Big Data Architecture for Building Energy Management Systems

M. Dolores Ruiz, Juan Gómez-Romero, Carlos Fernandez-Basso, and Maria J. Martin-Bautista

***Abstract*—The enormous quantity of data handled by building management systems are key to develop more efficient energy operational systems. However, the inability of current systems to take benefit from the generated data may waste good opportunities of improving building performance. Big Data appears as a suitable framework to sustain the management system and conduct future prospective analysis. In this article, we present a Big Data-based architecture for the efficient management of buildings. The different Big Data components are involved not only in the data acquisition phase, but also in the implementation of algorithms capable of analyzing massive data collected from very heterogeneous sources. They also enable fast computations that can help the generation of optimal operational plan generations to improve the building functioning. The proposed architecture has been effectively introduced in four different-purpose buildings, demonstrating that Big Data can help during the energy cycle of the building.**

***Index Terms*—Building energy management system (BEMS), Big Data, distributed computing, energy building.**

I. Introduction

BUILDING management systems (BMS) are computer-based systems controlling and monitoring the mechanical and electrical equipment of a building, like the heating, ventilation, and air conditioning (HVac), lighting, and power systems, although some other systems like fire or security control can be also included. From this arises the term Building energy management system (BEMS) to denote the components related to the energy management in the BMS. Typically, a BEMS comprises several software and hardware systems, which are usually very heterogeneous in purpose and implementation. These systems enable not only the management of several aspects of the building such as the scheduling of actions and real-time controls based on the building status, but also to be aware of the evolution of the building according to energy parameters. Then, the BEMS is key to efficiently manage building functioning attending to reducing operation costs and its energy consumption. In fact, it has been estimated that the systems controlled by the BEMS are responsible up to 40% of its energy consumption [1]. Therefore, improperly configured BEMS can be responsible for huge energy wastes. Even a small improvement in the control system can potentially save thousands of Euros a year and significantly reduce the CO_2 emissions. Thus, it is paramount to achieve the objective of energy reduction by improving the management of buildings. In fact, in the literature, different architectures and strategies have been proposed to improve energy consumption from different perspectives (see for instance [2] and [3]).

BEMS usually collects data about several parameters including both indoor and outdoor measurements. There is an increasing fashion to collect heterogeneous types of data that can be further employed to accurately plan and anticipate the building necessities, by exploiting, for instance, automatic control capabilities. However, the heterogeneity of buildings and their designs make it difficult to converge to an unique architecture working for the majority of already existing and implemented BMS in order to improve and optimize their energy savings and to identify the common trends and patterns that could lead to better operation plans.[1]

Thus, this article proposes a BEMS architecture based on the Big Data paradigm to efficiently manage the general daily operations of big facilities. This approach is divided into three different layers enabling the three key aspects in the operational phase of the building: data acquisition, control unit, and data analysis layers. The paradigm behind the Big Data is based on distributed programming using MapReduce functions. Platforms like Hadoop [4] or Spark [5] are available to develop Big Data algorithms and their libraries are getting larger in recent years offering a broad catalogue of data mining and machine learning (DM&ML) algorithms that can be used to discover meaningful information from massive data collections. The building energy field has been aware of this and there are more and more works using data science algorithms for facilitating the extraction of information [6].

Particularly, the aim of this article is threefold. First, to revise the existent literature about the use of Big Data techniques in the ambit of energy management systems (EMS). Second, to propose a Big Data-based architecture for BEMS. Lastly,

Manuscript received October 9, 2020; revised October 13, 2021; accepted November 17, 2021. Date of publication November 23, 2021; date of current version June 13, 2022. The work was supported in part by the Spanish Ministry of Science, Innovation, and Universities under Grant TIN2017-91223-EXP, in part by the FEDER Programme 2014–2020, in part by the Andalusian Regional Government under Grant A-TIC-244-UGR20, and in part by the European Union (Energy IN TIME EeB.NMP.2013-4) under Grant 608981. Paper no. TII-20-4705. *(Corresponding author: M. Dolores Ruiz.)*

The authors are with the Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain (e-mail: mdruiz@decsai.ugr.es; jgomez@decsai.ugr.es; cjferba@decsai.ugr.es; mbautis@decsai.ugr.es).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2021.3130052.

Digital Object Identifier 10.1109/TII.2021.3130052

[1]An operational plan (OP) is understood as a complete sequence of instructions (or setpoints) for the building equipment.

to implement and validate the architecture proposed in four different pilot sites in multipurpose buildings: a commercial, an office, a hotel, and an airport. All the development was made in the context of the European FP7 project Energy IN TIME aimed at developing a simulated environment system to efficiently manage the operation and maintenance of buildings.

The rest of this article is organized as follows. Section II reviews the existing EMS proposals focusing on the role of Big Data. Section III describes a catalog of Big Data technologies that can be employed in the design of a BEMS architecture. Section IV presents our proposal for a BEMS architecture based on Big Data. Section V describes how the system was implemented for the pilot site buildings showing the potential benefits of the proposed architecture. Finally, Section VII concludes this article.

II. Existing EMS Proposals and the Role of Big Data

EMS can be considered as the heart of the building processing since they are in charge of the management and orchestration of different building components. Among all the available EMS, we have distinguished three major groups.

A. Building Energy Management Systems

The first group we want to highlight is that of incorporating Big Data into the BEMS. Some theoretical proposals can be found in the literature but without real implementation in buildings. In [7], an EMS architecture is proposed to take advantage of the Internet of Things and Big Data technologies to allow personalization of processing mechanisms in order to adapt to users' behaviors. Ahad and Biswas [8] proposed a theoretical architecture for energy efficiency and secure handling (via a two-fish encryption technique) of massive data integrated in a distributed system. A level-based architecture is proposed in [9] for managing big amounts of data from several data sources using smart grid applications that can be accessed on request of processing components.

Another proposal using Big Data, implemented at the Smart Living Lab, Fribourg, Switzerland, is proposed in [10], where a system architecture called big building data is presented to enable the retrieval of historical data, visualization, and processing. The system is based on the concept of *virtual objects* and employs Apache Kafka for message queue processing of these objects and Apache Flink for a latter processing of these records.[2] Regarding data storage, HDFS format is employed for raw data and MySQL for virtual objects metadata.

