0 | 378 |
| **Experience** | 2 | 0 | 88 | 2 | 45 | 2 | 0 | 139 |
| **Hotel** | 4 | 0 | 0 | 92 | 40 | 11 | 9 | 156 |
| **Other tokens** | 15 | 30 | 15 | 0 | 7424 | 41 | 17 | 7542 |
| **Service** | 13 | 0 | 7 | 1 | 169 | 423 | 6 | 619 |
| **Staff** | 3 | 0 | 0 | 0 | 66 | 3 | 117 | 189 |
| **Total** | **499** | **359** | **111** | **98** | **7928** | **503** | **158** | **9656** |

**The model's performances for RE**

The task of RE proves to be more complex, since our model contains a relatively high number of relation types (14). The F1 score of 0.61 validates the model with a precision of 0.65 and a recall of 0.58. As in the case of NER, the recall is lower than the precision, thus further annotations could improve the performances. Figure 12 illustrates the indicators for each relation type.

| Relation Types | F1 | Precision | Recall | % of Total Annotations |
|---|---|---|---|---|
| benefits_from | 0.64 | 0.7 | 0.59 | 16% (743/4579) |
| books | 0.64 | 0.77 | 0.54 | 6% (254/4579) |
| communicates_... | 0.58 | 0.63 | 0.54 | 6% (261/4579) |
| employs | 0.57 | 0.74 | 0.47 | 2% (75/4579) |
| experiences | 0.55 | 0.56 | 0.54 | 6% (271/4579) |
| ⚠ **happens_in** | **0.4** | **0.39** | **0.42** | **16% (747/4579)** |
| has | 0.51 | 0.67 | 0.41 | 5% (216/4579) |
| ⚠ **impacts** | **0.49** | **0.52** | **0.47** | **9% (401/4579)** |
| ⚠ **is_responsible_f...** | **0.46** | **0.56** | **0.39** | **2% (108/4579)** |
| ⚠ **offers** | **0.41** | **0.46** | **0.37** | **11% (522/4579)** |
| receives | 0.59 | 0.64 | 0.55 | 18% (809/4579) |
| sells | 0.64 | 0.72 | 0.57 | 4% (172/4579) |

**Fig. 12.** The performance indicators for each relation type

As can be observed, eight of the relation types have the F1 score over 0.5 and the others a little below the threshold. All the relation types have a better precision than recall, except for "happens_in", which is also one of the most frequent annotations (16%), alongside "benefits_from" (16%) and ,,receives" (18%). On the other hand, the relation types ,,is_reponsible_for", "employs" and "sells" are rarely met in the training documents, therefore they may be not as relevant for the model as the others. The confusion matrix for RE (table 4) shows no confusions between any relation type and the others. This was expected, since any relation is defined by a parent and a child and there are no relations with the same parent-child combination. However, there is a relatively high percentage of confusions between correctly identified and false positive types of relations.

**Table 4.** The confusion matrix for RE

| **Relation Types** | **benefits_from** | **books** | **communicates_with** | **employs** | **experiences** | **happens_in** | **has** | **impacts** | **is_responsible_for** | **offers** | **receives** | **sells** | **Total** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **benefits_from** | 176 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 300 |
| **books** | 0 | 56 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 103 |
| **communicates_with** | 0 | 0 | 64 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 118 |
| **employs** | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 30 |
| **experiences** | 0 | 0 | 0 | 0 | 57 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 105 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **happens_in** | 0 | 0 | 0 | 0 | 0 | 110 | 0 | 0 | 0 | 0 | 0 | 0 | 265 |
| **has** | 0 | 0 | 0 | 0 | 0 | 0 | 33 | 0 | 0 | 0 | 0 | 0 | 80 |
| **impacts** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 64 | 0 | 0 | 0 | 0 | 136 |
| **is_responsible_for** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 36 |
| **offers** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 67 | 0 | 0 | 182 |
| **receives** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 172 | 0 | 311 |
| **sells** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 39 | 68 |
| **Total** | **252** | **73** | **101** | **19** | **101** | **283** | **49** | **124** | **25** | **147** | **269** | **54** | **2365** |

### 4.3. Results compared to other projects

Below the performances of the model described in the paper are compared to those of other models that use similar techniques and technologies for NLP. Six studies have been identified that use similar technologies for both model development and performance evaluation. This is a fact of great importance, because not having the same methodological approach, the comparison of indicators presented in the studies would not have been valid. Table 5 shows a comparison between these studies and the model described in the present paper.

**Table 5. Comparison of the performance of our model in relation to other studies**

| No. | Study | NLP based on ML tool | Domain | F1 Score for NER | F1 score for RE | Number of classes | Number of relations |
|---|---|---|---|---|---|---|---|
| 1 | Our model | Watson Knowledge Studio | Hotel | 0.79 | 0.61 | 6 | 14 |
| 2 | [26] | Watson Knowledge Studio | Cybersecurity | 0.81 | 0.58 | 18 | 33 |
| 3 | [27] | Watson Knowledge Studio | IoT security | 0.68 | 0.46 | 17 | 30 |
| 4 | [28] | Watson Knowledge Studio | Medical field | 0.49 | 0.19 | 5 | N/A |
| 5 | [29] | Watson Knowledge Studio | Shipping | 0.67 | 0.55 | 4 | 2 |
| 6 | [30] | Watson Knowledge Studio | Medical field | 0.73 | N/A | 5 | N/A |
| 7 | [31] | Stanford Named Entity Recognizer | Cybersecurity | 0.80 | N/A | 10 | N/A |

N/A = Not available

Similar comparisons are presented in papers [26] and [27]. In addition to the F1 scores for NER and RE, table 5 also takes into account the number of classes for each

of the two functionalities, as well as the software tool that was used to implement the models. All identified projects use the Watson Knowledge Studio service, except for the study conducted by [31] in which Stanford Named Entity Recognition was preferred.

The fields in which the models were applied vary: three projects with applications in cybersecurity, two in medical field, one in shipping and one in the hotel field (the model described in this paper).

The projects [26], [31], as well as our model present the highest F1 scores for NER, having very close performance indicators values. The highest value for the precision was obtained by [26] - 0.88, followed by [31] and our model, with the same precision of 0.83. All three studies have the recall lower than the precision.

Regarding the RE task, the model presented in this paper achieves the best performances (0.61), followed by [26] with the F1 score equal to 0.58. Once again, both projects have higher value for precision in comparison with recall, by 0.07 in the case of our model and by 0.10 in the case of the project [26].

Table 5 also illustrates the number of classes and the number of relations for each study. The more classes the NLP model is based on, the better the level of understanding of the field. On the other hand, the complexity of the model increases in direct proportion to the number of classes. In order to obtain a more detailed representation of the hotel industry, while maintaining the complexity of the model at an acceptable level, we chose to include a relatively small number of classes (6). However, we decided to divide four of them into subclasses, totaling 14 such subtypes, as presented in section 2.

The developed model presents high performances in relation to the studies identified in the specialized literature, aspect that highlights the relevance of the research.

## 5. Conclusions and future work

In the context of the ever-expanding volume of the Internet, automated data processing solutions are becoming increasingly necessary. In this paper, we study the current issues related to the processing of natural language with the help of artificial intelligence.

The paper describes detail the development and implementation processes of a NLP based on ML model specialized in the hotel field. The methods, techniques and technologies used are described. The model performs three types of tasks: NER, RE and SA. The performance scores were presented, along with a description of the testing methodology, the values of the indicators and how they were obtained. High values of F1 scores for both entity recognition and relationship extraction demonstrate the validity of the model. Also, it shows good performances in comparison with other similar projects.

In order for the model to be usable by any interested individual, a web application was developed. This allows visitors to automatically analyze hotel reviews. The application also contains a feedback form, where users can offer valuable insights and help improve the solution.

In addition to the scientific achievements presented, a limitation of our work was also identified. This is related to the use of commercial software tools (Watson Knowledge Studio and Watson Natural Language Understanding) as part of the solution. In the future we want to develop similar models, but using only free and open-source libraries and frameworks.

For future research, we would like to test new techniques, methods and technologies specific to the NLP field. It is also desired to develop new ML models for other fields, apart from the hospitality industry, as well as developing solutions for NLP in the Romanian language.

**References**

[1] Giacomo Del Chiappa, Carlota Lorenzo-Romero , and Maria-del-Carmen Alarcon-del-Amo, "Profiling tourists based on their perceptions of the trustworthiness of different types of peer-to-peer applications," *Current Issues in Tourism, 21.3*, pp. 259-276, 2018.

[2] Markus Schuckert, Xianwei Liu, and Rob Law, "Hospitality and tourism online reviews: Recent trends and future directions," *Journal of Travel & Tourism Marketing, 32.5*, pp. 608-621, 2015.

[3] K. L. Xie, C. Chen, and S Wu, "Online Consumer Review Factors Affecting Offline Hotel Popularity: Evidence from Tripadvisor," *Journal of Travel & Tourism Marketing. Advance online publication. doi:10.1080/10548408.2015.1050538* , 2015.

[4] David D'Acunto, Annamaria Tuan, and Daniele Dalli, "Are Online Reviews Helpful for Consumers?: Big Data Evidence From Services Industry,*"* *Exploring the Power of Electronic Word-of-Mouth in the Services Industry. IGI Global* , pp. 198-216, 2020.

[5] Fernando Iafrate, *From big data to smart data. Vol. 1.:* John Wiley & Sons, 2015.

[6] Aitor García-Pablos, Montse Cuadros, and Maria Teresa Linaza, "Automatic analysis of textual hotel reviews," *Information Technology & Tourism 16.1*, pp. 45-69, 2016.

[7] Yin Kang and Lina Zhou, "RubE: Rule-based methods for extracting product features from online consumer reviews," *Information & Management, 54.2*, pp. 166-176, 2017.

[8] Roshan Fernandes and Rio D'Souza GL., "Semantic analysis of reviews provided by mobile web services using rule based and supervised machine learning techniques," *International Journal of Applied Engineering Research 12.22*, pp. 12637-12644, 2017.

[9] P. S. Bhargav, G. N. Reddy, R. R. Chand, K. Pujitha, and A. Mathur, "Sentiment Analysis for Hotel Rating using Machine Learning Algorithms," *International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8.6*, 2019.

[10] C. H. Ku, Y. C. Chang, Y. Wang, C. H. Chen, and S. H. Hsiao, "Artificial Intelligence and Visual Analytics: A Deep-Learning Approach to Analyze Hotel Reviews & Responses," in *In Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[11] B. Cowan et al., "Named entity recognition in travel-related search queries," in *In Twenty-Seventh IAAI Conference*, 2015.

[12] C. Chantrapornchai and A. Tunsakul, "Information Extraction based on Named Entity for Tourism Corpus," in *16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2019, pp. 187-192.

[13] K. Zvarevashe and O. O. Olugbara, "A framework for sentiment analysis with opinion mining of hotel reviews," in *Conference on Information Communications Technology and Society (ICTAS)*, 2018.

[14] C. I. Lee, T. C. Hsia, H. C. Hsu, and J. Y. Lin, "Ontology-based tourism recommendation system," in *4th International Conference on Industrial Engineering and Applications (ICIEA)*, 2017, pp. 376-379.

[15] M. Y. Pai, D. C. Wang, T. H. Hsu, G. Y. Lin, and C. C. Chen, "On Ontology-Based Tourist Knowledge Representation and Recommendation," *Applied Sciences, 9(23)*, p. 5097, 2019.

[16] L. Afuan and N. Hidayat, "Ontology model for tourism information in Banyumas," in *AIP Conference Proceedings 2094, (1)*, 2019.

[17] C. Lohvynenko and D. Nedbal, "Usage of Semantic Web in Austrian Regional Tourism Organizations," in *International Conference on Semantic Systems*, 2019, pp. 3-18.

[18] M. Chaves and C. Trojahn, "Towards a multilingual ontology for

ontology-driven content mining in social web sites," in *Proceedings of the ISWC 2010 Workshops, Volume I, 1st International Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web*, 2010.

[19] D. H. Yoo and Y. M. Suh, "An Ontology-based Hotel Search System Using Semantic Web Technologies," *The Journal of Society for e-Business Studies, 13(4)*, pp. 71-92, 2008.

[20] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment analysis in tourism: capitalizing on big data," *Journal of Travel Research, 58(2)*, pp. 175-191, 2019.

[21] S. Banerjee and A. Y. Chua, "Trust in online hotel reviews across review polarity and hotel category," *Computers in Human Behavior, (90)*, pp. 265-275, 2019.

[22] Tiberiu-Marian Georgescu, Modelarea Bazată pe Volume Mari de Date. Securitate Cibernetică în Contextul Big Data (Eng. "Modelling Based on Large Volumes of Data. Cybersecurity in the Context of Big Data"). Ph.D. Thesis, The Bucharest University of Economic Studies, Bucharest, Romania, 2019

[23] Tiberiu-Marian Georgescu et al., "A Survey on Named Entity Recognition Solutions Applied for Cybersecurity-Related Text Processing," in *International Congress on Information and Communication Technology*, London, 2020.

[24] G. Acampora, V. Loia, C.S. Lee, and M. H Wang, *On the Power of Fuzzy Markup Language.*: Springer, 2013.

[25] CoNLL. (2018) Universal Dependencies. [Online]. HYPERLINK "https://universaldependencies.org/conll18/evaluation.html" https://universaldependencies.org/conll18/evaluation.html

[26] Tiberiu-Marian Georgescu, "Natural Language Processing Model for Automatic Analysis of Cybersecurity-Related Documents," Symmetry 12.3, 2020.

[27] T. M. Georgescu, B. Iancu, and M. Zurini, "Named-Entity-Recognition-Based Automated System for Diagnosing Cybersecurity Situations in IoT Networks," *Sensors, 19(15)*, 2019.

[28] Timothy, et al. NeCamp. (2017) Data Science For Social Good, University of Chicago. [Online]. "https://dssg.uchicago.edu/wp-content/uploads/2017/09/necamp.pdf"

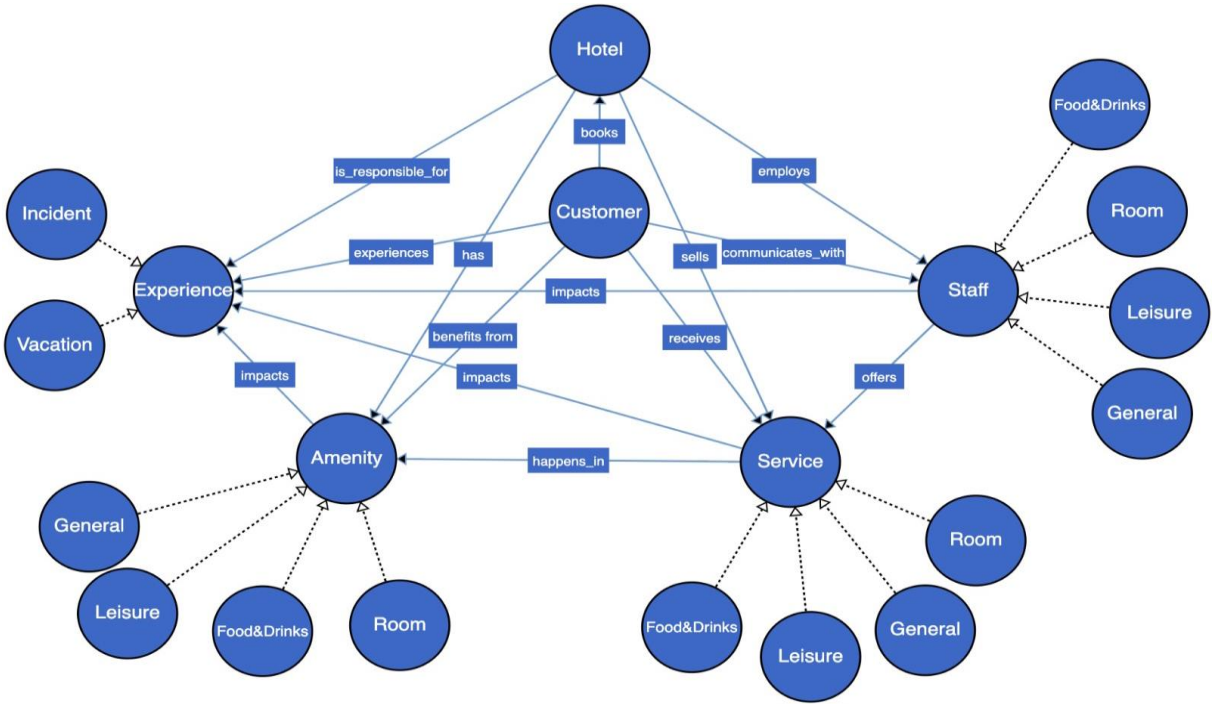
[29] J. E. H. Fritzner, "Automated Information Extraction in Natural Language (Master's thesis, NTNU)," Norwegian University of Science and Tehnology, Master's thesis 2017.

[30] L. Tonin. (2017) Digitala Vetenskapliga Arkivet. [Online]. "http://www.diva-portal.org/smash/get/diva2:1087619/FULLTEXT01.pdf"

[31] A. Joshi, R. Lal, T. Finin, and A. Joshi, "Extracting cybersecurity related linked data from text.," in IEEE Seventh International Conference on Semantic Computing (pp. 252-259), 2013, pp. 252-259.

## Annexes

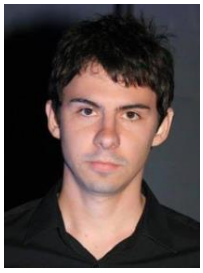
**Annex 1.** The extended version of the ontology, including the subclasses




**Annex 2.** The entities of the class Amenity

| Entity | Part of speech | Lemmas | | | |
|---|---|---|---|---|---|
| a la carte restaurant | noun | a la carte restaurant | a la carte restaurants | a-la-carte restaurant | a-la-carte restaurants |
| air conditioning | noun | air conditioning | air-conditioning | | |
| ambiance | noun | ambiance | | | |
| architecture | noun | architecture | | | |
| arrival area | noun | arrival area | arrival areas | | |
| atmosphere | noun | atmosphere | | | |
| balcony | noun | balcony | balconies | | |
| bali bed | noun | bali bed | bali beds | balinese bed | balinese beds |
| bar | noun | bar | bars | | |
| bathroom | noun | bathroom | bath | bath room | |
| beach | noun | beach | beaches | | |
| beach bed | noun | beach bed | beach bed | beachbeds | beach beds |
| beach chair | noun | beach chair | beach chairs | | |
| beach grill | noun | beach grill | | | |
| beachfront suite | noun | beachfront suite | beachfront suites | beach front suite | beach front suites |
| bed | noun | bed | beds | bedding | |
| bedding | noun | bedding | | | |
| blanket | noun | blanket | blankets | | |
| buffet restaurant | noun | buffet restaurant | buffet restaurants | | |

**Bogdan-Ştefan POSEDARU** graduated from the Faculty of Business and Tourism in 2012. He is currently a student at the Informatics Systems for the Management of Economic Resources Master program at the Bucharest University of Economic Studies. He worked as a freelancer in the IT field for 3 years. He was a member in various research projects, and since September 2018 he is one of the three co-founders of the start-up Chess Coders (https://chesscoders.com/). His main fields of interest are web technologies and natural language processing.



**Tiberiu-Marian GEORGESCU** graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2012. In 2015 he graduated the Informatics Systems for the Management of Economic Resources Master program. He completed his PhD program in Economic Informatics in September 2019 at the Bucharest University of Economic Studies. Currently, he is working as a Research Assistant in the Department of Economic Informatics and Cybernetics at the Bucharest University of Economic Studies. His main fields of interest are cybersecurity, machine learning and natural language processing.

