Energy Signature Analysis: Knowledge at Your Fingertips

Energy signature analysis: knowledge at your fingertips

Andrea Acquaviva*, Daniele Apiletti*, Antonio Attanasio*†, Elena Baralis*, Lorenzo Bottaccioli*,
Federico Boni Castagnetti‡, Tania Cerquitelli*, Silvia Chiusano*, Enrico Macii*, Dario Martellacci‡, Edoardo Patti*
* Dipartimento di Automatica e Informatica, Politecnico di Torino, ITALY
Email: {name.surname}@polito.it
†Istituto Superiore Mario Boella, Torino, ITALY
Email: attanasio@ismb.it
‡IREN Energia Torino, ITALY
Email:{name.surname}@gruppoiren.it

***Abstract*—Energy efficiency and energy consumption awareness are a growing priority for many countries. Among the large variety of methods proposed by energy scientists and professionals to evaluate building energy consumption, a widely adopted approach is the energy signature. Since the energy data easily scale towards very large datasets, the problem of characterizing energy efficiency through the energy signature from these huge data collections becomes challenging. This paper presents a distributed system, named ESA, for the collection, storage, and analysis of a large amount of energy-related data to keep continuously informed users on their energy consumption and building performance. ESA exploits a Big Data approach to perform a scalable and distributed computation of the building energy signature, which is exploited to forecast the expected power consumption for given contextual conditions in a specific time period. ESA characterizes monitored buildings through direct indicators designed to (i) evaluate the efficient use of the heating system by comparing latest observations with past energy demand in the same conditions, (ii) rank the overall building performance with respect to nearby and similarly characterized buildings. Experimental results on real energy consumption data demonstrate the effectiveness and the efficiency of the proposed distributed system to provide actionable knowledge at user fingertips for actors interacting with ESA.**

I. Introduction

Energy efficiency is a growing policy priority for many countries around the world, as governments seek to reduce wasteful energy consumption and encourage the use of renewable sources. The International Energy Agency (IEA) has estimated that in terms of primary energy consumption, buildings represent roughly 40% of total final energy consumption in most countries. The amount of this energy used for heating and cooling systems is about 60% in the residential sector and 45 % in the service one [1].

Innovative systems should be designed to continuously monitor a smart city environment and provide all stakeholders the tools to improve energy efficiency. To characterize energy consumption different methods have been proposed in the literature by energy scientists and professionals. Among them a widely adopted energy indicator is the energy signature used to estimate the total heat loss coefficient of a building. However, providing a real time computation and a comparative analysis of the energy signature to different user profiles (e.g., energy analysts, consumers, users living in the building) definitely calls for big data approaches due to a large volume of energy related data collected in a smart city scenario.

In this work, we aim at studying the total heat loss coefficient of a building (K_{tot}) by analyzing the power supplied by the heating system with respect to the difference between the internal temperature of the building (T_{in}) and external temperature of the ambient (T_{ex}). The estimation of the K_{tot} value exploits a linear regression of average power samples at different time granularity levels with respect to the average difference between T_{in} and T_{ex}. Differently from previous studies that addressed the energy signature [2], [3] as a simulation methodology, we propose a distributed system, named ESA, to efficiently compute the performance of every building in a city in near real-time and keep continuously informed users (e.g., energy manager, people living in the building) on their energy consumption and building performance. We believe that energy awareness can be enhanced with detailed, actual and actionable data as the ones provided by ESA.

ESA collects data from smart meters deployed in thousands of buildings in a major Italian city by IREN [4], and a large variety of data from different web services (e.g., meteorological data [5], contextual and topological features of the monitored buildings [6], [7]). ESA also collects and analyzes indoor climate conditions by means of temperature sensors installed in a subset of the monitored buildings. Since the collected energy data easily scale towards very large datasets, ESA exploits a Big Data approach to perform a scalable and distributed computation to characterize building performance through two indicators based on the energy signature. Both indicators compute the expected power consumption for given contextual conditions in a specific time period. However, the first one, named *intra-building indicator*, evaluates the building performance by comparing latest observations with past energy demand in the same conditions. The second one, named *intra-building indicator*, ranks the overall building performance with respect to nearby and similarly characterized buildings. Experimental results on a large volume of real energy consumption data show the effectiveness of ESA to effectively characterize building performance, and its optimal scalability as well.

This paper is organized as follows. Section II introduces the energy signature method, while Section III presents the overview of the Energy Signature Analysis system. Sec-

tions IV, V, VI describe the main layers of the ESA system. Section VII discusses the experimental results obtained on real data. Section VIII reviews existing work, and Section IX draws conclusions and presents future developments of this work.

II. The Energy Signature

The *energy signature* is a world wide recognized method for the analysis of building energy consumption. This method was developed in the 80's by American government after the oil crisis, it has been introduced in the European regulatory framework (EN 156036:2008) and was recognized at Italian level in UNI (11300:2008). The energy signature method has been used in many studies to extract the total heat loss coefficient of a building [2], [3], [8]. The latter is recognised as an interesting key energy indicator [2], [9] of a building.

