Knowledge management system for big data in a smart electricity grid context

Eugénia Vinagre

Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development
Polytechnic of Porto .
R. Dr. António Bernardino de Almeida, 431, P-4249-015 PORTO
PORTUGAL
E-mail: empvm@isep.ipp.pt

Tiago Pinto *

BISITE Research Centre
University of Salamanca
Calle Espejo 2, 30071, Salamanca, Spain
E-mail: tpinto@usal.es

Gil Pinheiro

Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development
Polytechnic of Porto .
R. Dr. António Bernardino de Almeida, 431, P-4249-015 PORTO
PORTUGAL
E-mail: gmvp@isep.ipp.pt

Zita Vale

Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development
Polytechnic of Porto .
R. Dr. António Bernardino de Almeida, 431, P-4249-015 PORTO
PORTUGAL
E-mail: zav@isep.ipp.pt

Carlos Ramos

Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development
Polytechnic of Porto .
R. Dr. António Bernardino de Almeida, 431, P-4249-015 PORTO
PORTUGAL
E-mail: csr@isep.ipp.pt

Juan Manuel Corchado

BISITE Research Centre
University of Salamanca
Calle Espejo 2, 30071, Salamanca, Spain
E-mail: corchado@usal.es

** Corresponding author*

Structured Abstract

Purpose – We have been witnessing a real explosion of information, due in large part to the development in Information and Knowledge Technologies (ICTs). As information is the raw material for the discovery of knowledge, there has been a rapid growth, both in the scientific community and in ICT itself, in the study of the Big Data phenomenon (Kaisler et al., 2014). The concept of Smart Grids (SG) has emerged as a way of rethinking how to produce and consume energy imposed by economic, political and ecological issues (Lund, 2014). To become a reality, SGs must be supported by intelligent and autonomous IT systems to make the right decisions in real time. Knowledge needed for real-time decision-making can only be achieved if SGs are equipped with systems capable of efficiently managing all the surrounding information. Thus, this paper proposes a system for the management of information in the context of SG to enable the monitoring, in real time, of the events that occur in the ecosystem and to predict following events.

Design/methodology/approach –The proposed system is based on the Apache Spark to provide in real-time a streaming and distributed processing. This knowledge management system architecture supports the development of enhanced data, information and knowledge analysis and management methodologies. This work proposes a novel data selection methodology that filters big volumes of data, so that only the most relevant and correlated information is used in the decision-making process in each given context.

Originality/value –New challenges arise with the upsurge of a Big Data era. Correlations between a huge data volume unstructured are often wrong when methods are dependent on data itself. It becomes more important than ever to know what to ignore and focus on what is important. It is in this scope that this paper gives its contribution. The proposed methodology searches correlations in data and only the most relevant data is used in each context. Data use is thus adapted to each situation, improving the forecasting process by reducing the data variability. The data filtering process also provides its contribution by reducing the forecasting execution time by using less, but more adequate data, in the training process.

Practical implications –Using the proposed methodology, training data can be chosen automatically accordingly to the data relevance and correlation for each problem, preventing the use of excessive and ambiguous data; and preventing an over-filtering of data that often comes from using only small amounts of highly correlated data while discarding information that could be relevant but whose value is not easily perceived. A case study is presented, considering the application of the proposed methodology. Results show that the data selection increases the forecasting effectiveness, as well as the computational efficiency of the forecasts, by using less yet more adequate data.

Keywords – Big Data, Data Analytics, Knowledge management, Smart grids

Paper type – Academic Research Paper

1 Introduction

Information is the raw material of knowledge discovery. Scientists and Companies use the data to discover new facts and solve many problems. And this new era of information defined as "Big Data" where the volume of data available grows exponentially, there is the potential to make faster advancements in many scientific disciplines and improve the profitability and success of many companies. Could this new information age boost the development of Smart Grids? Can it help solve many of your challenges and contribute positively to the efficient management of your complex processes?

A Smart Grid (SG) is defined as a complex electrical network that can intelligently integrate the actions of all users connected to it - generators, consumers and those that do both – to efficiently deliver sustainable, economic and secure electricity supplies.