**Florin-Valeriu PANTELIMON** graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2015. He is currently a student at the Informatics Systems for the Management of Economic Resources Master program at the Bucharest University of Economic Studies. He has been working as a software developer for several companies ranging from consulting companies, healthcare and clinical trials to international stock exchanges and game development companies. His main fields of interest are entrepreneurship, cloud computing and web development.

# Application for the efficiency improvement of the work process in an energy company

Mădălina-Florina DANALACHE, Simona-Vasilica OPREA
The Bucharest University of Economic Studies, Romania
danalachemadalina17@stud.ase.ro, simona.oprea@csie.ase.ro

*In modern society, electricity is necessary for the proper conduct of daily activities, becoming essential for life. All existing domains such as industry, transport, information technology, agriculture, economy, etc. requires energy resources to carry out the activities undertaken in optimal conditions. Thus, distribution companies invest enormously to transform networks into smart grids through a digitization effort that is constantly supported by modernization projects.*
*As an alternative with lower costs, an application can be implemented that helps consumers and employees regarding the the continuity of electricity supply, the quality of services provided to them and better communication between the consumer and the distribution operator.*
***Keywords****: electricity distribution company, consumer satisfaction , web application, prompt and quality services, automation of daily activities, easy reporting*

## 1 Introduction

Electricity is the most consumed source of energy in the world, becoming an essential part of modern life. According to the report made by A.N.R.E. in 2018 at national level, we can see the large number of consumers connected to the electricity network, more precisely 9.448.823, of which 5.170.629 are living in urban areas and 4.278.194 in rural areas. [1] Values have increased over the last few years.

The discovery of electricity led to the development of innovative technologies aimed to improve consumer satisfaction and companies efficiency from different domains of activity. Currently, the modernization of the energy domain involves the use of state-of-the-art technologies and equipment that can be controlled remotely: remote controlled separators, protection relays that allow the disconnection of the defective sector, smart meters which will benefit by the end of 2020 only 30% of the consumers. Due to the small percentage of beneficiaries and insufficient upgrades, there are still persistent problems such as: poor management of material stocks, long electricity outages, meticulous bureaucracy due to filling out paper forms, overloading telephone lines, significant waiting on a phone call to report an incorrect functioning of the electricity distribution network etc.

The main goal of this article is to present an alternative for electricity distribution companies, to which all users who have a device that allows internet connection have access. By using modern frameworks such as Laravel, Vue.js, the basic languages PHP, JS, CSS, HTML, the SASS preprocessor, the MySQL database but also the various APIs, I created an application that automates time and resource consuming physical operations by implementing features such as: online derangement reporting form, interactive map of derangements, digitized reports, prioritization of derangements, directions, calculation of travel distance, automatic sending of an SMS that notifies the consumer, calculation of monthly statistics at the company level etc. This innovation aims to improve the current level of the management system and increase performance at the company level.

## 2 Objectives

The developed application, digitizes within an electricity distribution company two areas: that of easy communication with consumers and the one of streamlining the work of network operators.

The first area developed will ensure the fulfillment of one of the objectives proposed by this type of companies more precisely, the customer loyalty and satisfaction.

Various surveys have shown that society is reluctant from various psycho-emotional reasons to call on strangers to report a problem, instead prefers to write a message on various applications or send an email. Another notable reason is the lack of time. The century of speed makes its presence felt and consumers do not have the patience to wait for the release of an operator to communicate the discomfort appeared.

To help these people and to increase their comfort, communication between consumers and distributors will be digitized through an online form with a series of standard questions to determine the type of inconvenience and whether certain measures have been taken before contacting the work team. For example, „Are there other consumers who do not have electricity?".

Before completing the form, the user can consult the map of existing derangements on the employees' work list. This will eliminate duplicate requests of fixing the derangement.

These functionalities do not only help consumers but also bring a benefit to dispatchers because during bad weather conditions or major electicity outages, they do not cope with the large number of consumers who want to report a problem. In those situations, consumers will choose to fill out the online form to reduce the waiting time. All this information from consumers will be transmitted in real time to the service team, without the need for a qualified person to inform employees by phone about the problems that have arisen.

The list of derangements to which the employee has access is ordered ascending according to the degree of importance (sick people, medium tension incidents, collective derangements, individual derangements), to their level of completion (available, taken over, started, completed), after the date of sending the report and depending on the distance of the service teams from the place where the derangement was reported.

Distribution operators may view additional consumer and derangement information and may request guidance to the location. An SMS notification system will be used to inform the person concerned that the derangement has been taken over and the intervention team is moving towards it, all this to eliminate the risk of her not being at home and also to make the work easier for employees because they no longer have to call her.

Completing a disturbance entails completing a questionnaire which, based on the answers, generates a report (observation note, movement report) which, by digitizing it, eliminates the large volume of handwritten documents and reduces the time to search or study them.

All the functionalities presented have methods behind which generate data for creating graphs and monthly statistics on working hours, employees performance, cost of materials, consumption of materials, time in which consumers ran out of electricity etc.

It can be seen that through all these automations brought to an electricity distribution company, the working time of the employees would be optimized, the derangements can be located more easily, the cost of traveling to them is reduced, the communication problems with the consumers are reduced, as well as those related to real-time inventory records would no longer exist.

All these economic problems will decrease significantly, and the company's profit may increase.

Although they are not visible in digital format, there are also improvements in the psycho-emotional level of the employees, the call center operators are not so overwhelmed and

the stress felt by them is considerably diminished, the work team no longer has so many manual tasks to do, being so much easier to view, complete and transmit data in the online environment .

## 3 Designing the Application

For designing the functionalities I used the relational database MySQL. Its logical schema is presented below, in **Fig. 1**.

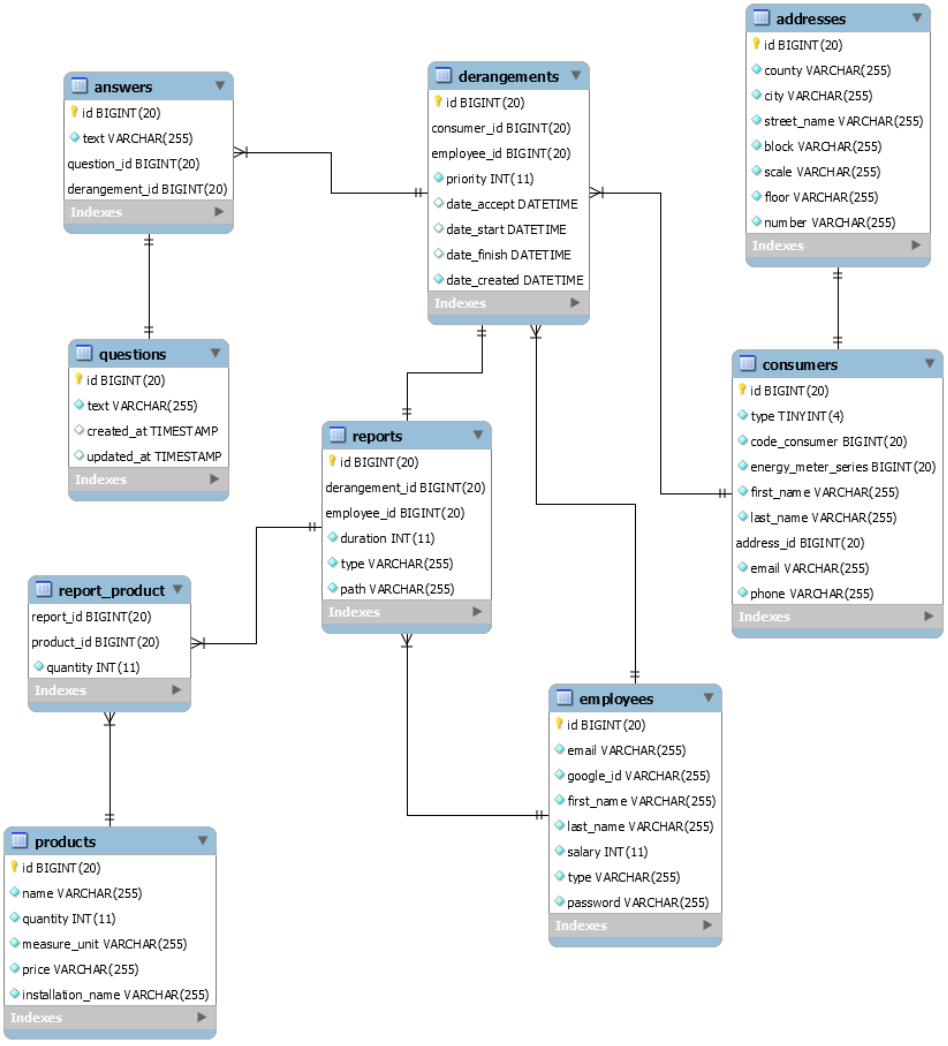



**Fig. 1** Logical Schema of the Database

The communication with the database is done through the Model component that belongs to the architectural model MVC (Model-View-Controller), which I will talk about in the following sections.

The user interface is an important part, because for most users this interface is not just a visual part of the application, but the entire computing system. I considered that the design of an interface should be as suggestive as possible and easy to use, as we can see in **Fig 2**. The readability of the functionalities, the minimization of the complexity, the colors used, the pictorial realism can be a plus for the users who will definitely return in case of need.

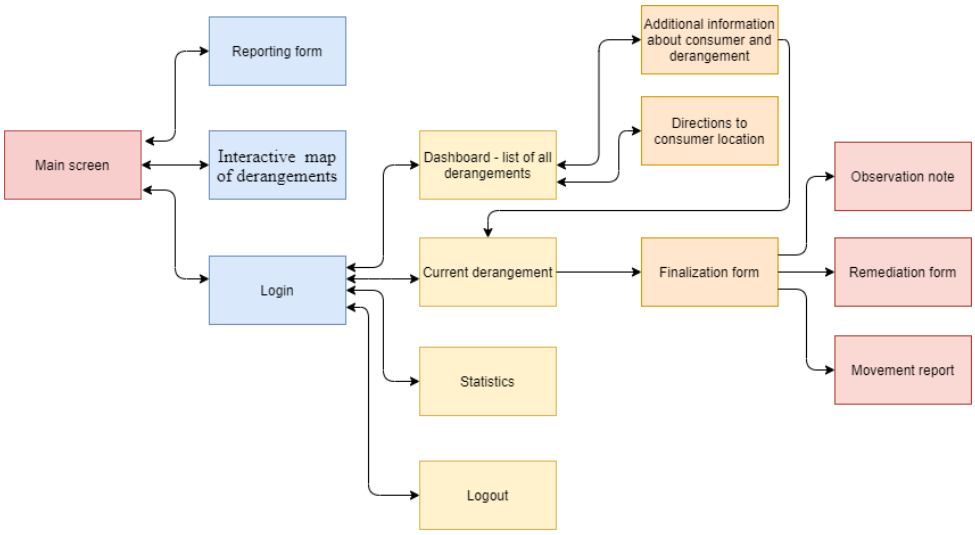



**Fig. 2** User Interfaces Diagram

## 4 Software Technologies

For the implementation of the web application I used **PHP**, **JS**, **CSS** and **HTML** as basic languages. They blend seamlessly with popular frameworks, such as PHP-based **Laravel** for server-side work (used especially for web route control and interaction with the **MySQL** database),**Vue.js** for the interface part, JS-based framework for modeling the information received from the server, but also the **SASS** preprocessor for using CSS in a much more organized way, useful for complex applications.

The web pages were written in HTML and stylized through CSS,based on the SASS preprocessor. It allows the use of a much more efficient and organized language than the common CSS. Once the application is compiled, the CSS file that is normally used is generated.

The dynamism and interactivity of the application is given by the complex visual effects created by using the JavaScript object-oriented programming language. The scripting language, PHP is mainly used on the back end of the website and has facilities such as: generating dynamic content, encrypting data, performing operations on data retrieved from the database, controlling user access etc. [2]

Using the PHP framework, Laravel, involves the use of the architectural model **Model View Controller (MVC)** transforming the code into a well-organized and very easy to manage, thus respecting the principle of software development DRY (Don't Repeat Yourself). Considering that this application is a complex one, this type of model is very useful because it isolates the logical part from the design part.

In order to achieve the highest possible level of performance of the developed application, I chose to use a second framework, Vue.js, integrated with Laravel. This is a progressive framework used to build user interfaces. [3] It is based on the concept of **Single-Page Applications (SPA)** which involves dynamically rewriting the current web page with new data from the web server, instead of loading a new page.

For database management I used MySQL, being currently the most popular open-source DBMS. I chose this management system because it is widely used together with the PHP programming language, on which the presented application is based and can be used standard SQL commands already known by me.

Developing the functionality of visualizing the derangements on the map and providing guidance involves the use of the **HERE**

**Maps API** for JavaScript. This is a set of interfaces that allow programmers to include in web applications interactive maps that can be seen on both mobile and desktop devices and can transform a partial address into a complete one with the possibility of determining the geolocation (latitude and longitude).
The functionality of automatically sending the notification SMS to the consumer is created using the **Vonage communications API**, which has a reduced cost of only 0.06 eurocents per SMS.
One last API used is **Pusher**. It provides real-time communication between applications, more specifically, between the report form and the employee dashboard and between the dashboard and the current derangement page. When a report is sent, the dashboard page is updated, adding according to certain criteria a fault in the appropriate section. The dashboard is updated again when a derangement changes its status. I used the Visual Studio Code development environment to write the source code, and XAMPP to host the website on my personal computer.

## 5 Methodology

As I mentioned in the previous sections, I used for my application the architectural model MVC (Model-View-Controller). The components of the MVC model are:

- Model- has the role of performing operations related to the database
- View- contains the records taken from the database and presented to the user.
- Controller- has the role of controlling the actions of a website.

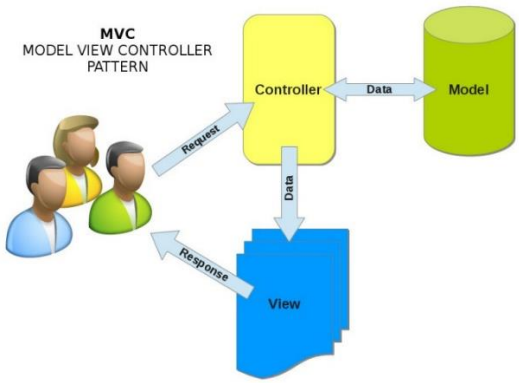



**Fig. 3** MVC architectural model [4]

As we can see in **Fig.** 3, the MVC mode of operation is simple: a user initializes a request, and the controller takes the received data and converts it to the meaning of the Model and sends it to it. The model connects to the database and retrieves the necessary information and then sends it back to the controller.
The controller processes the data received from the model and sends it to the View component. The latter, by compiling the data, generates an interface adapted to the user's requirements, prepared by a new interaction. As can be seen, the Controller is the intermediate component between View and Model, the two do not communicate directly with each other.

## 6 Application Interface Depiction

When accessing the site we are greeted by a friendly interface which by the way the buttons are placed denotes a high degree of accessibility to all age categories. The first page is an informative one whose purpose is to inspire confidence in consumers regarding the company's performance and the safety of the derangement reporting platform. (**Fig. 4**)

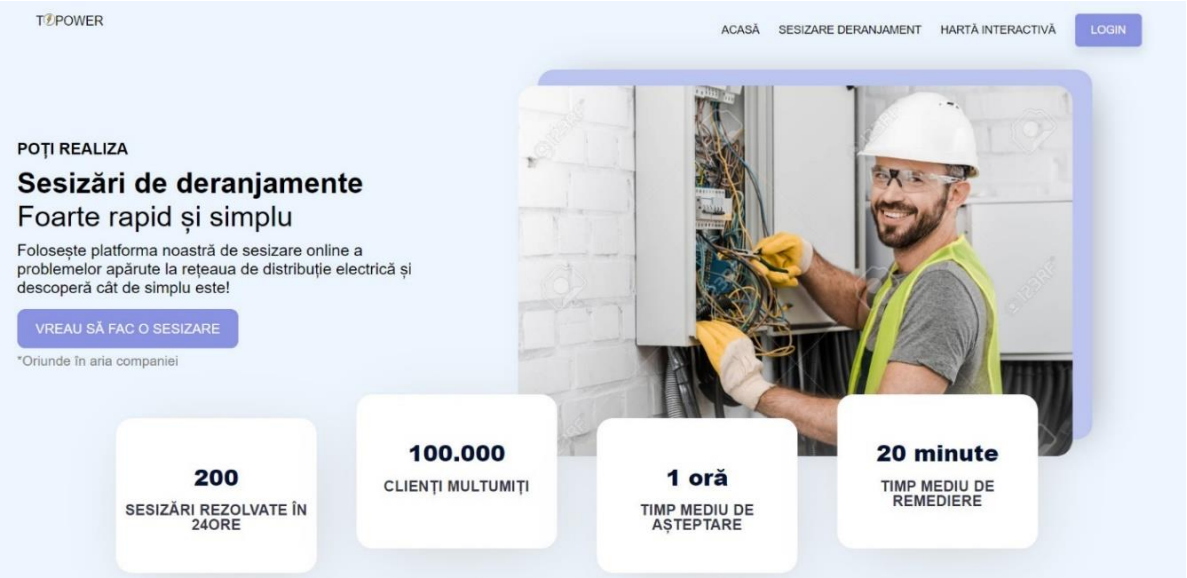



**Fig. 4** Main Screen

A feature available to the consumer is represented by the interactive map in which the current derangements are displayed through pins.

It should be mentioned that in the database I have stored only the consumer's address, and in order to display exactly his report on the map, I use an HERE Map function of transforming the address into geolocation. Additional information about them is displayed individually by pressing the desired pin. (**Fig. 5**)


**Fig. 5** Interactive map of derangements

The reporting form can be accessed through the central button "I want to make a report" or

from the navigation bar "Derangement reporting ". By pressing any button, a simplistic form will open, with 4 fields that have an average completion of approximately 2 minutes. (**Fig. 6**)

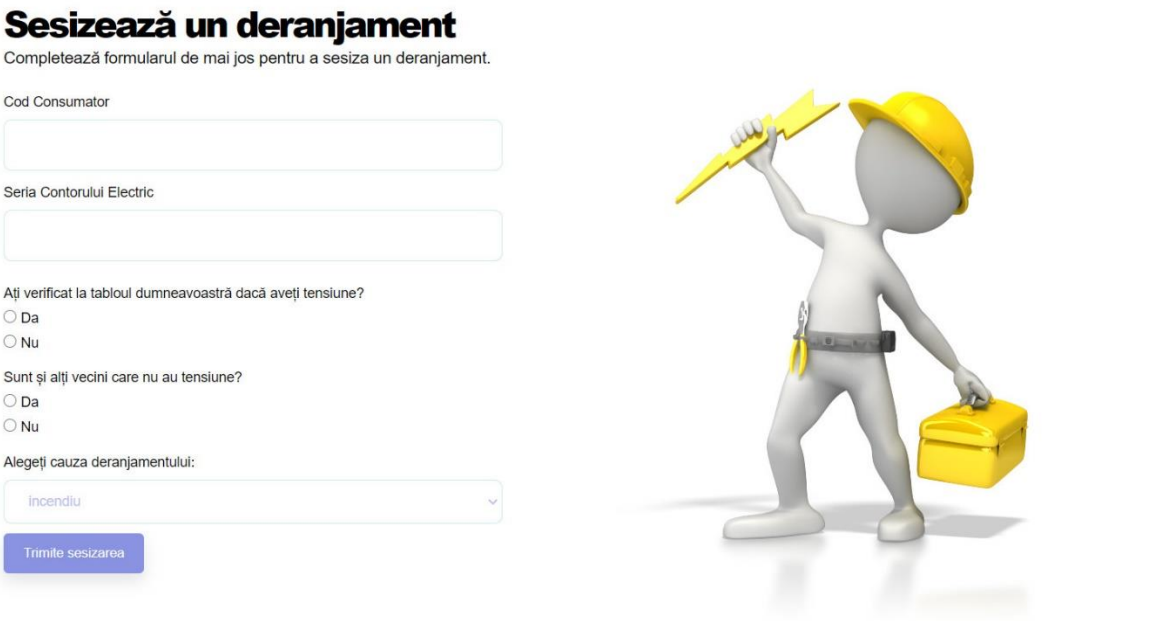



**Fig. 6** Reporting form

To avoid frauds, the consumer's identity is automatically verified in the database by filling in the consumer code or the meter series. A second verification is the visual one by the consumer and is represented by the display from the database of the address that matches the written numerical code. If there is already a derangement on that street in the database, a message will be displayed to avoid redundant requests.