Specifically, the *total heat gain* in a building (denoted as Q_{tot}) is expressed as $Q_{tot} = Q_{loss} + Q_{dyn}$ where Q_{loss} represents the ventilation and thermal losses and Q_{dyn} is the heat dynamically stored or released by the building. The term Q_{loss} is expressed as $Q_{loss} = K_{tot} \cdot (T_{in} - T_{ex})$ where T_{in} is the *internal* temperature of the building and T_{ex} is the *external* temperature of the ambient, while K_{tot} is the *total heat loss coefficient* of the building. The term Q_{dyn} takes into account the dynamic of the building. Since Q_{dyn} is related to the thermal inertia of the building, the estimation of its value may be a complex task. Q_{dyn} is expressed as $Q_{dyn} = C \cdot \frac{\delta T}{\delta t}$ where C is the thermal mass of the building, representing the building capability to release or store heat. When the Q_{dyn} value is approximated to zero, the steady-state analysis of the building efficiency can be performed [2], [3]. Specifically, the dynamic contribution Q_{dyn} decreases when energy data are analyzed at coarse granularity (as monthly, weekly) [9]. Instead, these effects are emphasized when finely-grained data are analyzed (as every 15 minutes)[2].

The total heat gain Q_{tot} in a building can also be expressed based on the contribution of four terms as in $Q_{tot} = Q_h + Q_{el} + Q_p + Q_{sun}$ where Q_h is the power supplied by the heating system, while Q_{el}, Q_p, and Q_{sun} represent the heat gains due to electricity usage (Q_{el}), people presence (Q_p) and solar radiation (Q_{sun}), respectively. The influence of random variables (as occupancy, wind, solar gains) and the heat gains due to the electricity usage can be neglected when coarsely-grained data are analyzed [3]. In this case, terms Q_{el}, Q_p, and Q_{sun} can be approximated to zero.

This study focuses on the steady-state analysis of the building efficiency. Consequently, energy data are analyzed at different coarse granularities to neglect both the dynamic contribution Q_{dyn} of and the influence of random variables. Furthermore, K_{tot} estimation is normalized to a single unit of volume, i.e., W/m^3. It follows that the total heat gain Q_{tot} in the building is equal to the power supplied by the heating system per unit of volume and to the ventilation and thermal losses (i.e., $Q_{tot} = Q_h = Q_{loss} = K_{tot} \cdot (T_{in} - T_{ex})$). The linearity of the model has been evaluated as done in [8].

III. The ESA System

Figure 1 shows the overall architecture of the ESA system for monitoring and analyzing the building performance in terms of energy efficiency. We focus our study on energy data

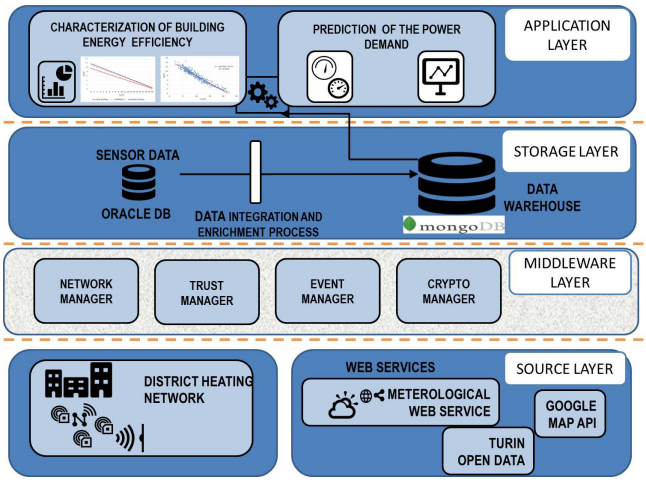

Fig. 1. The ESA system architecture.

collected by the *District Heating Network* (DHN) managed by IREN. IREN is a multi-utility company listed on the Italian Stock Exchange and operates in the sectors of electricity, thermal energy for district heating, gas, management of integrated water services as well as the collection and disposal of waste. As of March 2015, there are about 4 thousands monitored buildings by IREN in Turin, Italy, each generating about 2,000 data frames per day. Thus, ESA needs to manage a growing base of at least 8 million data frames per day.

As shown in Figure 1, the ESA architecture consists of four layers: (i) *Source Layer*, (ii) *Middleware Layer*, (iii) *Storage Layer*, and (iv) *Application Layer*. These layers are briefly described below and detailed in the following sections.

The *Source Layer* includes all the objects that provide source data to the ESA system such as smart meters and web services that countinuously provide data of interest. Energy consumption data are collected by means of a large number of smart meters (4,000 as of March 2015) deployed in Turin city (Italy) by IREN [4] to monitor thermal energy through the *District Heating Network*. Thermal smart meters collect data on different aspects of energy consumption (e.g., instantaneous power, cumulative energy consumption, water flows and corresponding temperatures) monitored roughly every 5 minutes. Furthermore, indoor climate conditions are monitored through indoor temperature sensors deployed in a subset of the monitored buildings. ESA exploits different web services to collect heterogeneous open data to feed a growing and enriched base of interesting data. Specifically, collected smart meter data are enriched with spatial and temporal information at different granularity levels as well as with various meteorological conditions available as open data sources.

Thermal energy data, indoor climate conditions and enriched data are managed by the *Storage Layer* component. It aims at (i) processing and storing all the collected smart meter data, (ii) enriching them with contextual and topological features related to the monitored buildings, and (iii) storing enriched and integrated data into a non-relational database to effectively support different complex analytics services.

The *Application Layer* component analyzes the collected data and produces useful feedbacks to the different users of the

ESA system. It addition it suggests ready-to-implement energy efficient actions or strategies. Different heterogeneous analytics services can be easily integrated in ESA. In this work we focus on evaluating and analyzing the building performance through the evaluation of energy signature. The analysis aims at (i) characterizing, evaluating and ranking building efficiency with respect to the nearby and similar buildings, and (ii) forecasting energy demand by means of the computation of the expected power consumption for a forecast outdoor temperature T_{ex} in a future time period. The first analysis is provided to both consumers and users living in the building, while the second one to both energy managers and energy analysts. Both analyses are performed in a near real-time fashion.

