Big Data Techniques For Renewable Energy Market

(Discussion Paper)

Michelangelo Ceci[2], Nunziato Cassavia[1], Roberto Corizzo[2], Pietro Dicosta[1], Donato Malerba[2], Gaspare Maria[3], Elio Masciari[1], and Camillo Pastura[3]

[1] ICAR-CNR
[2] UNIBA
[3] UNICAL
[4] GFM Integration
masciari@icar.cnr.it, nunzio@nemoweb.it, {michelangelo.ceci, donato.malerba}@uniba.it, {roberto.corizzo, pietro.dicosta}@gmail.com, {camillopastura,gaspare.maria}@gfmintegration.it

Abstract. The problem of accurately predicting the energy production from renewable sources has recently received an increasing attention from both the industrial and the research communities. It presents several challenges, such as facing with the high rate data are provided by sensors, the heterogeneity of the data collected, power plants efficiency, as well as uncontrollable factors, such as weather conditions and user consumption profiles. In this paper we describe *Vi-POC* (Virtual Power Operating Center), a project conceived to assist energy producers and, more in general decision makers in the energy market. In this paper we present the Vi-POC project and how we face with challenges posed by the specific application domain. The solutions we propose have roots both in big data management and in stream data mining.

1 Introduction

Renewable energy is actually a strategic sector for all the European countries, due to the strategic and urgent need of reducing pollution emission. In this perspective, Vi-POC provides a framework for collecting, storing, analyzing, querying and retrieving data coming from heterogeneous renewable energy production plants (such as photovoltaic, wind, geothermal, Sterling engine, water running) distributed on a wide territory. Moreover, Vi-POC implements an innovative system for real-time prediction of the energy production.

More in detail, it exploits big data technologies in order to effectively manage data coming from heterogeneous sources, possibly of different nature (i.e. photovoltaic plants collect data which are different from those collected by wind plants). Indeed, we decouple the model and the energy source in order to make the system flexible and scalable. Moreover, due to heterogeneity and high volume of data being analyzed, Vi-POC exploits suitable big data analysis techniques in order to perform a quick and secure access to data that cannot be managed with the traditional data management approaches.

In addition, Vi-POC is intended to refine (i.e. making more efficient, effective and reliable) raw predictions usually available from national electric authorities. High-precision prediction of energy needs will lead to two key advantages: 1) the definition of a better offer for the energy market and 2) the definition of an accurate purchase strategy.

The prototype could be useful both for main players of the energy market such as the distributors and smaller companies that act between offer (trailers) and request in the supply chain in order to build better supply planning for their customers. Moreover, the synergy between modern renewable energy production sites and advanced technologies for data storage and analysis allow a continuous monitoring of the production process. More in detail, we define suitable acquisition and storage models tailored for the production process in order to analyze data both in real time and batch mode with the goal of helping the strategy management in a crucial way.

2 Background

In a market organization of the energy business, the power contribution of single sources (especially renewable energy sources) becomes important in defining the price in the daily or hourly market: variations in the estimated generated power will influence the final clearing price [7]. For this reason, it is demanding to monitor the production and consumption of energy, both at the local and global level, store historical data and design new and reliable prediction tools. In the literature, researchers typically distinguish between two classes of approaches: statistical and physical. While in the latter, the basic idea is that of refining Numerical Weather Prediction (NWP) forecast by means of physical considerations about the site (e.g. obstacles, orography) [17][12], the former is based on models which establish a relationship between historical values and forecasted variables. These last approaches may take or not take into account NWP data.

Methodologically, there are approaches which are based on time series [10] and approaches that learn adaptive models [7]. In this respect, the adaptive models are generally considered to produce more reliable predictions, especially regarding to concept drift, but require a continuous training phase. For this reason, we resort to the approaches where adaptive models are proposed. For example, in [20], the estimation of the model parameters is based on an exponential weighted adaptive recursive least squares controlled by a forgetting factor. A different solution is proposed in [25], where a recursive method for the estimation of the local model coefficients of a linear regression function is proposed. In this case, the time dependence of the cost function is ensured by exponential forgetting of past observations. In [15] the author uses a stochastic gradient for online training of neural networks in wind power forecasting. Another work which uses neural networks is [6], where the authors train local recurrent neural networks of online learning algorithms based on the recursive prediction error. Bacher et al. [5] propose to forecast the average output power of rooftop PV systems by considering past measurements of the average power and NWP forecasts as inputs to an autoregressive model with exogenous input (ARX).

