# Improving Cities Sustainability through the Use of Data Mining in a Context of Big City Data

Carlos Costa and Maribel Yasmina Santos

*Abstract*— **Nowadays, cities consume more energy to fuel their day-to-day activities. With the rise of electrical devices we face more challenges associated with energy control and distribution. Apart from this, we also spend a lot of energy trying to either heating or cooling our homes. This paper illustrates an architecture to extract, load, transform, mine and forecast Big Data. This technological architecture makes use of a dataset containing electricity and gas consumption of homes distributed within multiple USA cities and states. The main purpose of our work consists in delivering to citizens a new form of self-monitoring their electricity and gas consumption, by comparing them to other homes within their cluster or state and by forecasting future energy consumptions. Moreover, the architecture also delivers to energy providers and cities a smarter overview of the energy landscape. This work uses simulated data from United States of America along with Hadoop, WEKA and Tableau to store and process Big Data, to produce clusters and time series forecasts, and to visualize information, respectively. The results reveal that, using this architecture, it is possible to produce accurate clusters of homes based on their energy consumption and it is also possible to forecast future electricity consumptions with a small margin of error.**

*Index Terms*— **Big Data, Clustering, City Sustainability, Smart City, Time Series Forecasting.**

## I. Introduction

Urban centers are growing and they seem to be the first choice for modern living, based on the fact that more than half of the population is living in urban environments [1]. With this phenomenon, various problems arise and cities need to adapt themselves to this trend.

In the last years we started hearing on a new concept, the concept of Smart Cities. Governments are facing more costs on labor, transportation, infrastructures, energy, and many other basic needs. Furthermore, citizens are behaving like natural consumers of government services and are now demanding more, regardless of the existing constraints [2]. Here is where Big Data comes along. Cities and their citizens generate vast amounts of data, with multiple degrees of complexity, at different speeds, from various sources, that does not conform to traditional technologies. This lead us to the general definition of Big Data [3]–[5].

The emerging need to make cities smarter, associated with the relatively recent concept of Big Data and the possibilities it brings, constitute the motivational basis for the development of this Big Data analytics architecture. It is able to process data from a city and, as we shall demonstrate, provide intelligent services, both for citizens and for the government or other stakeholders, through the use of data mining techniques such as clustering and time series forecasting [6]. Clustering is used to identify groups of homogeneous homes, with similar patterns in terms of energy consumption, enabling comparison and ranking, while time series forecasting is used to foresee future consumptions. The CRISP-DM model is used to conduct the data mining process, going through the phases of business understanding, data understanding, data preparation, modeling and evaluation [7].

The data used to validate the architecture is the “EPLUS TMY2 residential base” dataset [8], containing 238 files. Each file represents one year of electricity and gas hourly consumption, from a simulated home in a certain city in USA. Information about all the USA states was also extracted, containing all the USA state abbreviations, names, population and land area.

It is expected that the proposed architecture adequately support the intelligent monitoring and forecasting service, delivering refined visual data analyses. To validate the obtained results, the intra-cluster similarity (within cluster sum of squared errors) is considered, besides cluster variety, as well as a small error rate for the time series forecasting.

This document is structured as follows: Section II summarizes related work and describes the ways in which this work contributes to the state-of-the-art in this field. Section III illustrates the proposed technological architecture and gives an overview of the used dataset; Section IV describes the data preparation and mining process, including clustering and time series forecasting; Section V presents the data analysis and visualization, in order to redefine the energy bill and improve energy consumption monitoring. Finally, Section VI concludes with some remarks about the undertaken work and some guidelines of future work.

## II. Related Work

Within the scientific community there are already available some approaches to forecast energy consumption. Some of the works are mainly related with the energy price [9]–[11], while others address the research around energy loads [12], [13]. The common practice around these related works seems to be the mining of clusters before applying forecasting models. According to Alzate and Sinn [12], they

Manuscript received March 11, 2015; revised March 31, 2015. This work was supported by FCT – Fundação para a Ciência e Tecnologia, within the Project Scope: UID/CEC/00319/2013 and funded by the SusCity project, MITP-TB/CS/0026/2013.

Carlos Costa is with ALGORITMI Research Centre, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal (corresponding author, phone: +351-253-510308; fax: +351-253-510300; e-mail: a61555@alunos.uminho.pt).

Maribel Yasmina Santos is with ALGORITMI Research Centre, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal (e-mail: maribel@dsi.uminho.pt).

have achieved a 20% improvement in forecasting accuracy, using clustering before applying a forecaster.

Regardless of the used clustering techniques, such as K-Means [9], [11], Subtractive Clustering [10], Kernel Spectral Clustering [12] or Partial Clustering [13], as well as forecasting techniques, such as Neural Networks [11], Support Vector Machines [9], Adaptive Neuro-Fuzzy Inference System [10], PARX [12] or Fuzzy Inference [13], there seems to be a general common approach: use of clustering to improve efficiency of the forecasting model, either by adapting a model for each cluster or by using clustering as a feature extraction technique. These related works focus their results on improving data mining efficiency with state-of-the-art techniques, and in general the results show that the outcome is satisfactory. However, they are mainly focused on the data mining process and results, discarding not only the nature of the real world data that requires new storage and processing technologies but also the importance of the possible technological deployment, in order to deliver new services to citizens.