In this study, we focus on the layers described above. However, the *Middleware Layer* is also included in ESA. This layer enables the interoperability across heterogeneous data sources, both hardware and software, by creating a peer-to-peer network in which the communication between peers is trusted and encrypted. Thanks to the Event Manager unit, a publish/subscribe approach is also provided to increase the system scalability [10]. Such functionalities will support the management of an "energetic repository" for the city in which different actors (or peers) can publish and subscribe to energy data providers.

IV. Source layer

The Source Layer currently includes a large number of smart meters, installed in buildings and providing energy related data, and heterogeneous data retrieved from different web services (e.g, meteorological data provided by the Weather Underground web service [5]). ESA also collects indoor climate conditions by means of temperature sensors installed in a subset of the monitored buildings. However, any other data source can be easily integrated in ESA.

Remote measurements of energy consumption are collected by means of gateway boxes installed in the monitored buildings. Each gateway is an embedded device that manages all the sensors deployed in the building. It includes a GPRS modem with an embedded programmable ARM CPU. Each gateway has in charge the management of all the sensors deployed in its building. An ad-hoc software has been developed to execute the following activities: sensor management, GPRS communication, remote software update, data collection scheduling, and collected data sending to a remote server. Thermal energy is measured under different aspects, such as instantaneous power, cumulative energy consumption, water flows and corresponding temperatures. Furthermore, gateways also collect indoor temperature and the status of the heating system. Every 5 minutes, the gateways send a data frame to the Storage Layer (see Section V).

The Source Layer also collects the historical meteorological data from the Weather Underground web service [5], which gathers data from Personal Weather Stations (PWS) registered by users. For the city of Turin more than twenty PWS are available, reflecting with high accuracy the real conditions registered in monitored building neighborhood, as opposed to other services that provide estimated values with respect to a wider area.

Furthermore, ESA also retrieved contextual and topological features of the monitored buildings from different web services. Specifically, Google Maps APIs [6] have been exploited for geocoding building street addresses. Topological information about neighborhood names and district names have been retrieved from open data sources provided by the local public administration [7].

V. Storage layer

The *Storage Layer* provides a two-level database architecture for processing, storing, enriching data, and supporting complex analytics services. The first level, named *sensor data storage*, collects the sensor data continuously received from smart meters. Sensor data then are integrated with meteorological data and enriched with spatial and temporal information at different granularity levels. The enriched dataset is stored in the second level of the Storage Layer, named *data warehouse*, exploiting a non-relational schema-free horizontally-scalable database, MongoDB [11].

A. Sensor data storage

Gateways installed in buildings send the data frame including energy related data to the Storage Layer. Each data frame is assigned to one of four dispatchers to guarantee the system reliability. Each dispatcher delivers the frame to a cluster of computers including different processing servers where data are stored in a HDFS distributed file system. The dispatcher is able to recognize if the processing server has stored the frame correctly. In this case it sends an ACK to the gateway, which can then send the next data frame. Each processing server elaborates the received data and stores the result in an Oracle relational database. The logical model of the database includes the following three tables: (i) The *Building* table contains the main features characterizing each building such as address and volume; (ii) the *Sensor* table stores the list of sensors in each building and their characteristics (e.g., unit of measure, description, sensor type and model); (iii) the *History* table stores all the collected measurements.

B. Data enrichment and integration process

Data collected through the smart meters are integrated with meteorological information collected from the Weather Underground web service [5], which gathers data from Personal Weather Stations (PWS) registered by users. Each metereological measurement includes the air temperature (expressed in degree Celsius), the relative humidity (percentage), the precipitation level (mm), the wind speed (km/h) and the sea level atmospheric pressure (hPa). The date and time of each measurement is also included. For each PWS, we considered an average measurement frequency equal to 5 minutes. Collected weather data are pre-processed before the data integration phase because of different time and space intervals with respect to the energy-related timeline and the monitored building addresses. For example, weather data may be unavailable for a specific building address or instant of time, while the energy related data are instead available. The solution adopted in ESA supposes that (i) weather data timestamp is aligned to the closest timestamp available for the energy data. An approximated join is computed between weather data timestamp values and instantaneous power one. (ii) Weather

data associated with a specific building address are computed as a distance-based weighted mean of the values provided by the three nearest PWSs. The weight is inversely proportional to the distance from the PWS to the building address. Hence, three equally distant PWSs would have the same weight in determining the outdoor values of a given building.

Integrated data are enriched with additional contextual information acquired from external open data sources. More specifically, to analyze the *temporal distribution* of thermal energy/power, the following time granularities are considered: day, week, month, 2-month, 3-month, 6-month time periods. Moreover, each day is classified as holiday or working, and the measurement time is aggregated into the corresponding *daily time slot* (morning, afternoon, evening, or night). In Turin, heating systems are operated only from October 15th to April 14th, hence times periods outside this range were not considered.

To analyze the *spatial distribution* of thermal energy/power consumption, different space granularities are also considered beyond the building addresses. In addition, each *address* is mapped to the corresponding geographical *coordinates* (longitude and latitude degrees), *neighborhood*, and *city district* including that neighborhood. While the address is an information recorded for the monitored building, the geographical coordinates and both the neighborhood and district names corresponding to the address are added as additional contextual features to the repository. We exploited the Google Maps APIs [6] for geocoding street addresses. Furthermore, topological information about neighborhood names and districts are integrated in the repository as well. The latter have been retrieved from [7]. Topologies are used to graphically analyze the most significant spatial trends in thermal energy/power consumption data and were encoded in GeoJSON, which is a standard format for encoding a variety of geographic data structures.

