Solutions for Big Data Processing and Analytics in Context of Smart Homes

Adela BÂRA[1], Bogdan TUDORICĂ[2], Razvan Cristian MARALES[3]
[1]The Bucharest University of Economic Studies
[2]Petroleum - Gas University of Ploieşti
[3] METRO SYSTEMS Romania
bara.adela@ie.ase.ro, tudorica_bogdan@yahoo.com, razvan.marales@csie.ase.ro

Abstract. *The paper analysis Big Data storage and processing solutions applicable in the smart grid context in case of large volume of data coming from sensors (smart meters, IoT appliances). A flexible architecture for data management is proposed that consists in three layers: relational database (RB) tier, big data (DG) tier and data warehouse (DW) tier. A proof of concept implementation is also provided, using Elasticsearch and Kibana.*
Keywords: Big Data, Smart Home, IoT, Smart Metering, Data Management

1 Introduction

In the recent years, due to the development needs of certain big companies activating in the digital industry, the big data processing techniques were developed very fast. As presented in [1], the main types of big data processing techniques are:

- batch processing, which was initially accomplished by using Apache Hadoop as a big data analysis platform for big data. Apache Hadoop is based on the MapReduce concept, which was initiated by Google's batch processing programming model. The main functionality of Apache Hadoop is successfully done by dividing a very large data set into multiple smaller data sets, which are then processed in parallel; then, the reducer job get the results from all these smaller data sets. As the main disadvantages of using Apache Hadoop and MapReduce are represented by the fact that MapReduce cannot be used for real time sensor data or streaming data processing, and that the set of machine learning algorithms provided by Apache Hadoop is not enough to meet requirements for smart grid big data analysis. Considering this facts, the authors do not recommend Apache Hadoop as a good choice for big data analysis on smart grid systems. Another big data processing solution, as presented in [2], is Apache Spark. Apache Spark is defined as a general purpose cluster computing platform, which delivers flexibility, scalability and speed to meet the challenges of big data, even for the case of using big data in smart grid analysis. Apache Spark has many advantages: not only has the ability for batch processing, but it also is capable of doing iterative and streaming processing, it has more efficient machine learning algorithms and enhanced linear algebra libraries.
- stream processing, which involves processing each new instance of data, rather than waiting for the next batch, and then re-process everything – this way the data processing technique avoids unnecessary repetition of re-processing the data. This model of data processing enables Apache Spark to perform analytics on data with dynamic behaviour – this feature is very important in the situation when data comes continuously, in real time, and from different data sources.

- iterative processing, represents another category of big data processing, and is mainly used to solve problems that cannot be addressed with batch or stream processing. One of the main characteristics is that it can process all variety of data types frequently – in general, the model is time consuming due to their continuous operations of writing and reading which represents each iteration of the system. Once again, Apache Spark is the current leader for iterative processing because it has the power to process and hold data in memory across the cluster, and the data is written back only after the completion of the iterative process.

The concept of Smart Home involves an efficient management of demand response regarding consumers' behaviour, integrating different residential generation sources and electric appliances through smart meters, providing user friendly applications and real time billing systems. In this context, ICT solutions should be developed to support a real time integration of heterogeneous sources gathered from smart meters, intelligent appliances and micro-generation sources. Papers [3-9] provide an overview of smart home and demand response concept and functionalities: responsible for an efficient connection and exploitation of generation and consumption sources, providing automatic and real-time management of the grid, optimizing the level of reliability and improving the electricity providers' services which lead to energy savings and lower costs. An important aspect in developing ICT solutions for demand response is the design of a scalable architecture for data management. At the current moment, a wealth of data storage technologies is available, based on different ideas and with appropriate advantages and disadvantages:

- Direct storage in a data file. A multitude of competing formats are currently available for storing data in text or similar files, such as Comma Separated Values (CSV), eXtensible Markup Language (XML), JavaScript Object Notation (JSON) or Yet Another Markup Language (YAML);
- Storage in a classic relational database accompanied by a complete and user-controlled database management system (DBMS). Examples of such databases are: Oracle Database, MySQL, MS SQL Server, PostgreSQL, IBM DB2;
- Storage in a document store type NoSQL database characterized by schema-free organization of data. Records should not have a uniform structure and may have different columns. Data types of individual column values may be different for each record. Columns can have multiple values (arrays). Records can have a nested structure [10]. Examples of such databases are: MongoDB, Elasticsearch, CouchBase, CouchDB;
- Storage in a key-value store NoSQL database that can store pairs of keys and values, and serves to recover a value when the appropriate key is known. These databases are not normally suitable for complex applications but, in certain circumstances, provide specific benefits [10]. Examples of such databases are: Redis, Aerospike or Oracle NoSQL;
- Storage in a wide-column store NoSQL database that involves the storage of data in records with the ability to contain a very large number of columns