B. Home Energy Management Systems

The second group of systems, called home energy management systems (HEMS), are those, as the name indicates, in charge of managing house systems, mainly focused on creating optimal energy schedules adapted for consumers considering their profile and comfort, energy costs, environmental factors, etc. Although there are an abundant number of proposals for HEMS, the inclusion of Big Data techniques may not seem indispensable in such "small" systems.

[2]More information about Apache Kafka and Flink is presented in Section III.

However, the prototype study made in [11] for processing and analyzing the smart home data coming from HVac systems simulating a residential area, and the review made in [12] let us think that there are some avenues to also employ Big Data technologies in HEMS. In particular, in [11], a Big Data analytic software plus off-the-shelf business intelligence is employed to manage energy consumption and meet consumer demand. Beaudin and Zareipour [12] enumerated different proposals for HEMS focusing on the modeling approaches and their impact on HEMS. One of the main problems highlighted in this review are the limited computational capacities of household settings. In this regard, Big Data cloud computations may be a good solution to overcome this challenge.

C. Smart Grids Based Approaches

Smart grids can be seen as intelligent grids able to store, communicate, and make decisions [13]. Thus, they can be viewed as a next-generation information and management technology to enhance efficient and secure management of energy in buildings.

The role of Big Data in smart grids has arisen as necessary to manage the high volume of data generated by the smart buildings. Mayilvaganan and Sabitha [14] proposed a cloud-based architecture using Big Data for the pattern analysis of customers, although the proposal was not implanted in any building. A Spark platform was implemented in [15] for Big Data analysis combining batch and real-time data processing techniques for smart grid. Another proposal of Big Data analytic for smart grid was applied in [16] on two scenarios: a single-house and a smart grid to visualize the grid status and enable dynamic demand response with very promising results. Jaradat *et al.* [17] gave some recommendations and practices to explore the different applications of smart sensor networks managed and processed by Big Data technologies in the domain of smart power grid. In [18], an efficient three-factor user authentication scheme for a renewable energy-based smart grid environment is proposed. In [19], a review is also made from the perspective of analytic tools that Big Data offers for smart grids data analysis.

D. Discussion

It has been seen that the energy sector is currently aware of the advantages that Big Data technologies can offer when managing massive data coming from different management systems, although in the HEMS case it seems to be not necessary. Many of the works in the literature propose, but do not implement, those Big Data-based systems, so the real impact of using these technologies, how they behave, and how they can be integrated in a real EMS cannot be tested. In this work, we first enumerate the wide variety of Big Data technologies that can facilitate some parts of the EMS, and then we propose a BEMS architecture based on Big Data that has been applied in four different-purpose buildings, illustrating some of the Big Data applications that can be extended to other scenarios.

The impact of Big Data will be fully explored in the following sections from the perspective of enabling Big Data technologies that may help in the implementation of a BEMS architecture based on Big Data. In addition, there are other aspects that should be mentioned such as the scientific contributions of Big

Data when using different data sources (we refer to the analysis made in [20] or to the review made in [6]), how to handle multimedia data [21], or how to implement artificial intelligence tools in a Big Data architecture [22]. In addition, the Big Data framework can benefit from other enabling technologies such as the improved 5G communication structures to speed up the communications among the different components of the system while reducing the battery consumption. Particularly interesting is the transition from cloud computing to edge computing, consisting in moving the systems and software as nearer as possible to the place of use, gaining a reduction in time responses and the involved communication latency.

III. Enabling Big Data Technologies

Big Data offers a wide catalog of resources and technologies that can facilitate the management of the energy system. The main challenge is therefore to model an appropriate architecture taking into account the initial requirements of the energy system and the capabilities of the technologies employed to implement it. Given the wide variety of Big Data products, we have to think which of them are more adequate in the long term, whether they are open-source or proprietary software, if the product is mature enough, their security options, etc.

For instance, in household settings, Big Data cloud computations may be a good solution to increase the computational capacities without worrying about the system deployment or its maintenance. However, in big facilities, the building owners are more reticent to store their data in the cloud. In addition, some other challenges appear while managing different heterogeneous data that should be efficiently stored and accessed by different components of the BEMS.

In the following, we highlight some Big Data technologies classifying them attending to different aspects that should be taken into account when designing a BEMS architecture based on Big Data.

1) *Massive storage:* Big Data tools rely on NoSQL databases for storing data (for instance, HDFS, MongoDB, etc.). Compared to relational databases, NoSQL databases are more scalable and provide superior performance accelerating data access and storage since the data structure does not need to be defined in advance.
2) *Big data processing:* The main paradigm behind Big Data is known as MapReduce [23], a programming model for distributed computing. Its main advantage is its fault-tolerance achieved due to the distribution of computation in a redundant way, facilitating the deployment of algorithms to programmers. MapReduce computation is available in platforms such as Hadoop [4] or Spark [5].
3) *Streaming Big Data processing:* Streaming is meant for processing continuous data flows coming in real time during a period of time. Among the available stream processing platforms, we can distinguish those that enable a continuous running of processes, i.e., every time new data arrive, it can be immediately processed. In this group, we can enumerate Apache Storm, Flink, Kafka Streams, or Samza [24]. The other group, known as microbatching tools, collects incoming data together with its later processing in batches whose size can be changed according to programming necessities. In this group, we can find Spark streaming, Flink, or Storm-Trident [25].
4) *Service orchestration:* A typical Big Data pipeline is in charge of capturing, transforming, and storing data bringing them together. This should be weaved by a proper *conductor*, which orchestrates the process. This kind of service is provided by Akka tools. One of its main advantages is that it is distributable by design, i.e., every actor location is transparent and distributed by design. This enables a faster integration of different and heterogeneous sources in a distributed fashion. This type of architectures are available for instance in Hadoop-based distributions.
5) *System deployment:* In addition to the algorithms to be developed, it is important to count on an appropriate platform where all the components will be installed. This platform should be scalable (i.e., easy to append new components to the system), fault tolerant, and must enable the monitoring of the whole system functioning. Among the available ecosystems, we can distinguish the Apache Hadoop cluster or the Apache Mesos cluster. The market leader distributions for Hadoop are Cloudera, Hortonworks, and MapR. Depending on the BEMS necessities, the architecture designer will decide the one that better fits to the building requirements.