C. Data warehouse

The data collection from smart meters exploits an Oracle database due to the fixed and constant nature of those measurements. Instead, being enriched data significantly more variable and heterogeneous, their analysis requires a different technological solution. To this aim, enriched data are modeled into a document-oriented distributed data warehouse providing rich queries, full indexing, data replication, horizontal scalability and a flexible aggregation framework, including a distributed map-reduce engine. The current database empowering ESA analytics is MongoDB [11]. As soon as new sensor-collected data are available (i.e., within seconds), they are integrated with meteorological information, enriched with topological and contextual data, and added into a MongoDB sharded collection.

Following best practices in data warehouse design, data are de-normalized and redundant information is added to each record (document) to speedup read performance by avoiding join operations (which are not sopported by MongoDB). It results in fast querying operations and energy building signature computation. The implemented data warehouse provides horizontal scalability and data replication to increase read-scalability. Horizontal scalability is obtained by exploiting data sharding, i.e., storing documents across multiple distributed machines by dividing the collection and distributing its data over multiple servers, or shards. As the size of the data increases, ESA only needs to add more machines to scale and support the demand of a higher number of read and write operations. Each shard processes relatively fewer operations as the cluster grows, and the percentage of data that each server needs to store is reduced. MongoDB provides automatic sharding and the key design choice is the attribute whose values partition the collection documents (i.e., the shard key). In ESA the sharding is performed using a hash-based partitioning on the value of the building ID field. The choice of the shard key is motivated by the fact that the energy signature analysis is typically computed by grouping measurements per building. Since the number of buildings grows with the expansion of the ESA framework, the shard key is a natural scaling indicator. Hash-based partitioning has been chosen over the range-based partitioning approach to ensure that data are evenly distributed across the machines in the cluster, since no range queries are performed on the building identifier. Replication is obtained by exploiting MongoDB replica sets to provide redundancy and high availability. With multiple copies of data on different servers, replication avoids data loss from a single server failure. Currently, in ESA each replica set consists of a primary server, a secondary server and an arbiter. All writes go to the primary server, while the secondary server can be exploited to increase the read capacity at the cost of possible inconsistencies, which are easily tolerated at the application layer.

VI. Application layer

The *Application layer* provides different services to the actors interacting with the ESA system. In this paper we focus on the specific energy signature service, which exploits a Big Data approach to perform a scalable and distributed computation of the total heat loss coefficient K_{tot} estimation, and whose aim is twofold: to evaluate and rank building efficiency/performance over time, and to forecast the power demand.

To address the former objective, two indicators have been designed: (i) an intra-building indicator, which addresses the question of abnormal power consumptions given the current conditions with respect to past energy demand in the same conditions; to this aim, the most recent power consumption data for each building is compared to its own historical energy signature, thus identifying changes with respect to previously modeled energy behaviors of the same building; (ii) an inter-building indicator, comparing the building efficiency, given by its energy signature, with respect to nearby and similarly characterized buildings, where similarity takes into account spatial co-location, building size, and usage patterns, e.g., residential or office or public building.

The key intuition behind the designed indicators is based on exploiting the energy signature defined by K_{tot} to compute the expected power consumption for given contextual conditions in a specific time period. Contextual conditions can include any relevant attribute for the specific problem under investigation. In the current implementation, the difference between the outdoor T_{ex} and the indoor T_{in} temperatures, and the specific building characteristics (e.g., position, size, etc.) are considered as the key attributes defining a context. If the

given temperatures and time periods are the current ones (e.g., current outdoor temperature, now), and we consider the same building, then the intra-indicator is obtained, whereas using the energy signature of a group of similar buildings, with respect to the one under examination, leads to the inter-building indicator.

Finally, to reach the goal of forecasting the power demand, the same approach can be used, by exploiting the energy signature with a predicted value of outdoor temperature T_{ex} and a fixed value of target indoor temperature T_{in}, with the former obtained by weather forecasts, and repeating the computation for each future time period and each building of interest. Such estimation of future power demand helps district heating providers to better predict the energy demand.

To evaluate and rank building efficiency, its energy signature, defined by its total heat loss coefficient K_{tot}, is exploited. To this aim, the instantaneous power supplied by the heating system per unit of volume (Q_h) is correlated with the difference between the indoor temperature T_{in} and the outdoor temperature T_{ex}. The correlation is based on a linear regression of average power samples per unit of volume, aggregated at different time granularity levels. This process has been designed and developed as a cloud-based service on top of a MongoDB distributed cluster, and it is detailed in the following.

The analysis can be focused by filtering heating power consumption in a given date range t_{period} (e.g., a winter period, a month) and also in specific day time slots of interest t_{slot} (e.g., [5:00p.m-7:00p.m.], [10:00a.m.-7:00p.m.], [10:00a.m.-9:00p.m.]). Hence the time-specific energy signature will be relevant only for those subsets of time periods, both in the characterization and in the prediction applications. Focusing the energy signature by restricting the day and time periods helps in modeling different behaviors such as those in the steady state, in specific seasons, during office hours, etc.

The instantaneous power samples of interest are aggregated by computing the mean value in a given time window, t_{window} (e.g., hourly, daily, weekly). The resulting value indicates the mean power consumption over an hour, a day, or a week. While longer periods are more error-prone due to the large variance of the outdoor temperature T_{ex}, too short periods take into account the thermal inertia of the building, as discussed in Section II.