Sharma et al. [26] considered the impact of the weather conditions explicitly and used an SVM classifier in conjunction with a RBF kernel to predict solar irradiation. Bofinger et al. [8] propose an algorithm where the forecasts of an European weather

prediction center (of midrange weathers) were refined by local statistical models to obtain a fine tuned forecast. Other works on temporal modeling with applications to sustainability focus on motif mining. For example, Patnaik et al. [23] proposed a novel approach to convert multivariate time-series data into a stream of symbols and mine frequent episodes in the stream to characterize sustainable regions of operation in a data center. Finally, Chakraborty et al. [10] propose a Bayesian ensemble which involves three diverse predictors, that is, naïve Bayes, K-NN and sequence predicion (by means of a motif discovery algorithm).

However, most of the work referenced before operate in an offline mode. An exception is represented by [7], where training is performed online, according to the more general stream mining setting. The solution presented in this work is based on neural networks and the idea is that of adopting entropy concepts for their training. In particular, they combine Renyi's entropy with a Parzen window estimation of the error pdf as basis for training. Although results presented in such work are competitive, the presented approach does not consider the autocorrelation phenomenon according to which users (both energy suppliers and energy consumers) of the same type, located at close sites, tend to share similar properties on the basis of, for example, weather conditions in the area they are located. The previously referenced work [6] explicitly considers autocorrelation. However, as stated before, the solution the authors propose still works offline. Finally, an additional contribution of Vi-POC is in the use of big data technologies which, at the best of our knowledge, is not considered in the related work.

3 System Architecture

In this section we will describe *Vi-POC* (Virtual Power Operating Center) and we present in details its system architecture. Vi-POC is intended to fill an operational gap between the production layer and the management layer in the considered market.

Fig. 1 shows the role played by Vi-POC system in a real life energy market scenario. Vi-POC system collects data from different production plants operating by biomass, wind energy, solar energy or geothermal. These data are enriched by weather forecast data that will be exploited for prediction purposes. The system analyzes the collected data and provides predictive information that are used by trading centers in order to assist crucial activities as buying and selling energy. The availability of richer and more accurate information is a crucial requirement for companies operating on the energy market as they can obtain competitive advantages when performing an energy trade on the energy stock market or a bilateral trade.

Vi-POC achieves his goal by monitoring several production sites exploiting really different technology as mentioned above. Indeed, we gather information about the installed energy production modules, the manufacturers, the construction features and, (very important) their geographical position. Therefore, Vi-POC aims to create an automatized environment able to assist real-time monitoring of the distributed network of production sites. This activity leads to the production of huge amount of data that has to be properly analyzed. To give an idea, the three companies involved in the project (Sunelectrics s.r.l. (http://www.sunelectrics.it/), Iskra s.r.l. (http://www.rodonea.com/), GFM Integration (http://www.gfmintegration.com/)) collect data at intervals of 10 or 15

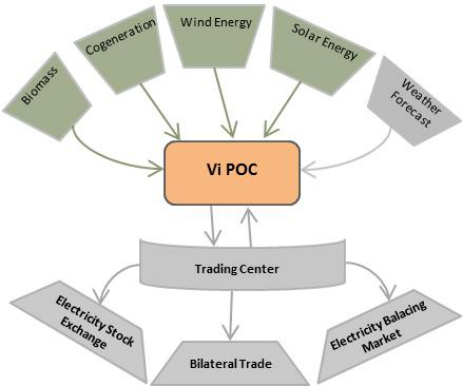

Fig. 1. Vi-POC operating scenario

minutes from more than 300 photovoltaic plants and wind power plants mostly located in south of Italy, for a total energy production of more than 50GW*year.

Vi-POC architecture is rather complex and includes several subsystems as depicted in Fig. 2. First of all, remote systems may not have a built-in monitoring systems or it could be not adequate, thus we need to install at the production site a remote agent called *Remote Terminal Unit* (RTU). This agent collects raw data generated by different devices depending on their operating mode. As data are collected they are pre-processed locally, then they are sent to the central system through one of the available communication networks made available. At the central site, the *Distributed Integration Layer* is devoted to collect data coming from remote RTUs. As mentioned above data are produced at high rates thus traditional data storage approaches may lead to inefficiency, thus we need to exploit a different paradigm, namely a Big Data approach that is most suited for dealing with high volumes and heterogeneous data. As data are organized according to the chosen model they are analysed by the Big Data Process Engine, that is responsible for processing data stored by the implemented prediction algorithms. Through this chain raw data are extracted, transformed and loaded in order to analyze them to create a value-added services that, through an appropriate *Service Integration Layer* (also referred as *Service Exposure*) will be available to the users that can profitably use them for improving their business.