Other related works already describe the smart meter data as Big Data, presenting some methods to visualize information and extract knowledge [14], [15]. Apart from that, there are some works being developed in order to study the importance of the storage and processing infrastructure [15]–[17], highlighting non-relational databases and Hadoop.

This work aims to demonstrate how we can process the recorded energy data through a technological Big Data analytics architecture, using clustering and time series forecasting techniques, not only to select the adequate forecasting models (Linear Regression, Neural Network, Support Vector Machines or Decision Tree) for each cluster, but also to enrich the visual analysis and final smart service, delivering a reinvented energy bill to citizens and a new form of monitoring and targeting energy consumption to governments and energy providers. Consequently, the presented results are focused not only on data mining success, but also on how we can change the consumer and provider experience, by delivering reinvented ways of presenting energy consumption. As we also aim to achieve a small clustering and forecasting error, this work makes use of Big Data technologies to validate the possible deployment in a real world application scenario and presents the final results in a rich visual analysis, in order to surpass the gap between a successful data mining application and a Smart City service.

## III. Data and Architecture Overview

All the introduced steps and technologies proposed in this work can be abstracted in a technological architecture that can be seen in Fig. 1, in order to understand what could be a starting point for future implementations of similar services.

The architecture makes use of multiple Hadoop components, such as: Hadoop Distributed File System (HDFS) to store raw files; PIG to process scripts in order to aggregate data; HBase to temporarily store PIG results; HIVE to act as a data warehouse, containing the final dataset to originate the visual analysis. Talend Open Studio for Big Data is responsible for all the data flow processes, directing data from HDFS and HBase to the local file storage and vice versa. Then, WEKA was used to build clustering and forecasting models. Talend Open Studio for Big Data uses the WEKA's Java library to integrate the models "on the fly" and store the results on HIVE. Finally, we are able to perform visual data analytics using Tableau.

In order to understand the dataset used in this work, its schema will be presented, as well as how all the 238 files were verified, extracted and stored in Hadoop.

Each file in this dataset contains data from a simulated home, with average characteristics, like 3 bedrooms and 1 or 2 bathrooms, taking into consideration the environment and climate in which it is inserted. There is one file per city in USA, representing the average hourly consumption of a home within that city. Fig. 2 helps in understanding the schema and content of the files.

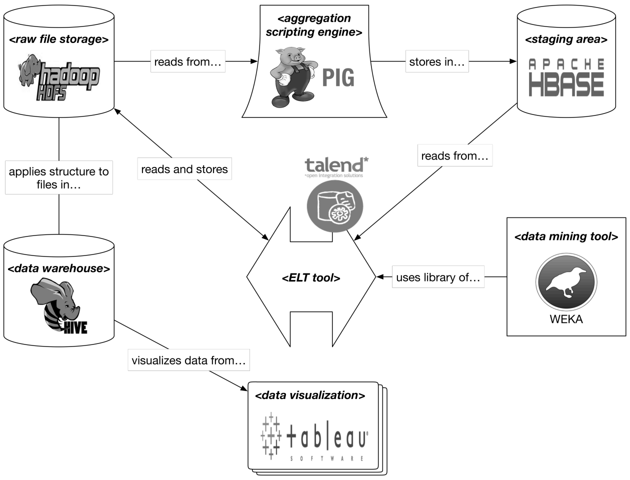
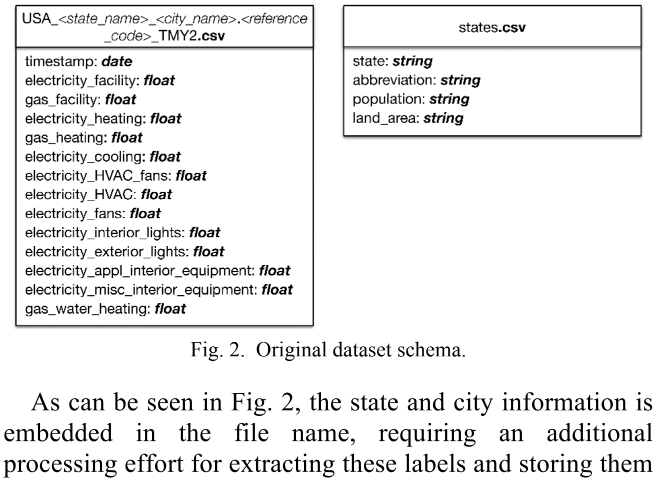


Fig. 1. Technological architecture overview.

Some classes of attributes can be identified: *general values of energy* (electricity facility, gas facility); *heating/cooling* (electricity heating, gas heating, electricity cooling, electricity HVAC fans, electricity HVAC, electricity fans, gas water heating); *lights* (electricity interior lights, electricity exterior lights); *interior equipment* (electricity appl interior equipment, electricity misc interior equipment). HVAC stands for "heating, ventilating and air conditioning", APPL means "appliances" and MISC means "miscellaneous".