In the new ecosystem of SG, all the players (i.e., power generation, transmission, distribution, customers, service providers, operations and markets) support their operations using a varied range of equipment that generate a large flow data. The last report issued by the European Union (Covig et al., 2014) refers numerous projects focusing the implementation of smart metering (SM). According to the same source, around 72 % EU customers are expected to be equipped with SM by 2020. The success of these projects launches an alert for the extensive amount of data generated in real time, that need adequate storage and analysis means to provide the development of dynamic services to better manage grid resources. Also, in the literature there are numerous references that characterize the type of data circulating on SG, at very high rates, as unstructured or, at most, semi-structured. Extrapolating this reality to the universe of the existent equipment and the foreseen roll outs by 2020 is easy to understand the challenge is now on the data management. However, all the data being generated can only be transformed into value if properly analysed to generate key knowledge in decision-making and the development of intelligent services able to dynamically manage the grid towards increasing sustainability, efficiency and safety. The value will be so much greater as the increasing ability to feed the ecosystem with data collected outside their own domain (e.g. atmospheric data, events, consumer behaviour, etc.), correlate and analyses them, not only to decide and predict, but also to discover something even imaginable.

There are numerous operations in the SG area that require data analysis (e.g., operability; cybersecurity and privacy; self-healing fault; demand response; competitive energy markets; auto configuration; resource optimization; real-time decisions; fore-casting; monitoring; etc.). Big data analytics is one of the biggest challenges in the BD domain. Traditional methodologies and algorithms are not prepared to run with large datasets (i.e. petabytes or more). Over the years, great efforts have been made on this issue and there are numerous references in the literature, of works and tools, for data analysis (based on batch and / or streaming) (M. Zaharia et al., 2016). Thus, we propose a system composed of several components, based on streaming and distributed processing provided by the Spark framework, to efficiently respond to one of the major challenges of SGs (i.e., real-time processing).

The paper is structured into 5 sections, with section 2 addressing Big Data evolution. Session 3 presents a proposed system for the management of information in the context of SG. A case study is presented in session 4 and the conclusions are presented in section 5.

2 Big Data evolution

In recent years, there has been an exponential increase of information generated and made available every day. Due to the rapid technological advancement (e.g. mobile devices, sensors, wireless communication, etc.) billions and billions of bytes are created every day. This phenomenon, referred to as Big Data, is characterized by 5 Vs (i.e. Volume, Velocity, Variety, Veracity, Value) (Chandarana and Vijayalakshmi, 2014):

- Volume - Volume refers to the vast amounts of data that is generated every second. We're not talking about gigabytes or terabytes, we're talking about a volume above the petabytes;
- Velocity - Velocity refers to the speed at which new data is generated and the speed at which it moves around;
- Variety - Variety refers to the different forms of data that we collect and use. Data comes in different formats, such as structured and unstructured (e.g. text, audio, video, images, log files, etc.);
- Veracity - Veracity refers to the trustworthiness of the data in terms of accuracy;

- Value - Value refers to the quantification of the benefits that can be extracted the data (when transformed into knowledge).

Each of these Vs represents real challenges (e.g. how to collect and transport a large volume of information; how to store this information, how to analyse and extract knowledge, how to ensure its security and privacy, how to process it in real time, etc.). Veracity represents one of the great big data challenges. If information is the raw material of knowledge, then wrong data can lead to wrong decision-making. And this, in some cases can be very critical and very dangerous. It is necessary to validate the veracity of the information, to make filters and to extract what really adds value.

The management of information with these characteristics (i.e. 5Vs) has aroused great interest in the scientific and business community. And much progress has been made. Figure 1 is illustrative, in relation to the growth of frameworks and tools, that have been available over the last years as a solution for the management of this type of information.