The application service can estimate K_{tot} by considering any combination of t_{period}, t_{slot}, and t_{window}, which are user-defined parameters. It will be up to the end-user presentation interface to choose the best indicators in any given context.

For each building, the instantaneous power values per unit of volume, and the difference between the indoor T_{in} and the outdoor T_{ex} temperatures are extracted from the MongoDB datawarehouse and aggregated over t_{window}. The result includes all the mean power values per unit of volume and the average difference T_{ex}-T_{in} for each t_{window}.

Given the mean power values (denoted as y) and the mean temperature difference values (denoted as x) we first compute $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$. Then the a and b terms of the linear equation $y = a + bx$ is computed as follows.

$$a = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

The b value corresponds to the total heat loss coefficient of a building, i.e., K_{tot}.

The MapReduce jobs of the estimation of K_{tot} were developed through custom JavaScript functions, and executed using the MongoDB MapReduce framework. For each document, the *map* function emits an object containing the values needed to compute the energy signature equation parameters for the related building. The *reduce* function is in a simple sum over all the records of the same building. Finally, a *finalize* function uses the aggregated results to compute the energy signature equation and returns the K_{tot} estimation for each building. Energy signatures can be aggregated over similar buildings by computing the average K_{tot} among them.

VII. Experimental results

We performed experiments on real data collected through sensors and smart meters from almost 2,000 buildings in Turin, Italy, over three years (2012-2014), with samples every 5 minutes. Energy data include the instantaneous power measured at sampling time in Watt (W) and the consumption since the last sample in Watt-hour (Wh). Indoor temperature values in Celsius degrees ($^{\circ}C$) are also fetched from the sensor data storage and synced with energy data. If different sensors for a single measurement are present (e.g., indoor temperature), an average value is considered for the whole building. The dataset has been integrated and enriched as discussed in Section V-B and loaded into the MongoDB datawarehouse. The complete collection in the MongoDB cluster contains the subset of data which have been used for the reported experiments, whose current total size is almost 300 GB.

Experiments address three issues: (i) the characterization of energy signatures for different buildings at different granularity levels (Section VII-A), (ii) the sensitivity and robustness of the energy signature method (Section VII-B), (iii) the horizontal scalability of the ESA analytic system with respect to the number of nodes in the cluster (Section VII-C).

Experiments were performed on a cluster of 8 nodes running MongoDB version 2.6.8 and configured as a sharded cluster, consisting of three different components. The nodes were assigned to each component as follows: (i) Up to 5 dedicated nodes (node4 to node8) were configured as the actual shards in charge of the data storage. (ii) One node (node2) was configured as *query router* (mongos) and were in charge of directing operations to the appropriate shards. (iii) Three nodes (node1 to node3) were configured as *Config servers* (mongod –configsvr) to store the cluster's metadata, such as the mapping of the data set to the shards. To efficiently split documents among shards, the building ID property was selected as shard key. Each cluster node is a 2.67 GHz six-core Intel(R) Xeon(R) X5650 machine with 32 Gbyte of main memory running Ubuntu 12.04 server with the 3.5.0-23 kernel. All the reported execution times are real times.

To evaluate the quality of the linear regression that estimates K_{tot}, the Standard Error of Regression (denoted as S) is exploited:

$$S = \sqrt{\frac{1}{(n-2)}[\sum (y - \bar{y})^2 - \frac{[\sum (x-\bar{x})(y-\bar{y})]^2}{\sum (x-\bar{x})^2}]}$$

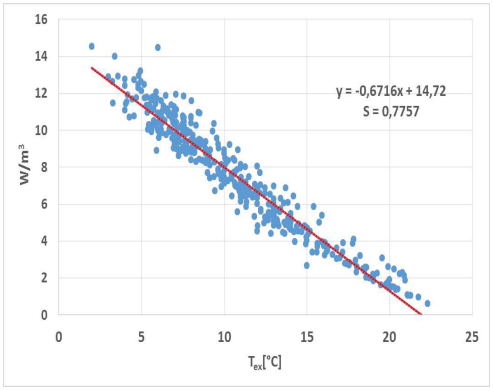

Fig. 2. Residential building, scatter plot of daily power consumption per unit of volume (W/m^3) with respect to T_{ex} ($^\circ C$).

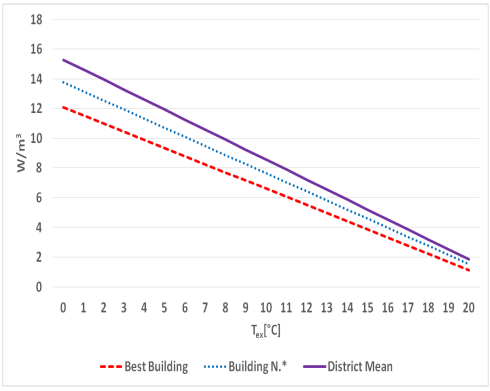

Fig. 3. Residential building, linear regression of daily power consumption per unit of volume (W/m^3) with respect to T_{ex} ($^\circ C$).

where $\bar{x}$ and $\bar{y}$ are the sample means, and n is the sample size. Small values of S identify a high accuracy of prediction because the 95% of predicted values will fall in a range of $\pm 2S$.

