A Data LifeCycle Model for Smart Cities

Amir Sinaeepourfard[1], Jordi Garcia[1], Xavier Masip-Bruin[1], Eva Marin-Tordera[1], Xuefeng Yin[2], Chao Wang[2]
(1) Advanced Network Architectures Lab (CRAAX),
Universitat Politècnica de Catalunya (UPC, BarcelonaTech),
Vilanova i la Geltrú, Barcelona, Spain
(2) College of Electronics and Information Engineering,
Tongji University, Shanghai, China
{amirs, jordig, xmasip, eva}@ac.upc.edu | {yinxuefeng, chaowang}@tongji.edu.cn

***Abstract*— Smart Cities are the most challenging and promising technological solutions for absorbing the increasing pressure of population growth, while simultaneously enforcing a sustainable economic progress as well as a higher quality of life. Several technologies are involved in a potential Smart City deployment, although data are the fuel to achieve the demanded and mandatory smartness. Data can be obtained from multiple sources, in large quantities, and with a variety of formats, therefore, an appropriate management is critical for their effective usage. Data life cycle models constitute an effective trend towards developing an integral and efficient data management framework, from data creation to data consumption and removal. In this paper we present the Smart City Comprehensive Data LifeCycle (SCC-DLC) model, a data management architecture generated from a comprehensive scenario agnostic model, tailored for the particular scenario of Smart Cities. We define the management of each data life phase, and describe its implementation on a Smart City with Fog-to-Cloud (F2C) resources management, an architecture that combines the advantages of both cloud and fog strategies.**

***Keywords*— Smart City, Fog to Cloud (F2C) computing Data Management, Resources Management, Data LifeCycle Model.**

I. INTRODUCTION AND MOTIVATION

It is expected that 70% of the world's population will live in cities and surrounding areas by 2050. Municipal managers have to devise new ways to manage and organize the city in order to mitigate the issues derived from such amount of population, while maintaining or even increasing the citizens' quality of life. Smart cities are the technological solutions designed, not only for absorbing the increasing pressure of population, but mainly for supplying better and more efficient services and processes, promoting a sustainable economic growth and, consequently, providing a higher quality of life to citizens [1, 2].

Smart cities involve different challenging technologies, and demand an exhaustive deployment of computing resources throughout the city (from sensors networks or mobile smart devices, to powerful data centers), all connected through several communication networks using different technologies (wireless sensor networks, 4G, WiFi, Bluetooth, etc.), and all together managed and coordinated by deploying sophisticated frameworks. However, beyond all technologies, the most precious resource for a city to become smart is data.

Data are the fuel for the Smart Cities technology. 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 contextual parameters, or some higher value knowledge extracted from complex data analysis. In fact, Smart Cities constitute the ideal scenario to generate abundant data from any kind of source, such as the own city's sensors, participatory sensing (for instance, sensors integrated in citizens' smartphones), data obtained from social media or any other third party application, surveillance cameras and devices, or any other city resource sensitive to contribute with additional information. For this reason, many efforts from academia and industry are being devoted to create and use data analysis algorithms in order to take advantage of this tremendous abundance of data; 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 this work, we propose a comprehensive DLC model in the context of a Smart City, which combines the advantages of the centralized and distributed management strategies: if a specific (or critical) data is required at real-time from a close location, it is obtained from the source (distributed); however if more complete data set is required, probably least recent, it is obtained from upper levels (thus with higher capacities), more centralized nodes. The model is named the Smart City Comprehensive DLC (SCC-DLC) model. Our proposal is comprehensive in the sense that it explicitly manages all data life cycle stages, from collection to removal, including storage and processing. In addition, other important features, such as data quality and data security, are considered. In our city scenario we assume an IT management policy based on the recently coined Fog-to-Cloud (F2C) computing paradigm [3], and propose a specific DLC model to efficiently manage complex data in the context of a Smart City environment.