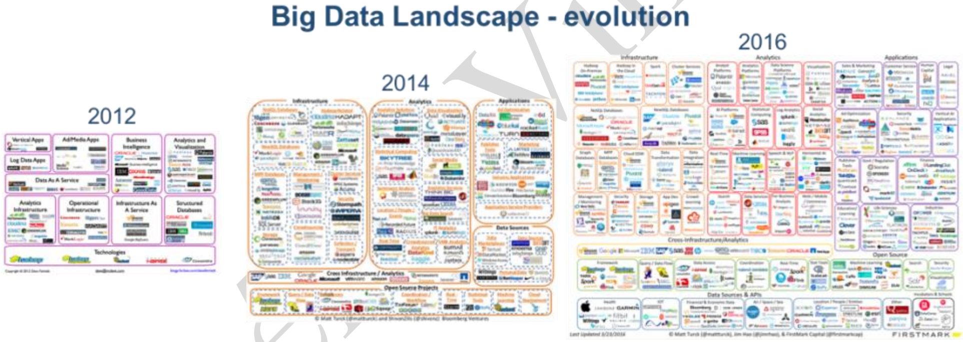


Fig.1 Evolution on Big Data - Landscape 2012 / 1016

Hadoop and Spark, are the most referenced frameworks in BD. The Apache Hadoop Framework was the first mainstream BD solution. It is based in Batch Processing, distributed file system HDFS (Hadoop Distributed File System), a programming model MapReduce and YARN (Yet Another Resource Negotiator) (Shvachko, 2010). Apache Spark is a set of tools and high level APIs for large scale distributed processing of data in-memory (M. Zaharia et al., 2016) . Currently, Spark is considered as the most active open source project in BD. Its speed advantages, allied with an out of the box integration of data manipulation using SQL like syntax, support for several storage systems, and ability to distribute machine-learning computation, have contributed to its success. However, today there is a greater tendency for hybrid systems to take advantage of the best of the two architectures (i.e. Apache Hodoop and Apache Spark), in order to ensure a better use of hardware resources. In fact, memory continues to be a rather expensive resource. On the other hand, it is necessary to reuse the investments made by the pioneering companies in the implementation of Big Data solutions. Not all decisions have to be made in real time. There are always decisions to be made next to real-time, medium and long term decisions, whatever the context of the business or research. There are more and more tools created to bridge the gap between these two universes and the latest versions of Apache spark already contemplate Hadoop integration.

3 Proposed system for information management in SG

In this section, we explain in detail the proposed system according to the three main branches of BD processing. The architecture of the proposed knowledge management system is represented in figure 2.

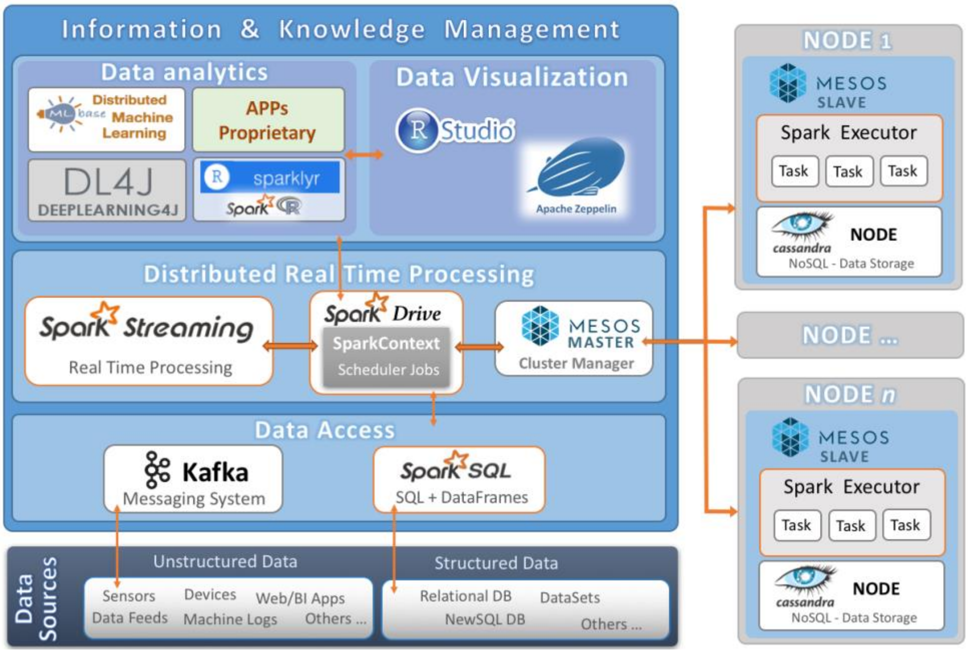


Fig.2 Knowledge Management System in the context of Smart grids

The proposed system is based on Apache Spark because the energy ecosystem is underloaded with real-time decisions. The implemented architecture consists of only 3 nodes (i.e. clusters) and is prepared to be easily expanded. The resource management and aggregation of tasks to each cluster is performed by Apache Mesos (Apache, 2017b), an open-source framework.