All these technologies can be deployed in private premises or in cloud available services. Amazon Lambda (or AWS Lamda) is one of the most known cloud services that allows the execution of a code in different languages (e.g., Node.js, Python, C#, or Java) and its integration with other AWS services. It is based on Amazon CloudWatch [26], which enables the monitoring and reaction to changes in the system. A similar service is offered by Microsoft with Azure Data Factory [27]. This service allows us to create data-driven workflows to orchestrate the storage and processing of data in other premises environment. This feature is very important when security and privacy of the data is to be preserved, and/or data are not stored in the cloud. Other big companies have started offering cloud services such as Google with the Google Cloud Platform [28] and Oracle with Oracle Cloud [29] offering access to their own Big Data and cloud storage services. Comparisons among some of these cloud services can be found in [27] and [30].

In the next section, we present our architecture proposal detailing its structure, components, and technologies employed, and in Section V, we describe some examples where this architecture has been successfully introduced in four distinct-purpose big facilities.

IV. Big Data-Based BEMS Architecture

In this section, we propose an efficient and fault-tolerant BEMS based on a Big Data architecture. The overall employed system architecture can be divided into different layers (see Fig. 1), each of which is in charge of a different part of the system: data acquisition and transportation, control unit, and data analysis plus decision-making. These categories are in turn

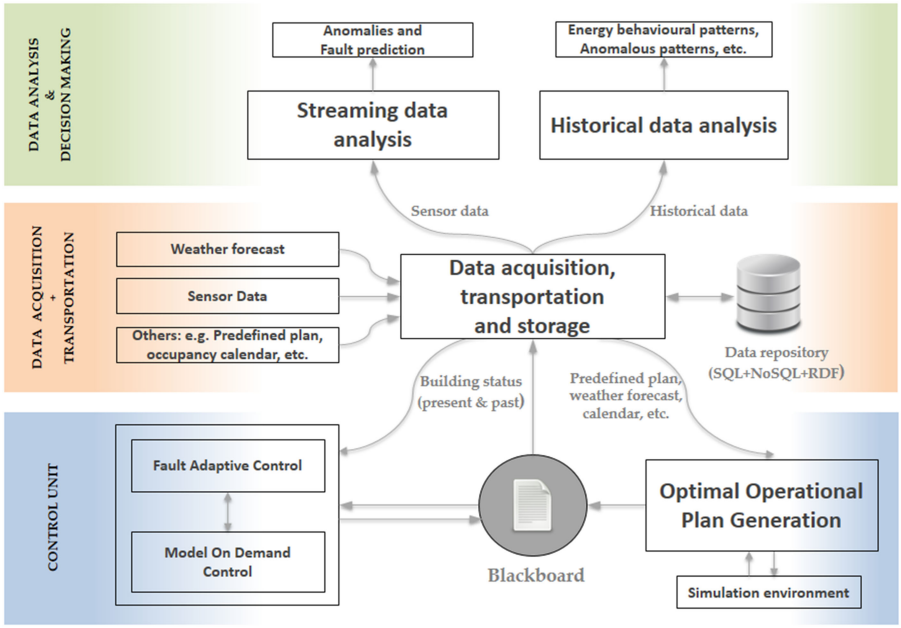

Fig. 1. Proposed BEMS architecture.

divided into various components responsible for making some specific tasks.

The overall system was deployed using Cloudera Manager distribution [31], which enables the installation and configuration through a web interface. Another important advantage is that the key management system is fault-tolerant, ensuring the good functioning of all modules. In addition, it provides tools for the monitoring of the different subsystems using the web interface and contains modules to securely encrypt data, in case it is necessary. Moreover, the current distribution includes support for HDFS and MongoDB for data storage, and Spark and Spark Streaming for data analysis, which incorporates the machine learning libraries containing useful algorithms that can be applied in the data analysis and decision-making layers.

A. Data Acquisition Layer

The data acquisition and transportation layer is in charge of managing all the data to be used in the rest of the modules. Its purpose is to connect and retrieve the available data from the locally installed BMS and then make them transparently available to the other modules. Depending on the building, in relation to its systems or uses, the needs and the amount of data are extremely different. Therefore, understanding the problems and aims is the first step to carry out. After this, it is necessary to determine, obtain, and clean the data, with the intention of keeping only those data needed for the rest of the components.

In this layer, storage of building information can be managed with different technologies. While NoSQL databases have shown advantages over the traditional relational models in terms of efficiency and flexibility, they fall short for organizing data coming from potentially very different sources and contexts. For this purpose, Semantic Web technologies provide standard languages (e.g., RDF and OWL) and vocabularies to address the challenges of data homogenization and interoperation. Particularly, the semantic sensor data (SSN) ontology [32] is the recommendation of the World Wide Web Consortium for describing sensors and their observations. The SSN ontology has been previously used for representing energy meter data and performance indicators [33], [34], and can be related to DAB-GEO [35], an upper level ontology for energy concepts [36]. Furthermore, sensor data models can be linked to other building data models, thus facilitating data integration and information retrieval. For example, building data of the building information model (BIM) can be leveraged into a semantic BIM [37] encoded in RDF/OWL, enabling cross-domain information integration, flexible querying, and imprecise parametric modeling [38], [39].

In the buildings where the proposed architecture was implemented, we mainly focused on the data managed in the pilot areas of the buildings. Weather forecasts (temperature, relative humidity, wind speed, precipitation, etc.) are continuously provided containing the future prediction of at least the following 24–48 h and they come from weather services available in each country. The occupancy calendar can be provided by the internal management of each building, e.g., the expected occupancy in a hotel, or can be collected from sensors, e.g., number of people working in an office. In this respect, based on the type of building, the calendar data may suffer variations at any time, which forces the adaptation of the control plan at any moment, such as, for instance, the cancellation of a room in a hotel and the delaying of a flight in an airport. The predefined OP is also stored and serves as a basis to generate the optimal OP for the subsequent days. The largest volume of data comes from the real-time data flows given by the sensors capturing the building status at different time frequencies. It is in this part where Big Data technologies play an essential role to efficiently manage

and store the data. Apart from this, there are other data that are handled by the data acquisition platform, which are mainly provided by the control unit layer: energy consumption, energy demand, OPs, equipment faults, blackboard history, etc.