The rest of this paper is summarized as follows. Section 2 describes some related work about data management in Smart Cities. Section 3 reviews some of our previous work for Smart Cities' resources management based on F2C computing, and discuss some basic considerations related to data management. In Section 4 we describe the SCC-DLC model, an adaptation of a comprehensive scenario agnostic DLC model into a Smart City scenario, and in Section 5 we show how the SCC-DLC model is managed in a Smart City F2C architecture. Finally, Section 6 concludes the contributions of this work and heads towards the following steps of our future research.

II. Related Work

Smart Cities' technology is a hot topic of current interest for the overall scientific community. There are multiple research directions and technologies related to resources management in the context of Smart Cities, such as Internet of Things (IoT), Internet of Everything (IoE), Service Oriented Architecture (SOA), and so on, which are summarized in [4]. Although most architectures have been proposed for resources management and organization, only few efforts are oriented explicitly on data management.

Most architectures designed with explicit data management schemes are centralized. This means that even though data is collected from different sources spread among the city (such as sensors, surveillance cameras, third party applications, external databases, etc.), data is accessible from a centralized site, usually in the cloud. For instance, in [5] Gubbi et al. propose a cloud centric vision for interaction between private and public clouds, later extended in [6] to propose an information framework for Smart City management. As shown in Fig. 1, the data flow is clearly specified, including four layers namely Data Collection, Data Processing, Data Management, and Data Interpretation. However, note that applications and services obtain the data from a centralized cloud computing platform.

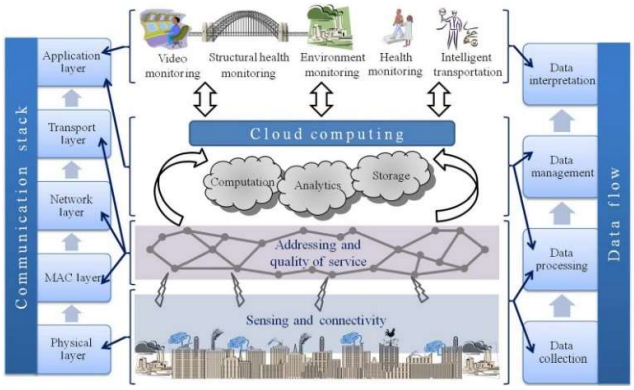

Figure 1. IoT architecture for Smart City [6].

In [7] Rathore et al. basically follow the same patterns but focus specifically on Big Data Analytics. This means that all collected data is preserved in the central cloud, and includes several additional data life cycle steps, such as data aggregation, data filtering, data classification, preprocessing, and decision making. In [8], Pena et al. also propose a Big Data centric framework for smart systems through Internet of Everything (IoE) but, basically, the model is similar to previous models in terms of data flow layers.

Oppositely, few architectures propose a distributed schema for resource allocation and management, using technologies such as Fog Computing [9] or Fog to Cloud Computing [3]; however, none of them has an explicit focus on data management and organization. One exception is found in [10] where Sarkar et al. explicitly address some issues related to data collection at fog level, and distributed temporal data storage also at fog level.

In this work we describe the adaptation of a comprehensive data lifecycle model in a Smart City with Fog to Cloud resources management. In [11] we surveyed most DLC models found in the literature and concluded that although each model is appropriate for its application domain, there is not any comprehensive DLC model according to a set of predefined 6 Vs challenges (namely Value, Volume, Variety, Velocity, Variability and Veracity). Later, in [12] we proposed the Comprehensive Scenario Agnostic DLC model (named COSA-DLC) as an efficient and global data management model to be easily applicable to any scenario. In this paper we present the adaptation of the COSA-DLC model to the specific scenario of a Smart City. The new model is named the Smart City Comprehensive DLC, or SCC-DLC model, for shorter.

III. Scenario Description

The proposed DLC model has been designed for efficient data management and organization in the context of a Smart City. We assume plenty of IT resources and data sources available in a modern city so, in this Section, we describe the particularities of the scenario considered in our model.

A. The city

In a modern city there is an ever unlimited amount of resources and technologies, including computing devices (from smartphones, computers in vehicles, embedded computers, to personal computers or more powerful data centers), other devices to generate data (sensors in the city, sensors in users' devices, surveillance cameras, and so on), communication networks (wired networks, such as Ethernet, optical fiber, or wireless technology, such as 4G, WiFi, RFID, Bluetooth, or any other ad hoc networking technology), and several management platforms to facilitate and optimize users' interaction with the Smart City.