3.1 Big Data Storage

The proposed system is being implemented in an existing microgrid infrastructure in GECAD (Vinagre et al., 2015), as shown in figure 3. GECAD microgrid lab includes 3 buildings on the campus of the Institute of Engineering of the Polytechnic of Oporto (ISEP / IPP), with generation (photovoltaic (PV) and wind power) and production. Each of the buildings is equipped with energy analysers that measure energy consumption (ie, consumption data acquisition every 10 to 15 seconds) in three load groups by rooms (i.e., Air Conditioning (HVAC), lighting and sockets). More details on GECAD microgrid can be found in (Gomes et al., 2016).

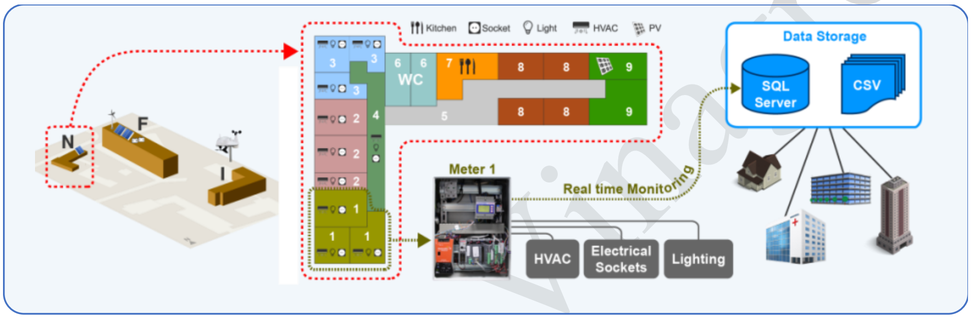


Fig.3 Data sources available in GECAD’s MicroGrid

This data sources are captured by the data access layer shown in figure 2. The data access layer is responsible for ingesting the data on the platform. This task is performed by the components: Apache Kafka (Apache, 2017a) (i.e. message queue optimized for the collection of unstructured data such as sensor data, analyzers, etc.) and SparkSQL (Apache Spark, 2017b) (i.e. a connector for structured data as Relational Database).

Kafka is collecting in real time data from the various analyzers represented in Fig. 3. The next phase will be to implement the Kafka connection with the ISEP Institute of Meteorology and the various sensors (i.e. presence sensors, temperature, humidity, brightness, etc.) which are currently being installed in GECAD buildings. SparkSQL allows access to historical data that is stored in SQL Server (e.g. consumption and production data, market price data, etc.).

The data collected in real time are being stored in Cassandra. Cassandra is a No-SQl database of column typology and easy integration into spark. According to the literature (Chebotko et al., 2015), Cassandra is an excellent solution for storing time series with good performance.

It is critical to compare multiple data stores intelligently and objectively so that sound architectural decisions can be made. We are evaluating storage performance with YCSB (Yahoo! Cloud Serving Benchmark) (Cooper et al., 2010). YCSB is an open standard for comparative performance evaluation of NoSQL data stores. Many factors go into deciding which data store to use for production applications, including basic features, data model, and performance characteristics on a given type of workload.

3.2 Big Data Analytics

Huge volumes of data, from the most varied natures, gathered from different sources, collected in different timings, often with high associated uncertainty, make the decision-making process a harsher task every day. Traditional methods of data analysis are not ready to deal with the speed at which data comes, nor with the amount of data that is necessary to be dealt with, nor with its variety. Big data analysis is a big challenge on Big Date and the real-time data analysis, that the energy ecosystem needs, is also a big challenge. We propose to the data analysis layer of the proposed system the following components:

- MLbase - Is an open source project developed for optimize and simplify building Machine Learning algorithms in a distributed setting (Sparks et al., 2013);
- DL4J - Deeplearning4j is an open-source, distributed deep learning project in Java and Scala (Skymind, 2017);
- Sparklyr - Provides an R interface to Spark's distributed machine learning algorithms. Allows also: filter and aggregate Spark datasets then bring them into R for analysis and visualization; create extensions that call the full Spark API and provide interfaces to Spark packages as MLlib (Apache Spark, 2017a) and easily establish connection with RStudio (Apache Software Foundation, 2017);
- Apps Proprietary - Applications, algorithms, etc. owners of GECAD.

A big problem in data analysis is the time spent with data preparation. It is estimated that approximately 80% of effort in data analysis is spent on data preparation and that only 20% of the time is spent in analysis methodology processing. Thus, the main goal of the analysis tools proposed in the system is to contribute successfully to address this problem. On the other hand, identifying forecasting as one of the great challenges of smart grids, we also want these tools useful in the search for methodologies capable of responding accurately to the prediction challenge.

