A Novel Architecture for Efficient Fog to Cloud Data Management in Smart Cities

Amir Sinaeepourfard, Jordi Garcia, Xavier Masip-Bruin, Eva Marin-Tordera
Advanced Network Architectures Lab (CRAAX),
Universitat Politècnica de Catalunya (UPC, BarcelonaTech),
Barcelona, Spain
{amirs, jordig, xmasip, eva}@ac.upc.edu

***Abstract*—Smart cities are the current technological solutions to handle the challenges and complexity of the growing urban density. Traditionally, smart city resources are managed with a cloud based solution, where sensors and devices are connected to provide a centralized and rich set of open data. The advantages of cloud based frameworks are their ubiquity, as well as an (almost) unlimited resources capacity. However, accessing data from the cloud implies large network traffic, high latencies usually not appropriate for real-time or critical solutions, as well as higher security risks. Alternatively, fog computing emerges as a promising technology to absorb these inconveniences. It proposes the use of devices at the edge to provide closer computing facilities and, therefore, reducing network traffic, reducing latencies drastically while improving security. In this work, we present a new framework for data management in the context of a smart city through a global fog to cloud resources management architecture. We show this model has the advantages of both, fog and cloud technologies, as it allows reduced latencies for critical applications while being able to use the high computing capabilities of cloud technology. As a first experiment, we estimate the network traffic in this model during data collection and compare it with a traditional real system.**

Keywords—Data Management; Fog-to-Cloud computing; Smart City; Data Aggregation; Resource Allocation; Distributed Data Management

I. INTRODUCTION

Today, it is estimated that around 54% of world's population is living in cities; however, it is expected that this percentage will rise up to 70% by 2050 [1]. This increase of urban density will impose higher requirements and demands in the city management and organization. In this context, smart cities become a promising as well as challenging technology to facilitate the sustainable development of cities. The goals of the smart cities are improving the citizens' quality of life by providing more advanced, sophisticated, but also efficient services, while fueling a sustainable economic growth. There are multiple research topics related to smart cities technology, starting from the sensors technologies, which generate a relevant part of the city information, the sensors devices organization, including technologies such as the Internet of Things (IoT), the Internet of Everything (IoE), other Service Oriented Architectures (SOA), or even a combination of these [2], using different communications technologies, such as wired Ethernet, or wireless WiFi, 3G/4G networks, or other ad-hoc low-power wide-area networks (LPWAN), and providing these data and resources available to create new and sophisticated services that ease life in the city according to the city's social, economic, and environmental models.

Traditional resource management architectures in smart cities rely typically on centralized cloud computing facilities. Advantages of cloud computing are the (almost) unlimited computing capacity, the cost efficiency (market scale) and the elasticity (pay as you go model) [3, 4]. However, moving all data and services to the cloud, which presumably may be far from the user, adds several inconveniences to this option such as high communication latencies, network overloading, but also increases the risks for failures and for security vulnerabilities [5-7]. Alternatively, fog computing is an emerging technology that contributes to reduce these inconveniences by using devices available at the edge. In this way, as the distance between the user and the computing resources is reduced, it does the communication latency, the network load, as well as the risks for connection failures and for security vulnerability.

In this scenario, notice that data are the essential fuel for smart cities development. They allow a city to become smart, instead of just automatized. This is rooted to the fact that data provide the required information for services to proceed according to the contextual state, or some higher value knowledge extracted from complex data analysis. In fact, the smart cities constitute an ideal scenario to generate abundant data from any type of source, mainly from the network of sensors deployed throughout the city, but also from the increasingly popular participatory sensing (for instance sensors integrated in citizens' smartphones), data obtained from social media or any other third party application, streams of data from surveillance cameras and devices, or any other city resource sensitive to contribute with valuable information. Managing and organizing efficiently all these diverse sources and vast volumes of data in such a context is a critical challenge for the success of smart cities.

In this paper we present a novel architecture for efficient data management in the context of a smart city, based on a fog to cloud distributed model for resources management. In

this research work we focus our attention to the data as the core resource in smart cities, and consider data during their whole life cycle, from data acquisition, including data processing and data preservation, up to the data destruction. The advantages of such a model is that it combines the advantages of both the cloud and the fog computing technologies, this is, keeping high performance capabilities for computational intensive applications and reducing communication latencies for real-time or critical services, while reducing network data traffic and enhancing fault tolerance and security protection. As part of this work, in this paper we also describe some basic data aggregation optimizations that can be easily implemented in our fog to cloud distributed architecture, and estimate the effects of the network data traffic reduction during the data collection stage.

The rest of this paper is organized as follows. Section 2 discusses some issues related to the data management, including the concept of the Data Life Cycle (DLC) model. Then, in section 3 different IT architectures for smart cities management are described, including the fog to cloud model over which our data management proposal is defined. So section 4 presents the details of the new architecture for data management in smart cities using the global fog to cloud distributed model, and discuss the advantages of this new approach. In Section 5 we describe some basic data aggregation optimizations to illustrate the potential of our proposal. And in section 6 we discuss some relevant related work about resource management, data management and data aggregation models. Finally, in section 7 we conclude this work and present our future research directions.

II. Data Management Issues

As stated in the previous section, data management and organization is a critical issue for the success of an effective smart cities management system. Many efforts from academia and industry are being devoted to create and use data analysis or data analytics algorithms in order to extract value from this tremendous abundance of data. In fact, in [8] we estimated that 8 GB of data could be generated every day in the city of Barcelona, only considering some basic public sensors' data (for instance, surveillance or traffic control cameras were not included in this report). However, not many researchers are paying attention to explicit data management strategies in the context of Smart Cities.

Data management involves all data life cycle phases, from production to consumption, including data collection, data archiving, data processing, data analysis, data analytics, or data removal, among others. Data LifeCycle (DLC) models constitute the main trend towards developing an integral data management framework, encompassing all data management stages, from data creation to data consumption. The main goals for a DLC model are to operate efficiently, to eliminate waste, and to prepare data products ready for end users matching the expected quality and security constraints. In previous research we proposed the Comprehensive Scenario Agnostic Data LifeCycle (COSA-DLC) model which was proved to be comprehensive, as it was designed as an efficient and global data management model to address the set of 6Vs challenges for big data management (namely Value, Volume, Variety, Velocity, Variability and Veracity), and scenario agnostic, as it is easily adaptable to any scenario or scientific environment [9]. Later, we adapted the COSA-DLC model to a smart city scenario, showing the ease of adaptation of our abstract DLC model. This new model is named the Smart City Comprehensive DLC, or SCC-DLC model, for shorter [6, 7]. The proposed SCC-DLC model has been designed for efficient data management and organization in the context of a smart city.

In Fig. 1 we show our vision of data life cycle in terms of main steps and data flow. We identify three major blocks, namely data acquisition, data processing, and data preservation. Data acquisition is one of the most important data related tasks in a smart city, because the more information gathered from the city, the more complete and sophisticated services can be provided (as long as these data are verified and with high quality). So the data acquisition block is the responsible to collect data, classify them, assess their quality, tag them, apply any eventual data aggregation technique, and prepare them for further usage. Data can then be processed or preserved. The data processing block is the responsible to transform data into information, knowledge, or any other higher value item, through complex analysis or analytical processes. This processed data can be consumed by end users or stored for future usage. And finally, the data preservation block is the responsible for data archiving, storing high quality data (curated in either the data acquisition or data processing blocks), and preparing them for publication, or for further processing phases.