All the acquired data should be stored for their latter usage by the different BEMS modules. In our case, the pilot sites data repositories are composed of a set of databases running under a common framework. This computational environment was implemented using different types of database management systems: relational, NoSQL, and RDF/OWL. In particular, to handle massive data, MongoDB was employed to store data coming from sensing devices, and results generated in the control unit layer and the data analysis layer. The processed data are stored in files using HDFS format which is the best format to then apply DM&ML methods in the data analysis layer.

B. Control Unit Layer

The control unit layer is in charge of managing the daily operation plan, aiming to leverage equipment control—particularly HVac. This is comprised of two different parts: the model predictive control (MPC) and the model on demand control (MODC). MPC provides the nominal optimal plan according to the weather forecast conditions and the MODC strategy aims at adapting the nominal control online with respect to the actual current system operating conditions, reducing the impact of deviations that may occur between actuals and forecasts on the ideal provided plan.

MPC generates the optimal OP that uses weather forecasts, user occupancy, energy tariffs, building status, and other information about the expected building operation conditions. All this information is employed to run simulations in order to find the optimal daily plan that adjusts to the comfort occupants' requirements with a lower energy consumption. This results in an OP, consisting of a list of OP setpoints to be applied at certain time periods, which is communicated to the operators using the blackboard (see Fig. 1) or automatically set to the devices in case they allow it. The simulation environment and its interaction with the MPC/MODC is extensively explained in [40] and [41]. Some results of the application of the control unit layer can be also explored in [40]–[42].

Pilot site operators have expressed their concerns about automatic setpoint sending. For this reason, we proposed a centralized remote control module as a proxy to present the setpoints to the expert users, who would confirm or discard the setpoints. It would also be possible to implement (instead, or in addition to this functionality) a list of critical and noncritical equipment, in such a way that only noncritical equipment could have setpoints automatically sent.

Associated to the MODC module, it is the fault adaptive control unit which controls if some indicators exceed some predefined thresholds in order to detect and isolate the possible errors happened during the building functioning. When the predicted conditions change or some faults/anomalies arise, the MODC takes place in order to fine tune the setpoints to correct the system behavior. In this case, the blackboard is again updated with the pertinent changes. All the setpoints written in the blackboard are eventually sent to the data acquisition layer for their storage.

C. Data Analysis and Decision Making Layer

Data analysis techniques have been employed during the last decades for different purposes within the energy field, among them, we can highlight the following:

1) energy demand prediction required for the efficient operation of a building;
2) building operation optimization;
3) monitoring operational status and failure detection of building equipment and networks verification;
4) analysis of the economic and commercial impact of user energy consumption;
5) energy fraud detection and prevention.

For this, a myriad of DM&ML techniques have been applied as it can be seen in the review made in [6]. This kind of processes entails the automatic analysis of registered data in a more human-friendly fashion, which will help, in a latter step, the decision-making process to recognize energy model adjustments, architectural design improvements, etc. Besides, the new venue of Big Data also serves for discovering new insights about buildings' energy behavior, enabling a faster analysis of the collected data when massive amounts have to be handled.

In our case, the Spark platform was used for the implementation of the data mining techniques to analyze both historical and real-time streaming data. In particular, Spark streaming was preferred for dynamically collecting data from sensors with different latency. We developed a procedure to discover trends and patterns relating different aspects of energy consumption and costs according to different time periods, as for instance: day-to-day, weekly, monthly, or seasonally consumptions. Stress that, among the collected data in the repository, there were very heterogeneous types of variables coming from sensors, weather, OPs, etc. whose values were taken in different time periods; some of them measuring the same phenomenon, having sometimes duplicated and/or missing values. For being able to adequately process the diverse kind of data, they were appropriately partitioned in adequate segments for the analysis, duplicates were removed, and a specific granularity was set for adjusting to the user expectations.

After all the preparation and data preprocessing, frequent itemset and association rule mining algorithms implemented in Spark were applied to obtain energy meaningful patterns [43]. For example, associations of the type *outdoor_temperature=(18 °C, 28 °C] → fresh_air_handling=(0%, 25%]* were found, meaning that there exists a relation between the *outdoor temperature* and the percentage of use of the *fresh air handling unit* in the proportion specified in the intervals. The main advantage of using Big Data tools is to enable a fast processing of such quantity of data collected during a period of time in a building, whilst classic association rule mining approaches do not enable this processing, aborting them in the majority of cases due to a memory overflow. Another advantage of using such Big Data framework is the fault-tolerant

TABLE I
DATA COLLECTED FOR THE FARO AIRPORT DURING ONE YEAR (2016)

Data type	# of variables	# of registers	Storage
General BEMS	7 897	40 799 048	9.1 GB
Quality	92	10 063 326	3.6 GB
Common areas sensors	8 498	16 727 178	4.2 GB
Flight	60	112 819	460 Mb
Consumption	75	673 311	154 Mb

mechanisms they have internally implemented (using data distribution and replication), which does not need to be handled by the programmer. Some of the problems solved by implementing data analysis techniques using Big Data can be found in the review [6] or in the work [43].

V. BIG DATA BEMS IN A REAL SETUP

The proposed architecture was implemented in four different pilot site buildings with different climates within the Energy IN TIME EU project including a commercial building, an office, a hotel, and an airport. For each of them, we are going to highlight some features of the implemented Big Data architecture focusing on the specific parts of each layer. This section does not intend to go into the details of each specific layer, which comprises another article on its own, but to illustrate the advantages that the proposed Big Data architecture offers.