3.1 Forecasting Module

Forecast may apply to a single renewable power generation system, or refer to the aggregation of large numbers of systems spread over an extended geographic area. Accordingly, different forecasting methods are used [24]. Forecasting methods also depend on the tools and information available to forecasters, such as data from weather stations and satellites and outputs from NWP models.

In this perspective, it becomes necessary to resort to data stream mining solutions which, on the basis of historical data of different nature, are able to forecast energy production for both small and large users.

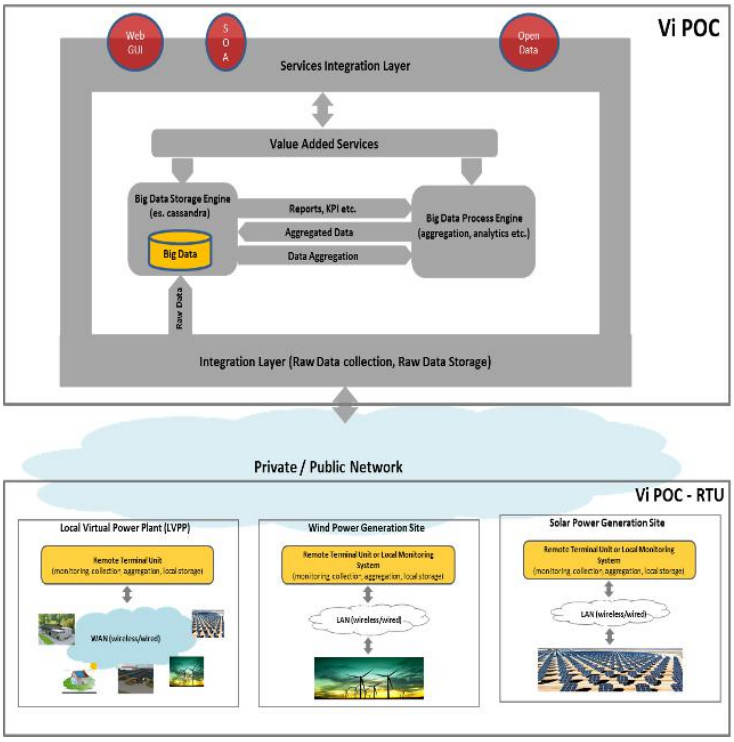

Fig. 2. Vi-POC architecture

In the literature, several data mining approaches have been proposed for power forecasting from renewable energy plants. Such algorithms, however, suffer from the non-adequate management of the autocorrelation phenomenon. Indeed, spatial proximity of sensors introduce autocorrelation in functional annotations and violate the assumption that instances are independently and identically distributed (i.i.d.), which underlines most data mining algorithms. Although the explicit consideration of these relations brings additional complexity to the learning process, we can also expect substantial benefits in predictive accuracy of learned classifiers [28][27].

As stated before, it has been recognized that physical (e.g. wind speed and solar irradiation) properties behavior exhibit a trail called concept drift which, in the vocabulary of data stream mining, it means that they change characteristics over time [7]. To take into account concept drift and, thus, deal with non-stationarity of data, either adaptive models [7] and time series [10] prediction algorithms have been applied. By focusing on the adaptive models, these require a continuous training phase which demands for a more complex underlying architecture, such as those of Big Data, in order to guarantee the possibility of storing and processing big volumes of data which arrive at high speed. Recently, several approaches that combine both the spatial and the temporal dimensions [13] from data produced as a stream, for power prediction from renewable energy plants have been proposed (see, for wind power prediction [11]). However, most of them do

not explicitly take the spatial autocorrelation phenomenon into account and, thus, do not exploit potential information coming from the spatial structure [9].

Finally, it is necessary to consider the noisy nature of the data. In fact, data that are collected by sensors can be never transmitted because of some (temporary) technical problems. This demands for solutions which work with missing data and missing labels,

Novelties of the methods that will be designed and implemented in Vi-POC are in:

- The possibility of considering the autocorrelation phenomenon. In this respect, weather conditions (e.g. quantity of rain, quantity of snow, wind direction and wind speed, temperature, solar radiation) are of fundamental importance for forecasting.
- The possibility of considering the real time and temporal nature of the data. This demands for efficient solutions which perform a quasi-real time analysis.
- The consideration of the noisy nature of the data. In fact, data that are collected by sensors can be never transmitted because of some (temporary) technical problems. This demands for solutions which work with missing data and missing labels, such as semi-supervised learning approaches[29].
- The semi-supervised learning setting also allows us to combine both the past history and forecasted values for weather conditions. This last aspect has been recently considered of fundamental importance of output power prediction [10].
- The possibility to work with large and diversified data. This demands for methods which work at different levels of granularity (both spatial and temporal). In fact, these methods are able to make the results of the analysis more easily understandable to non-experts.