USA_*<state_name>*_*<city_name>*.*<reference_code>*_TMY2**.csv**

timestamp: ***date***
electricity_facility: ***float***
gas_facility: ***float***
electricity_heating: ***float***
gas_heating: ***float***
electricity_cooling: ***float***
electricity_HVAC_fans: ***float***
electricity_HVAC: ***float***
electricity_fans: ***float***
electricity_interior_lights: ***float***
electricity_exterior_lights: ***float***
electricity_appl_interior_equipment: ***float***
electricity_misc_interior_equipment: ***float***
gas_water_heating: ***float***

states**.csv**

state: ***string***
abbreviation: ***string***
population: ***string***
land_area: ***string***

Fig. 2. Original dataset schema.

As can be seen in Fig. 2, the state and city information is embedded in the file name, requiring an additional processing effort for extracting these labels and storing them inside the file, for later analysis.

Working with Big Data requires to consider some characteristics the data might have: volume, variety, velocity, veracity and value [18]. In our demonstration case, taking into account that it is a proof-of-concept running only in one machine, the data had a considerable volume (238 files as mentioned, containing more than 8760 rows each, totaling more than 550 megabytes). Apart from that, in a real scenario, these data will come from various sensors networks and will be refreshed on an hourly basis. This dataset was used to test the architecture and intelligent service, because it brings with it the veracity of the simulation process [8] and the valuable information that can be extracted using data mining and visualization.

## IV. Data Preparation and Mining

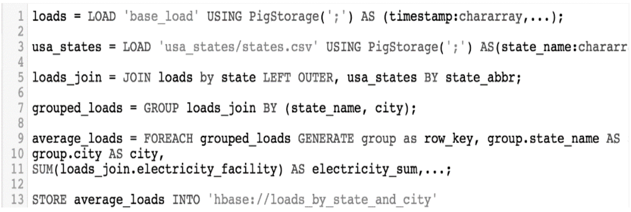
At this point is important to recall that Big Data quality is one of the most challenging steps, mainly due to its volume [19]. In this demonstration case, and after storing all 238 files in a comprehensive platform like Hadoop, after merging all the file in a unique one, containing all the available data, we were able to analyze data quality using Talend Open Studio for Data Quality. The data did not presented major flaws to consider in future transformations steps, mainly because it is simulated data that is not influenced by manual inserted problems. In a real world application it is also expected that the data do not present major flaws, due to the fact that it is extracted from sensors, using autonomous methods. Some files presented the value 0 in gas consumption. Later in this document it is explained how it will affect the development process.

The Data Mining task integrated a clustering exercise in order to segment homes by their electricity and gas consumption and a time series forecasting exercise to forecast future electricity consumptions. Combining these two techniques we deliver a method to compare a home's historic and forecasted consumption with other homes in its cluster and to compare consumptions between clusters.

### A. Clustering

To identify clusters from the dataset, a less detailed dataset was needed, namely, data grouped by state and city. To accomplish this, a grouping operation was performed using PIG, a high-level language embedded in Hadoop designated to perform data analysis.

In a general overview, all the 238 files were loaded, the energy dataset was joined with the file containing all states information from USA and all the data was grouped by state and city, calculating the sum of all grouped rows (Fig. 3).



```
loads = LOAD 'base_load' USING PigStorage(';') AS (timestamp:chararray,...);

usa_states = LOAD 'usa_states/states.csv' USING PigStorage(';') AS(state_name:chararr

loads_join = JOIN loads by state LEFT OUTER, usa_states BY state_abbr;

grouped_loads = GROUP loads_join BY (state_name, city);

average_loads = FOREACH grouped_loads GENERATE group as row_key, group.state_name AS
group.city AS city,
SUM(loads_join.electricity_facility) AS electricity_sum,...;

STORE average_loads INTO 'hbase://loads_by_state_and_city'
```

Fig. 3. PIG script for generating aggregated consumptions by state and city.

The results consist of annual energy consumption by state and city and were stored in HBase that acted as our staging area.

Because the data quality analysis demonstrated that there are not any null or blank values, only zeros were replaced by the global mean by state of the data used in the clustering process, previously stored in HBase (Fig. 4).

As far as constructing new data, the more detailed attributes of electricity consumption were aggregated in three categories: *heating/cooling*, *lights* and *interior equipment*. All other attributes remain the same. To format data, the float values were rounded to the nearest integer, due to data presentation purposes (Fig. 4).

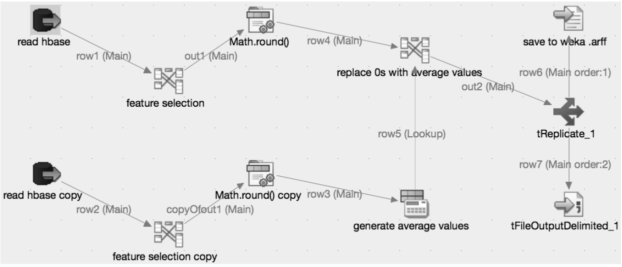


Fig. 4. Clean, construct and format data for clustering purposes.