A. Data Acquisition and Transportation in an Airport

The data acquisition and transportation layer was successfully implemented in every building, adapting it to the type of sensors, HVac machines, schedules, etc. The Faro airport is a particularly interesting case. For this building, all the data related to flights schedules were managed by the system. This airport, located in Faro, Portugal, is placed in a region with warm and arid summers, and cool and windy winters. It is comprised of mainly open spaces with big flows of people at certain times of the day (e.g., flight arrivals or departures). The main peculiarity of this case was the high variability of collected data, which have to be replaced every time a small or a big change occurs, since the programmed schedules varied a lot during the same day because of flight delays. In this regard, the proposed architecture, using the MongoDB system, was proactive and flexible to enable changing this type of features when necessary. For example, the MPC module had to adapt to the flight changes by recalculating the best operation plan in some places of the airport.

Table I summarizes the data collected in the Faro airport during one year in a pilot area comprised of 7208.92 m^2. The whole dataset comprises a total of 3 years with around 40 GB of data collected. Specifically, in the Faro airport, the databases were arranged in different parts according to the type of data (see Table I). The *General BEMS* database contains data from the general BMS such as the chiller/heat pump, boiler, water, and air treatment units, as well as the different zone setpoints that can be modified by the operator. The *Quality* database contains measurements from the wireless system controlling the quality conditions throughout the airport, such as the room temperature or the humidity, data coming from different air handling units in the pier feeding the boarding gates, and the alarm sensor data. The *Sensor database* from common areas is comprised of data coming from sensors located in the restaurant, luggage collection rooms, and exterior areas such as the bus gates. The *Flight* database collects all the information related to the arrivals, departures, and delays of flights, like, for instance, the origin, name of the company, type of airplane, number of passengers, boarding gates, arriving/departing time, etc. Finally, the *Consumption* database contains the information about the electrical consumptions of the different equipment taken into account in the control unit layer.

All these collected data were employed for their latter processing in the control unit layer. In this case, a reactive control was applied based on the data collected by the data acquisition layer. Particularly interesting was the application of *Flight database* in the control unit to improve the thermal energy and electricity consumption in the terminals according to the last minute information about arrivals, departures, and delays.

B. Control Unit in Different-Purpose Buildings

The control unit layer was successfully applied in every building. For illustrating purposes, we explain how the MPC was applied in different buildings in order to improve the energy savings of both cases (more details in [40]). The first building was a commercial and office building located in Helsinki, Finland, a region with harsh winters and comfortable summers. It is mainly comprised of offices with scheduled occupancy, and also with a public area with commercial and restaurant usage with varied flows of people. The second building was the Levi Panorama hotel, a ski resort in Kittilä, a region in northern Finland located north of the Arctic Circle within the Lapland region. It has a subarctic climate, with strong seasonal shifts, as well as polar night and midnight sun. The hotel activities are mainly concentrated during the winter period, having an occupation level higher than 90% during the ski season.

In order to evaluate the savings achieved by the control unit, the International Performance Measurement and Verification Protocol (IPMVP) [44] was applied. The IPMVP is based on calculating savings determined by comparing measured use or demand before and after the implementation of the BEMS, making suitable adjustments for changes in conditions.

The saving obtained is computed as the difference between the adjusted-baseline energy and the energy that was actually metered during the reporting period. This quantity is adjusted in order to segregate the energy effects of a savings program from the effects of other simultaneous changes affecting the energy using systems. In general, the following formula was applied:

$$\text{Savings} = \text{Baseline\_Energy} - \text{Post\_Install\_Energy} \pm \text{Adjustments}.$$

The demonstration activity in the commercial and office building started in April 2017 and was extended until the end of May 2017. The comparison of the heating consumption in the pilot area during some days is shown in Fig. 2. The energy savings achieved in the demonstration period compared to the

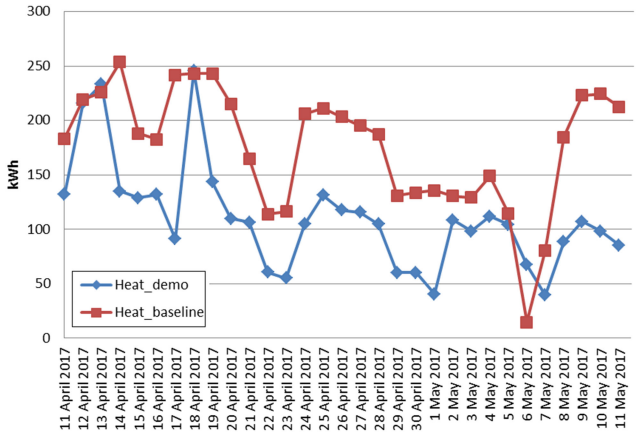

Fig. 2. Adjusted heat baseline versus real heat measured in the commercial and office building in Helsinki.

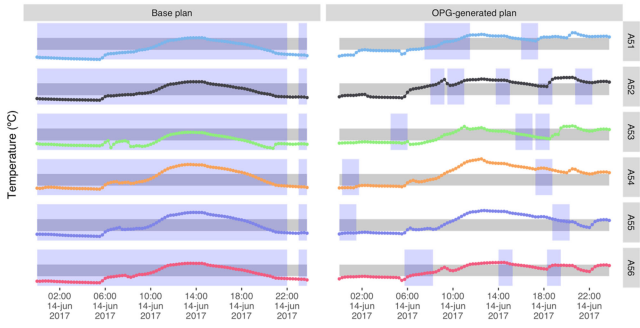

Fig. 3. Simulation of previous operation (base plan) versus OPG-generated plan in different boarding gates in the Faro airport.

baseline, that was studied with monitored data in 2016 with similar characteristics, were 26%.

Differently to the office building, in the Levi Panorama hotel, the operators did not allow us to give the full control to the system, but it was communicated to the operators through the blackboard (see Fig. 1). Thus, the proposed OPs were set by hand by the operators. The savings obtained were not very high (around 6%), because it was difficult to manage the rooms space due to the interaction of the clients.

In all these cases, the Big Data infrastructure enabled a faster computation of the MPC, which is crucial to provide the optimal operation plan among the exponential number of possible combinations.

For the case of the airport located in Faro, the control unit was focused on the optimization of HVac operation by using the information of flight data, occupancy, and boarding gates. The new BEMS was able to anticipate gate occupancy and expected comfort conditions under different operations, and consequently, optimize the setpoints and automatize control (see Fig. 3). The experiments performed in a simulated environment anticipated an overall reduction of 20% in energy consumption while guaranteeing users' comfort. However, the implementation during the month of August (peak season) achieved a 6% of savings.