All these aspects are generally taken into account separately in existing approaches. In our framework we intend to develop approaches which combine some of them and ultimately use all of them.

3.2 Big Data Analysis Module

In this section we present the architecture of our analysis engine. It effectively exploit the efficient column-oriented storage model described in previous section, the scalability of NoSQL systems and the OLAP server Mondrian. More in detail, the system depicted in Fig. 3 is composed by three modules:

1. Mondrian as OLAP Server
2. Hive as Query Executor on Hadoop MapReduce
3. HBase as NoSQL Data Storage

We choose to combine Mondrian and Hive in order to guarantee the distributed processing of queries across multiple nodes of our cluster composed of 20 (sixteen core each) nodes. We take also advantage of the usage of SQL as a common language for both the sub-systems. Moreover, the combined use of HBase and Hive allows us to overcome some speed limitations of Hive thus accelerating data access and querying. The latter is obtained by exploiting HBase main features such as vertical partitioning into Column Families, horizontal partitioning into Regions, replication, realtime access

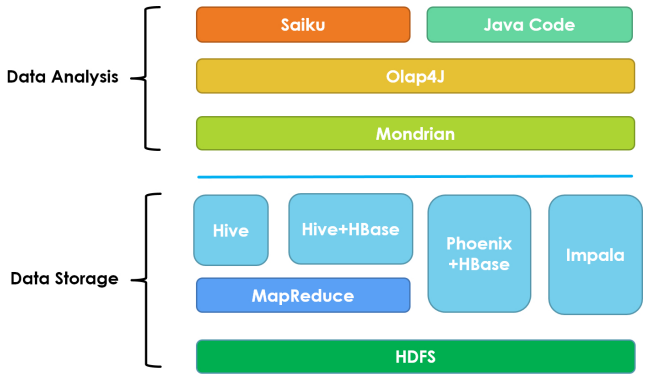

Fig. 3. Big Data Analysis Module architecture

and indexing. We point out here that the integration of these systems is far from being trivial as we need to take into account the different features of each module.

As regards the OLAP analysis, Mondrian issues SQL queries to Hive that translates them into MapReduce jobs, that access data stored in HBase through a JDBC driver. It performs the mapping of SQL commands to HBase commands (e.g. *get, put, scan, delete*).

In order to guarantee scalability, Mondrian propose two different solutions. The first solution is based on the so called *aggregate tables*. These tables contains pre-aggregate data. As an example, suppose that the system stores information about sales at hourly granularity level, but manager is interested to perform analysis having daily and weekly granularity level. We can create an aggregate table storing information at those granularity levels. The latter will result in a reduced size of data being returned when querying the repository. A second approach is based on caching. Indeed, Mondrian allows to cache schema, members, and segments (the objects used for aggregating data). This means that as data are queried they are materialized.

The proposed architecture provides a complete tool for Big Data analysis that allows users to specify queries in a simple way disregarding the complexity of the data acquisition and cleaning. Indeed, we performed preliminary tests on target power plants however due to privacy issues we are not currently able to show these early results.

4 Conclusion and Future Work

In this paper we presented a framework for end-to-end analysis of Big Data. Besides to general approaches for Big data management we designed a tool tailored for a key real life scenario, i.e. the renewable energy market. Our effort was on the effective integration of Pentaho and Hadoop environment in order to fully manage Big Data life cycle in the target scenario, from the data loading with Kettle to the data analysis with Mondrian. This makes Vi-POC suitable for the application of data stream mining algorithms which guarantee high-accuracy predictions. As current work we are working to the definition of a new query language for Mondrian and new predictive algorithms.

Indeed, we are investigating the possibility to focus on "hot" queries (e.g. count on a specific dimension that results particularly relevant) in order to achieve higher efficiency in response time.

Acknowledgements

We would like to acknowledge the support of the Italian Ministry of Education and Research through the PON-REC project Vi-POC - Virtual Power Operation Center (Grant number PAC02L1_00269).