Simply press of the "Send report" button triggers the appearance of a modal window aimed at notifying the consumer that his request is sent for solving. **(Fig. 7)** He has the possibility to print the confirmation to facilitate the retention of the generated id as well as the details about the report, important information in case of irregularities.

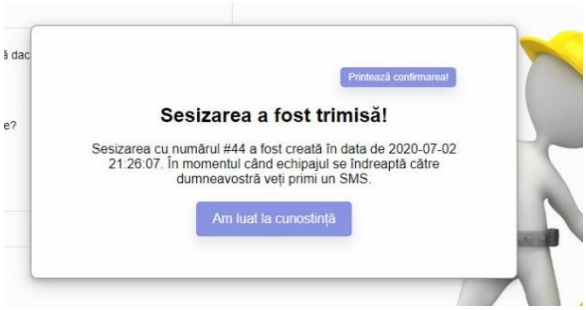



**Fig. 7** " Successful sent" modal screen

Authentication to additional app features can be done using the existing username in the database or through Google authentication, with the condition that the e-mail address is listed in the database. After authentication, the dashboard page opens and can be seen all the derangements that are less than 3 days old. After 3 days, only the completed ones will be deleted from the dashboard.

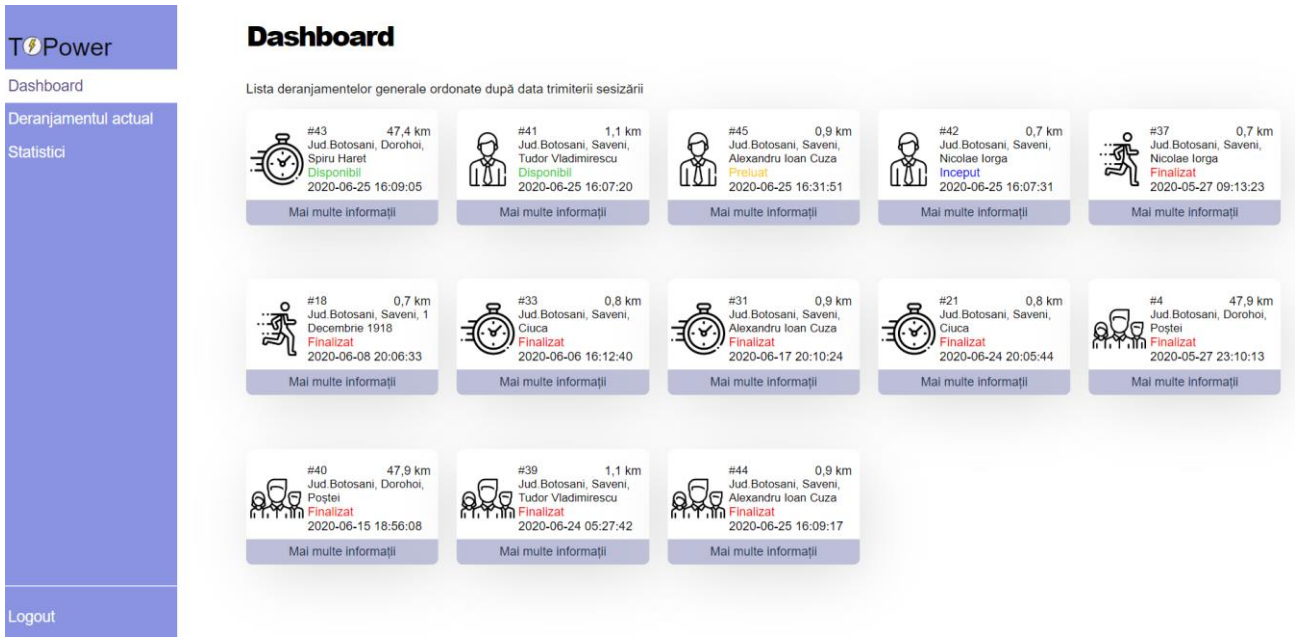



**Fig. 8** Dashboard

The priority of a derangement is generated based on the answers given by the consumer and is visually represented by a specific picture. (**Fig. 8**) In this way we can exemplify 4 types of priorities in the order of importance that the employee must give:

**Priority type 0 or sick people** – people who are dependent on equipment that need a continuous supply of electricity. In the database there is a record of sick people by that attribute "type", so when a consumer sends the form a check will be made immediately to generate this priority degree.

**Priority type 1 or medium voltage incidents** – the answer to the last question generates this priority. Options such as "fires", "fallen conductors", "broken poles", "other situations that present a risk of electric shock" are processed and set the priority of the derangement. The other drop-down options are ignored and other form questions will be considered.

**Priority type 2 or collective derangement** – after the number of reports or after the answer to the question 2 it is determined whether there is a significant number of consumers who do not have electricity.

**Priority type 3 sau individual derangement** – the answer to the question number 2 also determines whether the problem is an individual one, which is placed last in the ranking in the employees' dashboard.

If two derangements have the same status and the same priority, the employee is free to choose between the criterion of distance to the location or the date of sending the report.

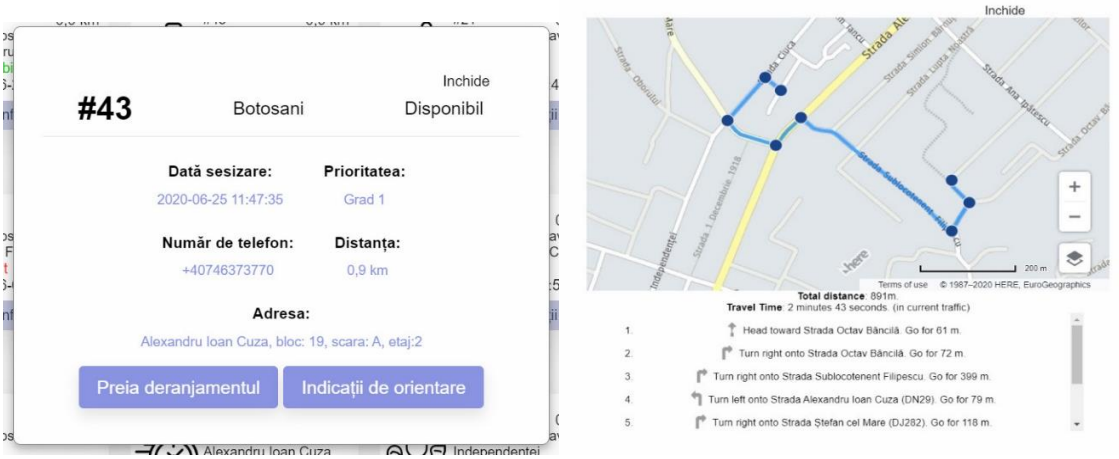



**Fig. 9** Additional information, Directions to consumer location

The employee can view additional information about the derangement and can ask for guidance if the address is unknown. (**Fig. 9**) Taking over a derangement ends up sending an automatic SMS to the customer, as well as constantly changing the derangement status, and through the web socket the status is updated in real time and in the dashboard. (**Fig. 10**)


**Fig. 10** Received SMS example

The completion of the derangement leads to a modal window represented in (**Fig. 11**), through which the employee answers to a maximum of 3 questions that establish what type of report he has to complete.

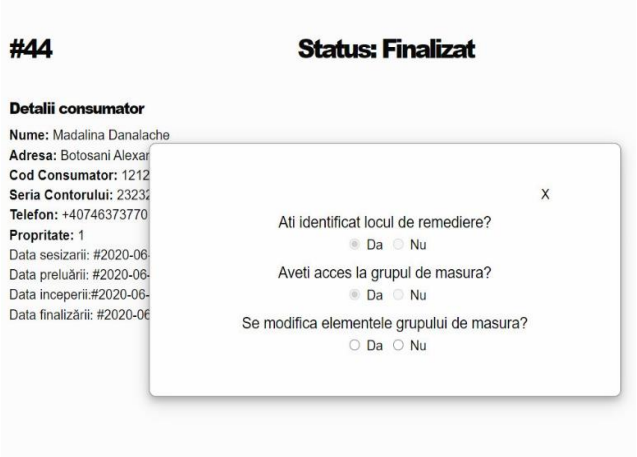



**Fig. 11** Finalization form

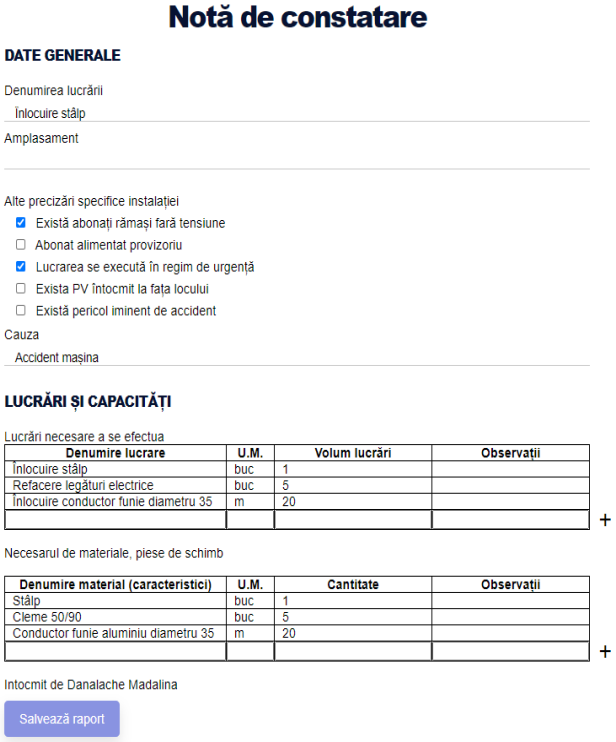
The 3 types of reports are:

**Observation note** – if the employee does not identify the place of remediation of the electricity network or does not have access to the measuring group, this type of report is completed with the works and materials necessary to complete the problem and will be sent to specialized teams. (**Fig. 12**)

**Remediation form** – the completion is determined by not modifying the elements of the measuring group, and through the database it will be possible to choose the name of the installation from a drop-down type element (LEA / LES / PT / FB / BMPM / BMPT) and the consumed materials corresponding to the

installation. These are dynamically subtracted from the database for a good management stocks.

**Movement report** – the change of the elements of the measuring group determines the completion of this type of report, to which are added, in addition to the previous report, two important tables: the characteristics of the meter and the elements of the sealed measuring group. They have inputs that can be easily filled in by employees. (**Fig. 12**)


Notă de constatare

DATE GENERALE

Denumirea lucrării

Înlocuire stâlp

Amplasament

Alte precizări specifice instalației

- [x] Există abonați rămași fară tensiune
- [ ] Abonat alimentat provizoriu
- [x] Lucrarea se execută în regim de urgență
- [ ] Exista PV întocmit la fața locului
- [ ] Există pericol iminent de accident

Cauza

Accident mașina

LUCRĂRI ȘI CAPACITĂȚI

Lucrări necesare a se efectua

| Denumire lucrare | U.M. | Volum lucrări | Observații |
|---|---|---|---|
| Înlocuire stâlp | buc | 1 | |
| Refacere legături electrice | buc | 5 | |
| Înlocuire conductor funie diametru 35 | m | 20 | |
| | | | |

+

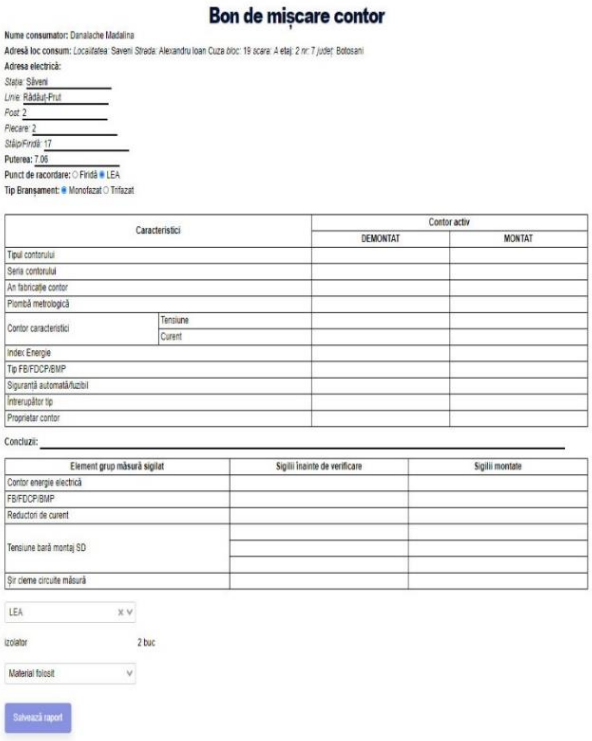
Necesarul de materiale, piese de schimb

| Denumire material (caracteristici) | U.M. | Cantitate | Observații |
|---|---|---|---|
| Stâlp | buc | 1 | |
| Cleme 50/90 | buc | 5 | |
| Conductor funie aluminiu diametru 35 | m | 20 | |
| | | | |

+

Intocmit de Danalache Madalina

Salvează raport


Bon de mișcare contor

Nume consumator: Danalache Madalina

Adresă loc consum: Localitatea: Saveni Strada: Alexandru Ioan Cuza bloc: 19 scara: A etaj: 2 nr: 7 județ: Botoșani

Adresa electrică:

Stația: Săveni

Linie: Rădăuți-Prut

Post: 2

Fiecare: 2

Stâlpi/Firidă: 17

Puterea: 7.06

Punct de racordare: ○ Firidă ◉ LEA

Tip Branșament: ◉ Monofazat ○ Trifazat

| Caracteristici | | Contor activ | |
|---|---|---|---|
| | | DEMONTAT | MONTAT |
| Tipul contorului | | | |
| Seria contorului | | | |
| An fabricație contor | | | |
| Plombă metrologică | | | |
| Contor caracteristici | Tensiune | | |
| | Curent | | |
| Index Energie | | | |
| Tip FB/FDCP/BMP | | | |
| Siguranță automată/fuzibil | | | |
| Intrerupător tip | | | |
| Proprietar contor | | | |

Concluzii: _____

| Element grup măsură sigilat | Sigilii înainte de verificare | Sigilii montate |
|---|---|---|
| Contor energie electrică | | |
| FB/FDCP/BMP | | |
| Reductori de curent | | |
| Tensiune bară montaj SD | | |
| | | |
| | | |
| Șir cleme circuite măsură | | |

LEA

Izolator 2 buc

Material folosit

Salvează raport

**Fig. 12** Observation note, Movement report

Compared to written reports, digital ones are much more accessible, some information coming automatically from the database.

The saving of each form is done by calling a PDF export function. The generated file is stored in the "Storage" folder within the application. After this last step, the employee can take over another derangement available in the dashboard.

By selecting the Statistics navigation link from the side menu, we will discover a page that is as visually appealing as it is simplistic in terms of access. The selection of a month determines the display of statistics that can be daily: the number of derangements resolved on that day, or monthly: the average time of electricity outage, the cost of used materials, the total cost of salaries, the number of derangements remedied in each type of priority, the average time to take over a derangement and the average travel time. All these statistics are represented by Pie, Bar, Line and Bubble charts. (**Fig. 13**)


**Fig. 13** Statistics

## 7 Future work

One of the main benefits of this application is the possibility to improve existing functionalities as well as the development of completely new ones.

The complexity of the functionalities can be extended by the possibility of processing large volumes of data:

- the existence of several counties under the tutelage of the company, divided into work centers
- employees to be divided into specialized teams
- the database may contain the consumer's electrical address, the characteristics of the existing meter, seals, etc.
- several types of installations and materials etc.

Another improvement is the creation of a private administrator account that can check the performance of employees (real-time location, travel speed, working hours, etc.)

## 8 Conclusions

The web Application for the efficiency improvement of the work process in an energy company is accessible from any device, with an attractive interface that has ease of use, created with modern frameworks, popular programming languages and facilities offered by APIs.

The usefulness of the application presented in

this article comes from the fact that it helps consumers in finding and noticing much easier failures in the electricity distribution network, as well as employees by eliminating standard work procedures performed manually.

I believe that the implemented functionalities are perfectly in line with the current requirements of the company and bring an extra novelty in the energy field. The main goal was to automate actions that normally generated "dead" times, all in order to satisfy consumers by providing quality services, to increase the company's performance and to maintain the continuity of electricity supply.

## 9 Acknowledgment

This paper presents the scientific results of the project "Intelligent system for trading on wholesale electricity market" (SMARTRADE), co-financed by the European Regional Development Fund (ERDF), through the Competitiveness Operational Programme (COP) 2014-2020, priority axis 1 – Research, technological develop-ment and innovation (RD&I) to support economic competitiveness and business development, Action 1.1.4-Attracting high-level personnel from abroad in order to enhance the RD capacity, contract IDP 37 418, no. 62/05.09.2016, beneficiary: The Bucharest University of Economic Studies.

## References

[1] A. N. D. R. Î. D. ENERGIE, "RAPORT NAŢIONAL 2018," ANRE, Bucureşti, 2019.

[2] "PHP Introduction," w3schools, [Online]. Available: https://www.w3schools.com/php/php_intro.asp. [Accessed 05 May 2020].

[3] "Introduction in Vue.js," vuejs, [Online]. Available: https://vuejs.org/v2/guide/. [Accessed 28 May 2020].

[4] M. Katalin, "Model-View-Controller," [Online]. Available: http://www.science.upm.ro/. [Accessed 15 May 2020].

[5] Micheu_Katalin, "Model-View-Controller," [Online]. Available: http://www.science.upm.ro/~traian/web_curs/Web_tech/lucr_stud/Micheu_Katalin.pdf. [Accessed 28 May 2020].

**Mădălina-Florina DANALACHE** (b. April 08, 1998) is a third year student of the Faculty of Economic Cybernetics, Statistics, and Informatics at the Bucharest University of Economic Studies. She has graduated Theoretic High-School "Dr. Mihai Ciucă" from Botoșani in 2017 with specialization in Mathematics-Informatics. She is familiar with Databases, C++, JavaScript, and the HTML, CSS languages that she masters at an advanced level in combination with modern frameworks such as Laravel and Vue.js.

**Simona-Vasilica OPREA** (b. July 14, 1978) received the MSc Degree through the Infrastructure Management Program from Yokohama National University, Japan, in 2007, the first Ph.D. degree in Power System Engineering from the Bucharest Polytechnic University in 2009, and the second Ph.D. degree in Economic Informatics from the Bucharest University of Economic Studies in 2017. She is involved in several national and H2020 European research projects as member or project manager. She is currently project director for a H2020 project entitled Multi-layer aggregator solutions to facilitate optimum demand response and grid flexibility (acronym SMART-MLA).

