SHAPE: A NEW BUSINESS ANALYTICS WEB PLATFORM FOR GETTING INSIGHTS ON ELECTRICAL LOAD PATTERNS

Diego Labate
Enel Distribuzione-Italy
diego.labate@enel.com

Paolo Giubbini
Enel Distribuzione-Italy
paolo.giubbini@enel.com

Gianfranco Chicco
Politecnico di Torino-Italy
gianfranco.chicco@polito.it

Mario Ettorre
Exeura-Italy
mario.ettorre@exeura.eu

ABSTRACT

This paper deals with an innovative Web software platform for Business Analytics applied to the load patterns sourced from the Enel network's smart meters. Using ad hoc developed advanced time series analysis, the SHAPE (Statistical Hybrid Analysis for load ProfilE) platform enables the Data Analyst to solve in a user-friendly Rich Internet Applications (RIA) important tasks such as hourly energy analysis, customer classification, load prediction, and non-technical losses detection. The platform, developed within a R&D project, represents a tangible example of Applied Research-Industry knowledge transfer. By means of a Web Service interface, the platform exposes and shares the implemented models with other Corporate or third-party applications. The SHAPE datawarehouse currently stores three years progressively updated of load patterns for about 100,000 measurement points at the customers' premises.

INTRODUCTION

In today's global world, generating new knowledge and turning it into innovative products and services is crucial to maintain and enhance industry competitiveness. Transforming the results of scientific research into new products is, however, a complex process involving a broad range of actors. Profitable results may come from University-Industry R&D interactions.

The transition from *supplier-centric* to *customer-centric* framework carried out by Enel after the liberalization of the electricity market needs not only an efficient and effective smart metering system, but also requires new business strategies and innovative services.

After 13 years from the first smart meter installed, Enel is today the utility with the largest Automated Meter Reading/Management (AMR/AMM) infrastructure in the world consisting of more than 31 millions of smart meters remotely managed in Italy. From the first Enel smart meter, the energy market has been subject to significant changes, including EU corporate unbundling of the energy sector. The electricity market actors have now to face new challenges, such as better customer services provision and tariff definition, energy quality supply enhancement, technical and non-technical losses reduction, medium- and low-voltage network update to deal with renewable energy producers (two-way active energy flows), as well as rationalization of the network investment costs. Each market participants needs new business strategies in order to meet stringent requirements. For this purpose, the problem of characterizing the load and predicting the consumption behavior has been recognized as relevant, and the technological improvement in the metering devices has leveraged various issues in load pattern data management. With the introduction in the recent years of AMR/AMM systems in many countries, there has been a growing interest in developing applications based on load pattern data, mainly due to the ability of the underlying electronic meters technology to record data at a relative low cost consumption at 1-60 minutes resolution, rather than monthly.

The knowledge on the shape of the electricity consumption has gained increasingly higher importance, with the aim of partitioning the consumers into a number of classes representing the actual usage of electricity during time, assisting enhanced load prediction or non-technical losses detection [1]-[4]. The current approaches refer to managing the data of a large amount (i.e., many thousands) of customers [5][6].

Thanks to the AMR/AMM system called "Telegestore", Enel is today potentially able to measure and collect remotely a large amounts of consumption patterns recorded on a 15 minute basis from more than 31 millions of customers. The extraction of hidden knowledge from this huge amount of data requires state-of the-art technologies based on Data Mining and Machine learning and approaches on Big Data [7].

In this context, Enel Distribuzione has recently started the R&D project "SHAPE" in collaboration with Politecnico di Torino as the scientific research partner and Exeura as the industrial software-development partner. The SHAPE project represents a tangible example of Research-Industry knowledge transfer. Taking the results of the Applied Research phase, the final objective of the project is the development of a Business Analytics Web platform as a flexible and extensible Data Mining suite tailored for the electrical domain, allowing Analysts build their own analytics workflow in a user-friendly Rich Internet Application (RIA). The platform consists of the modules :

1. Datawarehouse management
2. Basic load analysis
3. Customer load classification
4. Load prediction
5. Non-technical losses detection support