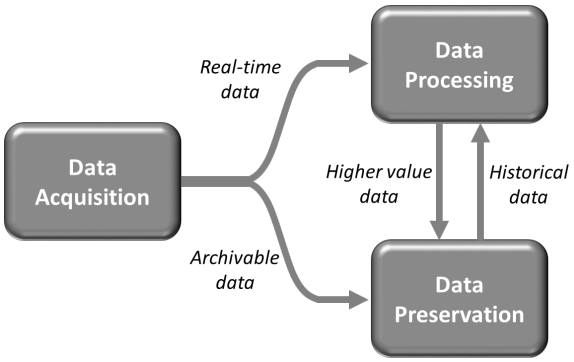


Fig. 1. Data flow in the data life cycle.

The data flow is as follows. Data are gathered into the system through the data acquisition block, which collects data from different sources. If data are required for immediate processing, they can be considered real-time data; otherwise, if they are prepared for storage, they can be considered achievable data. Note that either all or part of the processed data can also be preserved, and vice versa, i.e. these two data flows are not exclusive. When archived data from the data preservation block are used for processing, these are considered historical data. So the data processing block can use both real-time and historical data for processing. Finally, the results of data processing can be consumed by end users or stored back through the data preservation block: in this case, these data are considered to be higher value data.

In Fig. 2 we show the SCC-DLC model adapted to the smart cities scenario. As seen in the figure, each block is implemented through a set of phases to fulfil the required functionalities, as follows:

- The data acquisition block includes the data collection, data filtering (which performs some optimizations, such as data aggregation), data quality (aiming to appraise the quality level of collected data), and data description (tagging data with some additional information) phases.
- The data processing block encompasses the data process (which provides a set of processes to transform raw data into more sophisticated data/information), and data analysis (implementing some analysis or analytic approaches for extracting knowledge) phases.
- And the data preservation block consists of four phases which are the data classification (aiming to organize and prepare data for efficient storage), data archive (storing data for short and long terms consumption), and data dissemination (publishing data for public or private access).

Note that it is not necessary to implement any data quality phase in the data processing nor in the data preservation blocks because all data flowing to these blocks has previously been checked for quality in the data acquisition block. A complete description of each phase and its behavior can be found in [6].

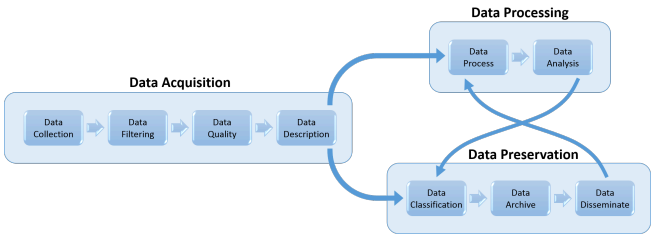


Fig. 2. The SCC-DLC model.

Finally, notice that all gathered data are accessible for smart cities' services consumption, usually through some sort of open access interfaces. In our proposal, we characterize data according to its age, ranging from real-time to historical data. For instance, real-time data is the one generated and just consumed, generally in critical very low latency applications. Such real-time data entails some implicit proximity constraints, because these data are difficult to be critical in remote services. Alternatively, data becomes historical (older data) as long it is accumulated and stored on files or databases. In this case, historical data can be considered to be farther away (even if originally close) because accessing data from cloud, for instance, requires higher latency. We also consider real-time critical data is requested in relatively small sizes, because very large volumes of data can hardly be considered for real-time. On the other hand, historical data can be requested in any, small or large data sets, and any type of fast or complex processing is expected to be done.

III. Architectures for Smart City Resources Management

Resources management in the context of smart cities can be approached from two main perspectives: centralized or distributed. In a centralized approach, a main data center (probably in the cloud) is the responsible to organize and manage all resources from the city, gathering all data generated by sensors at the edge (traffic monitoring, energy meters, noise detection, or air pollution control, among others) and transferring them through some sort of global wireless communication technologies such as 3G or 4G. In addition, processing facilities are also provided in the centralized data center, as long as it has very high computing and storage capabilities. For instance, Fig. 3 shows an architecture for smart city resources management based on cloud computing [10]. This model considers four layers, namely physical, network, cloud, and application layers. The physical layer includes all physical devices to obtain raw data from the city. The network layer provides support for sending the sensed data to the main cloud computing environment. The third layer is the cloud layer which is able to process, compute and analyze all raw data to meaningful information as initial feeds for any further services and applications. And the last layer is the service layer, which is ready to access data from the cloud layer and convert, interpret or combine each other for services and applications. In scenarios like this, there is no doubt about the almost unlimited computing capacities and the ubiquity of such resources; however, some limitations due to the physical distance between resources and services can be reported, such as network overloading, high communication latencies, and high probability of failures and security risks, as mentioned before [6, 7].

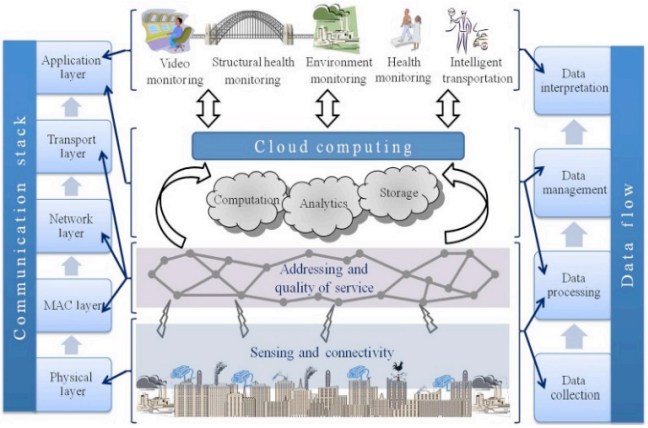


Fig. 3. Example of cloud based information framework for smart cities[1].

Alternatively, in a distributed architecture, the resources management can be performed by using different devices distributed among the city. There are many recent proposals for distributed resources management, including cloudlets, fog computing, and edge computing. As part of these, fog computing has emerged as a promising technology for resources management in the internet of things, by using the

[1] Figure extracted from the original paper in [10]

computational capabilities of the set of devices present in the edge [11, 12]. With this strategy, data do not have to be moved to a central (and far remote) data center (usually in cloud) and, as a consequence, the network traffic and latencies can be reduced, while increasing fault tolerance and security safety.

In this work, we assume a hierarchical fog to cloud distributed architecture, as described in [6, 13], where cloud computing is considered for deep storage and processing, and fog computing is considered for critical and real-time processing. The system can use each computing option according to the requirements of the particular service executed. Fig. 4 illustrates this architecture. The model is hierarchical and it can consider a variable number of levels; however, and for simplicity, the figure just shows a three layers' architecture.