that dynamically vary. The column names are also variable, the record keys are not fixed, and, since a record can have billions of columns, databases of this type can be seen as two-dimensional key-value matrices [10-11]. Examples of such databases are: Cassandra, HBase, Microsoft Azure Table Storage and Google Cloud Bigtable;

- Storage in a time series database optimized to manage time series data, each record being associated with a time stamp. Time series data can be produced by sensors, smart meters or RFIDs - IoT equipment, or can represent the values recorded in a high-speed stock exchange system. These databases are designed to efficiently collect, store and query different time series data. While these data can also be stored in other types of databases - from key-value stores to relational systems, specific challenges sometimes require specialized systems [11]. Examples of such databases are: InfluxDB, Kdb+, RRDtool, Graphite și OpenTSDB.

2 Big data solutions for Smart Home

2.1. Big Data processing solution

During the development of the SMARTRADE project, the success of an efficient data integration is related to the correct choice of the technology used to store and manage data. The nature of the project involves the use of a number of different data storage technologies, due to the stratified nature of the target application. It is assumed that the final prototype will contain at least the following levels of operation, virtually forming a stack of applications:

- Level 1 - Storage of data at the level of the IoT (Internet of Things) type systems that substantiate the acquisition and / or local centralization of data;
- Level 2 - The primary storage of data accumulated from the IoT systems and other sources, prior to any cleaning, aggregation, and analysis of data;
- Level 3 - Secondary data storage, used after data cleaning has been applied. Aggregation and analysis operations may be applied on the secondary storage data, whether we are talking about supervised analysis (common queries, but also classical statistical analysis), or unsupervised analysis (data mining operations);
- Level 4 - Tertiary storage of processed data / processing results, which is to be the integrated storage in a final consumption estimation application, resulting in competitive offers for the wholesale electricity market.

Each of the above four levels has distinct features that require the use of a particular storage technology among those listed in the previous section.

Storage at the level 1 of the application stack

The storage of data in IoT devices implies significant restrictions, both in terms of available storage space and usable processing capacity (doubled narrowed in turn by the limited performance of the processors used and the need for low energy consumption, which is encountered in many IoT devices). In [12] IoT devices are classified into four categories according to their complexity level:

- Extremely simple devices whose communications are limited to signalling their state in case of an event. Both the memory and the processing capacity of these devices are reduced to the minimum required for operation, and most of them do not

have proper storage memory. Some smart-metering devices from older generations could fall into this category;

- Slightly more complex IoT devices. The memory and processing capability of these devices is reduced, but it allows processing and storage not directly linked to the main function of the device. They also involve the transmission of the measured data in a single sense, but various forms of aggregation and representation of data are also possible. Many of the new generation smart metering devices belong to this category. For this IoT device type, some data storage is available at least for logging. It is also important to note that many such devices are proprietary devices and one can't interact directly with the software installed on them or with storage media and can't choose the data storage technology used. As a solution to this situation, the use of modular equipment and / or equipment fully or partially based on open technologies is recommended. In this category of devices, if it is possible to control the storage, it is most likely that only data files can be used [13] because either the address space is too small or the memory is not large enough to implement a query mechanism;
- Interactive devices capable of bi-directional communication. Additionally, such devices offer programmability and many of the capabilities of personal computing systems (multiple storage systems, multiple inputs and outputs, both analogue and digital). Use of data files and embedded relational or NoSQL databases is recommended for this category of devices [13];
- Intelligent devices with high processing and storage capabilities raised to the level they even allow running artificial intelligence applications. On this type of devices, it is possible to run non-embedded, relational or NoSQL databases [14].

Storage at the level 2 of the application stack

The method chosen for the storage of data accumulated from IoT devices depends specifically on the number of IoT devices from which data is collected and the data traffic density for each device (the frequency with which data is transmitted multiplied by the size of each transmission). Under a certain number of IoT devices from which data is collected and under a certain density of data traffic per device, it is perfectly possible to use any type of commonly used database, being preferable, in this context, to use relational databases. The motivation for this choice is given by the superior data query facilities offered by these products.