3.3 Big Data Visualization

For the visualization of Big Data, we propose the Apache Zeppelin. Apache Zeppelin is an open source web-based notebook. It provides integration in Apache Spark. Visualizations are not limited to SparkSQL query, any output from any language backend can be recognized and visualized. But the latest official release (v. 0.5) does not yet support the R programming language

To solve Zeppelin's R language problem, we propose to use RStudio. It is also a notebook and provides an interface for visualization.

4 Electricity market forecasting – a case study

This section shows an illustrative practical application of a study that can be achieved using the proposed system to manage different knowledge sources through the combination of diverse algorithms. In order to reduce the time spent in preparing the data for analysis, and in order to increase the relevance of the data that is used for dana analysis, a methodology based on data filtering process of electricity market data from the Iberian market - MIBEL (MIBEL, 2017) is applied. The predictions are undertaken by forecasting approaches, namely Artificial Neural Networks (ANN) (Pinto et al., 2012), and Support Vector Machines (SVM) (Pinto et al., 2016). The proposed data filtering process using the proposed big data approach searches for correlations in data, so that only the most relevant data is used in each context. This methodology uses a clustering algorithm (Jiawei et al., 2006), which creates sub-groups of data according to their correlation. The clustering process is evaluated so that the number of data sub-groups that brings the most added value for the decision-making process is found. The used data set contains hourly electricity market prices from MIBEL ranging from January 2008 onwards (until September 2014). Figure 4 illustrates the data separation using the clustering process, considering 4 clusters or data groups.

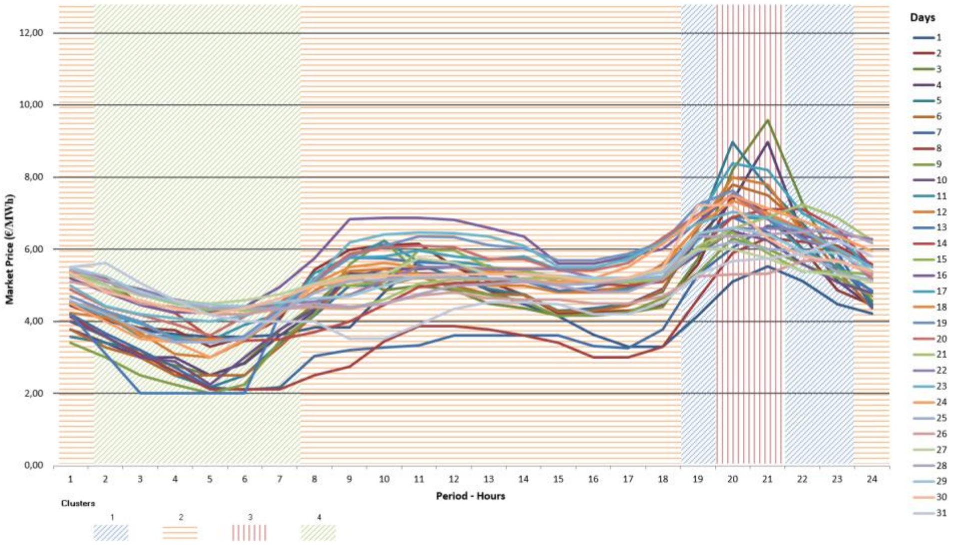


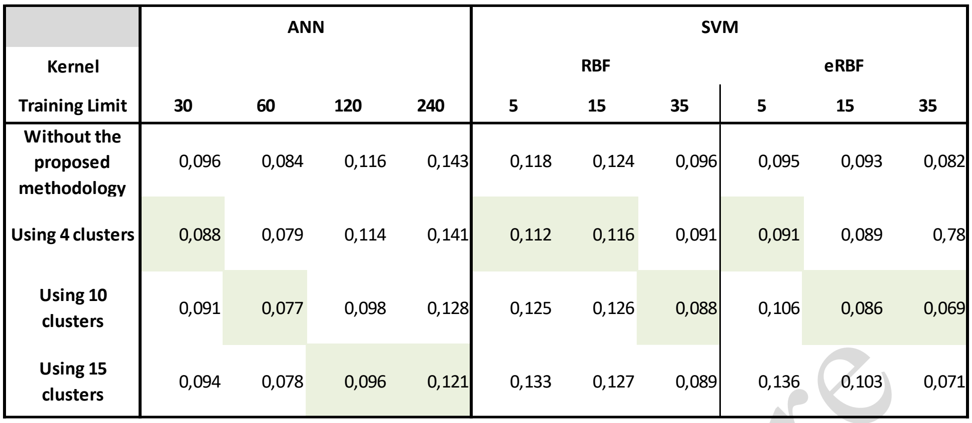
Fig.3 Data grouping resulting from the clustering process using: a) 4 clusters