As previously illustrated, all the attributes were included in data extraction and cleansing process. After doing some further analysis we maintained only the general consumptions on electricity and gas attributes, opting to exclude from the clustering process all the detailed consumptions (heating/cooling, lights and interior equipment). This decision was made after verifying the dispersion of the data, as some detailed consumptions did not present a significant level of dispersion. Apart from this, the general consumptions and the more detailed ones are correlated, and highly correlated attributes tend to influence some cluster techniques [20].

In this study, the clustering process is undertaken using the K-means algorithm, which requires the specification of an input parameter, k, representing the number of clusters. Once there is no indication of the appropriate number of clusters for this dataset, all the available data was iteratively used to produce clusters, incrementing the number of clusters to produce and recording each intra-cluster similarity error.

K-means is a well-known clustering model that partitions a dataset into k groups, selecting the cluster centers and iteratively refining them [21]. This was the only chosen model due to the simplicity to evaluate the results, using the intra-cluster similarity (within cluster sum of squared errors). The K-means model was built using WEKA's default parameters, changing only the number of clusters to produce, and using the Euclidean as distance function. The Fig. 5 shows the intra-cluster similarity for each clustering trial, each one with a different number of clusters.

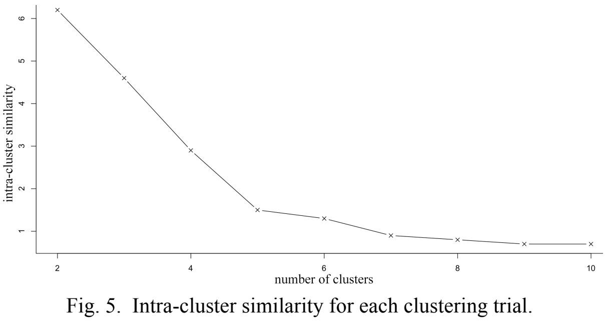
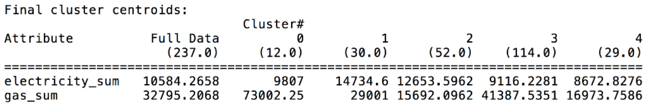


Fig. 5. Intra-cluster similarity for each clustering trial.

Using the L method in which the "*knee* is found in a number of clusters vs. clustering evaluation metric graph" [22], we are able to identify that 5 seems to be the ideal number of clusters for this dataset.

The model synthesis is shown in Fig. 6, pointing the centroids of the identified clusters and the number of cities in each one of them. The clustering model is built using the consumption of electricity and gas, grouped by state and city, as previously explained.



```
Final cluster centroids:
                                   Cluster#
Attribute             Full Data           0             1             2             3             4
                        (237.0)      (12.0)        (30.0)        (52.0)       (114.0)        (29.0)
=====================================================================================================
electricity_sum      10584.2658        9807       14734.6    12653.5962     9116.2281     8672.8276
gas_sum              32795.2068    73002.25         29001    15692.0962    41387.5351    16973.7586
```

Fig. 6. Clustering model synthesis.

### B. *Time series forecasting*

For the time series forecasting process, the results from the clustering process are used as input for a file containing the *state name*, *state abbreviation*, *city name* and *cluster number*, which result was joined with the original consumption dataset (Fig. 2). Another PIG script was coded to process that step and group the result by cluster number and timestamp (day and hour). This gives us the hourly consumption of each cluster during the entire year.

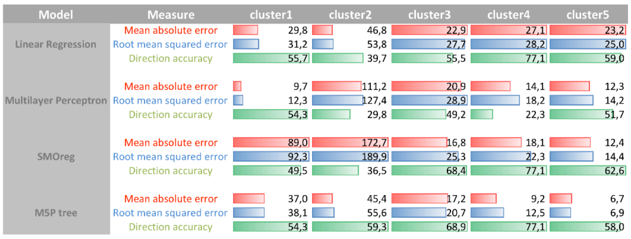
Regarding the time granularity, and as the dataset contained one year of energy consumption, not allowing for any seasonality analysis, the chosen time granularity was per week, being able to predict the next weeks of energy consumption, offering high value for monitoring and planning. After this process, the dataset to use in the forecasting process includes the following attributes: *cluster number, week and electricity and gas consumptions*.

Before starting the forecasting process, and as the first and last weeks of the dataset did not offer the full 7 days, we choose to discard them, improving the variance of the time series.

Previously in the clustering process, the detailed attributes had to be removed, leaving only the general electricity and gas attributes. However, because we are now dealing with time series, we have to remember that some values of gas consumption were zero, causing one of the clusters to have zero as center. Besides that, we observed that another cluster presents serious declines in gas consumption. Due to these two facts, this work will only do forecasts of electricity consumption.