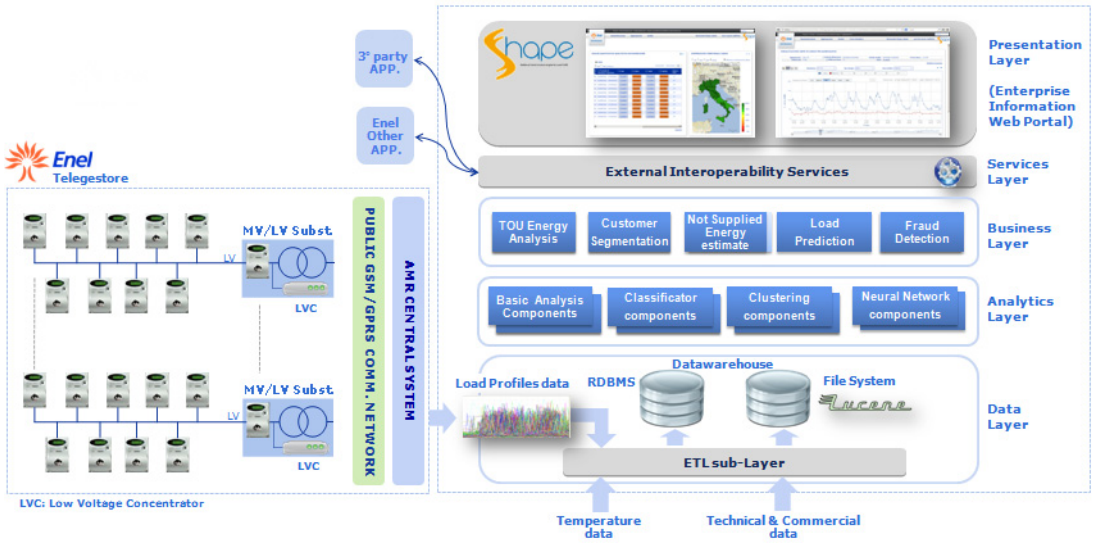


Figure 1 – SHAPE Architecture

Each module, in turn, is divided into sub-modules. At the moment of the paper submission, the development of modules 4 and 5 is ongoing and will be described in a next work. Currently the SHAPE datawarehouse (DW) stores information from about 100,000 Low Voltage (LV) smart meters, gathered starting from January 2011 and currently in progress, including about 1,000 balance measurement points installed within MV/LV substations. All information are treated for statistical purposes, in aggregate and anonymous form.

DATAWAREHOUSE & BASIC LOAD ANALYSIS

The SHAPE platform is provided with its own DW (Figure 1). All the necessary data such as daily load patterns, technical and commercial related information, temperature data, are provided from different Company systems data source. In the SHAPE context, a daily load pattern is a 96-point time series in which each point represents the energy in the last 15-minute interval. The patterns refer to active power consumption, active power production, and reactive power. Therefore, each year, for each customer's measurement point, SHAPE stores 3*96*365=105,120 measured data. Considering that the current number of measurement points already stored in the DW is expected to grow, retrieval and access this big quantity of data required efficient storage solutions also to enable a real-time experience to the user. Table 1 represents the amount of required storage space and number of points to be stored for a year-long load data.

Time Interval	#Smart Meters	Daily Load data	Amount of Data	Amount of Measures
1 Day	100K	1 KB	98MB	29Mln
1 Month	100K	1 KB	3GB	864Mln
1 Year	100K	1 KB	36GB	10.4Bln

Table 1 – Amount of Load data

Load pattern storage and retrieval

For load pattern storage, the File System implementation chosen is based on the open source Apache indexing technology Lucene™. Lucene is a full-text search library in Java which adds search functionality to an application or website by adding content to a full-text index. Lucene is able to achieve fast search responses because, instead of searching the text directly, it searches an index in a very efficient manner. In the SHAPE specific application a Lucene "Document", corresponds to a daily load pattern. A Document is a sequence of Fields. A Field is a *<name-value>* pair. In the SHAPE application, one field represents a daily load pattern stored as binary blob, other fields represent date, identifier, etc. Field values may be stored, indexed and/or analyzed. This kind of implementation has reduced retrieval and update time when the overlying functionality is requiring it. Other data such as technical and commercial information, have been stored in a Relational DBMS. Figure 2 shows the time comparison for querying data between 2 different Data Layer implementations. In a first solution named Database, 30 millions of daily load patterns have been stored using Microsoft SQL server 2008R2 (Intel® Xeon® CPU E7@2.00GHz 4 processor, 24 GBRAM, RAID 5 storage) with tables structured so that each 96-point load pattern is stored as a binary blob. In the second solution, the same load patterns have been indexed and stored in the server's file system by means of the structure provided by Lucene. Then, we measured the retrieval time of a growing set of load patterns, repeating the same specific query 5 times and then calculating the average time. We executed the same queries using the Lucene's own interrogation language. The results shown in Figure 2 demonstrate that the storage solution based on Lucene is more time-efficient in retrieving the information than the RDBMS solution.
Data cleaning operations on load patterns are carried out during periodic DW update, whereas missing values are handled by means of the "Load prediction" module. No

data dimensionality reduction is performed at this stage. At the end of each load patterns update (typically monthly), the application reports the details about the data quality stored in the DW and the errors detected and automatically corrected. By means of local storage support, SHAPE also enables the user to upload "spot" load patterns, to analyze them in a private area, without compromise the main DW.