From figure 4 it is visible that the clustering process is able to group data according to its similarity. It is visible a clear separation between the peak periods (where the market price is higher, namely hours 20 and 21); off-peak periods, between hours 2 and 7; mid-valued price periods, between hours 8 and 18, and hours 1 and 24; and near-peak periods, namely hours 19, 22 and 23

The forecasting methodologies, using the proposed methodology, only consider data of each sub-group when performing a forecast for a corresponding hour. E.g. if performing a forecast for the electricity market price of hour 21, when using 4 clusters, only data referent to the historic market prices in hours 20 and 21 are considered. Table 1 presents a comparison of the average Mean Absolute Percentage Error (MAPE) achieved by the forecasting process using the ANN and the SVM with and without the use of the proposed methodology, for a total of 2160 day-ahead forecasts (considering the individual forecast of each of the 24 hours of 90 days). Different amounts of historic data for training (training limits) are compared for both ANN and SVM. The SVM uses two different kernel functions: Radial Basis Function (RBF) and exponential Radial Basis Function (eRBF), with the following parameters, as detailed in (Pinto et al., 2016):

- *Kernel* RBF: σ (angle) = 6, ε-*insensitive* = 0, *limit* = ∞, *offset* = 1;

- *Kernel* eRBF: σ (angle) = 18, ε-*insensitive* = 0, *limit* = ∞, *offset* = 1;



	ANN				SVM					
Kernel					RBF			eRBF		
Training Limit	30	60	120	240	5	15	35	5	15	35
Without the proposed methodology	0,096	0,084	0,116	0,143	0,118	0,124	0,096	0,095	0,093	0,082
Using 4 clusters	0,088	0,079	0,114	0,141	0,112	0,116	0,091	0,091	0,089	0,78
Using 10 clusters	0,091	0,077	0,098	0,128	0,125	0,126	0,088	0,106	0,086	0,069
Using 15 clusters	0,094	0,078	0,096	0,121	0,133	0,127	0,089	0,136	0,103	0,071

Table 1 Comparison of ANN and SVM MAPE forecast error (%) when using the proposed methodology with different numbers of clusters, and when not using the proposed methodology

Table 1 shows that the forecast error is smaller when using the proposed methodology, for all considered cases. It is also visible that the use of a small number of clusters (less data separation, hence larger amounts of data used for the training process) work the best in the cases with smaller training limits. In these cases the number of considered historic days is small, therefore less data is used in the training process from the start (only the most recent, up-to-date data); for this reason a clustering process that does not add an excessive data limitation works better than the cases with a larger number of clusters. The data filtering with only a small number of data sub-groups is able to provide enough focus on the correlated used data, without limiting this amount excessively. On the other hand, when using larger training limits, the amount of considered data is by itself larger, hence a clustering process with more clusters works better. In this case, the stronger limitation of data makes sure that only the most correlated data is used, and since the length in time from which data is used is bigger, the filtering process does not result in an excessive data limitation, as occurs particularly in the case of SVM when using a training limit of 5 or 15 days, rather focusing on similar data throughout the time, and catching these particular data tendencies. In these cases the use of a large number of clusters degrades the forecasting process, making the results worse than without using the proposed methodology. From the analysis of the results presented in Table 1 it can be concluded that the ideal number of clusters, i.e. the level of data limitation, is dependent on the amount of data itself.

5 Conclusions

This work proposes a system for knowledge management in the context of SG to enable the monitoring, in real time, of the events that occur in the ecosystem and to predict following events. The proposed system is based on the Apache Spark to enable distributed processing in real-time. This knowledge management system architecture supports the development of enhanced data, information and knowledge analysis and management methodologies.