Testing time series forecasting using WEKA is very similar to other traditional data mining techniques, like classification or regression. To evaluate models the holdout method was used, leaving 20% of the dataset for testing purposes. There were 3 different metrics: Mean Absolute Error (MAE); Root Mean Squared Error; Direction Accuracy.

Four models were built and assessed (Linear Regression, Multilayer Perceptron, SMOReg and M5P tree), for each of the five clusters, using WEKA's default parameters. The maximum lag was set to 12 weeks and the number of time units to forecast was set to 8 weeks, meaning that the model will mainly look at the previous 12 weeks to forecast the next 8. Fig. 7 shows the obtained results.



| Model | Measure | cluster1 | cluster2 | cluster3 | cluster4 | cluster5 |
|---|---|---|---|---|---|---|
| Linear Regression | Mean absolute error | 29,8 | 46,8 | 22,9 | 27,1 | 23,2 |
| | Root mean squared error | 31,2 | 53,8 | 27,7 | 28,2 | 25,0 |
| | Direction accuracy | 55,7 | 39,7 | 55,5 | 77,1 | 59,0 |
| Multilayer Perceptron | Mean absolute error | 9,7 | 111,2 | 20,9 | 14,1 | 12,3 |
| | Root mean squared error | 12,3 | 127,4 | 28,9 | 18,2 | 14,2 |
| | Direction accuracy | 54,3 | 29,8 | 49,2 | 22,3 | 51,7 |
| SMOreg | Mean absolute error | 89,0 | 172,7 | 16,8 | 18,1 | 12,4 |
| | Root mean squared error | 92,3 | 189,9 | 25,3 | 22,3 | 14,4 |
| | Direction accuracy | 49,5 | 36,5 | 68,4 | 77,1 | 62,6 |
| M5P tree | Mean absolute error | 37,0 | 45,4 | 17,2 | 9,2 | 6,7 |
| | Root mean squared error | 38,1 | 55,6 | 20,7 | 12,5 | 6,9 |
| | Direction accuracy | 54,3 | 59,3 | 68,9 | 77,1 | 58,0 |

Fig. 7. Electricity forecasting evaluation.

The obtained measures indicate that the chosen models behave really well, taking into consideration that the errors are measured in kilowatts/hour (kw/h). This means that for each cluster, there is at least one model that can forecast the next 8 weeks of electricity consumption with a Mean Absolute Error (MAE) less than 16.8 kw/h, except for the cluster 2, whose lowest observed MAE is around 46. Moreover, there are models capable of forecast with a MAE of 6.7 kw/h (Fig. 8). Taking into consideration that values fluctuate between 120 and 412 kw/h, these tests reveal satisfactory results.

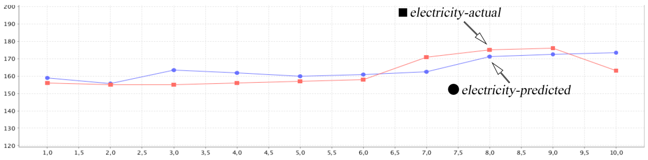
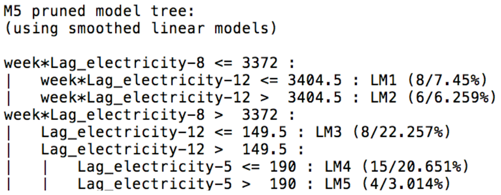


Fig. 8. Example of a tested model with a low MAE.

The clustered dataset was tested with all 4 models, with the goal of finding the best forecaster for each cluster and its corresponding cities. The underlying model of the 2nd, 3rd, 4th and 5th clusters is proven to be the M5P tree, while for the 1st cluster is the Multilayer Perceptron, as can be seen in Fig. 7. Consequently, a home will inherit the forecaster of its cluster. For example, if a New York home is in cluster 3, the forecaster that will be used to predict the electricity consumption will be the M5P tree.

As can be concluded, the model that successively proves to be the most suitable is the M5P tree, a decision tree whose leaves are linear models, as shown in Fig. 9.



```
M5 pruned model tree:
(using smoothed linear models)

week*Lag_electricity-8 <= 3372 :
|   week*Lag_electricity-12 <= 3404.5 : LM1 (8/7.45%)
|   week*Lag_electricity-12 >  3404.5 : LM2 (6/6.259%)
week*Lag_electricity-8 >  3372 :
|   Lag_electricity-12 <= 149.5 : LM3 (8/22.257%)
|   Lag_electricity-12 >  149.5 :
|   |   Lag_electricity-5 <= 190 : LM4 (15/20.651%)
|   |   Lag_electricity-5 >  190 : LM5 (4/3.014%)
```

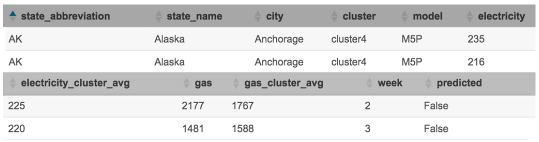
Fig. 9. M5P tree model for cluster 5.

Next section presents the analysis and visualization of the prediction models in a Smart City context.