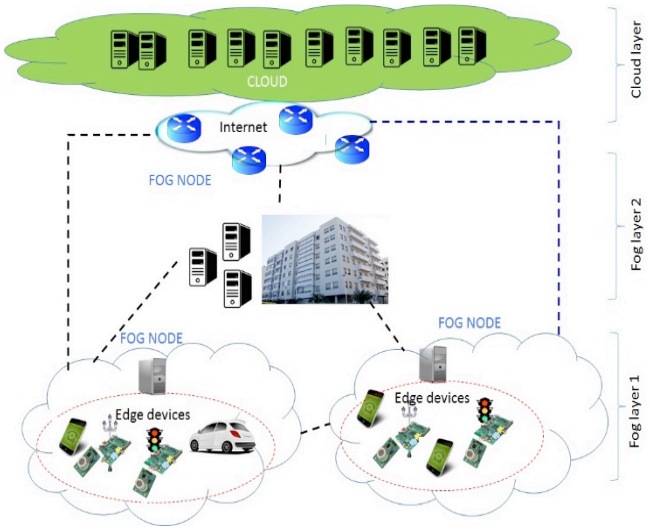


Fig. 4. Hierarchical fog to cloud resources management[2].

The fog layer 1 is formed by a set of fog nodes, each of which consists of a set of "close" devices (at the edge) that can coordinate to perform a joint processing, computing, or just storage, according to the combined capacity of its devices. The scope of a fog node can be tuned, but generally involves a set of devices within a limited area that can communicate each other. So the city is fully organized and divided by a set of fog nodes. The fog layer 2 is responsible for more complex and sophisticated computing, and a node at level 2 is in charge of a set of nodes at level 1, becoming a hierarchical structure. Any computation too large to be done at level 1 is moved upwards to level 2. And finally, the cloud layer is the highest level in the F2C architecture. It provides the highest computing capabilities, and it is the responsible of the whole set of fog nodes at level 2. Because of its almost unlimited computing capabilities, it will assume any computing task that cannot be performed in lower levels, although latency becomes much higher.

[2] Figure extracted from https://arxiv.org/abs/1611.09193.

IV. A New Architecture For F2C Data Management

The distributed hierarchical F2C resources management architecture provides an interesting framework for data management in the context of smart cities, according to our SCC-DLC model proposal. In this section we present a novel architecture for efficient fog to cloud data management in smart cities, consisting on the mapping of the SCC-DLC model onto the smart city F2C resources management architecture. Our model is illustrated in Fig. 5. Note that data acquisition is mainly performed at fog layer 1, as well as some basic data processing and data preservation actions. The fog layer 2 can enhance the data processing and data preservations capabilities of level 1 by providing higher computing capabilities. And finally, the cloud layer will be the responsible of a more complex and more sophisticated data processing over a much broader set of (presumably historical) data, as well as the responsible for permanent data preservation.

In the following subsections the functionalities of each data lifecycle block in this architecture are described and, then, we discuss the advantages of our model.

A. Data acquisition

Data acquisition is mainly performed at fog layer 1. In fact, all sensor devices (such as the sensors network deployed throughout the city, but also surveillance cameras or sensor data from smart phones) are part of fog nodes at this level according to their respective location. So most data are collected at fog layer 1. There can eventually be some additional data collected from web services or third party applications, and these will be collected at cloud level (where web services run), but these will be a small data set compared to the vast volumes of sensor generated data.

As long as the data are being collected, the following phases from the data acquisition block can also be performed at fog layer 1, where a reasonable amount of computing resources is available. For instance, the data filtering phase can apply filters to remove redundant data and can apply some data aggregation techniques to further reduce the amount of data to be managed. Data quality can also be implemented at this fog layer, assessing and guaranteeing higher data quality. And data description can be performed in order to tag data according to the city business model considered, for instance, timing information (creation, collection, modification, etc.), location positioning (city, country, GPS coordinates), authoring, privacy, and so on.

Data collected at fog layer 1 will be periodically moved upwards to layer 2, and data collected at layer 2 from a set of fog nodes at layer 1 will be combined and periodically moved upwards to the cloud level, which will collect the whole data set from the city. Note that data al fog layer 1 can be immediately used at this same level (real-time data), so there is not any need to move urgently these data to higher levels and, therefore, the frequency for the periodical upwards data movements can be strategically decided in order to accommodate it to the network traffic.

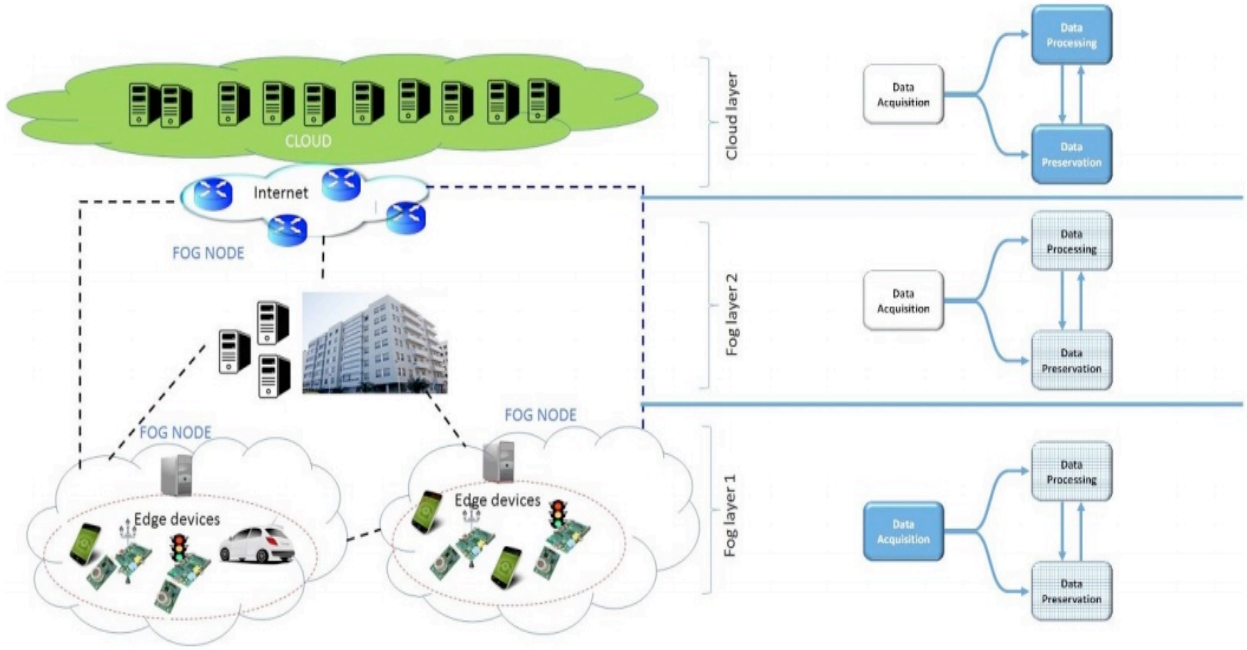


Fig. 5. Mapping of the SCC-DLC model onto the F2C architecture.