# The influence of art upon the feeling of life fulfilment

Iuliana COMAN
Iuliana.Coman.ARDS@gmail.com

*The aim of this paper is to analyse the influence of art upon the feeling of life fulfilment and the participation of art to the consolidation of the community.*
*The analysis is performed using data obtained from an experimental study on a sample of 120 persons with higher education in the south of Romania. Data were analysed using SPSS and Microsoft Excel and contain descriptive tables. The analysis took into consideration the comparison of the results obtained using a variety of statistical methods.*
*Following the analysis, it was concluded that the presence of art both influence the feeling of life fulfilment and contributes to the consolidation of the community.*
***Keywords:*** *Art, Society, Statistics*

## 1 Introduction

The hypothesis used in the development of this study is the major influence that art has on society, building, beyond the cultural identity that decisively defines a community, the necessary framework that ensures the progress both at the individual level and at the society level.
Starting from this hypothesis is also followed an assessment of the impact of art on the social context and on individual. The analysis aims both to obtain an image of the art impact and the identification of the dynamics that are manifesting in this market.
The influence of art upon the life of individuals and upon the life of society was the subject of several studies organized during last years. The capacity of art to influence the life of individuals was analysed in the study moderate by Sorensen, (2011), and the study presents the incontestable impact of art on researchers' life, in society and into individuals' life.
Pantano published her research (2011) revelling the influence of culture on the behaviour of consumers of local products in the Calabria area. The results of the study show the major influence which culture has on consumers behaviour. Bachleda and Bennani (2016) describe the impact of the psychological components on the visual arts, and the results show that the most powerful influence on the behaviour of visual art consumers are several traits of personality.
Croitoru and Becuț (2017), presents the tendency manifested at international level to evaluate the impact of art in the social and economic environment. Studies organized in France or Canada, demonstrate that art has a strong influence in society, developing the independence of individuals, expanding the capacity of knowledge, developing the understanding and the ability to act. Also, art participates in building social cohesion and developing society by encouraging civic participation.

## 2. Metodology

For this analyse, an experimental study was organised, on a sample of 120 people. The aim was to identify the influence of art on the population with higher education in the south of Romania. This target group was selected in since the high level of education implies a consistent exposure of this segment of the population to the interaction with art and that context allows a deeper analysis of the impact of art upon the population.
In defining the questionnaire, different studies and the statistical analyses carried out in recent years in the field of art was taken into consideration, allowing the identification of the main factors which are manifesting in this field.

Among the most important factors included in the questionnaire are those analysed in this paper: the presence of art in the lives of the respondents, the feeling of fulfilment in life, the material comfort and the participation of the art in strengthening the society.

**Statistical Methods**

The statistical methods used to analyze the associations in this paper are the following: Yule coefficient, Chi Square Method and Cramer's V test, Onicescu Informational Correlation.

**The coefficient of association Yule**

The coefficient of association Yule implies the drawing up of an association table for the variables of alternative type (YES / NO; F / M; etc.). The association table consists of two rows and two columns, in which at the end of the rows the values of the two associated characteristics are passed, and within the table the corresponding frequencies are passed.

***Table. 1. Example: Cross Table Considerations regarding Art Presence and Art Influence upon life fulfilment***

| | | Do you consider that Art was present in your life | | |
|---|---|---|---|---|
| | | Disagreement or partial accord | Total Acord | Grand Total |
| Do you consider that you have a fulfilled life | Disagreement or partial accord | 53 | 19 | 72 |
| | Total Acord | 19 | 29 | 48 |
| | Grand Total | 72 | 48 | 120 |

Source: Authors' own research

To determine the numerical value of the association coefficient indicating the existence and intensity of the connection, the coefficient of Yule is calculated according to the relation:

$Q= (n_{11}*n_{22} - n_{21}*n_{12})/(n_{11}*n_{22}+n_{21}*n_{12})$; $Q \in [-1,1]$

If: $Q = 0$ lack of association between $x_i$ și $y_i$

$Q \rightarrow 0$ weak association between $x_i$ și $y_i$

$Q \rightarrow \pm 1$ strong association between $x_i$ și $y_i$

$Q = \pm 1$ perfect association between $x_i$ și $y_i$

**Chi Square and Cramer's V Test – Assessment of the influence manifested between two variables**

Chi square test was one of the methods used for the analysis of the influences that manifesting in the world of art. (Coman, Mihaita, 2019). The method is used to identify the relationships between factors, is also called the Chi Square method, the association test, Chi, Hi or X2. The test was introduced by Karl Pearson (1857-1936) in 1900 and involves the verification of the hypothesis of association between the answers obtained in a questionnaire to the alternatives of a question and the verification of a certain data set that may follow a known statistical distribution.

The method include the definition of cross tables of the results intersecting the answers of two questions X with the alternatives Xi, where i = 1, ..., r placed as rows (r) of the table, and Y with the alternatives Yj, with j = 1,. ....., c placed in columns (c) of that table.

The questions considered as segmentation variables (independent, causal, extrinsic, exogenous) were placed in the columns of the table.
A concrete example of a cross Table was presented above - Cross Table Considerations regarding Art Presence and Art Influence upon life fulfilment.

The methodology for identifying the potential relationships include the next steps:
1. Formulation of the null hypothesis H0, which states that between the two variables-segmentation questions there is no causal link or association;
2. Choosing the significance level or threshold α and calculating the number of degrees of freedom of the table according to the formula (r-1) (c-1); based on these data, one assumes from the table of distribution $\chi 2$ its value, theoretically (index t);
3. Calculating the expected theoretical frequencies (expected, in case of a homogeneity test), according to the following formula:

$$\theta_{ij}=(\text{Total Line } i \times \text{Total Line } j) / \text{Total General} = T_{i.} * T_{.j}/T..$$

4. Calculation of $\chi 2$ (index c) using the formula:

$$\chi 2=\sum^{r}_{i=1}\sum^{c}_{j=1} [(x_{ij}- \theta_{ij})^2/ \theta_{ij}]$$

5. $\chi 2$ is compared with the one obtained from the distribution table $\chi 2$ as follows:
• if $\chi^2$ calculated > $\chi^2$ theoretical, the null hypothesis is rejected and therefore there is an association or potential relationship between the studied segmentation variables;
• if $\chi^2$ calculated < $\chi^2$ theoretical, the null hypothesis is accepted and therefore there is no association or potential relationship between the studied segmentation variables.
After identifying the existence of the association between the segmentation variables, the Cramer's V test is used to verify how strong the connection between the two variables is.

$$V= \sqrt{\chi^2/ [ (N)Min (r-1, c-1)]}$$

The scale of values that Cramer' V can have is the following:

| (0 – 0.10] | (0.10 – 0.3] | (0.3 – 0.5] | (0.5 – 0.7] | (0.7 – 1] |
|---|---|---|---|---|
| There is no association | weak association | Moderate association | Strong association | Very strong association |

**Informational Energy Onicescu**
The description of the method of calculating the Informational Energy, according to Rizescu and Avram (2014), and Mihaiță (1983), includes the aspects presented below. Informational Energy Onicescu comprises the amount of information generated by the diversity of a context. If we take in consideration a system S characterized by the states s1, s2, ..., sn having the corresponding weights p1, p2, ..., pn, where $\sum n\ i = 1\ pi = 1$.
The Onicescu Information Energy of the S system is calculated as the sum of the squares of the weights of the individual states:
$E_s=\sum^{n}_{i=1} pi^2$ .
The information energy values range from 1 / n to 1, $(1 / n \leq Es \leq 1)$.

The Information Energy reaches the value 1 / n when all the states have the same weight 1 / n (the uniformity of the system achieved) and 1 when one of the states of the system has a weight of 1 (and therefore all the other 0).

The information energy decreases in direct proportion with the increase of uniformity, or determination of systems. The information energy increases in direct proportion as the differentiation of the system increases.

The informational energy of a system composed of two or more independent elements is the product of their corresponding informational energies.

**Onicescu Informational Correlation**

Academician Octav Onicescu (1892-1983), is one of the greatest mathematicians of Romania with a remarkable international reputation. The basic concept of Onicescu information statistics is the Onicescu Information Correlation. This is a coefficient that has a remarkable property, namely its equality with 1 represents the identity of the distributions. In order to describe the Onicescu Information Correlation, the methodology presented by Oprea (2017), Mihaita and Capota (2005) or Onicescu (1971) were used.

If we consider two experiments A and B, characterized by a system with n events: A1, A2, ..., Year and B1, B2, ..., Bn, with the following probability distributions: p (A1) = p1, p ( A2) = p2,…, p (An) = pn; p (B1) = q1, p (B2) = q2, ..., p (Bn) = qn. The Onicescu Information Correlation between A and B, denoted IC (A, B) can be calculated with the following formula:

$IC(A, B)= \sum^{n}_{i=1} p_i q_i$

Information correlation can take values between 0 and 1.

$0< IC(A,B) <1$

The information correlation allows us to quantify the association between two event systems having common characteristics.

**The influence of the presence of art on the feeling of fulfillment in life.**

The analysis starts with the assessment of the association between these two factors using the Chi Square method.

***Table 2. Contingency table Presence of the Art - The feeling of life fulfilment***

| Presence of the Art /The feeling of life fulfilment | Disagreement or Partial Accord | Total Accord | Grand Total |
|---|---|---|---|
| Disagreement or Partial Accord | 53 | 19 | 72 |
| Total Accord | 19 | 29 | 48 |
| **Grand Total** | **72** | **48** | **120** |

Source: Authors' own research

***Table 3. Table Calculation Chi Square - Optimized Distributions Art Presence - The feeling of fulfilment in life***

| | Disagreement or Partial Accord | Total Accord | Grand Total | Hi $^2$ | C'S V |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 43.2 | 28.8 | 72.0 | 13.8 | 0.34 |
| Total Accord | 28.8 | 19.2 | 48 | | |
| **Grand Total** | | | | | |

Source: Authors' own research

Chi Square analysis reveals a moderate association, Cramer's V is 0.35 (between 0.3 and 0.5).

For the calculation of the Yule coefficient was constructed a table in which in addition to the contingency table was included a column with the Product of the two diagonals.

On the total line of this column, is included R = the ratio of the two products.

The Yule coefficient is calculated as Yule = (R-1) / (R + 1)

***Table 4. Calculation Table for Yule Coefficient - Association between Presence of Art - Feeling of Life fulfilment***

| Cause (Presence of Art) / Effect (Life Fulfilment) | Disagreement or Partial Accord | Total Accord | Total | Product of diagonals | Yule |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 53 | 19 | 72 | 1537 | |
| Total Accord | 19 | 29 | 48 | 361 | |
| Total | 72 | 48 | 120 | 4.26 | 62 % |

Source: Authors' own research

The table shows that the Yule coefficient also gives the same perspective - a moderate association between the two variables.

Going further, Onicescu Informational Correlation Coefficient is included. To calculate this coefficient a table is developed, starting from the table of contingency of the two variables two new columns.

In the first new column is included the Informational Energy of those two alternatives, on the total line of this column is included the sum of the two informational energies - representing the Informational Energy of the system.

In the second column is calculated the sum of the squares of the weights of the two segments generated by the effect variable, applied in turn in each of the two segments generated by the cause variable.

In the first line of the column is placed the square of the weights of the segments generated by the variable effect on the first segment generated by the variable cause, and in the second the square of the weights generated by the variable effect in the second segment generated by the variable cause.

In the last line of the column (total line), is calculated the Onicescu Correlation Coefficient as a ratio between the Informational Energy and radically from the product of the two values in line 1 and line 2 of the same column.

***Table 5. Table Calculation for the Coefficient of Correlation Onicescu - Association of the Present Art and the Feeling of life fulfilment***

| Cause (Presence of Art)/ Effect (Life Fulfilled) | Disagreement or Partial Accord | Total Accord | Total | Informational Energy | Kcor |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 53 | 19 | 72 | 0.29 | 0.61 |
| Total Accord | 19 | 29 | 48 | 0.16 | 0.52 |
| Total | 72 | 48 | 120 | 0.45 | 0.80 |

Source: Authors' own research

Onicescu Informational Correlation Coefficient also indicates the presence of an association between the two analysed variables.

A first conclusion can be reached: the presence of art has an influence upon the feeling of fulfilment in life.

**The influence of material comfort on the feeling of fulfillment in life**

The three methods described above will be applied for analysing the association between material strength and the feeling of life fulfilment.

***Table 6. Contingency table Material Comfort - The Sentiment of Fulfilment in life***

| | | Acord Comfort Material | | |
|---|---|---|---|---|
| | | Disagreement or Partial Accord | Total Accord | Grand Total |
| Do you consider that you have a fulfilled life | Disagreement or Partial Accord | 64 | 20 | 84 |
| | Total Accord | 8 | 28 | 36 |
| | Grand Total | 72 | 48 | 120 |

Source: Authors' own research

***Table 7. Calculation table for Chi Square - Optimized distributions Material comfort and Feeling of Life fulfilment***

| | Disagreement or Partial Accord | Total Accord | Grand Total | Hi Patrat | CRAMER'S V |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 50.4 | 33.6 | 84.0 | 30.58 | 0.505 |
| Total Accord | 21.6 | 14.4 | 36 | | |
| **Grand Total** | | | | | |

Source: Authors' own research

The table reveal that a strong association is present between material comfort and the feeling of life fulfilment. Cramer's V coefficient is in the range 0.5 - 0.7.

***Table 8. Yule Calculation Table - Material Comfort Association - The Sentiment of Fulfilment in Life***

| Cause (Material Comfort)/ Effect (Life Fulfilment***)*** | Disagreement or Partial Accord | Total Accord | Total | Product of diagonals | Yule |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 64 | 20 | 84 | 1792 | |
| Total Accord | 8 | 28 | 36 | 160 | |
| Total | 72 | 48 | 120 | 11.20 | 84% |

Source: Authors' own research

Yule coefficient also confirms the strong association between the two variables.

***Table 9. Table Calculation for the Coefficient of Correlation Onicescu - Material Comfort Association - Sentiment of Life Fulfilment***

| Cause (Material Comfort)/ Effect (Life Fulfilment*)* | Disagreement or Partial Accord | Total Accord | Total | Informational Energie | Kcor |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 64 | 20 | 84 | 0.37 | 0.80 |
| Total Accord | 8 | 28 | 36 | 0.06 | 0.51 |
| Total | 72 | 48 | 120 | 0.44 | 0.68 |

Source: Authors' own research

Also, the Onicescu Correlation Coefficient offers the same perspective: a strong association is identified.

A second conclusion could be the next: variables material comfort and the feeling of fulfilment in life are in a strong association, with a strong influence. More than this the influence identified is stronger than the influence between the presence of art and the feeling of life fulfilment.

**The influence of the presence of art in the consolidation of community**

The statistical metods presented were used for assesment of the asociation existing between the Presence of Art and the conviction regarding the capacity of art to particiapte to the Consolidation of the society.

**Tab 10. Contingency table Presence of Art - Consolidation of the society**

| | | Presence of art | | |
|---|---|---|---|---|
| | | Disagreement or Partial Accord | Total Accord | Grand Total |
| Consolidation of the community | Disagreement or Partial Accord | 41 | 9 | 50 |
| | Total Accord | 31 | 39 | 70 |
| | Grand Total | 72 | 48 | 120 |

Source: Authors' own research

***Tab 11. Table Calculation of Chi Square - Optimized Distributions Art Presence - Company Consolidation***

| | Disagreement or Partial Accord | Total Accord | Grand Total | Chi Square | CRAMER'S V |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 30.0 | 20.0 | 50.0 | 17.29 | 0.38 |
| Total Accord | 42.0 | 28.0 | 70 | | |
| **Grand Total** | | | | | |

Source: Authors' own research

According to the analyse an average association is revealed. In fact, was identified a stronger association comparing to the association between the Presence of art and the Feeling of Life Fulfilment and, a weaker association of it is compared with the association between the Material Comfort and the Feeling of Life Fulfilment.

***Table 12. Table for Calculation of Yule coefficient - Association between Presence of Art and Consolidation of the society***

| Cause (Presence of Art)/ Effect (Consolidation of the community) | Disagreement or Partial Accord | Total Accord | Total | Product of diagonals | Yule |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 41 | 9 | 50 | 1599 | |
| Total Accord | 31 | 39 | 70 | 279 | |
| Total | 72 | 48 | 120 | 5.73 | 70 % |

Source: Authors' own research

The calculation reveals the existence of an association. The value of the Yule coefficient is between the value of the coefficient calculated for the association Presence of Art - Life Fulfillment and the one calculated for the association of Material Comfort - Life fulfillment.

***Table 13. Calculation Table for the Coefficient of Correlation Onicescu - Material Comfort Association - Sentiment of Life Fulfilment***

| Cause (Material Comfort)/ Effect (Life Fulfilment | Disagreement or Partial Accord | Total Accord | Total | Informational Energy | Kcor |
|---|---|---|---|---|---|
| Disagreement or Partial Accord | 64 | 20 | 84 | 0.37 | 0.80 |
| Total Accord | 8 | 28 | 36 | 0.06 | 0.51 |
| Total | 72 | 48 | 120 | 0.44 | 0.68 |

Source: Authors' own research

The Coefficient of Correlation Onicescu take a value situated between those calculated for previous associations. The Onicescu correlation coefficient calculated for the association of Life fulfilment - Material comfort is found between the coefficient calculated for the association between the Presence of the art and Life fulfilment and the coefficient for the association of Material comfort - Life fulfilment.

**Conclusions**

Although at an intuitive level is considered that art has an influence on each one of us, or on society, in this paper these influences of art are analysed using statistical methodologies.

Art is present in the life of respondents, only 6% of respondents expressed their disagreement (total or partial) regarding the presence of art in their life. In the same time, most of respondents (81%) consider that art influence the life of respondents. The

analysis carried out went more in depth and confirmed that Presence of art manifest an influence upon the sentiment of Life Fulfilment. The results of the analysis revealed the existence of a moderate influence of the art's presence upon the feeling of life fulfilment.

91% of the respondents consider that art participate to the consolidation of the community and of the society. Also, the presence of art manifests an influence upon the conviction regarding the capacity of art to participate to the consolidation of the community, which is another conclusion of the analysis carried out.

Moreover, the influence of art upon the conviction regarding the consolidation of society has proved to be stronger if is compared with the influence of presence of art upon the feeling of life fulfilment.

It should also be mentioned that the analysis was carried out using three statistical methods (Chi Square Method and Cramer's V Test, Yule Coefficient and Onicescu Informational Correlation), all of them revealed the same influences and similar intensities of the analysed associations.

**References**

[1] Bachled, C. L., Asmae, B., (2016), Personality and interest in the visual arts, Arts and the Market, Emerald Publishing

[2] Croitoru, C., Becuț, A., Institutului National de Statistica (2017) Barometrul de Consum Cultural O radiografie a practicilor de consum cultural. 2017 Edition

[3] Coman, I., Mihaita, N. (2019) Factors influencing the impact of Art on the life satisfaction, Proceedings of the 8th INTERNATIONAL CONFERENCE SYNERGIES în COMMUNICATION Bucharest, ASE, Romania

[4] Rizescu, D., Avram, A., (2014) Using Onicescu's Informational Energy to Approximate Social Entropy, Procedia - Social and Behavioral Sciences · February 2014

[5] Mihaita, N., Stanciu-Capota, R., (2005) Relations statistiques fortes, cachees, fausses, et illusories Applications de la statistique informationelle - edition bilingue, Publishing House ASE

[6] Mihăiţă N. (1983) Onicescu Information Statistics în a Multiple Marketing Data Proces și ng Methodology, Economic Computation and Economic Cybernetic Studies and Research, review no. 2.