## V. Data Analysis and Visualization: Reinventing the Energy Bill

This section shows how the models that served as forecasters for each cluster are used to forecast data from homes within the corresponding cluster. This was accomplished by the integration of WEKA's Java library into Talend Open Studio for Big Data, allowing, for each home, to have not only historic data, but also forecasted data.

At this point all the historic, clustered and forecasted data were in HDFS, and it was necessary to give back some structure to it, forming a perfectly fitted dataset for analysis and visualization. All relevant attributes were joined in a HIVE table (Fig. 10), intended to store data in a structured form, which is suitable for visual analysis. The Hive table includes: *State_abbreviation* - abbreviation of the USA state name; *State_name* - name of the USA state; *City* - name of the USA city; *Cluster* - cluster number that the respective home belongs; *Model* - model used to forecast data from the respective home; *Electricity* - home's electricity consumption in that week; *Electricity cluster avg* - average electricity consumed by all homes in that cluster on that week; *Gas* - home's gas consumption in that week; *Gas cluster avg* - average gas consumed by all homes in that cluster on that week; *Week* - number of the week. Starts in week 2, ends on week 60. Weeks 53 to 60 contain forecasted values; *Predicted* - flag pointing forecasted or historic data.



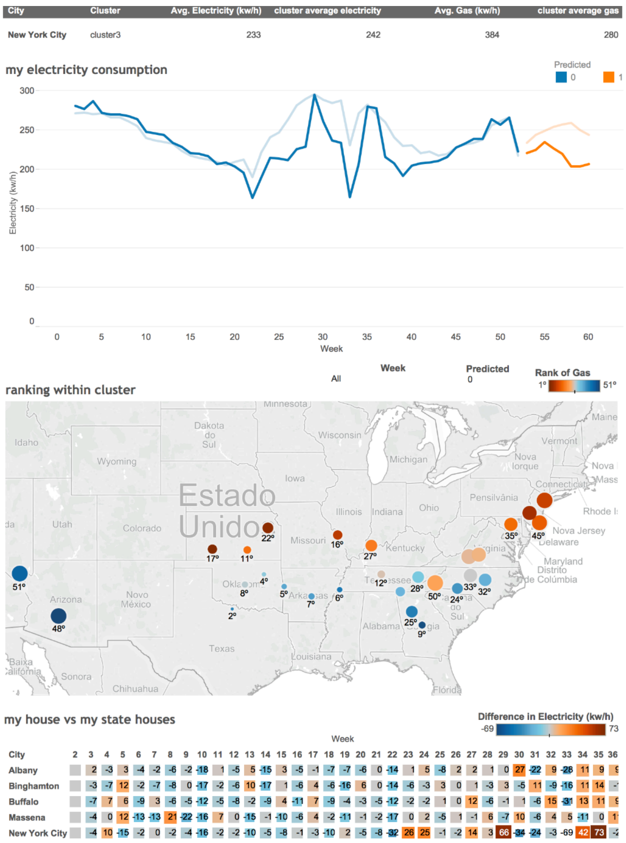
| state_abbreviation | state_name | city | cluster | model | electricity |
|---|---|---|---|---|---|
| AK | Alaska | Anchorage | cluster4 | M5P | 235 |
| AK | Alaska | Anchorage | cluster4 | M5P | 216 |

| electricity_cluster_avg | gas | gas_cluster_avg | week | predicted |
|---|---|---|---|---|
| 225 | 2177 | 1767 | 2 | False |
| 220 | 1481 | 1588 | 3 | False |

Fig. 10. Final dataset's sample for visual analysis.

Since the system is capable of processing energy consumption data and it is also capable of clustering and forecasting it, new perspectives on how to give feedback to homes can emerge. Fig. 11 shows a visual data analysis for one home in New York City. The analysis begins comparing the New York home against the cluster in which it is inserted, presenting the average energy consumptions. Then, the line chart overlaps the home electricity consumption (dark blue line) with the cluster consumption (light blue line), containing not only historic data but also predicted values (dark orange line for the home and light orange line for the cluster), resulted from the application of the time series forecasting model. Besides that, it is also possible to rank homes by their energy consumption, comparing a certain home to others within its cluster. That ranking can be illustrated in a geographical map. The last chart in Fig. 11 shows a heat map, with the aim of comparing the home in New York with other homes in the same USA state.