References

1. Big data. *Nature*, Sep 2008.
2. Data, data everywhere. *The Economist*, Feb 2010.
3. Drowning in numbers - digital data will flood the planet - and help us understand it better. *The Economist*, Nov 2011.
4. D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom. Challenges and opportunities with big data. A community white paper developed by leading researchers across the United States. Mar 2012.
5. Peder Bacher, Henrik Madsen, and Henrik Aalborg Nielsen. Online short-term solar power forecasting. *Solar Energy*, 83(10):1772 – 1783, 2009.
6. T. G. Barbounis and J. B. Theocharis. Locally recurrent neural networks for wind speed prediction using spatial correlation. *Inf. Sci.*, 177(24):5775–5797, December 2007.
7. R.J. Bessa, V. Miranda, and J. Gama. Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting. *Power Systems, IEEE Transactions on*, 24(4):1657–1666, Nov 2009.
8. S. Bofinger and G. Heilscher. Solar electricity forecast - approaches and first results. In *20th European PV conference*, 2006.
9. Daniel Borcard, Pierre Legendre, Carol Avois-Jacquet, and Hanna Tuomisto. Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, 85(7):1826–1832, July 2004.
10. Prithwish Chakraborty, Manish Marwah, Martin F. Arlitt, and Naren Ramakrishnan. Fine-grained photovoltaic output prediction using a bayesian ensemble. In Jörg Hoffmann and Bart Selman, editors, *AAAI*. AAAI Press, 2012.
11. Alexandre Costa, Antonio Crespo, Jorge Navarro, Gil Lizcano, Henrik Madsen, and Everaldo Feitosa. A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, 12(6):1725 – 1744, 2008.
12. U. Focken, M. Lange, and H.P. Waldl. Previento - a wind power prediction system with an innovative upscaling algorithm. In *European Wind Energy Conference and Exhibition*, 2003.
13. Mohamed Medhat Gaber, Arkady Zaslavsky, and Shonali Krishnaswamy. Mining data streams: A review. *SIGMOD Rec.*, 34(2):18–26, June 2005.
14. João Gama. *Knowledge Discovery from Data Streams*. Chapman and Hall / CRC Data Mining and Knowledge Discovery Series. CRC Press, 2010.
15. George Kariniotakis. *Contribution to the development of an advanced control system for the optimal management of wind-diesel power systems*. PhD thesis, Ecole nationale superieure des Mines de Paris, 1996.

16. A. Labrinidis and H. V. Jagadish. Challenges and opportunities with big data. *PVLDB*, 5(12):2032–2033, 2012.
17. L. Landberg. Short term prediction of power production of wind parks. *J. Wind Eng. Ind. Aerodynam.*, 80:207–220, 1999.
18. S. Lohr. The age of big data. *http://www.nytimes.com/2012/02/12/sunday-review/bigdatas-impact-in-the-world.html*, Feb 2012.
19. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, May 2011.
20. H. A. Nielsen, P. Pinson, L. E. Christiansen, T. S. Nielsen, H. Madsen, J. Badger, G. Giebel, and H. F. Ravn. Improvement and automation of tools for short term wind power forecasting. In *EWEC 2007, European Wind Energy Conference - Scientific Track, Milan, Italy*, 2007.
21. Y. Noguchi. Following digital breadcrumbs to big data gold. *National Public Radio, http://www.npr.org/2011/11/29/142521910/the-digitalbreadcrumbs-that-lead-to-big-data*, Nov 2011.
22. Y. Noguchi. The search for analysts to make sense of big data. *National Public Radio, http://www.npr.org/2011/11/30/142893065/the-search-foranalysts-to-make-sense-of-big-data*, Nov 2011.
23. Debprakash Patnaik, Manish Marwah, Ratnesh K. Sharma, and Naren Ramakrishnan. Temporal data mining approaches for sustainable chiller management in data centers. *ACM Trans. Intell. Syst. Technol.*, 2(4):34:1–34:29, July 2011.
24. Sophie Pelland, Jan Remund, Jan Kleissl, Takashi Oozeki, and Karel De Brabandere. Photovoltaic and solar forecasting. Technical report, IEA PVPS, 2013.
25. Pierre Pinson, Henrik Aa. Nielsen, Henrik Madsen, and Torben S. Nielsen. Local linear regression with adaptive orthogonal fitting for the wind power application. *Statistics and Computing*, 18(1):59–71, March 2008.
26. Navin Sharma, Pranshu Sharma, David E. Irwin, and Prashant J. Shenoy. Predicting solar generation from weather forecasts using machine learning. In *SmartGridComm*, pages 528–533. IEEE, 2011.
27. Daniela Stojanova, Michelangelo Ceci, Annalisa Appice, and Saso Dzeroski. Network regression with predictive clustering trees. *Data Min. Knowl. Discov.*, 25(2):378–413, 2012.
28. Daniela Stojanova, Michelangelo Ceci, Donato Malerba, and Saso Deroski. Using ppi network autocorrelation in hierarchical multi-label classification trees for gene function prediction. *BMC Bioinformatics*, 14:285, 2013.
29. Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.