B. *Data storage*

Data are generated at fog layer 1, but gradually moved upwards to the fog layer 2, and upwards to the cloud layer, where they will be permanently preserved. So, in fact, the F2C hierarchy acts as a reversed memory hierarchy, where data are created and the lowest cache level (fog layer 1) and moved upwards to main memory (cloud layer) instead of being created at main memory and moved to lower cache levels of the memory hierarchy.

Data generated at fog layer 1 will be temporarily stored at this level, allowing real-time applications an instant access to these data. The smart city business model can decide the amount of temporal data that can be stored at this level, as well as the frequency of updating to upper levels. Similarly, data gathered at fog layer 2, consisting of data received from several fog nodes at layer 1, will be temporarily stored at this level 2. This will make up a set of less recent data (as it has been received after some period of time) but from a broader area, comprising the combination of the respective fog nodes' areas at layer 1. Finally, data will be permanently preserved at cloud layer, unless any expiry time is defined.

The different phases included in the data preservation block will be mainly executed at the cloud level, where the permanent storage is performed. Note that these phases are not urgent and, therefore, their execution can be delayed to the time in which data are received to the cloud layer. This is the case of the data classification phase, responsible for classifying and ordering data before storing, and eventually implementing the appropriate techniques for data versioning, data lineage or data provenance. And the data dissemination phase, responsible for providing a user interface for public or private access to stored data, and responsible for implementing any protection, privacy or security policies according to the city business requirements.

C. *Data processing*

Data processing can be performed at any F2C layer, according to the requirements of the application or service. For instance, critical real-time services will be executed at fog layer 1 in order to have a faster access to the (just generated) real-time data. Note that accessing data locally inside the boundaries of a fog node is much faster than moving the data to a centralized cloud data center and, after, reading these same data from the cloud to the local node.

Alternatively, deep computing complex applications will be executed at the cloud layer. Note that *i*) in the cloud the computing resources are unlimited and, *ii*) the data set of a high performance computing application will presumably be very large and, therefore, be part of the historical data set stored at the cloud layer. Note that in this case, where computation requires very high capabilities, adding more latency to the first access to data will not be significant in the overall performance.

For the other applications, they will be executed at the lowest fog layer that provides the required computing capabilities and the lowest fog layer that contains the required data set. As a general rule, the closer the layer, the faster responses times. An additional consideration in this case is when the required data is not present in the current fog node at layer 1, but can be accessed from either a node at a higher layer or a neighbor fog node at the same layer 1. This option may eventually be considered and solved using some sort of cost model to estimate the effects of both cases and proceed according to the lowest cost.

D. Advantages of the F2C data management model

The most obvious advantages of this F2C data management model are that it can benefit from the combined advantages of both, the cloud and the fog computing technologies. This is, high computing and storage capabilities from the cloud layer, and reduced network traffic and communication latencies from the fog layers. However, some additional advantages can be obtained from this hierarchical and distributed model, as listed below:

- Real-time data accesses are much faster than in a centralized architecture. This higher speed is not only due to the reduced communication latencies, but due to the fact that accessing data from a centralized system requires the data to be moved first to the cloud, classified and stored there, and then moved back to the edge. So two times data transfer through the same path.
- By reducing the data transmission length, the security risks and the probability of communication failure are reduced as well and, additionally, privacy can be easily enhanced.
- By having the just collected data available at fog layer 1, the network load is drastically reduced because some applications will be able to access these data locally, avoiding several remote data accesses through the network.
- By having the just collected data available at fog layer 1, the transmission to the cloud can be delayed without any performance loss. This allows additional optimization implementations, such as:
  - Performing some data aggregation techniques to reduce the volume of data to be transmitted upwards, without any computational constraint, as data do no need to be sent immediately.
  - Adjusting the frequency of the data transmission in order to use the network in periods when the traffic load is low.
- Traditional centralized systems define a low frequency policy for data collection from sensors in order to reduce the total amount of data to be transmitted in the network. By having the real-time data available at fog layer 1, the data collection frequency can be increased at this level without overloading network load and, therefore, providing more precision and accuracy from the sensed data at no additional cost.

V. Optimizing Data Collection through Aggregation

In this section, we aim at providing some validation for our distributed data management strategy based on a F2C resources management architecture, by estimating the effects of some basic data aggregation techniques and comparing them with a real centralized cloud system, named Sentilo, which manages the municipal open data from the city of Barcelona [14].

A. Data Aggregation

Data Aggregation provides a splendid facility as part of data management to do some kind of processing for gathering, reducing, mixing, or presenting information somehow as a summary [15]. The main objective of data aggregation techniques is reducing the amounts of managed data, and can be obtained through diverse techniques, such as data combination, data redundancy elimination, data compression, bandwidth reduction or power consumption reduction, just to name a few.

Recently, data aggregation has been tailored with the concepts of data and information mining progression, business demands and human analysis techniques, where data must be explored, collected, and presented in a report-based and shortened format in their networks [16]. There are some different view to do data aggregation in theoretical and practical scenarios. Traditional views concentrated to specific network devices and resources such as Wireless Sensor Networks (WSN) to manage data aggregation approaches [17-19]. The other view extends the previous view to go beyond ubiquitous and distributed scenarios (instead of focusing on specific devices and network) such as big data [16], cloud and distributed computing [16, 20], web technologies [21, 22], or real-time systems [19].

In WSN environments, sensors are located closer to the regions of the measured phenomena. So, it is very obvious the data aggregation techniques and approaches provide some help in such environments to perform data redundancy elimination, delay reduction, data accuracy (data quality), data security (reliability), traffic management, network scalability and minimizing overhead (bandwidth usage, processing requirements and power and energy wastage) [16, 18, 23]. In [24], the authors propose more sophisticated aggregation algorithms by proposing some soft computing techniques based on artificial neural networks, genetic algorithms, fuzzy logic models, and particle swarm techniques.

In cloud computing environments, cloud computing provides (almost) unlimited, scalable as well as elastic resources. For this reason, cloud computing adopt some data aggregation approaches and techniques to produce high level and sophisticated final outcome. In [25], the authors provide a full data model from sensors nodes to cloud computing environments for a smart city scenario. This model has two main layers which are sensors nodes and cloud computing layers. The sensors nodes collect data from city and pass to the cloud computing layer. The cloud layer is responsible to data collection and aggregation, data filtering (including classification), and data processing (including preprocessing, processing, and decision making).

With respect to distributed data aggregation, a recent survey [20] presents a taxonomy for distributed data aggregation approaches. They propose two main taxonomies, named communication and computation. The communication taxonomy focuses on the communication aspects (including communication/routing strategy and network topology). It is divided into structured (including hierarchical and ring protocols), unstructured (including

flooding/broadcast, random walk, and gossip routing protocols) and hybrid data aggregation approaches. Alternatively, the computation taxonomy encompasses to decomposable functions (including hierarchic, averaging, and sketches basis and principles methods), complex functions (including digests basis and principles methods) and counting (including deterministic and randomized basis and principles methods) data aggregation approaches.