When a critical level is reached, which is mainly dependent, as highlighted above, on the number of devices from which data is collected and the density of data traffic per device (a number of other factors need to be taken into account, such as data communication system capabilities and temporal overlay of data transmissions from multiple IoT devices), the use of relational databases becomes undesirable, for reasons of limited performance.

The limitations given by the use of relational databases are quite well known. For example, MS SQL Server supports theoretically 32767 competing connections over a cluster of servers, but in real-world applications, their number is severely limited by the computing system used for the database server and the data communications system, reaching, for example, 50-200 competing connections per server. A similar situation is encountered in MySQL where the maximum recommended number of concurrent connections per server instance in version 5.5 is 500 (with an average of 100 competing connections) for a transactional usage scenario. The same situation can be seen in Oracle Database 12c - a maximum of 2000 concurrent

connections to a database instance. On the other hand, the initial arguments that led to the success of NoSQL solutions on the market, the volume of data circulated and its velocity (to which, subsequently, variety and veracity were added to form the 4 V's of NoSQL databases), are indeed visible even at a superficial glance. For this reason, for any data collection system that involves large amounts of data, a NoSQL database is the recommended solution. One has to determine only the NoSQL database subtype to be used, depending on the proposed application architecture - the three main types of NoSQL databases are available (document stores, key-value stores, and wide columns stores), but also some less used types can be taken into account. Among the less commonly used database types, for smart-metering applications, the databases specialized in the storage of time series are very important.

Storage at the level 3 of the application stack

For the secondary data storage, the same considerations, from the point of view of volume and velocity of data, that imposed the solution at the second level of the application stack, should be considered. Even though the number of data sources is significantly reduced, the data volume is probably the same order of magnitude as the one at the level 2 of the application stack.

Storage at the level 4 of the application stack

If a separate application is preferred, as a consumer of the results of the analyses performed on the data generated by the IoT devices, such an application is likely to be only a small application, limited both as complexity and as the volume of used data - it will not have as inputs the actual data which is generated, collected and analysed at the other levels, but the power generation and consumption profiles resulting from these data, in conjunction with other types of information (e.g. prices from electricity producers and transporters, etc.). For such an application, the need to use a NoSQL storage solution is no longer stringent, although it can be considered for various practical reasons. The implementation for a complete big data solution involves the combination of multiple data storage technologies, for the various operating levels specified above. None of the existing data storage technologies meet the requirements of all these levels and it is necessary to select a separate solution for each of them. Particular attention should be paid in this context, not only to the data solutions used per se, but also to the modalities found for the transfer of data between these solutions.

2.2 Big Data Analytics

In the paper [15] are presented the most important tools, adopted for forecasting big data, and the following techniques are mentioned: Factor models, Bayesian models and Neural Networks. The paper [16] also illustrates the importance of big data, by the fact that Big Data and predictive analysis goes hand in hand in the modern age, with companies focusing on obtaining real time forecasts using the increasingly available data. The importance of big data analytics derives the need of developing the existing data mining techniques to meet the requirements, as stated in paper [17]. The author states that there is a need for the adoption of powerful tools, such as Data Mining techniques, which can aid in modeling the complex relationships that are inherent in Big Data. To support the need for more powerful tools, the paper [18] states that the process of forecasting big data is very hard, and identifies one important reason why the process gets so complicated, as the traditional forecasting tools cannot handle the size, speed and complexity inherent in Big Data. These characteristics of big data are than completed in the paper [19], where the author also identifies another important factor of the problems arising in big data

when forecasting big data, as the lack of a structure in these data sets and the size.
The main challenges when using big data for forecasting were identified in the paper [15], as follows:

- the skills required for forecasting with Big Data and the availability of personnel – as presented in the paper [20], developing these skills can be a big challenge. The authors recommend that in order to overcome this issue, Higher Educational Institutes should upgrade the syllabuses to incorporate the skills necessary for understanding, analyzing, evaluating and forecasting with Big Data;
- the noise that is likely distorting the signal and the accuracy of the forecast is another extremely important challenge in Big Data forecasting as identified by [21];
- new techniques should be developed for filtering the noise in Big Data to enable accurate and meaningful forecasts;
- hardware and software requirements for forecasting big data are another great challenge of big data, as is also stated in the paper [19].