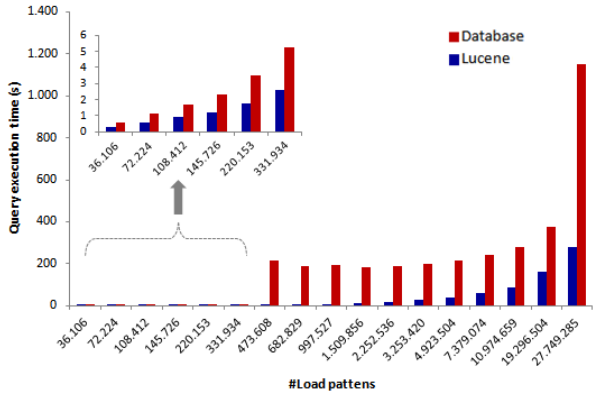


Figure 2 – Lucene vs. RDBMS time comparison

Load patterns' geographical coverage

One of the most important basic functionalities planned during the design of the platform has been to ensure the operator to obtain knowledge about geographical load pattern coverage, jointly with some related indexes. By proper interface, the user is able to visualize colored territorial nodes, as a function of amount of daily load patterns stored in the DW among the theoretical value in a given time period, defined as the product *measurement points*days* (IdCA index, Figure 3.a). In a different matrix component interface, the user can visualize details about daily patterns coverage (Figure 3.b). Suitable ad hoc coverage indexes have been developed.

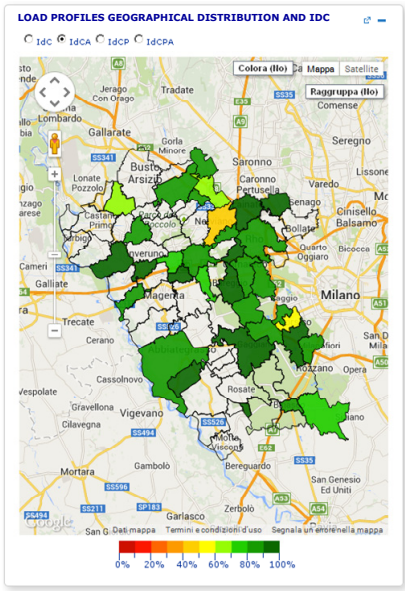
Other modules

Further basic analysis refers to these sub-modules:

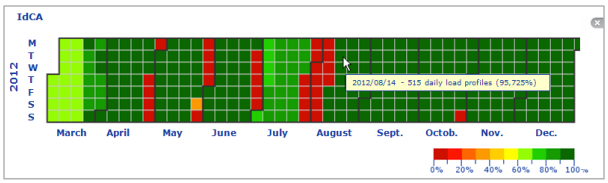
- Load geographical aggregator
- Load pattern viewer (Figure 4)
- Time-of-use energy geographical analysis
- MV/LV substation energy balance assessment

CUSTOMER CLASSIFICATION

Although classification and clustering are often mentioned in the same breath, they are based on different analytical approaches. In SHAPE, *customer segmentation* indicates the process of "classes discovery" which is generally carried out through clustering, and *customer classification* is the process in which, given a previous customers' classes partition, a classification model is induced through a supervised training phase based on classes labels. New customers, for which the outcome class is unknown, are presented to the model and scored accordingly.



a) spatial coverage



b) daily coverage

Figure 3 - Load pattern coverage near Milan in a given time period (a) and details per day (b)

Figure 4 – Load pattern viewer

Customer segmentation is the process of dividing customers into distinct, meaningful and homogeneous subgroups based on various attributes and characteristics. The type of segmentation used depends on the specific business objective. For the SHAPE's objectives one of the first task implemented was to reveal, for the first time in Italy, the customers' consumption classes at the National level based on consumption behavior. The Enel's Italian LV customer base was sampled taking into account the characteristic variable *annual energy*. Through a pre-sampling test we had established that the *annual energy* was quite heterogeneous. Therefore, in order to reduce the cost of the survey (load patterns extraction) and at the same

time increase the accuracy, a stratified sampling technique [8] has been developed and implemented, rather than simple sampling. The consumption points (customers) have been grouped in H = 10 macro-categories (strata) according to their commercial categories, as follows:

1. Residential Households
2. Transport
3. Agriculture
4. Industry
5. Commercial
6. Public Lighting
7. General Building Services
8. Heat Pumps
9. Non-Residential Households
10. Generation (Producer/Prosumer)

To represent the variability inside a single stratum, the annual energy consumption has been used as the representative feature. Starting from the number of customers N_h belonging to each stratum h = 1,…,H and from the mean value μ_h and standard deviation σ_h of the representative feature (obtained from the Company's databases), the statistically significant total number of points n has been determined for a given confidence probability (99%) with associated multiplier k of the estimated standard deviation and per cent amplitude d% of the confidence interval referring to the mean value μ of the representative feature, as follows:

$$n(d\%,k)=\frac{\left(\sum_{h=1}^{H} N_h\cdot\sigma_h\sqrt{\frac{N_h}{N_h-1}}\right)^2}{\sum_{h=1}^{H}\frac{N_h^2\sigma_h^2}{N_h-1}+\left(\frac{d\%}{100\,k}\sum_{h=1}^{H}N_h\mu_h\right)^2}$$

For example, with 99% confidence probability (k = 2.58) and d% = 5% the total number $n(d\%,k)$ is about 17,000. Correspondingly, the statistically significant number of point n_h for each stratum h = 1,…, H has been obtained. In addition, each macro-category can be partitioned into sub-categories on the basis of the reference power.

For each macro-category, different clustering procedures have been executed using the n_h measurement points on the complete load patterns for one year, with the aim to reveal the best inner partition of each macro-category into classes and obtain the class-related *typical load profiles*. Rather than consider seasons as separate periods, load seasonality has been assessed by means of a data-driven dedicated algorithm. The results indicate 21 main customers classes and 82 typical active energy load profiles associated to distinct periods of the year. A decision tree classification algorithm has been trained from the 82 typical load patterns. A classification model has assessed, calculating for each macro-category the related classification performance.

Customer Segmentation and Classification is a task repeatable by means of the SHAPE Workflow interfaces. Customers can be segmented by choosing active consumption, active production, or reactive energy. Assisted by a suitable geographical map, the Analyst starts creating a new experiment and selecting any measurement points aggregation. The Analyst can then create a Customer Segmentation workflow supported and boosted by a variety of ready-to-run plug-ins and algorithms tailored for energy analytics, some of which are described in Table 2.

At the end of the Workflow configuration, SHAPE executes the selected algorithms and visualizes the results, as well as adequate measures, in order to assess the revealed model and classes (Figure 5). For each step in Table 2, multiple choices can be made. In this case SHAPE generates a number of "runs", obtained by the Cartesian product of the different choices and comparing then the different results.

Workflow Step	I° level parameterization	II° level parameterization
1.Customers & time period selection	Aggregation or Macro-category	Optional statistical significance test
2.Pre-processing options	Normalization	Contract power, Mean, Max, Min/(Max-Min)
	Scale reduction	1…24 hours, 1 week, etc.
	load pattern mean	Weekday, Saturday, Holiday, Season, Month, etc.
	Low consumption load pattern filter	Filter out the load patterns with more then a given percentage of data below a given threshold
3.Clustering algorithms	Hierarchical (with different linkage criteria), Follow the Leader, K-means, Support Vector Clustering, hybrid methods.	Specific parameters for the selected algorithm
4.Partition adequacy measures	CDI - Clustering Dispersion Indicator, DBI - Davies-Bouldin Index, MDI - Modified Dunn Index, MIA - Mean Index Adequacy, SI - Scatter Index, SMI - Similarity Matrix Indicator, WCBCR - Ratio of Within Cluster sum of squares to Between Cluster variation	
5.Classification	Model selection and parameterization	