A. Characterization of energy signatures

Figures 2 and 3 show the energy signature of a random building in the Turin area. The chosen t_{window} values are 24 hours and 7 days, hence the analysis considers the daily mean power values per unit of volume with respect to the daily mean outdoor temperature[1]. The analysis has been performed by considering as t_{period} the latest full Italian heating season (at the time of writing), from October 15th, 2013 to April 14th, 2014. To focus the analysis on the steady state, t_{slot} has been set to the time range from 5 to 9 pm.

Figure 2 focuses on the daily t_{window} scatter plot and its resulting regression (red line) to estimate K_{tot}. A low S value of 0.78 is obtained, whereas the estimated value of K_{tot} is 0.67, which is a good result in terms of energy performance, as we will see later in the experiments. Even without knowing which ranges of K_{tot} are good or bad, and we cannot suppose all actors of the ESA system will be so skilled, providing a performance comparison with similar buildings helps in defining a perspective at a glance.

Figure 3 shows the linear regression by aggregating mean power values over daily t_{window} for the considered building (dotted line). Figure 3 also shows the energy efficiency of the considered building (dotted line) with respect to the energy signature of (i) the most efficient building (dashed line) and (ii) the average power profile (solid line), considering all buildings in the corresponding district. Such comparative information, suitably presented to each actor of the ESA system, allows to rank the buildings within districts, immediately putting in perspective the initial value of K_{tot}=0.67: even if it is a generally good value in terms of energy efficiency, as we will see later, being better than its district average, the best performing building in the same district is far better. In such a situation, an end-user can consider to adopt energy-aware structural improvements to reach and improve over the best performing building.

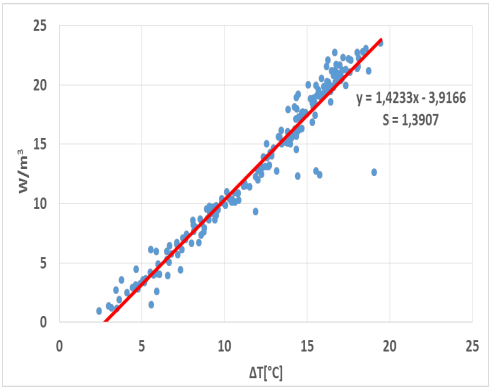

Fig. 4. School building, scatter plot of daily power consumption per unit of volume (W/m^3) with respect to T_{in}-T_{ex} difference ($^\circ C$).

Figure 4 shows the daily power consumption per unit of volume of a Turin school building, where a sensor network has been deployed for real-time monitoring of indoor temperature. The analysis has been performed by considering t_{period} from the beginning of the heating season 2013-2014 in Turin, on October 15th, 2013, to the latest data available at the time of writing, February 28th, 2015. Since many indoor temperature sensors are deployed in different rooms of the school building, the instantaneous indoor temperature T_{in} has been computed as the mean temperature value for each timestamp. The analysis has been performed by considering values in t_{slot}=$[5:00p.m. - 9:00p.m]$ and data are aggregated over daily t_{window}. Similarly to the regression of a residential building presented in Figure 2, also the school building K_{tot} estimation (red line) from the linear regression has a low S value (1.39). The estimated K_{tot} value is 1.42, indicating a poor energy performance, at least when compared with residential buildings.

B. Sensitivity and robustness of the energy signature method

The two main user-defined parameters of the ESA system are the aggregation period t_{window} and the t_{slot} day-time filter. Hence, we present an evaluation of the robustness of the proposed energy signature analysis to such parameter settings.

[1]In most residential buildings, the indoor temperature is not monitored through a sensor network, hence we considered in the analysis a fixed value of 20$^\circ C$, since it is the typical value set by local regulations. Being fixed T_{in}, the charts report T_{ex} only.

Table I shows the K_{tot} estimation computed through the linear regression and the S values indicating how the linear regression is able to correctly model the studied phenomenon. Furthermore, the comparison between the best and the average building allows to see the results from a different perspective.

Focusing on S, the first evidence is that the longer the t_{window}, the better the linear regression. For each t_{slot}, S values are lower for weekly t_{window} than daily t_{window}. S values for hourly t_{window} are much higher than the rest. This general behavior is expected and stems from the data smoothing effect of considering averages over longer periods of time, which hides outliers or temporary exceptional behaviors. Even if the equation $\forall t_{slot}, t_1 \geq t_2 \Rightarrow S_{t_{window}=t_1} \geq S_{t_{window}=t_2}$ holds true in all reported cases but one, we can see that the S values for hourly t_{window} are more sensitive to the t_{slot} selection. In particular, both the 5:00-7:00pm and the 5:00-9:00pm t_{slot} ranges yields similar results in terms of S values (of course, periods also overlap), whereas the 6:00am-10:00pm t_{slot} range for hourly t_{window} has extremely high S values: 4.61 for the best building, 7.41 for a random building. From such values, we can note that (i) the 6:00am-10:00pm t_{slot} range is generally the best fit for the linear regression, thanks to the longer period facilitating steady state modeling of the heating system and limiting the dynamic and thermal inertia effects; (ii) the hourly t_{window} often leads to unsatisfactory K_{tot} estimations, due to the poor fit of the linear regression. The exception to these findings is the 5:00-7:00pm t_{slot} for hourly t_{window}, which has a low S value (0.88) with respect to the average hourly model behavior (S always above 1.11). Finally, we can note that all combinations of parameters that have a low S (from Table I, lower than 1.2) lead to a coherent and stable K_{tot} estimation: for each fixed t_{slot}, the best building K_{tot} estimation delta is always lower than 0.02, and the random building is always lower than 0.01.