Paper [22] presents statistical significance as a major threat in big data forecasting accuracy, as is very possible to make false discoveries from Big Data. Supporting this idea, in the paper [23] is observed that given the quantity of data to be processed, and forecasted, is a very complex task to differentiate between randomness and statistically significant outcomes;
Architecture of algorithms and the Data Mining techniques have been designed to handle data of much smaller sizes, as opposed to the size of Big Data is also considered a high challenge and in the paper [24] is contained a detailed evaluation of challenges, associated with the application of Data Mining techniques to Big Data;
In paper [15], the authors also consider Big Data itself a challenge for big data forecasting, because of its main characteristics: evolves and changes in real time; includes unstructured data; highly complex in structure.

2.3 Configuration of the Big Data Management solution

The proposed data management solution will be developed as a scalable and customized framework using open solutions based on cloud computing. The data framework will contain several tiers that can be customized. These tiers will allow complete customization, using different technologies for interconnecting smart meters and sensors (*SM/IoT tier*); fast processing and real-time analyses through big data (*BG-Tier*); high-reliable, low-latency, secure and performant data transmission and management with relational databases (*RB-Tier*); historical and advanced analytics through data warehouse (*DW-Tier*). These tiers are illustrated in Figure 1.

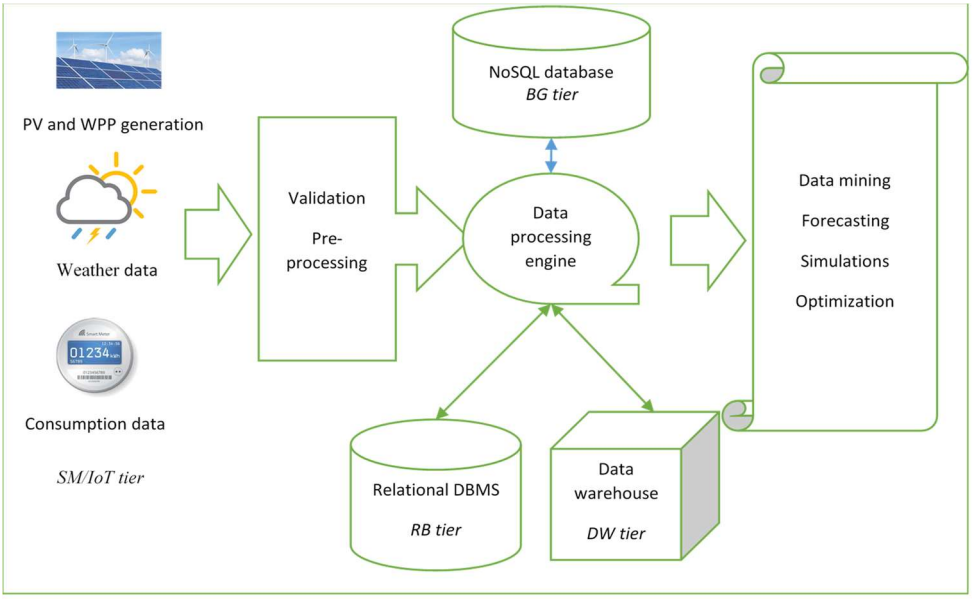

Fig. 1. Data management solution for SMARTRADE project

3. Proof of concept implementation

For the proof of concept we implemented a scalable system using different technologies. The system is able to retrieve data about electricity consumption from residential consumers and send it to the data processing engine, where the data is validated and stored in a persistent database. Based on the final data, the system can perform various calculations that will support the purpose of forecasting and optimisation.

For simulations we used a TP LINK HS110 smart plug connected to electric appliances, which is able to provide information such as the current consumption, power or voltage in real-time. The smart plug is communicating through WiFi with a Raspberry PI 3 mini-controller. The mini-controller is running on a Raspbian Linux distribution having a NodeJs application inside that is querying data about energy consumption from the smart plug at a fixed interval and distributes it outside the house to a messaging queue. Each time the NodeJs application is making a query and receives a valid response, it is seen in the system as an event. The event is basically created in the house and it propagates in the system until it is stored in the database.

We decided to use the messaging queue and the publish/subscribe pattern to ensure the speed and scalability of our application. From the message queue, the information is consumed by a Java application that is deployed in a public cloud. The application has the role of pre-processing the information, check the consistency and validate it. In the end, if the data is valid, it gets stored in an Elasticseach database on a specific index.