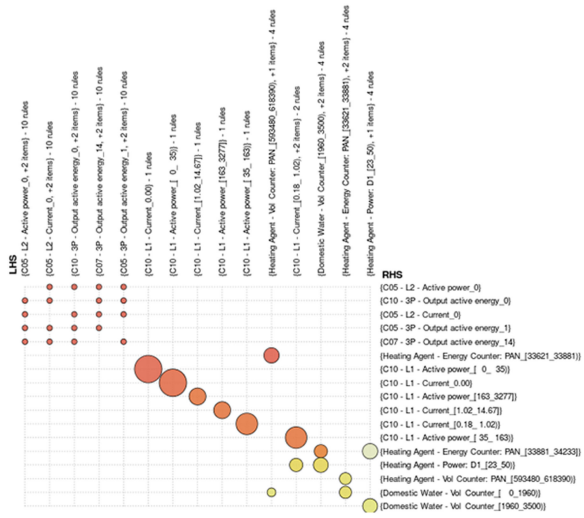

Fig. 4. Some association rules obtained for the period: January 4–14 in the office building in Bucharest.

This situation was helpful to conduct a subsequent analysis of the building operation, which drew two conclusions: 1) the HVac system of a few boarding gates (with higher solar irradiance) was insufficient to cover the cooling demand and 2) in busy days with frequent boarding gate changes, the control unit should apply a less aggressive optimization policy. This analysis would not have been possible without the new capabilities for data acquisition and analysis of the new BEMS.

C. Data Analysis in an Office Building

Data collected in the acquisition layer can be exploited in the control unit layer for improving energy management of the buildings, and they can be useful, as in this case, in the data analysis layer. In this regard, data analysis and decision support tools can help to support and improve other aspects within the energy functioning of the building, such as to examine the state and the fails of equipment and networks, to analyze the economic and commercial impact of users' energy consumption, or to detect and prevent energy fraud [6].

For exemplary purposes, we present the application of a data mining technique for an office building in Romania. This building is located in Bucharest, Romania, a city with warm summers, and very cold and dry winters, and it is comprised of offices with constant flows of people and scheduled occupancy. In our case, we applied association rule mining techniques to discover useful and hidden correlations that are not so direct for building operators.

A set of interesting variables were selected among all the collected data in the building. Then, they were preprocessed, filtered, and aggregated into meaningful intervals in order to obtain more expressive patterns. The analysis was made in different time periods: ten days, one month, and one quarter during the winter season, obtaining some interesting patterns relating the energy consumption and some specific sensors. In Fig. 4, we

can see some of the extracted relations in a ten-day period. The patterns are association rules of the form LHS $\rightarrow$ RHS, where LHS and RHS stand for left-hand-side and right-hand-side of the rule, respectively, and they represent the relation between both parts of the rule. In addition, it is measured how frequent is the presence of both parts of the rule in the data, and how strong the relation between them is. The higher the size of circles, the higher the support value of the rule (i.e., more frequent is the joint presence of LHS and RHS in the analyzed data), and the darker the color (i.e., closer to orange) the higher the strength of the rule.

In Fig. 4, we can observe some direct relations between power sensors (current and active power) that do not provide meaningful information, since they are directly related. However, some interesting rules have been discovered relating the heating with the domestic water and some energy consumptions, which justify the use of central heating in winter months, to heating the water above its initial temperature. Other type of information relating the temperature with the use of heating equipment and windows is represented in the following rule:

$$\{\text{WindowsPAN} = \text{on}, \text{Windows PAS} = \text{on}\} \rightarrow$$
$$\{\text{Output temperature warm}, \text{Setup PAS} = \text{off},$$
$$\text{Setup PAN} = \text{off}, \text{Temperature} = \text{comfort}\}.$$

This rule was obtained from the data collected during summer, and it says that we reach a comfortable temperature when heating equipment is not working and the windows are open. More details and other results can be found in [45] and [46].

VI. Future Challenges

The previous examples show how Big Data can help in different processes involving several aspects of the building management like the data gathering and analysis and also the operative plan generation. Among its main advantages is their application for managing massive quantities of data with a proven fault-tolerant configuration for their posterior processing.

However, as it has been shown, in the case of the Panorama hotel, the system could not be fully applied in some aspects due to its dynamic nature because the majority of its functioning relies on the client side. Buildings with this kind of structure like hotels, houses, and some office buildings, can also benefit from the proposed BEMS architecture, taking advantage of the data acquisition and data analysis layers to study the energetic footprint of the building in order to adapt the OP depending on the season, the day of the week, the holidays, the weather forecast, etc.

Remark that for the integration of the Big Data infrastructure in a BMS, the buildings have to be adequately adapted providing them with sensors and measurement units to gather the sufficient data to their posterior analysis, proposing a control unit according to the building necessities. In addition, the building has to provide a server infrastructure or rent a cloud service to install the Big Data architecture that will be connected to the flow of data generated by the building.

The main challenges are therefore not only the building adaptation to this type of systems, but also the choice of adequate DM&ML tools capable of analyzing the huge amount of data. Here, we have presented the use of association rules but other DM&ML algorithms can be applied according to user preferences.

VII. Conclusion

The proposed BEMS was effectively introduced in four different purpose buildings with some particular variations, demonstrating that Big Data can help during the energy cycle of the building. Therefore, Big Data offered a suitable framework for an efficient and fault-tolerant building energy management due to the robustness of Spark systems. It is the heart of the energy system including not only the data acquisition phase, but also the implementation of systems capable of discovering new insights from data collected from the building metering joint with weather, occupancy, or OP data. It also enables fast computations that can help the generation of optimal OP generations, as it can be seen in the control layer of the proposed architecture.

This work opens the door to new BEMS implementations taking benefit of Big Data platforms, as well as the use of the DM&ML available algorithms that the scientific community is developing in the last years. The main objective is to move building operators closer to the use of these new tools that can effectively help to manage and also process the big volumes of data generated by the sensors and other building information, like weather forecasts and occupancy schedules, taking advantage of distributed computing, as it has been shown in the four different buildings where the architecture was applied.