		Best Building		Building N.*		District Mean
t_{window}	t_{slot}	K_{tot}	S	K_{tot}	S	K_{tot}
Weekly	6:00am-10:00pm	0.46	0.35	0.53	0.47	0.55
Weekly	5:00-7:00pm	0.51	0.67	0.72	0.55	0.74
Weekly	5:00-9:00pm	0.54	0.57	0.68	0.68	0.68
Daily	6:00am-10:00pm	0.46	0.64	0.53	0.55	0.54
Daily	5:00-7:00pm	0.53	1.02	0.72	0.90	0.73
Daily	5:00-9:00pm	0.55	0.62	0.67	0.77	0.67
Hourly	6:00am-10:00pm	0.36	4.61	0.49	7.41	0.51
Hourly	5:00-7:00pm	0.52	1.16	0.71	0.88	0.73
Hourly	5:00-9:00pm	0.53	1.11	0.64	2.38	0.64

TABLE I. t_{window} AND t_{slot} SENSITIVITY AND ROBUSTNESS.

C. Performance evaluation

We evaluated the scalability of the proposed architecture by measuring the speedup achieved for different numbers of shards in the MongoDB cluster (from 1 to 5 nodes). The MongoDB *chunk size* parameter that determines the sharded data balance among the nodes was left to its default value of 64 MB, being already more than three orders of magnitude smaller than the total data collection size of almost 300 GB, and thus leading to finely-grained balanced shards. The chosen shard key for the experiments is the *Building ID* field.

Figure 5 reports the speedup achieved with the 2,000 building data set. The black line represents the (positive side of the) ramp function, under which the computing speedup would be identical to the number of shards or, equivalently,

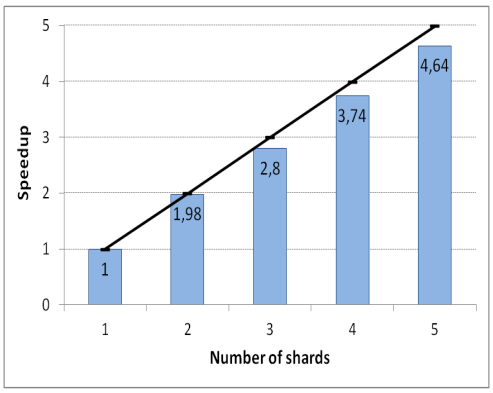

Fig. 5. Speedup on a 300GB-sized dataset (2,000 buildings).

the computing time for a cluster would be equal to that employed by a single shard divided by the number of shards in the cluster. The achieved results show that the algorithm scales roughly linearly with the number of nodes and the speedup approximately corresponds to the number of cluster nodes. From a design perspective, we track the almost optimal speedup back to the choice of the sharding key: it keeps data locality among map, reduce, and re-reduce iterations, since most of the energy signature computation is on a per building basis, and only district and city averages or ranks involve different nodes (but at that point the data set is so aggregated that it is extremely small).

VIII. Related Work

Important research activities have been carried out to use database management systems, exploratory data mining techniques, and statistical tools in the field of storage and analysis of energy data to evaluate the efficiency of buildings. The proliferation of sensor networks for monitoring indoor and outdoor environmental parameters has brought to the facility managers huge archives of measures with temporal and spatial references. Research contributions on these large data volumes have been carried out for: (i) supporting data visualization and warning notification [12]; (ii) efficient storing and retrieval operations based on NoSQL databases [13]; (iii) characterizing consumption profiles among different users [14], [15]; (iv) identifying the main factors that increase energy consumption (e.g., floors and room orientation [16], location [15]). Differently from the above research works, this paper proposes an integrated and distributed system able to collect a large volume of energy related data and efficiently compute two key indicators based on the energy signature method.

A parallel effort has been devoted to designing and implementing systems based on Big Data technologies to provide different cloud-based analytics services. Proposed solutions are general purpose [17] or tailored to a given application domain, such as thermal energy consumption [18], residential energy use [19], renewable energy [20], air pollution levels [21]. Authors in [17] highlight the key features that should be included in an analytics cloud service. Thereafter, the conceptual architecture of a big data analytics service provisioning platform in the cloud (CLAaaS) is presented. The platform will provide on demand data storage and analytics services through customized user interfaces and will apply Service

Level Agreements (SLAs) to provide controlled access to software and data resources. The components of the proposed platform are grouped into three functional categories: service management, workflow management and data management. In this paper we presented a distributed architecture similar to [17] but tailored to the energy signature service. The implemented system has been also evaluated on real data.

The work in [18] tries to point out the key features of an Energy Management System, to support frequent pattern discovering on event streams. A Data Stream Management System (DSMS) is used, to better suit the typical queries of real-time EMSs on time-varying data streams. Differently from [18] the ESA system exploits a non-relational schema-free database to efficiently support different and more complex energy analytics services.

Finally, a small subset of the experimental dataset for ESA has already been exploited by authors in [22]. Such previous work has completely different target and analysis approach, and a substantially different architecture (the only similarity lays in the datawarehouse design). [22] exploits four levels of data (energy, publication, social, and smart data net), whereas the current paper has only two data levels; the target of [22] is the power consumption analysis, whereas the current work aims at energy efficiency.

IX. Conclusions

This paper presented the design and implementation of the ESA system to provide different energy analytic services exploiting a Big Data approach. The three main layers of ESA, i.e., source, storage and application layers, have been thoroughly presented and discussed in the context of the specific energy signature service. Experimental results on real data show the effectiveness and the efficiency of the ESA system in exploiting the energy signature analytic service to evaluate and rank building efficiency and energy performance over time, and to forecast the power demand.