To have a cleaner view on the data in the system, we integrated Kibana in our database. Kibana is a Elasticsearch visualization tool that extends the database capabilities by providing new features such as UI, charts and graphs, real-time monitoring and alerts based on queries. To understand the data in real-time or during a certain period of time better, we implemented Kibana visualizers. As an example, Figure 2 presents two charts regarding energy consumption - the average (left-side) and the sum (right-side) of energy consumption monitored in one day. Using the visualizers, we were able to find the consumption patterns of the consumer. In Figure 2 two important peaks during the day are spotted - one in the morning and the other one in the evening.

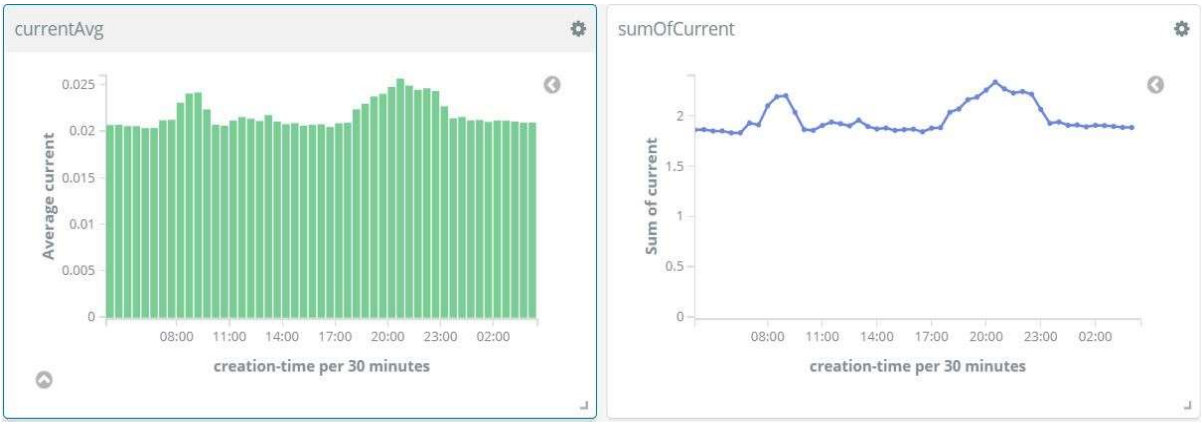

Fig. 2. Kibana visualization of one day energy consumption

In the proof of concept we implemented two tiers described in the paper, the SM/IoT tier and DB tier. In the future we are planning to introduce also the other two tiers in order to develop a complete scalable prototype.

4. Conclusions

The integration of IoT electric appliances in current development of the Smart Home concept require a scalable data management solution. The paper presented an overview of the Big Data solutions that are suitable for demand response management and proposed a scalable architecture implemented in the SMARTRADE project that consist in several tiers for data storage and processing. Two of these tiers are partially implemented during the project and some results were presented in the paper as a proof of concept.

Acknowledgement

This paper present 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 development 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 ID P_37_418, no. 62/05.09.2016, beneficiary The Bucharest University of Economic Studies.

References

[1] R. Shyam, G. Bharathi, K. Sachin, P. Prabaharan, K. Soman Pa - "Apache Spark a Big Data Analytics Platform for Smart Grid", Procedia Technology, 21, 171-178, 2015

[2] M. Zaharia, C. Mosharaf, J. Michael, S. Scott, I. Stoica - "Spark: cluster computing with working sets", Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, 10-10, 2010

[3] W. Wang, Z. Lu, "Cyber security in the smart grid: survey and challenges", Comput. Netw., 57, 13, 44-71, 2013

[4] M. McGranaghan, D. H. L. Schmitt, F. Cleveland, E. Lambert, "Enabling the integrated grid: leveraging data to integrate distributed resources and customers", IEEE Power Energy Mag., 14, 83-93, 2016

[5] V. Agarwal, L. H. Tsoukalas, "Smart grids: importance of power quality", Proceedings of first international conference on energy-efficient computing and networking, 13-15, April 2010, Berlin, 136-143, 2010

[6] S. M. Amin, "Smart grid: overview, issues and opportunities. Advances and challenges in sensing, modelling, simulation, optimization and control", Eur. J. Control, 17, 547-567, 2011

[7] Y. Yan, Y. Qian, H. Sharif, D. Tipper, "A survey on smart grid communication infrastructures: motivations, requirements and challenges", IEEE Commun. Surv.