A case study is presented, showing the potential of the proposed system. In specific, an electricity market forecasting approach using a data-filtering approach, is analyzed, making use of the data collected in real-time by the platform, and applying a combination of algorithms. Results show that the novel data selection methodology that filters big volumes of data, so that only the most relevant and correlated information is used in the decision-making process in each given context, is able to provide better forecasting quality, by integrating different knowledge sources and facilitating the combination of distinct data analysis algorithms.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 641794 (project DREAM-GO) and from FEDER Funds through COMPETE program and from National Funds through FCT under the project UID/EEA/00760/2013.

References

Apache. (2017a). Apache Kafka. Retrieved April 3, 2017, from https://kafka.apache.org/

Apache. (2017b). Mesos. Retrieved April 3, 2017, from http://mesos.apache.org/

Apache Software Foundation. (2017). sparklyr. Retrieved April 3, 2017, from http://spark.rstudio.com/index.html

Apache Spark. (2017a). MLlib. Retrieved April 3, 2017, from http://spark.apache.org/mllib/

Apache Spark. (2017b). Spark SQL. Retrieved April 3, 2017, from http://spark.apache.org/sql/

Chandarana, P., and Vijayalakshmi, M. (2014). Big Data analytics frameworks. *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, 430–434. https://doi.org/10.1109/CSCITA.2014.6839299

Chebotko, A., Kashlev, A., and Lu, S. (2015). A Big Data Modeling Methodology for Apache Cassandra. *2015 IEEE International Congress on Big Data*, 238–245. https://doi.org/10.1109/BigDataCongress.2015.41

Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R., and Sears, R. (2010). Benchmarking cloud serving systems with YCSB. *Proceedings of the 1st ACM Symposium on Cloud Computing - SoCC '10*, 143–154. https://doi.org/10.1145/1807128.1807152

Covig, C. F., Ardelean, M., Vasiljevska, J., Mengolini, A., Fulli, G., and Amoiralis, E. (2014). *Smart Grid Projects Outlook 2014*. *JRC Science and Policy Reports*. https://doi.org/10.2790/22075

Gomes, L., Lefrançois, M., Faria, P., and Vale, Z. (2016). Publishing real-time microgrid consumption data on the web of Linked Data. In *Power Systems Conference (PSC), 2016 Clemson University* (pp. 1–8). IEEE.

Jiawei, H., Kamber, M., Han, J., Kamber, M., and Pei, J. (2006). Data Mining: Concepts and Techniques. *San Francisco, CA, Itd: Morgan Kaufmann*, 745. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Kaisler, S., Armour, F., and Espinosa, J. A. (2014). Introduction to Big Data: Challenges, Opportunities, and Realities Minitrack. *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*, 728–728. https://doi.org/10.1109/HICSS.2014.97

Lund, H. (2014). Renewable Energy Systems, Renewable Energy Systems–A Smart Energy Systems Approach to the Choice and Modeling of 100% Renewable Solutions. Academic Press,.

MIBEL. (n.d.). MIBEL - Iberian Electricity Market Operator. Retrieved January 20, 2017, from http://www.omie.es/

Pinto, T., Sousa, T. M., Praça, I., Vale, Z., and Morais, H. (2016). Support Vector Machines for decision support in electricity markets′ strategic bidding. *Neurocomputing*, *172*, 438–445.

Pinto, T., Sousa, T. M., and Vale, Z. (2012). Dynamic artificial neural network for electricity market prices forecast. In *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on* (pp. 311–316). IEEE.

Shvachko, K. (2010). The Hadoop Distributed File System. *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 1–10.

Skymind. (2017). Deeplearning4j: Open-source, Distributed Deep Learning for the JVM. Retrieved April 3, 2017, from https://deeplearning4j.org/

Sparks, E. R., Talwalkar, A., Smith, V., Kottalam, J., Pan, X., Gonzalez, J., … Kraska, T. (2013). MLI: An API for distributed machine learning. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1187–1192. https://doi.org/10.1109/ICDM.2013.158

Vinagre, E., Gomes, L., and Vale, Z. (2015). Electrical energy consumption forecast using external facility data. In *Computational Intelligence, 2015 IEEE Symposium Series on* (pp. 659–664). IEEE.

Zaharia, B. Y. M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., … Gonzalez, J. (n.d.). P56-Zaharia. https://doi.org/10.1145/2934664

Zaharia, M., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., … Venkataraman, S. (2016). Apache Spark: a unified engine for big data processing. *Communications of the ACM*, *59*(11), 56–65. https://doi.org/10.1145/2934664