In this work we will apply some basic aggregation techniques as an example to show the facility of our model to use efficiently such kind of optimizations. The data aggregation techniques explored are:

- Redundant data elimination: In this technique we focus on providing a basic yet effective solution to easily reduce the amount of duplicated data collected from the sensors layer. For example, in case of weather measurement, each sensor sends the current temperature measurements, but this type of data is prone to repetitions, so eliminating them may easily reduce such amount of data.
- Compression: As data is collected and transmitted to an upper level delayed, there are some options to accumulate a reasonable amount of data and compute compression, in order to obviously reduce the amount of data transfer.

B. Experimental results

In a previous work [26], we estimated the amounts of data that can be generated in the future (and therefore transmitted through the network to the main cloud data center) in the city of Barcelona, through their data management platform, named Sentilo [14]. In this paper, we will compare these figures with the estimated data that should be transmitted using a F2C data management model as the one described in the previous section.

Barcelona is located in the north east part of Spain. The city has an approximate area of 100 km^2 and has a population of almost 1.62 million people. The city of Barcelona city has ten main districts which covers a total of seventy-three sections [27]. The city is furnished with urban equipment, such as 150.000 lampposts, 40,000 garbage containers. We measured that the future smart city of Barcelona should be covered with 320,925,019 physical sensors, and that they which would produce around 8 GB of data per day [26]. In this measurement, we only focus on five sensor categories of information and services, as defined by the Sentilo platform, namely energy, noise, urban, garbage and parking.

In our experimental results we have explored the data aggregation (redundant data elimination) and data compression approaches to show how our data management model can be easily optimized with respect to the estimated data in the future smart city of Barcelona [26]. In our approach, we described that the data classification (and therefore the data aggregation and data compression methods) can be applied at fog layer 1. According to the current distribution of districts and sections in Barcelona, we estimate that our fog layer 1 can be covers with 73 fog nodes, which is matched with the number of sections in Barcelona. In this case, our fog node covers almost 1 km^2, which is a reasonable fog node size. In addition, the fog layer 2 can be defined as 10 main nodes which are matched with the number of district in Barcelona.

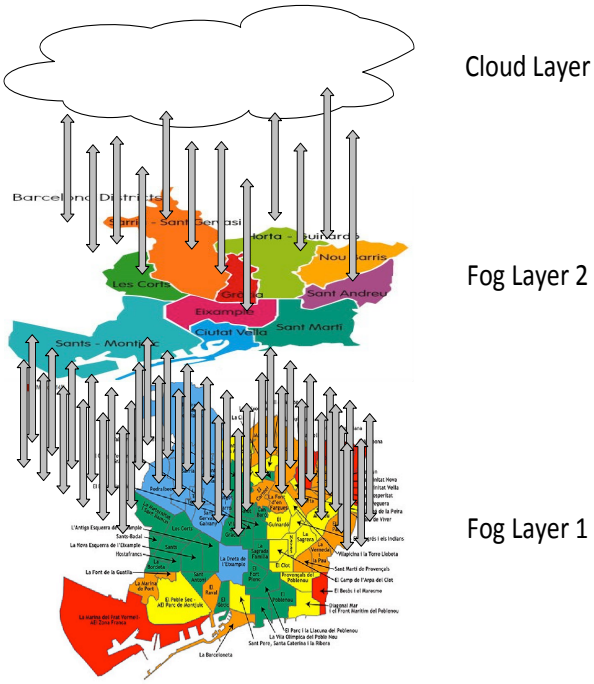


Fig. 6. Representation of the F2C data management in Barcelona.

The data classification phase classifies and organizes all data collected from the different categories of sensors. In our use case, Sentilo provides five categories of information and services which are energy, noise, urban, garbage and parking. Each category is divided into different types of information. For instance, the energy category contains electricity meter, external ambient conditions, gas meter, internal ambient conditions, network analyzer, solar thermal installation, and temperature. The noise category includes has three different types of information. The urban category encompasses to air quality, bicycle flow, people flow, traffic and weather. The garbage category has container glass, container organic, container paper, container plastic, and container refuse. And finally, the parking category has only one type of information.

The two basic data aggregation techniques explored will be implemented at fog layer 1, as explained in previous sections. They are redundant data elimination and compression. Many other data aggregation techniques could be easily applied in this architecture; however, these two basic techniques are enough to illustrate the facility and effectiveness of such optimizations in our model.

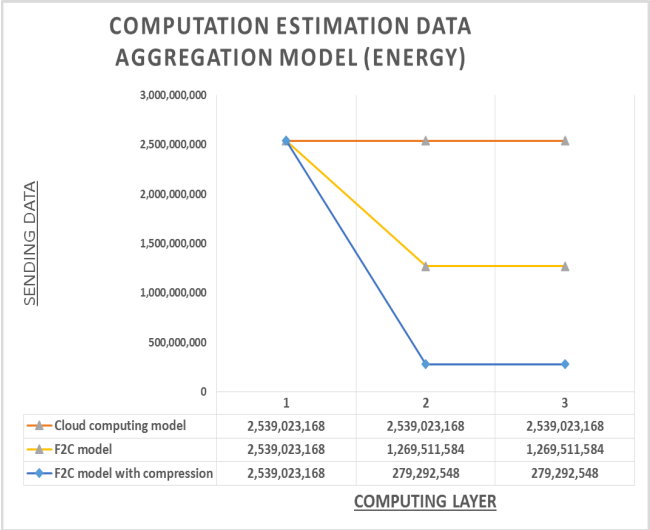
We have applied the aforementioned optimization techniques to the Barcelona smart city as our use case,

extracting real data from the Sentilo platform and computing the amounts of redundant data that can be removed at each measurement, and during one day. Regarding our observation, each category of information generates different magnitudes of redundant information. The results of this firs aggregation technique are shown in Table 1. As we observed, the redundant data for energy, noise, garbage, parking and urban is around 50%, 75%, 70%, 40%, and 30% respectively. This means that we have almost fifty percent efficiency at fog layer 1 after applying data aggregation in the energy category, which has reduced the data traffic from 2, 5 GB to 1, 2 GB per day. The noise category generates almost 0.6 GB which will be reduced to almost 0.1 GB at fog layer 2. The garbage category reduces from almost 0.3 GB to almost 0.1 GB after using data aggregation. The parking category reduces from almost 0.3 GB to almost 0.2 GB. And finally, the urban category has reduced the amount of data from 4, 7 GB to 3, 3 GB.