[7] Oprea, M., (2017) An Overview on the Contributions of the Academician Octav Onicescu to the Informational Statistics and Further Developments, International Conference on Virtual Learning VIRTUAL LEARNING – VIRTUAL REALITY.

[8] Pantano, E., (2011), Cultural factors affecting consumer behavior: a new perception model, Euromed journal of business, 2011

[9] Sorensen, D., (2013), Advancing Fields of Knowledge, Harvard Web Publishing.

**Iuliana Coman**
PhD Candidate
CSIE, ASE, Bucuresti

Iuliana Coman graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Academy of Economic Studies in 1997.

Her experience includes over 20 years of expertise in Marketing Management in important companies, leaders in the Romanian market – Dacia Renault Nissan, Altex Romana, Mediafax, Toyota Romania, GfK Romania.

# Exploiting stack-based buffer overflow using modern-day techniques

Stefan NICULA[1], Razvan Daniel ZOTA[1]
[1]The Bucharest University of Economic Studies
niculastefan13@stud.ase.ro, zota@ase.ro

*One of the most commonly known vulnerabilities that can affect a binary executable is the stack-based buffer overflow. The buffer overflow occurs when a program, while writing data to a buffer, overruns the buffer's boundary and overwrites adjacent memory locations. Nowadays, due to multiple protection mechanisms enforced by the operating systems, the buffer overflow has become harder to exploit. Multiple bypassing techniques are often required to be used in order to successfully exploit the vulnerability and control the execution flow of the analysed executable. One of the security features designed as protection mechanisms is Data Execution Prevention (DEP) which helps prevent code execution from the stack, heap or memory pool pages by marking all memory locations in a process as non-executable unless the location explicitly contains executable code. Another protection mechanism targeted is the Address Space Layout Randomization (ASLR), which is often used in conjunction with DEP. This security feature randomizes the location where the system executables are loaded into memory. By default, modern-day operating systems have these security features implemented. However, on the executable level, they have to be explicitly enabled. Most of the protection mechanisms, like the ones mentioned above, require certain techniques in order to bypass them and many of these techniques are using some form of address memory leakage in order to leverage an exploit. By successfully exploiting a buffer overflow, the adversary can potentially obtain code execution on the affected operating system which runs the vulnerable executable. The level of the privilege granted to the adversary is highly depended on the level of privilege that the binary is executed with. As such, an adversary may gain elevated privileges inside the system. Most of the time, this type of vulnerability is used for privilege escalation attacks or for gaining remote code execution on the system.*

***Keywords:*** *stack buffer overflow, return-oriented programming, libc attack, exploiting buffer overflow, stack protection mechanisms, address memory leak*

## 1 Introduction

The stack-based buffer overflow is one of the most commonly known vulnerabilities and it still one of the most exploited vulnerabilities that are affecting software and operating systems [1]. A successful exploitation of this vulnerability can sometimes be difficult to achieve and modern operating systems nowadays have protection mechanisms in place in order to prevent such issues from being exploited. These protections can also be implemented at the binary level in order to increase its security level. However, certain techniques can be used in order to bypass these prevention mechanisms but all the techniques described do need auxiliary information in order to be implemented. The study presented will be focused on Intel architecture x86, being more targeted around the Linux operating system internals and having as the main scope achieving code execution on the underlying operating system. The choice for Intel x86 architecture was being made by taking in consideration the significant difference between the x64 and x86 regarding calling conventions, general stack frame usage and registers. The x64 is far more complex compared to the x86 counterpart.

The paper will approach the exploitation of a stack-based buffer overflow by analysing the current exploitation techniques available, different protections implemented at the operating system level and on binaries. The

paper continues with the analysis of bypass techniques for the aforementioned protection mechanisms, a case study that applies some of those concepts, statistics about current exploit numbers and conclusions.

## 2 Exploitation prevention mechanisms

The buffer overflow has been an active research topic through the history of Computer Science and multiple aspects have been addressed in order to prevent different exploitations. We can encounter multiple protection mechanisms that prevent overflow from occurring or react once the overflow happens [2].

Data Execution Prevention (DEP) is implemented at the binary level and dictates the execution privilege on a memory location. This protection prevents malicious code from being executed directly from the buffer value by allowing only specific memory locations to have execution privileges. Only certain memory blocks have execution privileges if they explicitly request so [3].

Address space layout randomization (ASLR) is implemented by the binary or by the operating system. This protection mechanism randomizes the memory address of the binary and external libraries each time it gets executed. As such, every attack which is based on static known values will fail [4].

Stack canaries/cookies assure that the stack data is not corrupted or overwritten from untrusted user-supplied data. This method works by placing a small randomly chosen value inside the program stack, in memory, just before the stack return pointer. Because the buffer overflow is writing stack memory from lower to higher address, the return pointer will be overwritten and thus the stack canary will also be modified [5].

Partial or full RELRO removes all the dynamic linked functions and ensures that the Global Offset Table (GOT) is read-only. By making GOT entries read-only, an adversary can no longer overwrite external function call addresses to a controlled stack memory address [6].

Position Independent Executable (PIE) is an optional feature that can be used at compile-time which makes the executable behaviour as a dynamic external library at linking and loading time. This feature adds more randomization in the linking and loading process. A note here is that ASLR predates PIE and ASLR does not require PIE to be enabled [7].

These protection mechanisms can prevent the exploitation of a buffer overflow and can further limit an adversary's possibilities. For most of these mechanisms, an auxiliary vulnerability that can obtain a memory leak address is mandatory in order to bypass them.

## 3 Exploitation techniques and protection bypasses

Exploitation techniques can vary greatly depending on each buffer-overflow case; as a result, a full exploit payload will be subjective and customized depending on the environment, the software targeted and the operating system internals. A series of protection mechanisms are presented by every single layer mentioned. From these, some have evolved into security best practices implementation, while others are still struggling to get traction. Nonetheless, we can identify some common mechanisms that can be encountered on a normal environment setup enforced with the latest default protections. This general classification will be detailed in the next sub-menus, approaching runtime protections on memory level enforced by the underlying operating system and protections implemented on the binary level.

### 3.1 Bypassing DEP and ASLR

Some of the most common identified protection mechanisms are the DEP and ASLR. The Data Execution Prevention mechanism is implemented at the binary

level. This protection mechanism allows only specific stack frames to have execution privileges. This translates in the fact that data written in arbitrarily chosen stack-frames cannot be referenced by the instruction pointer to be executed. A good example would be a buffer overflow vulnerability that can be exploited in order to point the instruction pointer to a specific address inside the stack which is controlled by our buffer input. In this scenario, even though we have control on the Instruction Pointer, we cannot execute data that is being held in the stack frame which we overwrite. In order to bypass this protection, a technique called Return Oriented Programming (or ROP chain) can be used. By using this technique, which can also be referred to as Return to libc attack [8], we can bypass the DEP protection [9] by re-using code already present in the exploited binary. Sometimes, the studied binary does not have all the needed instructions inside its base-code in order to fully exploit an existing buffer overflow. As such, the libc attack can be used. Inside the libc library, we can re-use a variety of instructions to fulfil the scope of exploiting the vulnerability. To use the libc code-base, we need to further leak an address inside the targeted binary. This can be achieved, for example, by chaining a format string vulnerability affecting the vulnerable binary.

In regards to the format string vulnerability, this particular one is sometimes crucial in exploiting a vulnerable binary that has ASLR protection on. This is mainly due to the fact that ASLR protection is implemented by the binary or by the operating system [10]. In modern operating systems, the ASLR protection is implemented by default. As such, all the external libraries linked to the targeted binary are having randomized address values. However, the binary itself can opt to have the ASLR protection in-place. By doing this, the binary will randomize its instruction addresses and memory maps; each time the binary is executed. In order to defeat this protection mechanism, a vulnerability (information leak) such as a format string can be used in order to leak a base address that can be further used by the developed exploit. Another method of defeating this popular protection mechanism is to use a potential buffer overflow together with calling a function that uses stdout in order to print results. This can be further used by leaking GOT and PLT addresses in order to reveal libc base addresses [11]. Using the obtained libc base address from the function memory address leak, we can pinpoint the exact version of the library used by the executable. In this way, all the other function references can be calculated based on the initial libc version.

### 3.2 Return Oriented Programming and Return to libc attack

We can particularly note the concept of gadgets in a ROP chain. Gadgets are a set of instructions that serve our purpose of manipulating the executable in order to achieve our scope. Gadgets are pieces of code from the executable, commonly found in the loaded external libraries but can be found in the local binary code as well[12]. By using them, an adversary can do a variety of actions such as invoking syscalls while keeping the execution flow by always returning inside a stack controlled memory address. The need of RET opcode is mandatory for gadgets in order for us to keep the execution flow. Certain gadgets require different parameters that should be placed on the stack accordingly in the payload, before the function call. We know that functions are receiving parameters from the stack and because we control the stack using our buffer overflow, we can pass arguments to the called functions. In this way, we can create a chain of multiple gadgets that will provide the capabilities of executing code on the underlying operating system through the usage of the binary affected by a buffer overflow.

By comparing a typical Windows ROP chain with a Linux ROP chain, we can identify two different approaches that are quite common. For the Windows environment, oftentimes, the ROP chain's purpose is to make the stack executable and basically disable DEP using Windows API calls such as VirtualProtect, VirtualAlloc or NtSetInformationProcess whereas, for the Linux counterpart, the technique usually relies on executing directly a system command. The magic gadget from C, which basically is a code residing in the libc library that when called it's opening a shell, is typically the goto exploitation technique when using a ROP chain on a Linux binary. Of course, there are also alternatives for disabling DEP as well, on the Linux side, for example, the ret2mprotect.

### 3.3 Magic gadget C

As mentioned previously, one gadget that can be used to exploit a buffer overflow using ROP chain is the so-called C/C++ Magic gadget. Almost all of the libc libraries contain a version of the magic gadget. Basically, this gadget is used for ROP chaining and consists of some code residing in the libc which, when executed, opens a shell. The magic gadget code has to either call execve or issue the corresponding syscall directly. In our case, /bin/sh is set as a first argument. [13]

### 3.4 GOT overwrite

Another exploitation technique is defined by using the Global Offsets Table to overwrite function entries in order to execute malicious code [14]. This attack method can be avoided by implementing RELRO which basically removes all the dynamic linked functions and ensures that the GOT is read-only [15]. An example of a successful GOT rewrite would be overwriting a libc address with a local stack-frame address that contains malicious code. This can be prevented by making the GOT read-only at the initial launch of the binary file.

## 4 Memory leak using stdout functions

Given the constraints, aforementioned that can be applied to a specific executable, successful exploitation of a stack-based buffer overflow requires a certain memory address leak. This can be achieved in multiple ways. One of the most common techniques is finding and exploiting a format string vulnerability which basically allows us to leak values from the stack. Format string vulnerability is a type of vulnerability which allows an adversary to control the format of the printed output. [16]

Another technique that I will discuss in the next chapter is related to using certain C/C++ functions that manipulate stdout in order to leak entries from the Global Offset Table (GOT). The GOT contains the direct address of the function inside the external libraries. At compile time, that address is unknown, the dynamic linker will populate the entry when the binary is executed and the loading and linkage routines are executed. Inside the studied binary, the Procedure Linkage Table (PLT) is holding the trampoline address value to the GOT. By invoking a stdout from using the PLT address to the GOT address reference, we can obtain the actual address of a function from the loaded external library. [17]

## 5 Case study example

The previous chapters enumerated a series of exploitation techniques that can be implemented in order to bypass specific protection and prevention mechanisms. Let's look at the following code snippet example which is vulnerable to stack-based buffer overflow:

```
int main(){
        char local_var [60];
        puts("Enter some input:");
        fflush(stdout);
        fgets(local_var,700,stdin);
        return 0;
```

}

Inside the main function, we can note the initialization of the "local_var" variable which is of type char vector of size 60. We can note the usage of the "puts" function which we will later abuse in order to obtain an address memory leak. The "fgets" function call is receiving a stream input of max size 700 from the standard input and stores the input inside our previously declared local variable. Since there is no boundary check on the "local_var", the user can provide a size larger than the allocated size 60 of the buffer, thus resulting in a buffer overflow scenario.
For this particular exploit, we will use gdb peda and pwntools as an example of automating certain tasks and for the ease of use that these tools are bringing to the table [18]. The binary will be dynamically compiled targeting i386-x86 architecture for the Linux platform so the compiled analyzed binary will be an ELF file designed for x86 architecture.
By checking the security feature of the binary after the compilation, we can observe the following:

ASLR: OFF
CANARY: disabled
FORTIFY: disabled
NX: ENABLED
PIE: disabled
RELRO: partial

By investigating the checksec output from gdb peda, we can observe that this is a classic buffer overflow example that can be exploited using ROP chain gadgets. We can note that the binary does not have local ASLR enabled and we also note the lack of stack canaries. Even if the ASLR is disabled for the binary, the external libraries used are subject to randomization due to the ASLR enforced by the operating system [19]. We can also note that the NX (not executable) feature is enabled. By studying this particular case, we can note that the buffer overflow can be exploited but it will require a ROP chain in order to achieve code execution. That's because the NX privilege is enabled which does not allow us to redirect the execution flow in our controlled buffer but the missing stack canaries protection means that next execution instruction can be overwritten with our chosen address[20].
Considering that the binary is compiled with dynamic libraries, we will require a memory leak to obtain a base address from the libc library. Since we have no ASLR enabled on the binary level, we can search for the address of the puts function which is a stdout function. By invoking the puts using the PLT address of the puts function from the GOT, we can obtain the puts address inside the actual libc library. We want puts to call itself on the Global Offset Table which will give us the address of the puts in the binary that changes every single time [21]. We can obtain the binary PLT address of the puts by using objdump on the compiled binary.
When an external function such as puts is called, and example of a function trace call would be the following:

puts@plt 0x400476 -> puts@got 0x601018 -> puts@libc linked address

Because the program is dynamically linked, the external libraries such as libc are resolved using PLT and GOT. The way the function trampoline works helps in this situation, the GOT entry for the puts function holds the dynamically resolved address for that specific function. The PLT contains the function trampolines to the GOT structure table. The function _dl_runtime_resolve will resolve GOT entries with the correct value for the puts function from libc.
After obtaining the puts address from the libc, we can calculate the base address of the libc itself. We need this information in order to create ROP gadgets based on libc. The library address is modified every time the binary is executed so we need to calculate its

base address in order to correctly reference other code snippets from inside the libc.
From the libc base address, we can use any specific gadget from the libc library which will provide us a reverse shell. One particular gadget described earlier is the Magic Gadget which uses a series of code snippets to execute syscall as execve into /bin/sh.[22]

## 6 Statistics

By analyzing the publicly available indexing measurements, we can draw some interesting conclusions based on the data collected. For example, **Fig. 1** and **Fig. 2** are showing the overall trend of the stack-based buffer overflow CVE list taken from the MITRE website from 2005 to 2019. Interestingly, the number of CVEs appears to remain constant and even increases starting from 2016 while also taking into consideration the peak reports recorded in 2007. We need to keep in mind that these are reported CVEs and they do not necessarily have a publicly available Proof of Concept or full exploit. For this data, we can refer to the exploit-db website where we can identify that all-time 321 entries are related to a stack-based buffer overflow. That means, only a handful of vulnerabilities from the ones reported annually also have publicly available exploits as well.

| Year | Count | Year | Count | Year | Count |
|---|---|---|---|---|---|
| 2019 | 114 | 2014 | 125 | 2009 | 276 |
| 2018 | 215 | 2013 | 107 | 2008 | 282 |
| 2017 | 173 | 2012 | 106 | 2007 | 394 |
| 2016 | 120 | 2011 | 144 | 2006 | 198 |
| 2015 | 134 | 2010 | 160 | 2005 | 129 |

**Fig. 1.** Stack buffer overflow CVE entries count according to MITRE

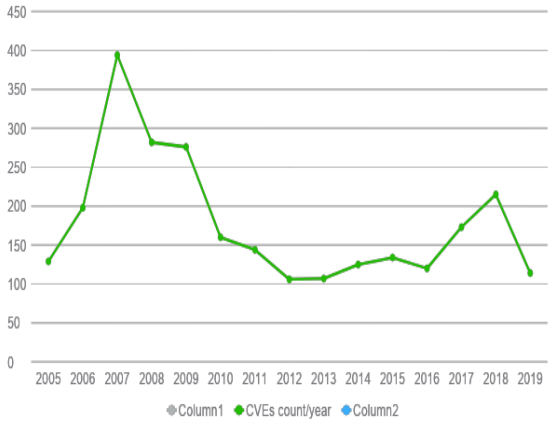



**Fig. 2.** Stack buffer overflow CVE entries flow chart according to MITRE

## 7 Edge cases and limitations

Compared to heap-based buffer overflows, the stack overflows can be considered much simpler yet they can present some interesting edge cases that are making the exploitation much harder.
Some buffer overflows could potentially be more situational than others. A good example would be the need of a partial overwrite of the EIP that is very unlikely although certain exploits do use this technique in order to bypass randomizations [24].
Limitations on exploitation can also include bad characters. Although they do not prevent the finding of the primitive buffer overflow, they are however hardening or sometimes even preventing full exploitation of the vulnerability, taking into consideration the other protection mechanisms in place as well.
Although extreme edge cases can be quite rare, full exploitation to bypass all the limitations requires a certain amount of analysis and dedication. Shellcodes can be generated while taking in consideration the bad characters as well however, the primary drawback is the size of the shellcode after the bad characters are applied. Usually, the shellcode size can exponentially get bigger with the increased number of characters to be avoided, sometimes even being impossible to generate position-independent code with too many bad characters in the blacklist [25].

Depending on the tested software, some common bad characters to be taken into consideration are 0x00, 0x0D, 0x0D and 0xFF. These characters should generally be avoided when building an exploit payload.

## 8 Windows vs Linux buffer overflow

On a Windows-based environment we can note specific particularities and situations when discussing stack-based buffer overflow exploitation. There are mainly two important differences that we note, we have the standard buffer-overflow that overwrites the saved returned pointer from the stack and we also have the SEH (Structure Exception Handling) based buffer overflows.

In a Windows-based software, if no explicit exception handlers are presented in the source code of the application, every thread will have an assigned handler and custom specific handlers will be added as an optional addition.[26] These values will be pushed onto the stack for each function and it will represent the pointers for treating different exceptions such as dividing by zero.

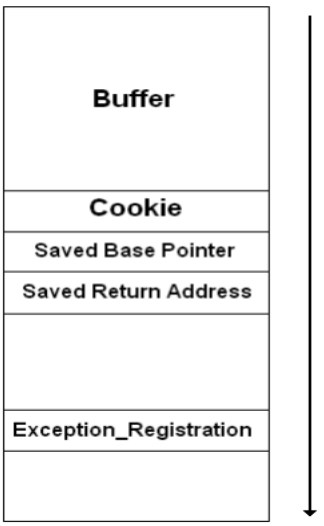



**Fig. 2.** Overwriting the stack with SEH entries

In a stack-based buffer overrun scenario, what would usually happen would be that the entries for the Structure Exception Handler will be overwritten by our buffer, resulting in a particular case where the return pointer is no longer our main EIP pivoting mechanism like in Linux. The SEH structure will no longer contain pointers inside their own exception handling routines but rather contain values overwritten by our buffer. This will cause the operating system to follow those values and consider them as valid addresses which would normally point to code paths that would resolve the exception.