References

[1] M. Waseem *et al.*, "Building energy metering and environmental monitoring—A state-of-the-art review and directions for future research," *Energy Buildings*, vol. 120, pp. 85–102, 2016.
[2] K. Kaur, T. Dhand, N. Kumar, and S. Zeadally, "Container-as-a-service at the edge: Trade-off between energy efficiency and service availability at fog nano data centers," *IEEE Wireless Commun.*, vol. 24, no. 3, pp. 48–56, Jun. 2017.
[3] N. Kumar, S. Misra, J. J. P. C. Rodrigues, and M. S. Obaidat, "Coalition games for spatio-temporal Big Data in Internet of Vehicles environment: A comparative analysis," *IEEE Internet Things J.*, vol. 2, no. 4, pp. 310–320, Aug. 2015.
[4] T. White, *Hadoop: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2012.
[5] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, *Learning Spark: Lightning-Fast Big Data Analysis*. Sebastopol, CA, USA: O'Reilly Media, 2015.
[6] M. Molina-Solana *et al.*, "Data science for building energy management: A review," *Renewable Sustain. Energy Rev.*, vol. 70, pp. 598–609, 2017.
[7] M. Marinov, P. Vitliemov, and E. Popova, "Towards Big Data and Internet of Things as key aspects of energy efficiency," *TEM J.*, vol. 6, no. 3, pp. 427–435, 2017.
[8] M. A. Ahad and R. Biswas, "Request-based, secured and energy-efficient (RBSEE) architecture for handling IoT Big Data," *J. Inf. Sci.*, vol. 45, no. 2, pp. 227–238, 2019.
[9] E. L. Lydia, A. K. Rebecca, R. M. Vidhyavathi, and K. Sumathi, "Handling of Big Data with a novel solution architecture on smart grids," *Int. J. Pure Appl. Math.*, 118, no. 7, pp. 367–371, 2018.
[10] L. Linder, D. Vionnet, J.-P. Bacher, and J. Hennebert, "Big building data—A Big Data platform for smart buildings," *Energy Procedia*, vol. 122, pp. 589–594, Sep. 2017.

[11] A. R. Al-Ali, I. A. Zualkernan, M. Rashid, R. Gupta, and M. Alikarar, "A smart home energy management system using IoT and Big Data analytics approach," *IEEE Trans. Consum. Electron.*, vol. 63, no. 4, pp. 426–434, Nov. 2017.
[12] M. Beaudin and H. Zareipour, "Home energy management systems: A review of modeling and complexity," *Renewable Sustain. Energy Rev.*, vol. 45, pp. 318–335, 2015.
[13] M. L. Tuballa and M. L. Abundo, "A review of the development of Smart Grid technologies," *Renewable Sustain. Energy Rev.*, vol. 59, pp. 710–725, Jun. 2016.
[14] M. Mayilvaganan and M. Sabitha, "A cloud-based architecture for Big-Data analytics in smart grid: A proposal," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, 2013, pp. 1–4.
[15] R. Shyam *et al.*, "Apache spark a Big Data analytics platform for smart grid," *Procedia Technol.*, vol. 21, pp. 171–178, 2015.
[16] A. A. Munshi and A.R. Yasser,, "Big Data framework for analytics in smart grids," *Electric Power Syst. Res.*, vol. 151, pp. 369–380, 2017.
[17] M. Jaradat *et al.*, "The Internet of Energy: Smart sensor networks and Big Data management for smart grid," *Procedia Comput. Sci.*, vol. 56, pp. 592–597, 2015.
[18] M. Wazid, A. K. Das, N. Kumar, and J. J. P. C. Rodrigues, "Secure three-factor user authentication scheme for renewable-energy-based smart grid environment," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3144–3153, Dec. 2017.
[19] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transp. Res. Part C: Emerg. Technol.*, vol. 86, pp. 580–596, Jan. 2018.
[20] C. Fan, D. Yan, F. Xiao, A. Li, J. An, and X. Kang, "Advanced data analytics for enhancing building performances: From data-driven to Big Data-driven approaches," *Building Simul.*, vol. 14, pp. 3–24, 2021.
[21] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, M. Maasberg, and K. R. Choo, "Multimedia Big Data computing and Internet of Things applications: A taxonomy and process model," *J. Netw. Comput. Appl.*, vol. 124, pp. 169–195, 2018.
[22] V. Marinakis, "Big Data for energy management and energy-efficient buildings," *Energies*, vol. 13, pp. 1–18, 2020.
[23] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.
[24] C. Prakash, "Spark streaming vs Flink vs Storm vs Kafka streams vs Samza: Choose your stream processing framework," *Medium*, 2018. Accessed: Feb. 06, 2019. [Online]. Available:https://medium.com/@chandanbaranwal/spark-streaming-vs-flink-vs-storm-vs-kafka-streams-vs-samza-choose-your-stream-processing-91ea3f04675b
[25] S. Chintapalli *et al.*, "Benchmarking streaming computation engines: Storm, Flink and Spark streaming," in *Proc. Int. Parallel Distrib. Process. Symp. Workshops*, 2016, pp. 1789–1792.
[26] Amazon Web Services, "AWS cloud watch," 2019. Accessed: May 31, 2019. [Online]. Available: https://aws.amazon.com/cloudwatch/
[27] S. Klein, "Azure data factory," in *IoT solutions in Microsoft's Azure IoT Suite*. New York, NY, USA: Apress, 2017, pp. 105–122.
[28] Google, "Google cloud platform," 2019. Accessed: May 31, 2019. [Online]. Available: https://cloud.google.com/
[29] Oracle, "Oracle cloud (Big Data)," 2019. Accessed: May 31, 2019. [Online]. Available: https://cloud.oracle.com/bigdata
[30] M. Tepakidareekul, "Serverless platform comparison: Google cloud function vs. AWS Lambda," *Medium*, 2018. Accessed: Sep. 20, 2020. [Online]. Available: https://medium.com/@manus.can/serverless-platform-comparison-google-cloud-function-vs-aws-lambda-8e060bcc93b4
[31] Apache Software Foundation, "Cloudera Manager," 2020. Accessed: Jul. 23, 2020. [Online]. Available: https://www.cloudera.com/products/product-components/cloudera-manager.html
[32] Semantic Sensor Network Ontology. Accessed: Oct. 08, 2021. [Online]. Available: https://www.w3.org/TR/vocab-ssn/
[33] S. Dey, D. Jaiswal, R. Dasgupta, and A. Mukherjee, "Organization and management of semantic sensor information using SSN ontology: An energy meter use case," in *Proc. 9th Int. Conf. Sens. Technol.*, 2015, pp. 468–473.
[34] E. Corry, P. Pauwels, S. Hu, M. Keane, and J. O'Donnell, "A performance assessment ontology for the environmental and energy management of buildings," *Automat. Construction*, vol. 57, pp. 249–259, 2015.
[35] J. Cuenca, F. Larrinaga, and E. Curry, "DABGEO: A reusable and usable global energy ontology for the energy domain," *J. Web Semantics*, vol. 61/62, 2020, Art. no. 100550.
[36] A. Hogan, "The Semantic Web: Two decades on," *Semantic Web*, 11, no. 1, pp. 169–185, 2020.
[37] P. Pauwels, S. Zhang, and Y.-C. Lee, "Semantic Web technologies in AEC industry: A literature overview," *Automat. Construction*, vol. 73, pp. 145–165, 2017.
[38] J. Gómez-Romero, F. Bobillo, M. Ros, M. Molina-Solana, M. D. Ruiz, and M. J. Martín-Bautista, "A fuzzy extension of the semantic building information model," *Automat. Construction*, vol. 57, pp. 202–212, 2015.
[39] I. Huitzil, M. Molina-Solana, J. Gómez-Romero, and F. Bobillo, "Minimalistic fuzzy ontology reasoning: An application to building information modeling," *Appl. Soft Comput.*, vol. 103, 2021, Art. no. 107158.
[40] J. Gómez-Romero *et al.*, "A probabilistic algorithm for predictive control with full-complexity models in non-residential buildings," *IEEE Access*, vol. 7, pp. 38748–38765, 2019.
[41] J. Gómez-Romero, M. Molina-Solana, M. Ros, M. D. Ruiz, and M. J. Martin-Bautista, "Comfort as a service: A new paradigm for residential environmental quality control," *Sustainability*, vol. 10 no. 9, pp. 2071–1050, 2018.
[42] A. Conserva *et al.*, "Energy IN TIME project: Summary of final results," in *Proc. 12th Conf. Sustain. Develop. Energy, Water Environ. Syst.*, 2018, pp. 4–8.
[43] C. Fernandez-Basso, A. J. Francisco-Agra, M. J. Martin-Bautista, and M. D. Ruiz, "Finding tendencies in streaming data using Big Data frequent itemset mining," *Know.-Based Syst.*, vol. 163, pp. 666–674, Jan. 2019.
[44] Efficiency Valuation Organization. Accessed: Sep. 29, 2020. [Online]. Available: https://evo-world.org/en/products-services-mainmenu-en/protocols/ipmvp
[45] C. Fernandez-Basso *et al.*, "Extraction of association rules using Big Data technologies Carlos," *Int. J. Des. Nature Ecodyn.*, vol. 11, no. 3, pp. 178–185, 2016.
[46] C. Fernandez-Basso *et al.*, "A fuzzy mining approach for energy efficiency in a Big Data framework," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 11, pp. 2747–2758, Nov. 2020.