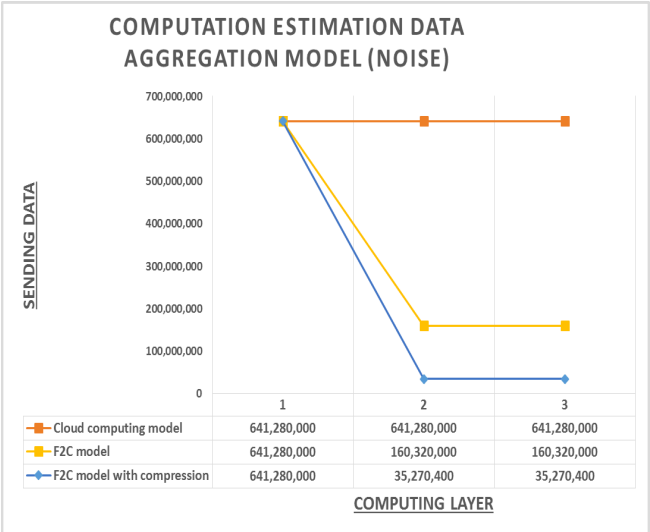
Finally, we can apply the data compression technique after using data aggregation techniques in order to further reduce the amounts of data to be transferred to higher layers. The Zip format, for instance, is one solution provided by PKWARE in 1989 [28]. In this experiment we have used the Zip format in our model to perform compression at fog layer 1. We have measured that 1.26 GB (1,360,043,206 bytes) have been compressed to 0.281 GB (295,428,463 bytes), achieving a format factor of almost 78% of efficiency. For this reason, as shown in Fig. 7, the amount of data in the energy category has been reduced from 2.5 GB to 0.27 GB after compression. The noise data has been reduced from 0.64 GB to 0.03 GB after applying data compression ///.In the garbage category, the amount of data has been reduced from 0.36 GB to 0.07 GB. The total parking data has decreased from 0.32 GB to the 0.07 GB after doing data compression. And finally, the total amount of urban data has shifted from 4.7 GB to 1.03 GB.

TABLE I. REDUNDANT DATA AGGREGATION MODEL

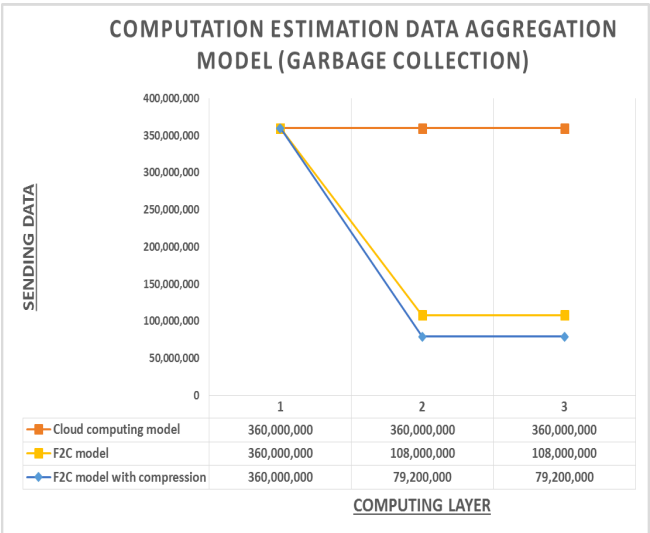
Category of inforamtion	Type of information	Number of sensors	Computing model									
			Cloud		Fog to Cloud (F2C)							
			Sending data (byte)									
			Total amount of data		Total amount of data per transaction				Total amount of data per day			
			by each sensor	Cloud layer	by each sensor	Fog layer 1	Fog layer 2	Cloud layer	by each sensor	Fog layer 1	Fog layer 2	Cloud layer
Energy monitoring	Electricity meter	70,717	22	1,555,774	22	1,555,774	777,887	777,887	2,112	149,354,304	74,677,152	74,677,152
	External ambient conditions	70,717	22	1,555,774	22	1,555,774	777,887	777,887	2,112	149,354,304	74,677,152	74,677,152
	Gas meter	70,717	22	1,555,774	22	1,555,774	777,887	777,887	2,112	149,354,304	74,677,152	74,677,152
	Internal ambient conditions	70,717	22	1,555,774	22	1,555,774	777,887	777,887	2,112	149,354,304	74,677,152	74,677,152
	Network analyzer	70,717	242	17,113,514	242	17,113,514	8,556,757	8,556,757	23,232	1,642,897,344	821,448,672	821,448,672
	Solar thermal installation	70,717	22	1,555,774	22	1,555,774	777,887	777,887	2,112	149,354,304	74,677,152	74,677,152
	Temperature	70,717	22	1,555,774	22	1,555,774	777,887	777,887	2,112	149,354,304	74,677,152	74,677,152
	Total number	495,019	374	26,448,158	374	26,448,158	13,224,079	13,224,079	35,904	2,539,023,168	1,269,511,584	1,269,511,584
Noise monitoring	Noise	10,000	22	220,000	22	220,000	55,000	55,000	768	7,680,000	1,920,000	1,920,000
		10,000	22	220,000	22	220,000	55,000	55,000	31,680	316,800,000	79,200,000	79,200,000
		10,000	22	220,000	22	220,000	55,000	55,000	31,680	316,800,000	79,200,000	79,200,000
	Total number	30,000	66	660,000	66	660,000	165,000	165,000	64,128	641,280,000	160,320,000	160,320,000
Garbage Collection	Container glass	40,000	50	2,000,000	50	2,000,000	600,000	600,000	1,800	72,000,000	21,600,000	21,600,000
	Container organic	40,000	50	2,000,000	50	2,000,000	600,000	600,000	1,800	72,000,000	21,600,000	21,600,000
	Container paper	40,000	50	2,000,000	50	2,000,000	600,000	600,000	1,800	72,000,000	21,600,000	21,600,000
	Container plastic	40,000	50	2,000,000	50	2,000,000	600,000	600,000	1,800	72,000,000	21,600,000	21,600,000
	Container refuse	40,000	50	2,000,000	50	2,000,000	600,000	600,000	1,800	72,000,000	21,600,000	21,600,000
	Total number	200,000	250	10,000,000	250	10,000,000	3,000,000	3,000,000	9,000	360,000,000	108,000,000	108,000,000
Parking Spot	Parking	80,000	40	3,200,000	40	3,200,000	1,920,000	1,920,000	4,000	320,000,000	192,000,000	192,000,000
	Total number	80,000	40	3,200,000	40	3,200,000	1,920,000	1,920,000	4,000	320,000,000	192,000,000	192,000,000
Urban Lab monitoring	Air quality	40,000	144	5,760,000	144	5,760,000	4,032,000	4,032,000	13,824	552,960,000	387,072,000	387,072,000
	Bicycle flow	40,000	22	880,000	22	880,000	616,000	616,000	3,168	126,720,000	88,704,000	88,704,000
	People flow	40,000	22	880,000	22	880,000	616,000	616,000	3,168	126,720,000	88,704,000	88,704,000
	Traffic	40,000	44	1,760,000	44	1,760,000	1,232,000	1,232,000	63,360	2,534,400,000	1,774,080,000	1,774,080,000
	Weather	40,000	120	4,800,000	120	4,800,000	3,360,000	3,360,000	34,560	1,382,400,000	967,680,000	967,680,000
	Total number	200,000	352	14,080,000	352	14,080,000	9,856,000	9,856,000	118,080	4,723,200,000	3,306,240,000	3,306,240,000
Total number		1,005,019	1,082	54,388,158	1,082	54,388,158	28,165,079	28,165,079	231,112	8,583,503,168	5,036,071,584	5,036,071,584