Table 2 – Customer Segmentation and Classification Workflow parameters

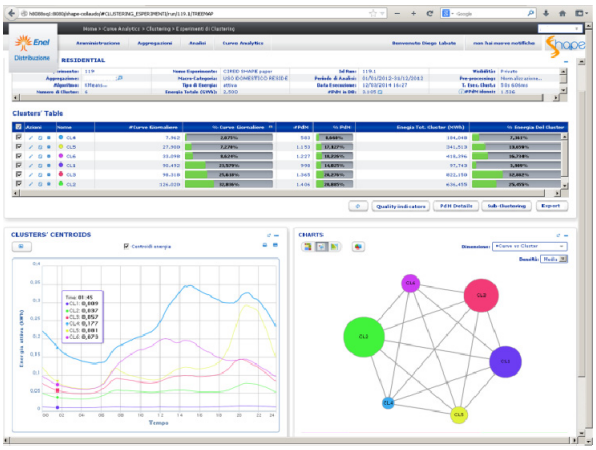


Figure 5 - Segmentation results viewer

PARALLEL COMPUTING TECHNIQUES

The Analytics layer in SHAPE exploits many-core programming techniques that are particularly effective in time series mining area, thanks to inherently high data dimensionality (number of time series on which computation takes place, as well as number of samples per each time series). Most of the implemented algorithms have been carefully crafted using OpenCL™ technology in order to distribute work among nVIDIA® GPU's computation unit while keeping work-item queues always full during both data pre-processing and mining phases. This required also automatic tuning of split level (number of many-core threads working on a local group in order to produce a result) and vectorization strategy, chosen by SHAPE in order to maximize performance and local memory allocation of underlying OpenCL-compliant device.
In tasks where many-core programming is not advisable due to insufficient data dimensionality or independence among data chunks, SHAPE uses work decomposition matched with a fork-join pool to keep all server's CPUs busy during computations.

CONCLUSION AND FUTURE WORK

Transforming the results of scientific research into new products is a complex process involving a broad range of actors. The Business Analytics platform implemented in the SHAPE project is the result of a successful Research-Industry knowledge transfer. The results delivered by the SHAPE project are of strategic importance, as they will provide decision support to current business processes and identification of new ones. The knowledge of the typical patterns of the aggregated energy consumption/production brings new understanding, directly available in SHAPE, on how the loads should be estimated in network applications. The economic benefit of understanding and predict customer loads will decrease investment costs by better matching long-term planning needs, will assist technical losses evaluation and allow better estimation of the economic losses resulting from service interruptions. The Non-technical losses module will support the verification of energy frauds and metering anomalies. The basic analyses are also suitable for directly using the real load patterns, enabling the calculation of time-dependent energy balances. The summary information on the consumption will be useful for more effective management of the electricity distribution network in synergy with existing initiatives in the Smart Grids field.
The SHAPE platform is ongoing at the moment of paper submission. The next modules to be released concern *Load prediction* and *Non-technical losses detection*, in which new findings in applied scientific research (to be reported in the future) are being implemented. Future refinements may include extensive analysis of the impact of prosumers' contributions on the network-related variables (voltage, current, losses, reliability).
On the side of Industrial software development, the enhancements deal with design and implementation of an architecture ready to manage Big data for a long time period with quick response time and high level of reliability and scalability, distributing both data and computations by means of appropriate technology.

ACKNOWLEDGMENTS

The authors wish to thank Lorenzo Gallucci, Antonio Notaristefano and Federico Piglione for their insights and implementation of the procedures.

REFERENCES

[1] G.Chicco, R.Napoli, P.Postolache, M.Scutariu and C.Toader, Customer Characterization Options for Improving the Tariff Offer, IEEE Trans. on Power Systems, vol.18, no.1, February 2003, pp.381-387.
[2] G.Chicco, R.Napoli, F.Piglione, M.Scutariu, P.Postolache and C.Toader, Emergent Electricity Customer Classification, IEE Proc. Gener. Transm. and Distrib., vol.152, no.2, March 2005, pp.164-172.
[3] L.M. Saini, and M.K. Soni, Artificial neural network based peak load forecasting using Levenberg-Marquardt and quasi-Newton methods, IEE Proc. Gener., Transm. and Distrib., vol.149, no.5, Sept. 2002, pp. 578–584.
[4] S.S.S.R. Depuru, L. Wang, and V. Devabhaktuni, Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft, Energy Policy, vol. 39, no. 2, February 2011, pp. 1007-1015.
[5] M. Koivisto, P. Heine, I. Mellin, and M. Lehtonen, Clustering of Connection Points and Load Modeling in Distribution Systems, IEEE Trans. on Power Systems, vol. 28, no. 2, February 2013, pp. 1255–1265.
[6] S. Ramos, J.M. Duarte, F.J. Duarte, Z. Vale, P. Faria, A data mining framework for electric load profiling, Proc. IEEE PES Conference on Innovative Smart Grid Technologies Latin America (ISGT LA), São Paulo, SP, Brazil, 15–17 April 2013.
[7] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q, Zhou, and V. Prasanna, Cloud-Based Software Platform for Big Data Analytics in Smart Grids, Computing in Science & Engineering, vol. 15, no. 4, July/August 2013, pp. 38-47.
[8] J. Neyman, On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, Journal of the Royal Statistical Society, Part IV, 1934, pp. 558-606.