M. Dolores Ruiz received the degree in mathematics and the Ph.D. degree in computer science from the University of Granada, Granada, Spain, in 2005 and 2010, respectively.

She held teaching positions with the University of Jaén, Jaén, Spain, University of Granada, and University of Cádiz, Cádiz, Spain. She is currently with the Approximate Reasoning and AI Research Group, Computer Science Department, University of Granada. She has been a part of the organization committee and has organized several special sessions about data mining in international conferences and has participated in more than ten R&D projects. Her research interests include data mining, information retrieval, correlation statistical measures, sentence quantification, and fuzzy sets theory.

Juan Gómez-Romero received the B.Sc. degree in computer science and the M.Sc. and Ph.D. degrees from the University of Granada, Granada, Spain, in 2004, 2006, and 2008, respectively.

He was a Lecturer with the Applied Artificial Intelligence Group, Universidad Carlos III de Madrid, Madrid, Spain, from 2008 to 2013, and a Research Associate in the EU FP7 Project Energy IN TIME with the University of Granada, from 2013 to 2017. Since 2019, he has been an Associate Professor with the Computer Science and Artificial Intelligence Department, Universidad de Granada. His research interests include machine learning for control optimization and simulation of power systems.

Dr. Gómez-Romero is currently the Principal Investigator of the projects PROFICIENT: Deep learning for energy-efficient building control and DeepSim: Deep learning of building simulation models.

Carlos Fernandez-Basso received the degree in computer science, the M.Sc. degree in data science, and the Ph.D. degree in computer science from the University of Granada, Granada, Spain, in 2014, 2015, and 2020, respectively.

He is currently a Postdoctoral Fellow with Causal Cognition Lab, University College London, London, U.K. He was a Lead Developer in the EU FP7 Project Energy IN TIME in the topics of building simulation and control, data analytics, and machine learning, and in the COPKIT Project in the topics of cybercrime, Big Data, and machine learning. From 2016 to 2018, he collaborated with the Data Science Institute, Imperial College London, London, U.K., where he has carried out research stays.

Maria J. Martin-Bautista received the degree in computer science and the Ph.D. degree from the University of Granada, Granada, Spain, in 1995 and 2000, respecticely.

She has been a Full Professor with the Department of Computer Science and Artificial Intelligence, University of Granada, since 1997. She has participated in more than 20 R&D projects and has supervised several research technology transfers with companies. She has supervised several Ph.D. thesis and has authored or coauthored more than 100 papers in high-impact international journals and conferences. Her current research interests include Big Data analytics in data, text and web mining, intelligent information systems, and knowledge representation and uncertainty.

Prof. Martin-Bautista is a Member of the Intelligent Data Bases and Information Systems Research Group. She was also a Program Committee Member for several international conferences.