Tutor., 15, 5-20, 2013
[8] SV. Oprea, B. Tudorică, A. Belciu, I. Botha "Internet of Things, Challenges for Demand Side Management", Informatică Economică, vol. 21, no.4/2017
[9] A. Florea, V. Diaconita, I. Dorobat, Business process modeling for sales processes automation, 15th EBES International Conference, 8-10 January 2015, Lisbon, Portugal, 2015
[10] S. Edlich, List of NoSQL databases". NoSQL - Your Ultimate Guide to the Non-Relational universe. Available http://nosql-database.org/ [December, 2017]
[11] DE-Engines.Com, Ranking of Relational DBMS. Available: https://db-engines.com/en/ranking/relational+dbms [December, 2017]
[12] B. Jones, The 4 Device Types in the Internet of Things, from a Data Perspective. Available: http://www.datasciencecentral.com/profiles/blogs/the-4-device-types-in-the-internet-of-things-from-a-data [December, 2017]
[13] B. Cole, XML - the embedded industry's not so secret weapon? Available: https://www.embedded.com/electronics-blogs/cole-bin/4217831/XML-the-embedded-industrys-not-so-secret-weapon [March 2018]
[14] G. Carvalho, Raspberry as a dedicated server? The result is amazing. Available: https://www.copahost.com/blog/is-it-possible-to-run-a-web-server-in-a-raspberry-pi-3-as-a-dedicated-server/ [March 2018]
[15] H. Hassani, E. Sirima Silva - "Forecasting with Big Data: A Review", Ann. Data. Sci., 2(1), 5-19, DOI 10.1007/s40745-015-0029-9, 2015
[16] D.Bernstein - "Big data's greatest power: predictive analysis" (2013) http://www.equest.com/cartoons/cartoons-2013/big-datas-greatest-power-predictive-analytics/ [March 2018]
[17] H. R. Varian - "Big data: new tricks for econometrics", J. Econ. Perspect, 28, 2, 3-28, 2014
[18] S. Madden - "From databases to big data", IEEE Internet Comput., 16, 3, 4-6, 2012
[19] D. Arribas-Bel - "Accidental, open and everywhere: emerging data sources for the understanding of cities", Appl. Geogr., 49, 45-53, 2014
[20] L. Einav, J. D. Levin - "The data revolution and economic analysis", Working Paper No. 19035, National Bureau of Economic Research, 2013
[21] N. Silver - "The signal and the noise: the art and science of prediction", Penguin Books, Westmins, 2012
[22] S. Lohr - "The age of big data" (2013) http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html [March 2018]
[23] B. Efron - "Large-scale inference: empirical Bayes methods for estimation, testing and prediction", Cambridge University Press, Cambridge, 2010
[24] H. Hassani, G. Saporta, E. S. Silva - "Data mining and official statistics: the past, the present and the future", Big Data, 2, 1, BD1-BD10, 2014

Adela BÂRA (b. October 11, 1978) is Professor at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Economic Cybernetics in 2002, holds a PhD diploma in Economics from 2007. She is the author of 15 books in the domain of economic informatics, over 50 published scientific papers and articles (among which over 20 articles are indexed in international databases, ISI proceedings, SCOPUS and 16 of them are WoS indexed). She participated as team member in 5 research projects and has gained as project manager 4 research grants, financed from national research

programs. Domains of competence: Database systems, Big data, Business Intelligence, Decision Support Systems, Data Mining & Analitics, Artificial Neural Networks.

Bogdan TUDORICĂ (b. July 10, 1976) is Lecturer at the Petroleum-Gas University of Ploiesti, Faculty of Economic Sciences. He has graduated the Faculty of Letters and Sciences of Petroleum-Gas University of Ploiesti in 1998 and holds a PhD diploma in Economic Informatics from 2015. He is author or co-author of 2 books in the domain of economic informatics, over 20 published scientific papers and articles (among which over 17 articles are indexed in international databases or ISI proceedings). He participated as team member in 3 research projects and 10 development projects. Domains of competence: Database systems, Big Data, Computer networks, Programming, Data Security.

Razvan Marales, Java Developer. METRO SYSTEMS Romania with interests in IoT, Cloud and Microservices. Currently, he is PhD student at the Bucharest Unoversity of Economic Studies, in the fied of Economic Informatics.