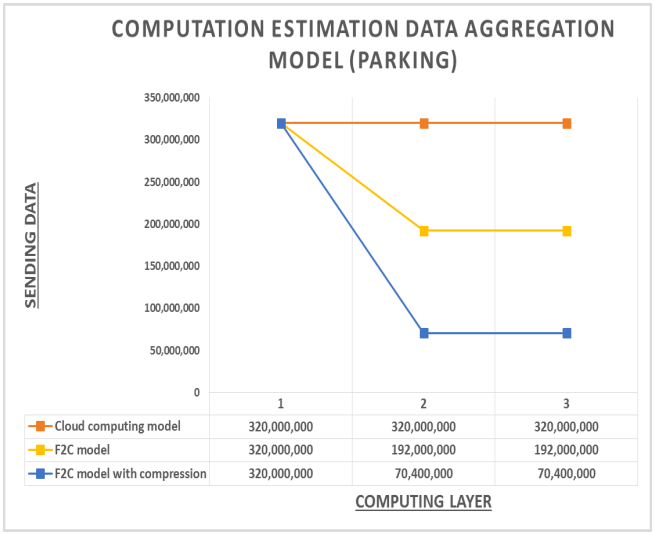
(a) Energy



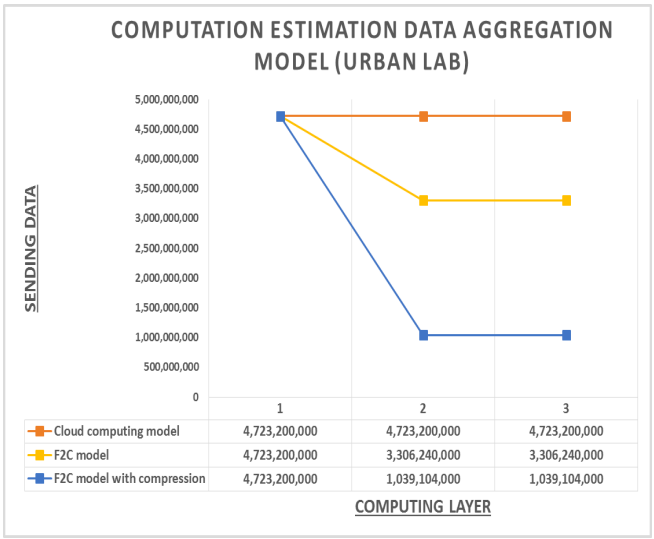
(b) Noise



(c) Garbage collection



(d) Parking



(e) Urban Lab

Fig. 7. Redundant data aggregation model

VI. Related Work

There are several efforts to handle traditional data management technologies, which mostly focus on the concept of Relational Database Management (RDBMS) and the more recent Extract-Transform-Load (ETL) process, for modeling data life stages in the concept of data warehousing environments [29-31]. Plus, the big data paradigm constrained additional difficulties to the traditional data management systems in the recent decades [29, 32]. The Data LifeCycle (DLC) models represent one great solution to focus on planning, organization and management of data beyond any specific technology, system and software, from creation to consumption [33-35]. Several DLC models generated for specific scenarios (like smart city [10, 36]), sciences [34, 37-40] and environments (like big data [29,

32]) have been proposed by many researchers in academia and industries.

With respect to resource management in the smart city environment, there are different trends. In one hand, the centralized view (cloud computing) believes that all physical resources must send the sensed data to the cloud computing data center through the communication network. In this context, the cloud computing environments aim to collect, aggregate and convert data to meaningful information [10, 41]. On the other hand, the alternative option is the distributed view that is used fog computing technology [7, 12]. Fog computing goes beyond the physical devices for further processing and preservation. In addition, authors in [13] propose a F2C computing that combines the cloud computing (centralized view) model with the fog computing (distributed view) model. Although there is few related work about distributed data management, it is not yet mature enough. So in our model we argue that data can be organized and managed at the fog layers (including data preservation and data processing) while using the deep computational performance of the cloud layer.

With respect to data aggregation, most related work concentrate to perform data aggregation in cloud environments [19, 25]. This option applies most sophisticated data optimization (such as data aggregation, data filtering and so on) at the cloud level. However, some reference encouraged the researchers to go beyond data aggregation in real-time services and distributed systems [16, 19]. For this reason, we conclude that there is not any work that proposes data aggregation in the edge of networks in the context of smart cities.

VII. Conclusion

In this paper we have presented a novel architecture for data management in smart cities based on a distributed hierarchical fog to cloud resources management system. The novelties of this approach are diverse:

- This model has been designed to be comprehensive, this is, considering all 6Vs challenges defined in the context of complex big data management;
- This model considers all data life cycle phases, from data production to data consumption, from data acquisition to data processing, data preservation, and an eventual data elimination;
- This model considers any available resource in the city to be part of the system, thus benefiting from the natural advantages of both technologies, cloud and fog computing.

The advantages of this architecture are also numerous. The most obvious advantage is that high computing and storage capabilities from the cloud layer can be combined with reduced network traffic and communication latencies from the fog layers, while enhancing fault tolerance and security and privacy protection. However, by providing such a hierarchical and distributed model, some interesting additional advantages rise:

- Real-time data accesses are much faster than in a centralized architecture;
- The network load is drastically reduced because many data can be accessed and used locally;
- Several aggregation techniques can easily be applied to further reduce the volume of data to be transferred through the network;
- The data transmission frequency can be adjusted in order to use the network in periods of low traffic;
- The data collection frequency from sensors can be increased at no additional cost, thus allowing higher precision and accuracy.

We have also explored the effectiveness of this architecture by exploring two basic data aggregation techniques, which are redundant data elimination and data compression, and compared to a real cloud based system from the smart city of Barcelona. We have shown by applying redundant data elimination that, in some cases, the data reduction rate reaches 75%. Additionally, by applying data compression, the data reduction rate increases to up to 78%. Although many other data aggregation techniques could be easily applied in this architecture, these two basic techniques are enough to illustrate the facility and effectiveness of such optimizations in our model.

As part of our future work we will explore more options related to data aggregation, and continue developing other data life cycle phases of our model, including data quality, data processing, data analysis, data storage, and data dissemination.

Acknowledgments

This work was supported by the Spanish Ministry of Economy and Competitiveness and by the European Regional Development Fund, under contract TEC2015-66220-R (MINECO/FEDER) and by the Catalan Government under contract 2014SGR371 and FI-DGR scholarship 2015FI_B100186.

References

[1] Department of Economic and Social Affairs of the United Nations. (2014). *World Urbanization Prospects, the 2014 revision*. Available at: https://esa.un.org/unpd/wup/.

[2] C. Kyriazopoulou, "Smart city technologies and architectures: A literature review", in *4th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS)*, 2015, pp. 1-12.

[3] X. Ouyang, D. Irwin, and P. Shenoy, "SpotLight: An Information Service for the Cloud," in *36th IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2016, pp. 425-436.

[4] X. Hu, A. Ludwig, A. Richa, and S. Schmid, "Competitive Strategies for Online Cloud Resource Allocation with Discounts: The 2-Dimensional Parking Permit Problem," in *35th IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2015, pp. 93-102.

[5] S. Kannan, A. Gavrilovska, and K. Schwan, "Cloud4Home--Enhancing Data Services with@ Home Clouds," in *31st IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2011, pp. 539-548.