We are currently extending the ESA system with an ad-hoc social platform where users are pro-actively engaged in the act of generating data related to their perception of thermal comfort, as well as useful feedbacks on thermal energy consumption of the buildings where they live or work. The social platform will also show to users both inter and intra-building indicators in an informative fashion.

Acknowledgments

The research leading to these results has partially received funding from the Piedmont Region under the POR FESR 2007/2013 n. 281-79 (EDEN Project).

References

[1] IEA, "Energy efficiency indicators," 2014. [Online]. Available: /content/book/9789264215672-en

[2] L. Belussi and L. Danza, "Method for the prediction of malfunctions of buildings through real energy consumption analysis: Holistic and multidisciplinary approach of energy signature," *Energy and Buildings*, vol. 55, pp. 715–720, 2012.

[3] J. Vesterberg, S. Andersson, and T. Olofsson, "Robustness of a regression approach, aimed for calibration of whole building energy simulation tools," *Energy and Buildings*, vol. 81, no. 0, pp. 430 – 434, 2014.

[4] IREN website, *http://www.gruppoiren.it/index.asp. Last access on March 2015*.

[5] Weather Underground, *Weather Underground web service. Available at http://www.wunderground.com/ Last access: March 2015*.

[6] Google Maps, *Available at http:/maps.google.it Last access: March 2015*.

[7] Turin GeoPortal, *Available at http://www.comune.torino.it/geoportale/ Last access: March 2015*.

[8] C. Ghiaus, "Experimental estimation of building energy performance by robust regression," *Energy and buildings*, vol. 38, no. 6, pp. 582–587, 2006.

[9] S. Danov, J. Carbonell, J. Cipriano, and J. Martí-Herrero, "Approaches to evaluate building energy performance from daily consumption data considering dynamic and solar gain effects," *Energy and Buildings*, vol. 57, pp. 110–118, 2013.

[10] E. Patti, A. Acquaviva, M. Jahn, F. Pramudianto, R. Tomasi, D. Rabourdin, J. Virgone, and E. Macii, "Event-driven user-centric middleware for energy-efficient buildings and public spaces," *IEEE Systems Journal*, 2014.

[11] K. Chodorow and M. Dirolf, *MongoDB: the definitive guide*. O'Reilly Media, 2010.

[12] D. Wijayasekara, O. Linda, M. Manic, and C. Rieger, "Mining building energy management system data using fuzzy anomaly detection and linguistic descriptions," *Industrial Informatics, IEEE Transactions on*, vol. 10, no. 3, pp. 1829–1840, Aug 2014.

[13] J. van der Veen, B. van der Waaij, and R. Meijer, "Sensor data storage performance: Sql or nosql, physical or virtual," in *Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on*, June 2012, pp. 431–438.

[14] O. Ardakanian, N. Koochakzadeh, R. P. Singh, L. Golab, and S. Keshav, "Computing electricity consumption profiles from household smart meter data," in *EDBT/ICDT Workshops'14*, 2014, pp. 140–147.

[15] S. Depuru, L. Wang, V. Devabhaktuni, and P. Nelapati, "A hybrid neural network model and encoding technique for enhanced classification of energy consumption data," in *Power and Energy Society General Meeting, 2011 IEEE*, July 2011, pp. 1–8.

[16] C. Filippn and S. F. Larsen, "Analysis of energy consumption patterns in multi-family housing in a moderate cold climate," *Energy Policy*, vol. 37, no. 9, pp. 3489 – 3501, 2009, new Zealand Energy Strategy.

[17] F. H. Zulkernine, P. Martin, Y. Zou, M. Bauer, F. Gwadry-Sridhar, and A. Aboulnaga, "Towards cloud-based analytics-as-a-service (claaas) for big data analytics in the cloud," in *IEEE International Congress on Big Data, BigData Congress 2013, June 27 2013-July 2, 2013*, 2013, pp. 62–69.

[18] D. Anjos, P. Carreira, and A. P. Francisco, "Real-time integration of building energy data," in *2014 IEEE International Congress on Big Data, Anchorage, AK, USA, June 27 - July 2, 2014*, 2014, pp. 250–257.

[19] C. Wang, M. de Groot, and P. Marendy, "A service-oriented system for optimizing residential energy use," in *IEEE International Conference on Web Services, ICWS 2009, Los Angeles, CA, USA, 6-10 July 2009*. IEEE, 2009, pp. 735–742.

[20] S. Lu, Y. Liu, and D. Meng, "Towards a collaborative simulation platform for renewable energy systems," in *IEEE Ninth World Congress on Services, SERVICES 2013, Santa Clara, CA, USA, June 28 - July 3, 2013*. IEEE Computer Society, 2013, pp. 9–12.

[21] L. G. Rios and J. A. I. Diguez, "Big data infrastructure for analyzing data generated by wireless sensor networks," in *2014 IEEE International Congress on Big Data, Anchorage, AK, USA, June 27 - July 2, 2014*, 2014, pp. 816–823.

[22] A. Acquaviva, D. Apiletti, A. Attanasio, E. Baralis, F. B. Castagnetti, T. Cerquitelli, S. Chiusano, E. Macii, D. Martellacci, and E. Patti, "Enhancing energy awareness through the analysis of thermal energy consumption," in *Proceedings of the Workshops of the EDBT/ICDT 2015*, ser. CEUR Workshop Proceedings, P. M. Fischer, G. Alonso, M. Arenas, and F. Geerts, Eds., vol. 1330. CEUR-WS.org, 2015, pp. 64–71.