The analysis described above exemplifies how innovative and intelligent Smart Cities services can be, and in this particular case, how clustering and time series forecasting can be joined to form a visual analysis, that contributes significantly for citizens to have a more controlled consumption experience, monitoring historical and predicted data and comparing their home with others in their cluster or state.



| City | Cluster | Avg. Electricity (kw/h) | cluster average electricity | Avg. Gas (kw/h) | cluster average gas |
|---|---|---|---|---|---|
| New York City | cluster3 | 233 | 242 | 384 | 280 |

| City | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Albany | | 2 | -3 | 3 | -4 | -2 | -6 | -2 | -18 | 1 | -5 | 5 | -15 | 3 | -5 | -1 | -7 | -7 | -6 | 0 | -14 | 1 | 5 | -8 | 2 | 2 | 1 | 0 | 27 | -22 | 9 | -28 | 11 | 9 | 9 |
| Binghamton | | -3 | -7 | 12 | -2 | -7 | -8 | 0 | -17 | -2 | -6 | 10 | -17 | 1 | -6 | 4 | -6 | -16 | 6 | 0 | -14 | -6 | -3 | 3 | 0 | 1 | -3 | 3 | -5 | 11 | -9 | -16 | 11 | 14 | -1 |
| Buffalo | | -7 | 7 | 6 | -9 | 3 | -6 | -5 | -12 | -5 | -8 | -2 | -9 | 4 | -11 | 7 | -9 | -4 | -3 | -5 | -12 | -2 | -2 | 1 | 0 | 12 | -6 | -1 | 1 | -6 | 15 | -31 | 13 | 11 | 9 |
| Massena | | -4 | 0 | 12 | -13 | -13 | 21 | -22 | -16 | 7 | 0 | -15 | -1 | -10 | -6 | 4 | -2 | -11 | -1 | -2 | -17 | 0 | 2 | -5 | 1 | -1 | 6 | -7 | 10 | -6 | 4 | 5 | -11 | 0 | 1 |
| New York City | | -4 | 10 | -15 | -2 | 0 | -2 | -4 | -16 | -2 | -2 | -10 | -5 | -8 | -1 | -3 | -10 | 2 | -5 | -8 | -32 | 26 | 25 | -1 | -2 | 14 | 3 | 66 | -34 | -24 | -3 | -69 | 42 | 73 | -2 |

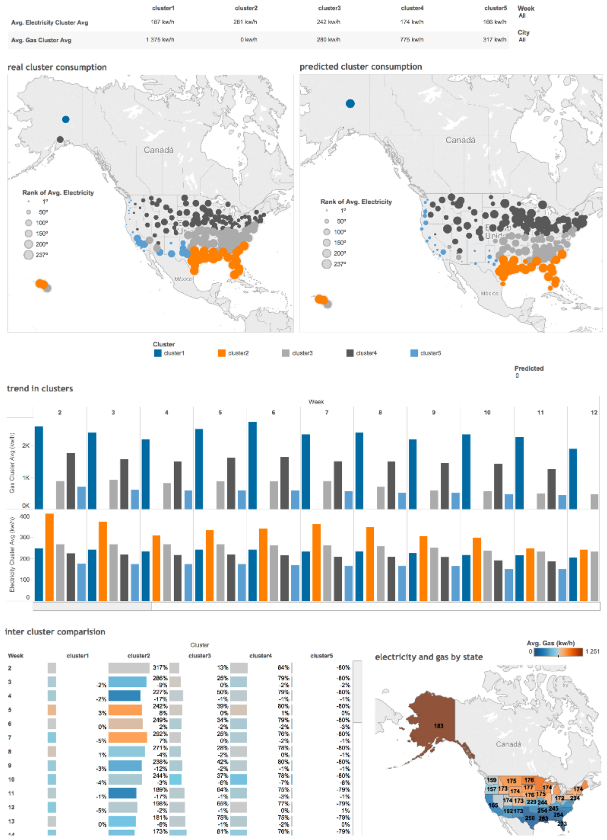
Fig. 11. Reinvented energy bill.

Traditional energy bills do not give a clear overview of our energy consumption. They, of course, tell us how many energy we spent in a certain period, but if we want to make our cities smarter and put architectures like this in real world applications the energy bill could: Tell how our home compares to our cluster average, and how it will compare in the future if we keep spending energy the way we did; Illustrate in a geographical map what is our home's energy ranking, using historical or predicted data; Show us how we compare to homes in the same state, on a weekly basis. These are only some of the possible examples.

Another perspective that this study provides on data is useful for energy providers and the government. In today's world, governments are having difficulties in managing resources, and energy is one of them. Fig. 11 shown a reinvented energy bill, while in Fig. 12 a new form of monitoring and targeting energy can be seen, with the ability to compare clusters average values and observe predicted changes in the clusters energy ranking.

Apart from that, the electricity and gas trend of consumption can also be analyzed on a weekly basis and an inter cluster comparison can be established, comparing each cluster consumption and their respective evolution from previous weeks.

Smart cities government can manage their resources much more easily, and energy providers can target and distribute energy based on a more panoramic view, such as the aggregation by clustering. Apart from that, they can predict how much energy will be needed during the next two months and how clusters ranking will change. All of these techniques lead them to a richer decision making process.



| | cluster1 | cluster2 | cluster3 | cluster4 | cluster5 |
|---|---|---|---|---|---|
| Avg. Electricity Cluster Avg | 187 kw/h | 281 kw/h | 242 kw/h | 174 kw/h | 166 kw/h |
| Avg. Gas Cluster Avg | 1 375 kw/h | 0 kw/h | 280 kw/h | 775 kw/h | 317 kw/h |

Fig. 12. New form of monitoring and targeting energy consumption.