When dealing with a SEH-based buffer overflow, a popular exploitation technique is the *pop/pop/ret* instruction set [27]. Due to the alignment on the stack for the EXCEPTION_REGISTRATION structure and the pointers associated with it, the overrun scenario often requires to pop-in two values of the stack and returning directly into our user-supplied shellcode. However, this is not always the case, depending on the situation, a SEH based buffer overflow could require multiple stack alignment moves in order to reach a known code cave.

Finally, another notable difference would be the ROP chaining creation process. Similar to the Linux case, after we take control over the EIP, we need to rely on built-in code or user-supplied shellcode in order to execute custom code on the machine. However, in our case, each Windows has a different DLL version even for the same build number, there can be differences in terms of Windows expansions, modifying the DLL version and offset itself. A hard-coded value can be used for the same deploy of Windows version but ultimately, the best approach would be a combination of memory information disclosure of a DLL base address followed by offset calculation to reach the needed function gadgets.

## 9 Conclusions

A stack-based buffer overflow can be exploited in multiple ways depending on a number of variables. First of all, the allocated buffer size can play a huge role in

choosing the right way to exploit the issue. In the previous case study shown, the buffer size was generous and allowed us enough space to inject various addresses and use multiple techniques without worrying too much about the memory space. If the buffer size was restricted to a small limited number of characters, additional steps would be required to successfully exploit the vulnerability. For example, an additional input buffer may have been required to redirect the execution flow into it however, that hypothetical input needed to be, again, controlled by the adversary. Some techniques used to get around the limited buffer size promote the usage of environmental variables that are loaded by the binary when executed. A memory leak address is needed in order to obtain such details. The second problem raised is related to the protection mechanisms that are preventing a straight forward exploitation technique. We cannot redirect the execution flow directly into our defined buffer due to DEP. Also, the address to libc functions is randomized each time we execute the binary given the ASLR protection enabled. We also have a disabled RELRO which should allow us some opportunities to overwrite the GOT-PLT entries inside the memory blocks. In order to bypass the aforementioned protections, a memory address leak is mandatory in order to obtain an address so we can further calculate our needed function addresses. A stack-buffer overflow cannot be exploited stand-alone, it can be situational and certain memory leak vulnerabilities are required given the protection mechanisms encountered in the process. A very important role is how to understand the internals of a program as well as properly identifying and using external libraries loaded by the executable in order to achieve code execution.

Buffer overflows are still emerging, active and real threats. Yearly, this specific vulnerability can be encountered in multiple CVEs reported on popularly known software [23]. In order to successfully exploit them, certain techniques are required in order to bypass common protection mechanisms. Nevertheless, these vulnerabilities are still found in solutions that have a high level of maturity in terms of security best practices and implementations. We should not overlook nor undermine their potential risk, even though modern-day systems are implementing multiple protection mechanisms in order to try and prevent such attacks.

**References**

[1] National Institute of Standards and Technology. ICAT Metabase.http://icat.nist.gov/

[2] Erick Leon, Stefan D. Bruda, Counter-measures against stack buffer overflows in GNU/Linux operating systems., The International Workshop on Parallel Tasks on High Performance Computing, Procedia Computer Science 83, 2016, Volume 83, pages 1301 – 1306

[3] A detailed description of the Data Execution Prevention (DEP) feature in Windows XP Service Pack 2, Windows XP Tablet PC Edition 2005, and Windows Server 2003, (Jul.2017) https://support.microsoft.com/en-us/help/875352/a-detailed-description-of-the-data-execution-prevention-dep-feature-in, retrieved Dec.2018

[4] Address Space Layout Randomization, (Mar.2003), https://pax.grsecurity.net/docs/aslr.txt, retrieved Feb. 2015.

[5] Buffer overflow protection, (Jun.2018), https://en.wikipedia.org/wiki/Buffer_overflow_protection#Canaries, retrieved Jan.2019

[6] Hardening ELF binaries using Relocation Read-Only (RELRO), (Jan.2019), https://www.redhat.com/en/blog/hardening-elf-binaries-using-relocation-read-only-relro, retrieved Jan.2019

[7] Position Independent Executables (PIE), (Nov.2012), https://access.redhat.com/blogs/766093/posts/1975793, retrieved Jan.2019

[8] Return-to-libc Exploit, (Feb.11), https://medium.com/@nikhilh20/return-to-libc-exploit-aa3fe6fb0d69, retrieved Mar.2019

[9] Bypassing DEP with ROP (32-bit), (Dec.2017), https://bytesoverbombs.io/bypassing-dep-with-rop-32-bit-39884e8a2c4a, retrieved Mar.2019

[10] Bypassing ASLR - Part I, (May 2015), https://sploitfun.wordpress.com/2015/05/08/bypassing-aslr-part-i/, retrieved Dec.2018

[11] Yan Fen, Yuan Fuchao, Shen Xiaobing, Yin Xinchun, Mao Bing, A New Data Randomization Method to Defend Buffer Overflow Attacks, International Conference on Applied Physics and Industrial Engineering, Physics Procedia 24, Volume 24, Part C, 2012, pages 1757-1764

[12] Bruce Dang , Practical Reverse Engineering: x86, x64, ARM, Windows Kernel, Reversing Tools, and Obfuscation, Wiley Publishing, 2014

[13] The magic gadget, (Sep.2016), )https://github.com/m1ghtym0/magic_gadget_finder, retrieved Apr.2019

[14] How to hijack the Global Offset Table with pointers for root shells, (Apr.2006), https://www.exploit-db.com/papers/13203, retrieved Apr.2019

[15] Ryan "elfmaster" O'Neill, Learning Linux Binary Analysis, Packt, 2016

[16] Format String Exploitation-Tutorial, https://www.exploit-db.com/docs/english/28476-linux-format-string-exploitation.pdf, retrieved Apr.2019

[17] P. Silberman and R. Johnson, A Comparison of Buffer Overflow Prevention Implementations and Weaknesses, presentation at Black Hat USA, Caesar's Palace, Las Vegas, NV, USA (Jul. 2004).

[18] Eldad Eilam, Reversing: Secrets of Reverse Engineering, Wiley Publishing, 2005

[19] Sahel Alouneh, Mazen Kharbutli, Rana AlQurem, Stack Memory Buffer Overflow Protection Based on Duplication and Randomization, The 4th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, Procedia Computer Science 21, 2013, pages 250 – 256

[20] G. Duarte. Epilogues, Canaries, and Buffer Overflows, (Mar. 19 2014), http://duartes.org/gustavo/blog/post/epilogues canaries-bufferoverflows/ , retrieved Feb. 2015.

[21] Ryan "elfmaster" O'Neill, Learning Linux Binary Analysis, Packt, 2016

[22] Smashing the Stack, (Apr.2014), http://phrack.org/issues/49/14.html, retrieved Oct.2018

[23] Mitre CVE Buffer Overflow search result, https://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=Buffer+Overflow, retrieved May.2019

[24] Kai Jander, Lars Braubach , Alexander Pokahr, Practical defense-in-depth solution for microservice systems, International Journal of ubiquitous systems and pervasive networks (JUSPN), volume 11, issue 1, 2019, pages 17-25

[25] Madjid Kara, Olfa Lamouchi, Amar Ramdane-Cherif, Software quality assessment algorithm based on fuzzy logic, International Journal of ubiquitous systems and pervasive networks (JUSPN), volume 8, issue1, 2017, pages 01-09

[26] Defeating the Stack Based Buffer Overflow pevention mechanism of Microsoft Windows 2003 Server, BlackHat Asia 03(Sept.2003), https://www.blackhat.com/presentations/bh-asia-03/bh-asia-03-litchfield.pdf, retrieved Aug.2018

[27] The need for a POP POP RET instruction sequence, (Oct.2010),

https://dkalemis.wordpress.com/2010/10/27/the-need-for-a-pop-pop-ret- instruction-sequence/, retrieved Oct.2019

**Stefan NICULA** graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2016 and followed a Master's degree in IT&C Security at the same university. He is a threat researcher and pentester with over 5 years of experience. His areas of expertise are in penetration testing, malware analysis, reverse engineering, and exploitation techniques, with a passion for Windows internals, vulnerability research, exploit development, and mitigation techniques. At present he is pursuing a PhD in Information Security at the Bucharest University of Economic Studies, focusing on heap memory exploits on browsers, Windows kernel vulnerabilities and fuzzing Windows API functions. Current publications and public presentations held by Stefan are covering areas such as IoT security evaluation and Windows binary exploitation, latest malware trends and recent developments in the exploit development field.

Răzvan Daniel ZOTA has graduated the Faculty of Mathematics – Computer Science Section at the University of Bucharest in 1992. He has also a Bachelor degree in Economics and a postgraduate degree in Management from SNSPA Bucharest, Romania. In 2000 he has received the PhD title from the Academy of Economic Studies in the field of Cybernetics and Economic Informatics. From 2010 he is supervising PhD thesis in the field of Economic Informatics, part of the Doctoral School of Economic Informatics in the Bucharest University of Economic Studies.

# Big Data Analytics in Smart Grids

Filip FEDELEȘ, Ionuț ȚARANU

*Data analytics are now playing a more important role in the modern industrial systems. Driven by the development of information and communication technology, an information layer is now added to the conventional electricity transmission and distribution network for data collection, storage and analysis with the help of wide installation of smart meters and sensors.*

*Big data has a potential to unlock novel groundbreaking opportunities in the power grid sector that enhances a multitude of technical, social, and economic gains. The currently untapped potential of applying the science of big data for better planning and operation of the power grid is a very challenging task and needs significant efforts all-around. As power grid technologies evolve in conjunction with measurement and communication technologies, this results in unprecedented amount of heterogeneous big data sets from diverse sources.*

***Keywords****: big data analytics, smart grid*

## 1 Introduction

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights. With today's technology, it's possible to analyze your data and get answers from it almost immediately – an effort that's slower and less efficient with more traditional business intelligence solutions.

Big data is a term used to describe massive amounts of information (**Figure 1**) that frequently occurs in the form of unstructured data sets that cannot be analyzed with standard database software.

The energy industry has worked with big data for years, regularly processing significant amounts of information produced on an intra-hourly basis.

Markets settle on metered data that measures power in five-minute increments. Utilities use supervisory control and data acquisition (SCADA) systems. Investors and planners run models with full representation of each generating unit, transmission load flow and hourly dispatch. Although other industries are relatively new to big data, they are finding innovative ways to use it. Applying these innovations to the energy industry promises to be transformative.

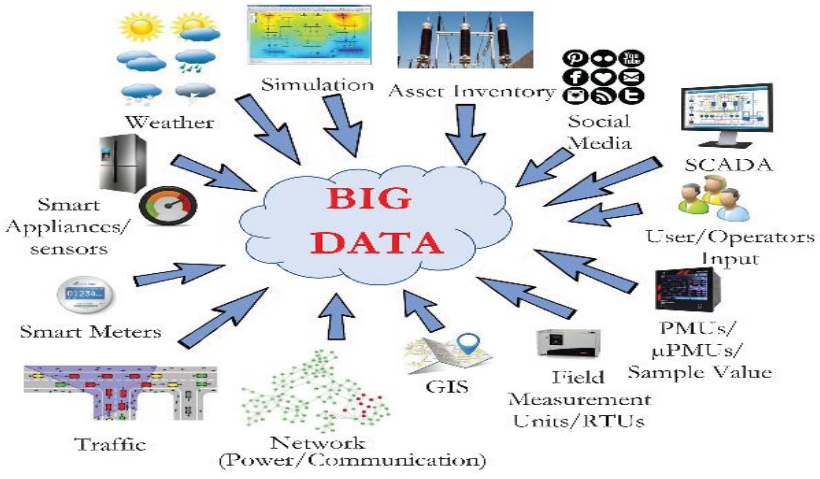



**Fig. 1.** Big Data ecosystem

## 2. Characteristics of Big Data

Big Data refers to the large, diverse sets of information (**Figure 2**) that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered. Big Data often comes from multiple sources and arrives in multiple formats.

Big Data can be categorized as unstructured or structured. Structured data consists of information already managed by the organization in databases and spreadsheets; it is frequently numeric in nature. Unstructured data is information that is unorganized and does not fall into a pre-determined model or format. It includes data gathered from social media sources, which help institutions gather information on customer needs.

The presence of sensors and other inputs in smart devices allows for data to be gathered across a broad spectrum of situations and circumstances.

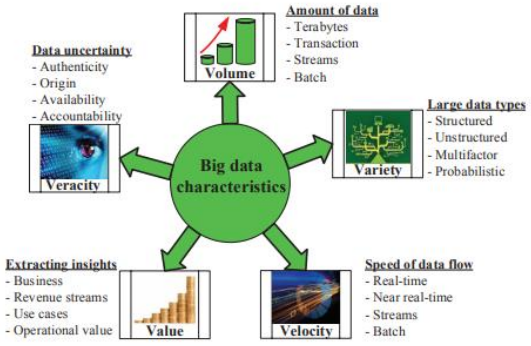



**Fig. 2.** Big Data characteristics

## 3.Characteristics of Smart Grids

Smart Grids comprise a broad mix of technologies to ptimize electricity networks, extending from the end user to distribution and transmission.

Not only can better technologies for monitoring, control and automation stimulate the development of new business models, they can unlock system-wide benefits including reduced outages, shorter response times, deferral of investments to the grids themselves and distributed energy resource integration.

At the end-user level, smart grids can enable demand flexibility and consumer participation in the energy system, including through demand response, electric vehicle (EV) charging and self-produced distributed generation and storage.

Demand flexibility can increase the overall capacity of the system to integrate variable renewables while accelerating the electrification of heating, cooling and industry at a lower cost. Deploying a physical layer of smart-grid infrastructure – underpinned by smart meters – can help unlock these benefits.

Electric power plants are generally dispatched so that the plants with the lowest operating costs (baseload plants) come on first, followed by more expensive plants when load increases, and finally, the most expensive plants during times of peak load1 . Very little electricity is stored for future use because storage is typically too costly. For this reason the marginal cost of supplying electricity is much higher during times of peak load. However, most electricity consumers are charged the same price for every kWh they consume. This is economically inecient as the prices consumers pay do not reflect the true costs of production. Advanced electricity pricing refers to a broad range of approaches and pricing programmes that try to make consumer prices more accurately reflect real-time production costs so that customers shift consumption toward times when electricity is less expensive. Advanced pricing can also shift consumption to times when RE is available. Three representative advanced pricing schemes are described further.

Electricity usage typically peaks around the same time every day in a given area. The simplest method of discouraging electricity use during peak times is to institute a time-of-use (TOU) price schedule, under which electricity is least expensive when loads are low (typically at night) and most expensive during peak times (usually afternoons).

Customers paying TOU rates may adjust loads manually or use building or home energy management systems (BEMS/HEMS) to control their loads. TOU pricing schemes may vary with the season but are generally set far in advance. This means TOU pricing does not help much on the few days per year when load approaches its annual peak. TOU pricing programmes are becoming common. TOU pricing is typically advantageous for solar PV, which produces power during the daytime, when the price is usually high.

Wind plant power forecasting has become a priority for grid operators as utility-scale wind plants have come to make up a significant portion of grid capacity in some areas. With wind penetrations around 25%, studies have shown that wind forecasting can save tens to hundreds of millions of dollars per year in operating costs over several states in the U.S. (Lew, et al., 2011). When NWP power forecasts for regional aggregations of wind plants are compared to actual wind power output for those aggregations, error rates of 5% are typical. Error rates for single locations are two to four times higher. Current day-ahead NWP error rates are not expected to drop significantly. Wind plants may also use very short-term (millisecond scale) wind nowcasting to optimize power output by dynamically adjusting the pitch of turbine blades (Madrigal, 2010). Light Detection and Ranging (LIDAR) and Sonic Detection and Ranging (SODAR) wind sensors located on turbines are used for this purpose. This technology is experimental.

## 4. Big Data applications in power distribution systems

The carbon emission reduction and sustainability of environment are the driving force and construction purpose of smart grid, which is designed in a decentralized structure. The employment of distributed generator units in modern power distribution system now provides an effective means for the utilization of widespread renewable energy such as wind and solar energy. These emerging microgrids are vital for the expectation of a low-carbon society. Moreover, the close distance between the generator and loads in microgrid improves the reliability of power delivery and reduces the power transmission loss. The ability to operate in an island mode also protects the load from damages caused by power system including voltage fluctuation, frequency deviation, etc.

Distribution automation (DA) is a concept of smart grid which focuses on the operation and system reliability at the distribution level. A successful DA has the capability to localize and isolate the faults in distribution system with a reduced restoration time and improved customer satisfaction. Under the concept of DA, increasing volume of operational data have been collected from supervisory control and data acquisition (SCADA) or advanced metering infrastructure (AMI) for state monitoring and fault diagnosis.

Thanks to the development of ICT technology in power systems, a huge volume of data can be collected via AMI and communication infrastructures. Power system operating data, weather information and log data of relay protection devices are processed as the input of a one class classification system, which is a data-driven model of fault phenomena based on a hybridization of evolutionary learning and clustering techniques. This fault recognition system is validated in the medium voltage power grid in Rome. The traditional statistical methods such as linear discriminant analysis (LDA) and logistic regression are discussed for mining the relation between power system faults and the features extracted from raw data.

Big data applications in distribution system planning can be divided into two categories

- Short term operations
- Long term planning studies

Short term applications are detection of energy theft, outage detection, peak load monitoring, customer consumption behavior modeling, special load and renewable forecast, distribution system visualization,

state estimation and distribution system planning, in which the first three applications are qualified to be very short term applications. Applications in Long term system planning studies include modeling customer consumption behavior under various incentives and pricing structures, transformation of distribution system planning process.

## 5. Implementation on a cloud computing platform

Cloud computing can be deployed as the infrastructure layer for big data systems to meet certain infrastructure requirements, such as cost-effectiveness, improved accessibility, and scalability. Based on the requirements of the proposed framework, Infrastructure as a Service (IaaS) clouds are appropriate to use to implement the smart grid big data framework. Cloud service providers such as, Amazon AWS and Google can be utilized to build a cluster that will host the framework. In this implementation, a Google cloud platform cluster with six machines is used.

As smart grid data increases exponentially in the future, utilities must envision ever-increasing challenges on data storage, data processing, and data analytics. Even though many electric utilities have realized that deployment of big data analytics is a must and not a choice, for future business growth and efficient operation, implementation of big data analytics in utility framework is lagging. Therefore, there is a need of comprehensive study to investigate current challenges, value proposition to stakeholders (e.g., consumers, utilities, system operators), operational benefits, and potential path forward to deploy big data analytics in power grids.

The high volume data gather in smart grid is similar in size and characteristics to the concept of big data. Big data is defined as data with high volume, velocity, and variety. The sampling frequency from perception devices can make the data size very large. Data velocity reflects the required speed for collecting and processing the data. Hence, big data management and processing techniques (hardware, software, algorithms, AI, etc) can be borrowed and applied in the domain of IoT. In addition, some applications of smart grid can perform their tasks only at specific time a day, such as weather forecasting and one-day ahead of time energy distribution, which can be performed at the night of every day. However some other applications perform their tasks all day round, such as real-time applications that monitor the power grid components. This is needed to speed up energy outage recovery process and real-time response to emergent behaviors in power demands. Even with today's development in big data processing techniques, managing of data in the smart power grid poses new challenges that are based upon the criticality of power systems, real-time response, proactive solutions, accurate predictions, and security. Hence, we address first the question of where to store the smart grid big data.