[6] A. Sinaeepourfard, J. Garcia, X. Masip-Bruin*, et al.*, "A data lifeCycle model for smart cities," in *IEEE International Conference on ICT Convergence (ICTC)*, 2016, pp. 400-405.

[7] T. V. N. Rao, A. Khan, M. Maschendra, and M. K. Kumar, "A Paradigm Shift from Cloud to Fog Computing," *International Journal of Science, Engineering and Computer Technology,* vol. 5, p. 385, 2015.

[8] A. Sinaeepourfard, J. Garcia, X. Masip-Bruin*, et al.*, "Estimating Smart City sensors data generation," in *The 15th IFIP Annual Mediterranean Ad Hoc Networking Workshop*, 2016, pp. 1-8.

[9] A. Sinaeepourfard, J. Garcia, X. Masip-Bruin, and E. Marín-Torder, "Towards a comprehensive data lifecycle model for big data environments," in *The 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies 2016 (BDCAT)*, 2016, pp. 100-106.

[10] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, "An information framework for creating a smart city through internet of things," *IEEE Internet of Things Journal,* vol. 1, pp. 112-121, 2014.

[11] B. Tang, Z. Chen, G. Hefferman*, et al.*, "A hierarchical distributed fog computing architecture for big data analysis in smart cities," in *The Fifth ASE International Conference on Big Data*, 2015, p. 28.

[12] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, 2012, pp. 13-16.

[13] X.Masip, E.Marín, A.Jukan, G.J.Ren, and G.Tashakor, "Foggy clouds and cloudy fogs: a real need for coordinated management of fog-to-cloud (F2C) computing systems," *Jounal of IEEE Wireless Communications Magazine,* 2016.

[14] *Sentilo*. Available online at: http://www.sentilo.io

[15] P. Patil and U. Kulkarni, "Delay Efficient Distributed Data Aggregation Algorithm in Wireless Sensor Networks," *International Journal of Computer Applications,* vol. 69, 2013.

[16] N. Karthick and X. A. Kalrani, "A Survey on Data Aggregation in Big Data and Cloud Computing," *International Journal of Computer Trends and Technology (IJCTT),* vol. 17, pp. 28-32, 2014.

[17] J.-Y. Chen, G. Pandurangan, and D. Xu, "Robust computation of aggregates in wireless sensor networks: distributed randomized algorithms and analysis," *IEEE Transactions on Parallel and Distributed Systems,* vol. 17, pp. 987-1000, 2006.

[18] S. Sirsikar and S. Anavatti, "Issues of Data Aggregation Methods in Wireless Sensor Network: A Survey," *Journal of Procedia Computer Science on Elsevier,* vol. 49, pp. 194-201, 2015.

[19] T. He, L. Gu, L. Luo*, et al.*, "An overview of data aggregation architecture for real-time tracking with sensor networks," in *20th IEEE International Parallel & Distributed Processing Symposium*, 2006, pp. 8.

[20] P. Jesus, C. Baquero, and P. S. Almeida, "A survey of distributed data aggregation algorithms," *IEEE Communications Surveys & Tutorials,* vol. 17, pp. 381-404, 2015.

[21] T. Knap and J. Michelfeit, "Linked Data Aggregation Algorithm: Increasing Completeness and Consistency of Data," *Journal provided by Charles University,* 2012.

[22] T. Knap, J. Michelfeit, and M. Necasky, "Linked Open Data Aggregation: Conflict Resolution and Aggregate Quality," presented at the Proceedings of the 2012 IEEE 36th Annual Computer Software and Applications Conference Workshops, 2012.

[23] S. Chhabra and D. Singh, "Data Fusion and Data Aggregation/Summarization Techniques in WSNs: A Review," *International Journal of Computer Applications,* vol. 121, 2015.

[24] H. R. Dhasian and P. Balasubramanian, "Survey of data aggregation techniques using soft computing in wireless sensor networks," *Journal of IET Information Security,* vol. 7, pp. 336-342, 2013.

[25] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the Internet of Things using Big Data analytics," *Journal of Computer Networks on Elsevier,* vol. 101, pp. 63-80, 2016.

[26] A. Sinaeepourfard, J. Garcia, X. Masip*, et al.*, "Estimating Smart City sensors data generation current and future data in the city of Barcelona," presented at the The 15th IFIP Annual Mediterranean Ad Hoc Networking Workshop, 2016, in press.

[27] *Ajuntament de Barcelona*. Available at: http://www.bcn.cat/estadistica/catala/index.htm

[28] PKWARE. *APPNOTE*. Available: https://support.pkware.com/display/PKZIP/APPNOTE.

[29] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *Journals & Magazines on IEEE Access,* vol. 2, pp. 652-687, 2014.

[30] S. Henry, S. Hoon, M. Hwang, D. Lee, and M. D. DeVore, "Engineering trade study: extract, transform, load tools for data migration," in *IEEE Conference on Design Symposium, Systems and Information Engineering*, 2005, pp. 1-8.

[31] S. Kurunji, T. Ge, B. Liu, and C. X. Chen, "Communication cost optimization for cloud Data Warehouse queries," in *IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*, 2012, pp. 512-519.

[32] F. L. F. Almeida and C. Calistru, "The main challenges and issues of big data management," *International Journal of Research Studies in Computing,* vol. 2, 2012.

[33] A. Levitin and T. Redman, "A model of the data (life) cycles with application to quality," *Journal of Information and Software Technology on Elsevier,* vol. 35, pp. 217-223, 1993.

[34] W. K. Michener and M. B. Jones, "Ecoinformatics: supporting ecology as a data-intensive science," *Journal of Trends in ecology & evolution,* vol. 27, pp. 85-93, 2012.

[35] J. Rüegg, C. Gries, B. Bond-Lamberty*, et al.*, "Completing the Data Life Cycle: using information management in macrosystems ecology research," *Journal of Frontiers in Ecology and the Environment,* vol. 12, pp. 24-30, 2014.

[36] M. Emaldi, O. Peña, J. Lázaro, and D. López-de-Ipiña, "Linked Open Data as the fuel for Smarter Cities," in *Modeling and Processing for Next-Generation Big-Data Technologies*, ed: Springer, 2015, pp. 443-472.

[37] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, and C. De Laat, "Addressing big data challenges for scientific data infrastructure," in *IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*, 2012, pp. 614-617.

[38] W. Lenhardt, S. Ahalt, B. Blanton, L. Christopherson, and R. Idaszak, "Data management Lifecycle and Software Lifecycle management in the context of conducting science," *Journal of Open Research Software,* vol. 2, 2014.

[39] J. Starr, P. Willett, L. Federer, C. Horning, and M. L. Bergstrom, "A collaborative framework for data management services: the experience of the University of California," *Journal of eScience Librarianship,* vol. 1, p. 7, 2012.

[40] L. Hsu, R. L. Martin, B. McElroy, K. Litwin-Miller, and W. Kim, "Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities," *Journal of Geomorphology,* vol. 244, pp. 180-189, 2015.

[41] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Journal of Future Generation Computer Systems on Elsevier,* vol. 29, pp. 1645-1660, 2013.