## VI. Conclusion

This paper presented a Big Data analytics architecture, contemplating data storage, processing, mining and visualization.

The architecture presented in this paper is adequate to support the storage and processing of all data. However, this is a high-level presentation of the same, and some variables are not contemplated yet, such as the infrastructure to run the Hadoop cluster, the ways of extracting data for other applications like mobile apps or Open Data Platforms, and security concerns.

The data mining components of the architecture showed interesting results, since a rich clusters variety was achieved, as the different electricity and gas consumptions averages of each cluster demonstrated. Also, K-means algorithm was able to achieve a small intra-cluster similarity. Apart from this, such a low error rate on almost every tested forecaster was a successful outcome. Finally, the visual data analysis merged all the results in a refined user experience, in order to successfully validate the architecture and its underlying energy monitoring service.

For future work it is worth noting the variables that are not yet contemplated in this technological architecture, such as the infrastructure required to run the Hadoop cluster, integration with mobile apps or Open Data Platforms, and security concerns. Also the fact that this work does not consider the seasonality of the consumption and this should be integrated in future works. To accomplish that, a dataset with multiple years of consumption needs to be used.

## References

[1] I. Vilajosana, J. Llosa, B. Martinez, M. Domingo-Prieto, A. Angles, and X. Vilajosana, “Bootstrapping smart cities through a self-sustainable model based on big data flows,” *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 128–134, Jun. 2013.

[2] J. Hedlund, “The Smart City: Using IT to Make Cities More Livable.” Dec-2013.

[3] M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” *Mob. Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[4] K. Krishnan, *Data Warehousing in the Age of Big Data*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2013.

[5] P. Zikopoulos and C. Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, 1st ed. McGraw-Hill Osborne Media, 2011.

[6] J. Gama, *Knowledge Discovery from Data Streams*. 2010.

[7] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, “CRISP-DM 1.0: Step-by-step data mining guide.” 2000.

[8] N. Clark, “Commercial and Residential Hourly Load Profiles for all TMY3 Locations in the United States,” *Catalog*, 2013. .

[9] Y. Xie, H. Zheng, and L.-Z. Zhang, “Electricity price forecasting by clustering-LSSVM,” in *Power Engineering Conference, 2007. IPEC 2007. International*, 2007, pp. 697–702.

[10] H. Zhou, X. H. Wu, and X. G. Li, “An ANFIS model of electricity price forecasting based on subtractive clustering,” in *2011 IEEE Power and Energy Society General Meeting*, 2011, pp. 1–5.

[11] F. Azevedo and Z. A. Vale, “Forecasting Electricity Prices with Historical Statistical Information using Neural Networks and Clustering Techniques,” in *Power Systems Conference and Exposition, 2006. PSCE ’06. 2006 IEEE PES*, 2006, pp. 44–50.

[12] C. Alzate and M. Sinn, “Improved Electricity Load Forecasting via Kernel Spectral Clustering of Smart Meters,” in *2013 IEEE 13th International Conference on Data Mining (ICDM)*, 2013, pp. 943–948.

[13] Y.-D. Gu, J.-Z. Cheng, and Z.-Y. Wang, “An fuzzy forecasting algorithm for short term electricity loads based on partial clustering,” in *2011 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2011, vol. 4, pp. 1560–1565.

[14] D. Alahakoon and X. Yu, “Advanced analytics for harnessing the power of smart meter big data,” in *2013 IEEE International Workshop on Intelligent Energy Systems (IWIES)*, 2013, pp. 40–45.

[15] 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,” *Comput. Sci. Eng.*, vol. 15, no. 4, pp. 38–47, Jul. 2013.

[16] M. Arenas-Martínez, S. Herrero-Lopez, A. Sanchez, J. R. Williams, P. Roth, P. Hofmann, and A. Zeier, “A Comparative Study of Data Storage and Processing Architectures for the Smart Grid,” in *2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2010, pp. 285–290.

[17] M. Mayilvaganan and M. Sabitha, “A cloud-based architecture for Big-Data analytics in smart grid: A proposal,” in *2013 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2013, pp. 1–4.

[18] P. Chandarana and M. Vijayalakshmi, “Big Data analytics frameworks,” in *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, 2014, pp. 430–434.

[19] M. Paryasto, A. Alamsyah, B. Rahardjo, and Kuspriyanto, “Big-data security management issues,” in *2014 2nd International Conference on Information and Communication Technology (ICoICT)*, 2014, pp. 59–63.

[20] E. Mooi and M. Sarstedt, “Cluster Analysis,” in *A Concise Guide to Market Research*, Springer Berlin Heidelberg, 2011, pp. 237–284.

[21] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, “Constrained K-means Clustering with Background Knowledge,” in *In ICML*, 2001, pp. 577–584.

[22] S. Salvador and P. Chan, “Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms,” in *16th IEEE International Conference on Tools with Artificial Intelligence, 2004. ICTAI 2004*, 2004, pp. 576–584.

This paper was modified in February 28, 2016, to update supporting institutions.