The increasing number in services and capabilities of cloud computing make it a good candidate to host SCADA systems. Cloud computing is a model that enables a convenient on-demand access to a shared pool of computing resources such as network, storage, servers, applications, and services. Cloud computing enterprises deliver their services to end users in three models namely, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). IaaS provides end users with operating systems, storage, network, and database services deployed within the cloud. PaaS provides end users with capabilities to deploy their applications such as programming languages and libraries that are available within the cloud. SaaS cloud provides a ready to use application for end users.

## 6. Key Challenges for Big Data Analytics

**Table 1**

| Challenges | Possible Impact | Potential Solution |
|---|---|---|
| Data Volume | Need of increased storage and computing resources | Dimensionality reduction, Parallel computing, Edge computing, Cloud computing, pay-per use |
| Data Quality | Lack of complete information, misleading decision | Probabilistic and storchastic analysis, data cleaning (e.g. dealing with missing values, smooth out noises, outliers, and inconsistent data) |
| Data Security | Vulnerable to malacious attack, compromise consumer privacy and integrity, mislead operational decision and financial transactions | Data anonymization (e.g. data aggregation, data encryption, P2DA) |
| Time Synchronization | Mislead operational decision, wrong interpretation of data, bad diagnostic of past events | Synchronize devices based on same radio clocks or satellite receivers |
| Data Indexing | Computational complexity, long processing time | Deploy new indexing techniques such as R-trees, Btrees, Quad-trees |
| Value Proposition | Non-acceptance by stakeholder, delay deployment of big data | Quantifying both technical and economic values to key stakeholders, namely consumer, system operator, utility. |
| Standards and Regulation | Interface challenges among various computing, storage, and processing platforms, delayed deployment | Regulatory entity define guidelines about data sharing/exchange, and standards should technically ensure regulatory aspects |

**Data volume**

Before we start to build any data processes, we need to know the data volume we are working with: what will be the data volume to start with, and what the data volume will be growing into. If the data size is always small, design and implementation can be much more straightforward and faster. If the data start with being large, or start with being small but will grow fast, the design needs to take performance optimization into consideration. The applications and processes that perform well for big data usually incur too much overhead for small data and cause adverse impact to slow down the process. On the other hand, an application designed for small data would take too long for big data to complete. In other words, an application or process should be designed differently for small data vs. big data.

This large amount of data exceeds the amount of data that can be stored and computed, as well as retrieved. The challenge is not so much the availability, but the management of this data. With statistics claiming that data would increase 6.6 times the distance between earth and moon by 2020, this is definitely a challenge.

Some of the newest ways developed to manage this data are a hybrid of relational

databases combined with NoSQL databases. An example of this is MongoDB, which is an inherent part of the MEAN stack. There are also distributed computing systems like Hadoop to help manage Big Data volumes.

**Data Quality**

Veracity, one of the most overlooked Big Data characteristics, is directly related to data quality, as it refers to the inherent biases, noise and abnormality in data. Because of veracity, the data values might not be exact real values, rather they might be approximations. In other words, the data might have some inherent impreciseness and uncertainty. Besides data inaccuracies, Veracity also includes data consistency (defined by the statistical reliability of data) and data trustworthiness (based on data origin, data collection and processing methods, security infrastructure, etc.). These data quality issues in turn impact data integrity and data accountability.

While the other V's are relatively well-defined and can be easily measured, Veracity is a complex theoretical construct with no standard approach for measurement. In a way this reflects how complex the topic of "data quality" is within the Big Data context.

Data users and data providers are often different organizations with very different goals and operational procedures. Thus, it is no surprise that their notions of data quality are very different. In many cases, the data providers have no clue about the business use cases of data users (data providers might not even care about it, unless they are getting paid for the data). This disconnect between data source and data use is one of the prime reasons behind the data quality issues symbolized by Veracity.

Data veracity, in general, is how accurate or truthful a data set may be. In the context of big data, however, it takes on a bit more meaning. More specifically, when it comes to the accuracy of big data, it's not just the quality of the data itself but how trustworthy the data source, type, and processing of it is. Removing things like bias, abnormalities or inconsistencies, duplication, and volatility are just a few aspects that factor into improving the accuracy of big data.

Unfortunately, sometimes volatility isn't within our control. The volatility, sometimes referred to as another "V" of big data, is the rate of change and lifetime of the data. An example of highly volatile data includes social media, where sentiments and trending topics change quickly and often. Less volatile data would look something more like weather trends that change less frequently and are easier to predict and track.

The second side of data veracity entails ensuring the processing method of the actual data makes sense based on business needs and the output is pertinent to objectives. Obviously, this is especially important when incorporating primary market research with big data. Interpreting big data in the right way ensures results are relevant and actionable. Further, access to big data means you could spend months sorting through information without focus and a without a method of identifying what data points are relevant. As a result, data should be analyzed in a timely manner, as is difficult with big data, otherwise the insights would fail to be useful.

**Data Security**

When producing information for big data, organizations have to ensure they have the right balance between utility of the data and privacy. Before the data is stored it should be adequately anonymized, removing any unique identifier for a user. This in itself can be a security challenge as removing unique identifiers might not be enough to guarantee the data will remain anonymous. The anonymized data could be cross-referenced with other available data following de-anonymization techniques.

When storing the data, organizations will face the problem of encryption. Data can't be sent encrypted by the users if the cloud needs to perform operations over the data. A solution for this is to use "Fully Homomorphic Encryption" (FHE), which allows data stored in the cloud to perform operations over the encrypted data so new encrypted data will be created. When the data's decrypted, the results will be as if the operations were carried out over plain text data. So the cloud will be able to perform operations over encrypted data without knowledge of the underlying plain text data.

A significant challenge while using big data is establishing ownership of information. If the data's stored in the cloud, a trust boundary should be established between the data owners and the data storage owners.

Adequate access control mechanisms are key in protecting the data. Access control's traditionally been provided by operating systems or applications restricting access to the information - this typically exposes all the information if the system or application is hacked.

A better approach is to protect the information using encryption that only allows decryption if the entity trying to access the information is authorized by an access control policy.

An additional problem is that software commonly used to store big data, such as Hadoop, doesn't always come with user authentication by default. This makes the problem of access control worse, as a default installation would leave the information open to unauthenticated users.

Big data solutions often rely on traditional firewalls or implementations at the application layer to restrict access to the information. The main solution to ensuring data remains protected is the adequate use of encryption. For example, Attribute-Based Encryption can help in providing fine-grained access control of encrypted data.

Anonymizing the data's also important to making sure privacy concerns are addressed. It should be ensured that all sensitive information is removed from the set of records collected.

Real-time security monitoring is also a key security component for a big data project. It's important organizations monitor access to make sure there's no unauthorized access. It's also important threat intelligence is in place to guarantee more sophisticated attacks are detected and the organizations can react to threats accordingly.

For example, many big data solutions look for emergent patterns in real time, whereas data warehouses often focused on infrequent batch runs. How do these different usage models impact security issues and compliance risk?

In the past, large data sets were stored in highly structured relational databases. If you wanted to look for sensitive data such as health records of a patient, you knew exactly where to look and how to access the data.

Removing any identifiable information was also easier in relational databases. Big data makes this a more complex process, especially if the data is unstructured. Organizations will have to track down what pieces of information in their big data are sensitive and then carefully isolate this information to ensure compliance.

Another challenge with big data is that you can have a big variety of users each needing access to a particular subset of information. This means the encryption solution you choose to protect the data has to reflect this new reality. Access control to the data will also need to be more granular to ensure people can only access information they are authorized to see.

## Conclusion

In this paper we explained every separate concept for big data, smart grid and cloud computing and how we can get all of them to work together for optimal end results.

We discussed the implementation of cloud energy storage devices, and cloud data storage mechanisms for the smart grid architecture. Using cloud computing applications, energy management techniques in smart grid can be evaluated within the cloud, instead of between the end-user's devices. This architecture gives more memory and storage to evaluate computing mechanism for energy management, and cost-

**Acknowledgments**

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CCCDI – UEFISCDI, project title "Multi-layer aggregator solutions to facilitate optimum demand response and grid flexibility", contract number 71/2018, code: COFUND-ERANET-SMARTGRIDPLUS-SMART-MLA-1, within PNCDI III.

**References**

[1] H. Hu, Y. Wen, T.-S. Chua, X. Li, Toward scalable systems for big data analytics:
a technology tutorial, IEEE Access 2 (May) (2014) 652–687

[2] P. Mirowski, S. Chen, T. K. Ho, C.-N. Yu, Demand forecasting in smart grids, Bell Labs Tech.J. 18 (4) (2014) 135–158.

[3] Xinghuo Yu, C. Cecati, T. Dillon, Simões, The New Frontier of Smart Grids, , M.G., *IEEE Industrial Electronics Magazine* (2011)

[4] S. Callahan. Big data: The future of energy and utilities (https://www.rdmag.com/article/2015/10/big-data-future-energy-and-utilities)

[5] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, A break in the clouds: Towards a cloud definition, J. ACM SIG-COMM Comput. Commun. Rev., vol. 39, no. 1, pp. 50–55, (2014).

[6] P. V. Krishna, S. Misra, D. Joshi, and M. S. Obaidat, Learning automata based sentiment analysis for recommender system on cloud, J.IEEE Int. Conf. Comput, Inform Telecommun Syst., (2013), pp. 1–5.

[7] Yang Zhang, Tao Huang & Ettore Francesco Bompard, Big data analytics in smart grids: a review (https://energyinformatics.springeropen.com/articles/10.1186/s42162-018-0007-5)

[8] J.N. Bharothu, M. Sridhar, and R.S. Rao, "A literature survey report on Smart Grid technologies", Proc. 2014 International Conference on Smart Electric Grid (ISEG), pp. 1-8.

**Ionuț ȚARANU** (b. April 28, 1975) has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies in 1999. He followed a master's degree in Databases for Business Application, within the same faculty. As founder of STIMA SOFT, Ionut is an experienced professional in custom software development, with focus on implementation and management of complex software solutions.

**Filip FEDELEȘ** (b. November 29, 1982) has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies in 2006. He followed a master's degree in Public Management, within the same faculty. As a senior developer at STIMA SOFT, Filip is interested in data structures, new technologies and developing complex software solutions.

# The Digital Transformation and Disruption in Business Models of the Banks under the Impact of FinTech and BigTech

Oona VOICAN
The Bucharest University of Economic Studies, Romania
oona.voican@yahoo.com

*The explosive development of artificial intelligence, machine learning and big data methods in the last 10 years has been felt in the financial-banking field which has subjected to profound changes aimed at determining an unprecedented increase in the efficiency and profitability of the businesses they carry out. The tendencies of applying the concepts coming from AI, together with the continuous increase of the volume, complexity and variety of the data that the banks collect, store and process have acquired the generic names of FinTech, respectively BigTech. Five main areas exist where FinTech and BigTech can provide improvements in business models for the banks: introducing specialized platforms, covering neglected customer segments, improving customer selection, reduction of the operating costs of the banks, and optimization of the business processes of the banks. We will present some of these improvements, and then we will show how the business models of the banks dramatically transform under the influence of these changes.*
***Keywords:*** *Artificial intelligence, BigTech, Business models, Digital transformation, FinTech.*

## 1 Introduction

The development of the digital economy in the last period affects all sectors of activity, including the banks and financial sector. Within it, banks have a special role, as engines of economic growth in all sectors of the economy. A new way of thinking is imposed on the banks through the transition to the digital economy. This process, however is not devoid of complexity.

On the one hand, the financial sector and banks are causing huge changes in the way companies and other economic organizations operate and, on the other hand, this sector is undergoing dramatic changes that make the traditional concepts of banks themselves need redefining. This interdependence between the changes to which the two basic components are subject, the companies, on the one hand, and the financial-banking sector, on the other, is best illustrated by the evolution of the business models of the banks. Consequently, the financial system includes a new component - **FinTech.**

This fundamentally changes the financial market, and technological innovation blurs the boundaries between financial products and services and subjects either authorized to offer its or actually involved in providing them [1]. As a result, new competitive forces and dynamics emerge and develop, expanding their influence on the market shares held by different financial intermediaries.

The main purpose of this paper is to analyze the strategies adopted by the different types of new entrants to the financial market (FinTech, BigTech and the usual financial firms) in the provision of banking services and to examine how digitization affects the business models of banks and other providers of banking services.

Most commonly used today in the literature is the working definition of the Financial Stability Board (FSB) for FinTech as a "technologically activated financial innovation, which could result in new business models, applications, processes or products with an associated material effect on financial markets and institutions and in the provision of financial services." [2].

The focus of this definition is on analyzing the implications on the effects of FinTech in banks and other financial-banking

institutions. It should be noted that the term FinTech refers to a wide range of software innovations applied in both existing banks and new entrants or in the process of being modernized.

A first observation made in the literature regarding the meaning of the various definitions of FinTech, as they are now understood, shows that it is perceived in relation to an innovative service, a new business model (which can be used by a regular bank or by a certain company) or a new software program used within a start-up that determines a strong imitative current in the financial industry.

A second observation is that some definitions distinguish between innovation and disruption, the first leading to the improvement of pre-existing rules in an existing regulatory framework, while the second leading to the emergence and rapid development of new rules that subsequently determine dramatically changing the framework of existing regulations.

**BigTech** is a generic name that is used for large technology companies active worldwide, having a relative advantage in using digital technology. BigTech companies are usually providers of web services (search engines (SEO), social networks, e-commerce etc.) for end-users on the internet and/or IT platforms or develop and maintain an infrastructure (storage and processing capabilities data) for which other regular companies provide products or services.

Like FinTech companies, BigTech companies are usually automated and use agile processes for software developments, giving them multiple possibilities to quickly adapt their systems and services to the needs and wants of users. BigTech companies have global business operations and have an extremely large customer base. They own and use a vast amount of information about their clients to provide them with the most appropriate financial services. In this way, the BigTech companies gain a decisive competitive advantage over the competition, for example compared to the traditional banks, in the provision of financial services. Many banks, financial institutions and Fintech companies develop partnerships with BigTech companies, eventually becoming relevant financial service providers.

Relevant examples of BigTech companies are Google, Amazon, Facebook and Apple, known as the GAFA group. Similarly, the BAT group refers to three of the largest technology companies in China, namely Baidu, Alibaba and Tencent. In addition, some well-known traditional companies, such as Microsoft and IBM, are gradually becoming more relevant to the international financial system and may be included among BigTech companies.

## 2 Literature review

Business models are, in the last two decades, a highly popular subject of analyzis, especially because they can best explain the differences that are observed between the results of different firms [3]. Theoretical and empirical studies suggest that the choice of suitable business models can lead to sustainable competitive advantages and higher financial performance for the companies that use them [4].

Initially, the notion of business model was mainly used in the field of management studies [5]. During that period, a business model was understood as a business strategy, which was based on the balance sheet ratios and income accounts associated with them.

The own studies on banking business models appear for the first time in the early works of [6] and [7].

The structural business model is placed between the strategic and operational level in the architecture of the companies [8]. While the strategic level determines the milestones and the long-term source of a competitive advantage, the operational level shows what needs to be done to achieve the long-term objectives [9].

Unlike the operational level, the business model offers a less detailed and holistic perspective on the company. It focuses mostly on basic logic steps to be followed for creating and increasing the value of the firm [4]. During the time, the business model has become a tool for analyzing companies, but also a subject that can be approached separately to be designed and innovated.

Studies on business models frequently use numerous well-established research branches, such as strategic management, information systems and management of innovation and change at firm level. Each of these areas of approach provides a different perspective on business models. Thus, in literature we retrieve different formulations, reflecting a kaleidoscope of views on what should actually represent a business model. We find such business models where emphasis is placed on the annual activities and transactions of the enterprise with stakeholders [5], on its resources [10] or its choices and results [9].

However, many of the definitions address business models not only as a lot of specific elements, but also from the perspective of how they are interconnected and mutually influenced [11]. This comes mainly from the fact that the multitude of elements included in the business model is a particular combination of elements that allow a firm to create and increase the value obtained in a unique way.

The objectives pursued by studies on business models can be diverse.

According to [12] three major reasons for building a business model are identified: a method of classifying existing business types, a way of analyzis for academic studies and a recipe for practitioners who want to reproduce and innovate successful models. However, as shown in recent literature, the approaches so far do not converge towards a common theoretical framework [5], nor can we find a dominant approach [11].

Regarding the business models of banks, many papers in recent years have proposed a classification of them into three broad categories, namely: significant, intuitive and based on quantitative characteristics of classified banks. These approaches can be classified based on the methodologies used as follows:

***1) Studies that use clustering to classify banks based on a set of characteristics***.

The clustering methodology used uses an algorithm that assigns elements (in this case banks) to the clusters so as to minimize the distance between the elements of a single cluster and to maximize the distance between the average/median/centroid of the formed clusters. Distance is a predefined metric that shows how similar or different the elements are, based on certain variables. These works are mostly based on the agglomerative clustering hierarchy method described by [13] or the partial clustering method based on the algorithm of Viki and Kiers [14].

***2) Studies that use qualitative approaches to classify banks according to the business model.***

These studies introduce a predefined classification of the business model, based on the activities, financing and legal structure of the banks. Banks are then assigned to each of these categories based on expert judgment.

Starting with the first group of papers using the clustering methodology [14], explore which variables are relevant for defining a banking business model and provide preliminary evidence on the importance of business model analyzis in banking regulation and supervision.

As described in [15] and [16] it was introduced the active-passive approach and used a hierarchical clustering method based on Ward's theory [13] to identify groups of banks with similar balance sheet characteristics. In order to take these factors into account collectively, without

over-representing any particular factor, indicators were used that constitute the defining characteristics of the activity/financing of a business model in banks in terms of assets and liabilities.

As indicated in [17], the authors obtained five types of banks based on a cluster hierarchy algorithm: concentrated retail, type I diversified retail (more trading assets and bank loans), type II diversified retail (based on mainly on debt), wholesale banks and investment-oriented banks. Business practitioners have designed and used countless business models capable of describing and developing a business. Examples of such business models are Canvas [18], Wirtz's business model, Siemens Business Model Framework BizMo [19] or the St Galen Management Model (SGMM).

Doleski study [20] highlights a few common elements of these business models, elements that are most influenced when disruptive factors appear, such as business digitization.

The first element, the normative framework, represents the normative dimension of a business model, the elements of value and strategy, the strategic dimension. The client, the market, the revenues, activation elements, processes, partners and finances are allocated to the operational point of view.

The aforementioned elements represent the generic basic modules of a business model and appear in principle in all models, regardless of sector and activity. Taken together, these elements are the core of the integrated iOcTen business model [20].

One of the best-known business models is Canvas, developed by [8], which offers a visual representation of the business models used by strategic management. The Canvas model offers a common language, so that those who evaluate and develop different business investment alternatives can easily communicate between them and with other specialists. This model became famous especially after Apple used it to invest in innovative products that were able to become dominant in the market.

The Canvas business model structures internal processes and activities into nine categories, each of which represents a building block for the realization of goods and services. The nine categories refer to four major aspects of a business, namely: customers, supply, infrastructure and financial viability. The nine categories considered are the following:

a. Customer Segments;
b. Value Proposals;
c. Channels;
d. Customer Relationship;
e. Revenue Streams;
f. Key Resources;
g. Key Activities;
h. Key Partnership;
i. Cost Structure.

Each category, in turn, is divided into components which are then completed in a grid containing the above categories. The name of the model comes precisely from the similarity between the final frame obtained and a canvas having different segments of different colours.

In recent years, however, the Canvas model has been replaced by other more evolved models, for example with the iOcTen business model, which, unlike the previous model, has the advantage that it allows the integration of the ten defined characteristics, taking into account permanently the changes and transformations that take place both inside and in the environment of the company [20].

Figure 1 shows the basic components of this business model:

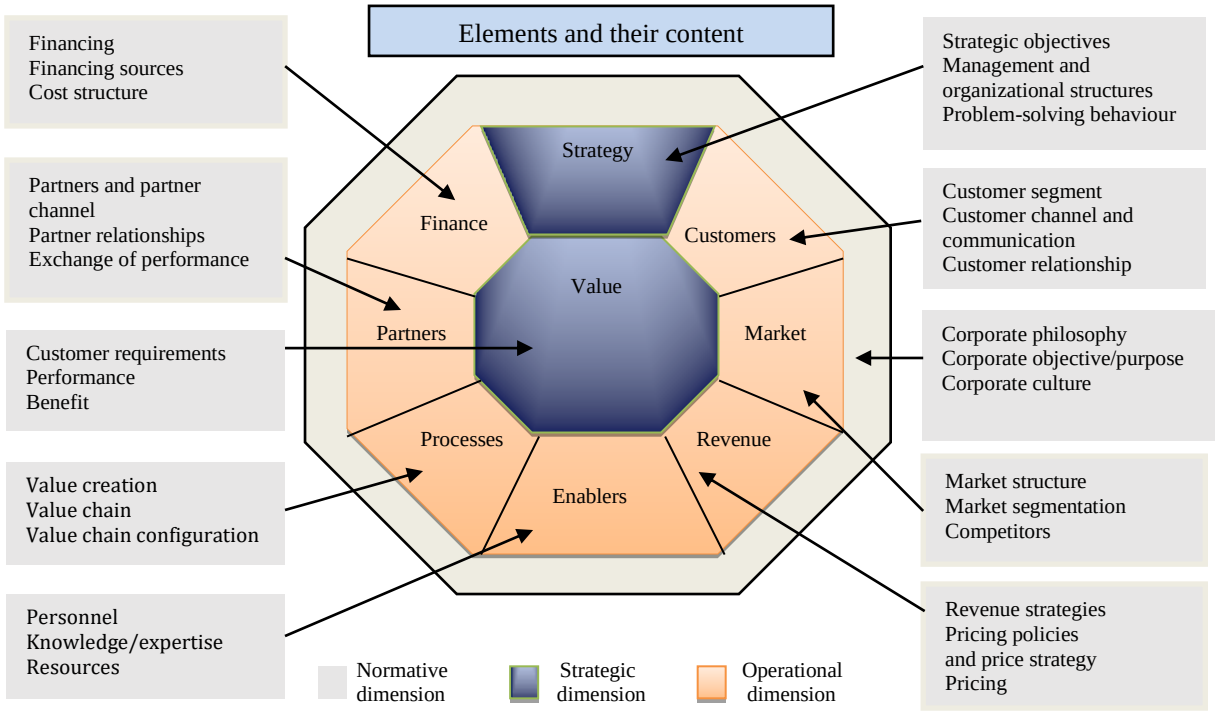



**Fig. 1.** Components of the iOcTen model [20]

The existence of three elements or concepts is observed: the normative dimension, the strategic dimension and the operational dimension.

The normative dimension refers to the main concepts used by the model: customer, market, revenue, enablers (processes), processes, partners and finances.

The strategic dimension refers to the strategic objectives, the organizational and management structures and the problem-solving behavior. At the same time, at this level, the Philosophy of the corporation, the Purpose/objectives of the corporation and the Culture of the corporation are defined.

Operational concepts are associated with each normative concept. For example, the concept of Value includes Customer requirements, Performance and Benefit. The concept of Customer is operationalized, at the operational level, by the Customer Segment, the Channel towards the customers and the communication with them, respectively the Customer Relationship. Market concept is operationally associated with Market Structure, Market Segmentation and Competitors. The concept of Revenue is described by the following components: Revenue strategies, Pricing policy and price strategy and Pricing.

The concept of Enablers includes: Personnel Knowledge/expertise and Resources. The concept of Processes is described by: Value creation, Value chain and Value chain configuration. Regarding the concept of Partners, it is described with the help of the operational components Partners and partner channel, Partner relationships and Exchange of performance. Finally, the concept of Finance consists of the operational concepts of Financing, Financing sources and Cost structure.

At the level of a company, each operational concept is materialized and described with the help of advanced operations, which represent the concrete modalities that are

used in each case separately to meet the requirements of the concepts used [21].

## 3 Methodology

The rapid and destabilizing impact of digital technologies requires different methods that take into account the dynamic nature of the forces of change in the present era. Business models that focus on permanently improving the efficiency of the value chain are no longer sufficient. Under new conditions, customers may not simply respond to bank signals and change their decision-making and business behavior. In this context, the methods of identifying digital disruptions and their causes become very important to ensure that decisions are made to counteract them.

In order to identify the real and potential disturbances, a method is used that offers the possibility to identify and analyze in depth the disturbing factors. This method uses three major components. The first component starts from the concepts of first order and second order impact, which are related to the disruption interruption already started, which initially helps the existing business structure, but then continues to undermine it. The second important component is the disturbance assessment matrix (DAM), designed to combine several disturbing factors to identify the disturbed area. The third component is the disturbance impact assessment (DIE), a detailed analyzis tool which determines the intensity of the impact of the disturbance on the main stakeholders in the respective business [22].

Digital disruption takes place in successive stages whereby the initial benefits obtained by digitization are subsequently lost through continuous digitization, which changes the nature of the pre-existing service or product to establish a new state of normality. These stages can be regarded as first and second order disturbances. The concept of the impact order is well known from the business model developed by [23].

They said: „reinventing industries, replacing products or services, creating new digital businesses, reconfiguring the provision of value and rethinking value propositions" [23].

This classification of the impact of digitization of the first and second order is done taking into account the following aspects:

Figure 2 shows the disturbance evaluation matrix (DAM):

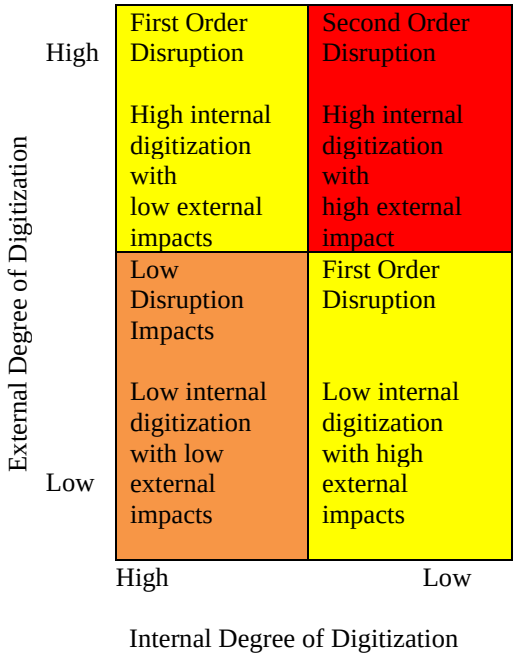



**Fig. 2.** The disturbance evaluation matrix (DAM) [20]

Figure 2 gives a graphical description of the method and evaluation of the characteristics of a potential change in the position of a firm or banks due to the digitization. The impact grid is a 2 x 2 matrix that provides a framework for an initial classification of digital disturbances. The X axis measures the internal degree of digitization, while the Y axis measures the external degree of digitization. The axes also refer to the four groups (presented in Table 1) with internal factors that are investors and producers and

external factors that are the consumers and the public.

**Table 1.** Disruption assessment matrix (DAS) [22]

| Domains/ Stakeholders | Legal, political | Economic and business | Health and Wellness | Cultural, ethical and moral | Techno-logical | Environmental |
|---|---|---|---|---|---|---|
| **Owners** | | | | | | |
| **Investors intern/extern** | | | | | | |
| **Stock markets** | | | | | | |
| **Producers** | | | | | | |
| **Employees** | | | | | | |
| **Managers** | | | | | | |
| **Consumers** | | | | | | |
| **…** | | | | | | |

The Disturbance Assessment Matrix (DAM) provides a means of visualizing assessment, scenario planning and a set of competitive Porter's analyzes [24].

In the disturbance assessment matrix (DAM), are represented six domains that determine people's actions and behavior:

- Legal and regulatory,
- political;
- Economic and business;
- Health and wellness;
- Cultural/social/ethical and moral;
- Technological;
- Environmental.

The stakeholder classification is made according to their relationship with the production (supply) or consumption of goods and services. They are viewed according to two categories: internal and external.

The Digitalization Impact Evaluation (DIE) framework (Table 2) provides a segmentation of the population to assess the impact of each domain on each stakeholder group.

Stakeholders are regarded as persons belonging to the categories considered. The purpose is to allow a clearer and more open framework of analyzis that takes into account any given digital development.

the different facets of organizational disruption. The grid can be viewed as a The impact assessment of digitization (DIE) is used together with the DAM as an initial step to determine the degree of impact that a digital innovation will have on the various stakeholder groups.

**Table 2.** Digitalization impact evaluation (DIE) [22]

| Impact | No impact | Impacts profitability | Impacts growth | Impacts existing scale | Severe disruption Impacts survival |
|---|---|---|---|---|---|
| **Owners and Investors** | | | | | |
| **Reduce the risk of capital loss** | | | | | |
| **Improve portfolio performance** | | | | | |
| **Producers** | | | | | |
| **Increase profitability** | | | | | |
| **Enhance customer relationships** | | | | | |
| **Customers** | | | | | |
| **Speed of delivery** | | | | | |
| **Lower cost** | | | | | |
| **Public** | | | | | |
| **Improve Environmental Performance** | | | | | |
| **Reduce carbon output** | | | | | |

## 4 Results and discussions

The disruption caused by the FinTech and BigTech in almost all areas of finance and banking creates a pressing need for new and radical solutions. This is also true for business models.

Let us present below some of the most important disruptions that have been manifested so far in the banking field and how they are reflected in changing the structure of the business models:

### a) Introduction of specialized platforms

To the extent that FinTechs focus on offering specific services and increasing the diversity of specialized lending opportunities, they can offer better portals that are successful among their customers. Such a portal can offer, in addition to regular payment transactions, a variety of other services at the clients' request, such as currency exchange or transfer of certain amounts of money to third parties. As a rule, FinTech focus on the most profitable banking services, such as managing personal accounts or SMEs, which represent about 50% of the profits made by banks [25].

FinTech companies offer a variety of electronic payment solutions. FinTech's development has facilitated the circulation of virtual and crypto money, which was impossible using only traditional banking systems. All of these solutions are integrated into decentralized blockchain architectural structures, which have a large extension among the different categories of clients, ensuring a particularly high level of communications security. Blockchain, a decentralized and distributed file network structure, will increasingly affect financial institutions in the future. Blockchain can make things more efficient in the financial services industry. Financial institutions will use blockchain for smart contracts, digital payments, identity management, and trading shares. Since fraud and identity theft cost financial institutions billions of dollars annually, blockchain has the potential to save the industry from experiencing these significant losses.

### b) Covering neglected customer segments

Traditional banks usually exclude a significant part of the population. Based on 2019 International Finance Corporation data, 45-55% of SMEs worldwide do not have an overdraft allowance, but would benefit from one, while 21-24% although they have accessed loans, they are in very limited of different regulations. Worldwide, there are another $ 2.4 trillion in credit applications. In the US, 44% of SME lending applications were rejected. In the UK there is a funding gap of up to £ 59 billion. In addition, about 30% of British SMEs have failed to obtain the funding they have tried to obtain.

By offering alternative lending services, Fintech companies offer innovative approaches to segments such as subprime lending, which are not served by traditional banking services. Financing syndicate Kabbage, Lendio, OnDeck and Swift Capital are some of the players who have made use of the 20% decrease in bank loans to SMEs after the 2007 and 2008 crisis. Loans for larger companies increased by about 4% in the same period. FinTech offers have grown rapidly by serving segments that are left out of traditional corporate banking channels - such as the subprime category for small businesses. By adopting FinTech and BigTech technologies, traditional banks can apply digital lending models to assess the creditworthiness of customers whose financial profiles do not comply with traditional lending rules.

### c) Improve customer selection

The data held by banks are as valuable as gold, especially all data relating to transactions. The concept of Big Data has shown its full value when it began to be

used in business models of banks. By combining internal data sources with other external sources, commercial banks are improving their pricing strategies, which can sometimes have very complex structures. This richness and variety of information, along with their complexity can also be used to decide how to classify customers and which products are most recommended for each type of activity they carry out. Data processing can also help in making credit decisions. Some of the new FinTech technologies in the US use over 2000 different data points from over 100 sources to make a credit decision.

**d) Reduction of the operating costs of commercial banks**

Fintech can create services, platforms and products with a small budget. Since these products are mostly made on virgin land, there are no inheritance issues, so pre-existing IT systems would have legal licensing issues. The various innovations suggested by FinTech can be used to reduce the operating costs of the respective banking system. For example, in the front-line space, virtual assistants can improve customer services, thus reducing labor costs, smart, electronic identification and optical character recognition solutions.

Standardizing and digitizing banking processes allows FinTech to sell products and services at substantially lower prices. In some cases, solutions have been used for international transactions that cost about 75% less than on traditional channels. In addition, companies such as Ripple, a large dairy manufacturer and distributor, use block chain technologies to speed up processing times for payment settlement and reduce transaction costs. Using digitized payment and block chain options, commercial banks not only increase the efficiency of their operations, but also save billions of dollars per year.

**e) Optimization of the business processes of the banks**

The increasingly widespread use of FinTech technology enables the automation of banking transactions, and customers can make payments using mobile phones and tablets (mobile banking). Therefore, bank employees need to play more a consultant role rather than conducting transactions directly.

According to Citigroup, the number of US bank personnel is expected to decrease by 30% by 2025, from 2.6 million to 1.8 million people today. In Europe, personnel employed in the banking industry are expected to decrease by 38%, from 2.9 million to 1.8 million. FinTech offers digital procedures that can improve the business processes of banks. For example, advanced search engines combined with automatic decision-making procedures lead to rapid credit approval.

Platforms such as the one developed by Earthport allow the settlement of international payments in real time. Finally, a platform like the one used by Ripple can encourage direct international payments, significantly reducing costs and downtime. The optimization of banking processes is one of the most disruptive changes that will affect the entire banking industry, but also its interdependence with the rest of the industries.

## 5 Conclusions

Developing suitable business models for the banks is today a top priority. Business models create the conceptual framework for the systematic realization of digital business ideas and open up new business areas, therefore, they represent the practically available toolkit for the implementation of advanced initiatives.

The FinTech makes it impossible to run a business based on conventional operations, the strictly traditional approaches used so far prove to be outdated. The advent of BigTech technologies would be both a blessing and a curse for banks and financial institutions, as

it opens up new opportunities with one hand and threatens to kill existing business models with the other [5].

Greater focus on well-individualized services enables FinTech and BigTech companies to provide efficient services with new technologies, including those based on artificial intelligence. Artificial intelligence algorithms play an increasing role in determining customer scoring, identifying bank frauds, or segmenting customers.

**References**

[1] EBA (2017). EBA Report on the Leverage Ratio Requirements under Article 511 of the CRR, EBA/Op/2016/13.

[2] FSB (2019). FinTech and market structure in financial services: Market developments and potential financial stability implications.

[3] A. Afuan and C. Tucci, "Internet Business Models and Strategies – Text and Cases," 2nd ed., Boston, McGrow Hill, 2003.

[4] R. Amit & C. Zott, "Value Creation in e-business", Strategic Management Journal, Vol. 22, No. 6/7, pp. 493 – 520.

[5] C. Zott, R. Amit and L. Massa, "The Business Model: Recent Developments and Future Research," Journal of Management, Vol. 37, 2011.

[6] A. Mehra, Resource and market-based determinants of performance in the US Banking industry. Strategic Management Journal, v.17, n.4, apr.1996.

[7] D. Amel and S.A. Rhodes, "Strategic Groups in Banking," The Review of Economics and Statistics, Vol. 70, pp. 685 – 689, 1999.

[8] A. Osterwalder, "The Business Model Technology a Proposition in a Design Science Approach," PhD Thesis, University of Lausanne, Swizerland, 2004.

[9] R. Casadesus-Masanell and J.E. Ricart, "From Strategy to Business Models and onto Tactics," Long Range Planning, vol. 43, pp. 195- 215, 2010.

[10] B. Demil and X. Lecocq, "Business model evolution: In search of dynamic consistency", Long range planning, 43 (2/3), 227-246, 2010.

[11] R. Casadesus-Masanell and F. Zhu, "Business Model Innovation and Competitive Imitation: The Case of Sponsor-Based Business Models," Strategic Management Journal 34, no. 4, pp. 464–482, 2013.

[12] C. Baden-Fuller and M.S. Morgan, Business Models as Models. Long Range Planning, 43, 156-171, 2010.

[13] J. H. Ward, Jr, "Hierarchical grouping to optimize and objective function," Journal of the American Statistical Association, 58(301), pp. 236-244, 1963.

[14] M. Viki and H. L. A. Kiers, "Factorial K-means analyzis for two-way data," Computational Statistics & Data Analyzis, 37(1):49-64, 2001.

[15] R. Ayadi, Arbak and W. De Groen, "Business Models in European Banking: A Pre-and Post-Crisis Screening," SSRN Electronic Journal, 2011.

[16] R. Ayadi and W. De Groen, "Banking Business Models Monitor 2014: Europe" (October 14, 2014). CEPS Paperbacks, 2014. Available at SSRN: http://ssrn.com/abstract=2510323.

[17] R. Ayadi, "Bank Business Models in Europe: Why Does it Matter for the Future of Regulation and Resolution?" (July 15, 2016). International Research Centre on Cooperative Finance Policy Paper Jul 2016. Available at SSRN: https://ssrn.com/abstract=2829168 or http://dx.doi.org/10.2139/ssrn.2829168.

[18] A. Osterwalder and Y. Pigneur, "Business Model Generation," London School of Economics, April, 2010.

[19] J. Mütze and A. Gerloff, Customer Value Co-Creation: Gemeinsam die Chancen der Digitalisierung, Realisierung Utility 4.0 Band 1. Springer Vieweg, 2019.

[20] O. D. Doleski, Integrated Business Model. Applying the St. Gallen Management Concept to Business Models, Springer Essentials. Wiesbaden: Springer Gabler, 2015.

[21] T. Kaiser, O.D. Doleski, "Advanced

Operations. Best Practices for the Focused Establishment of Transformational Business Models," Springer Vieweg, 2020, ISBN 978-3-658-27585-3.
[22] B. Stewart, R. Schatz and A. Khare, "Making Sense of Digital Disruption Using a Conceptual Two-Order Model," In: Khare A., Stewart B., Schatz R. (eds) Phantom Ex Machina. Springer, Cham, 2017.
[23] G. Westerman, D. Bonnet & A. McAfee, "Leading Digital: Turning Technology into Business Transformation", Harvard Business Press, 2014.
[24] M.E. Porter, "The Five Competitive Forces that Shape Strategy," Harvard Business Review, Jan., 86 (1), pp. 79 -83, 2008.
[25] L. Wewege, and M.C. Thomsett, "The Digital Banking Revolution. How FinTech Companies are Transforming the Retail Banking Industry through Disruptive Financial Innovation", De Gruyter, 3rd Ed., 2020.

**Oona VOICAN** has graduated from the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies. She followed a master's degree in Business Analysis and Enterprise Performance Control with the thesis *Data mining methods used in the banking system*. She worked 15 years in the banking system and now she is a senior adviser at the Ministry of Transportation, Infrastructure and Communications. She is currently a PhD Student in the field of Economic Informatics, mostly interested in Cybernetics, Data Mining, Big Data and Business Intelligence