In our envisioned scenario we consider a Fog-to-Cloud (F2C) resources management framework, i.e., a framework that combines both Cloud Computing and Fog Computing technologies. It is widely known that Cloud Computing is a technology that provides ubiquitous and (almost) unlimited resources on demand. It consists on a powerful data center physically located at any part of the world, but easily accessible through Internet. The main limitation of cloud computing is latency, due to its physical distance and the management complexity of such vast amounts of data [13]. Alternatively, Fog Computing [1, 9] is a technology that combines resources at the edge and provides a computing cluster very close to the user or the application. The

computing capacity of the fog is obviously much lower than that of the cloud; however, it provides several advantages, such as resources are local and therefore latency is much lower, issues such as privacy and veracity can be managed more effectively at fog level, network load reduction due to preventing data to be forwarded up to cloud, to name a few.

For this reason, F2C [3] has been proposed to make the most out of combining fog and cloud technologies. As shown in Fig. 2, the F2C architecture is a hierarchical model where devices at the edge are clustered in different Fog Nodes according to their physical location, ending up in a set of Fog Nodes spread along the city (layer 1 of the hierarchy). Each set of layer 1 Fog Nodes are grouped and managed by a more powerful (layer 2) Fog Node, hence building a hierarchical structure of nodes. The node at the highest level corresponds to the cloud. When applied to a Smart City context, the main objective of a F2C resources management strategy must be to organize all available resources in the city, providing a common pool of computing, storage and networking services, and managing them from the fog to the cloud in the most appropriate way according to both their operational features and the services demands.

It is worth mentioning that even though F2C computing is still in an embryonic stage –many design decisions and some technological challenges are yet to be addressed–, the main baselines described above are sufficient to understand the data management model proposed in this paper.

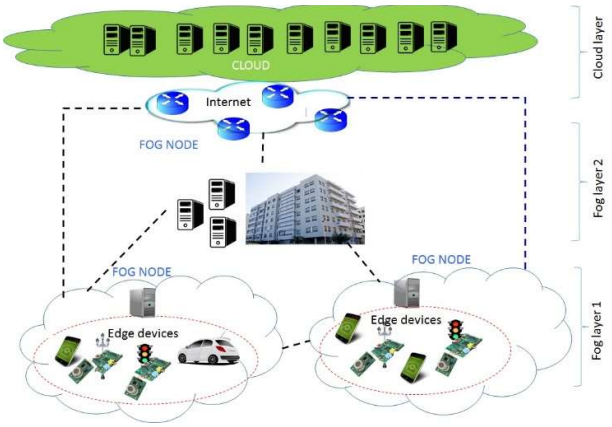

Figure 2. Fog to Cloud (F2C) computing architecture

B. The data

As recognized by many authors, data is the fuel for Smart Cities. Any service willing to provide smartness to a city requires some type of information about it. For this reason, it is important to have access to as many data which covers as much geographical area, in order to ease the development of advanced and sophisticated services.

Data can be obtained in a city from a high variety of sources, including city sensors (public sensors deployed by some city government department), participatory sensing (sensors in mobile users' devices, such as wearables, smartphones, vehicles), private sensing (sensors deployed by private organizations), different kind of cameras (surveillance, cameras in drones), or some other additional information from web services, including information systems from the public administration, users' data from social networks, or other data from local third party applications or corporate databases. In [2] we estimated that in the city of Barcelona 8 GB of data could be generated every day, only considering public sensors' data. This vast amount of data must be collected and managed, and provided for easy and, perhaps, open users' access.

Collected data are accessible for Smart Cities' services usage, 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 low latency applications. Note that real-time data entails some proximity constraints, because the further the data is generated, the more time is required to obtain it, especially in the absence of a direct connection where several intermediate nodes (or platforms) must be crossed. Alternatively, data becomes historical (older data) as long it is accumulated and stored on larger files or databases. In this case, historical data can be considered to be further away (even if physically close) because accessing data from cloud, for instance, requires higher latency. We also consider real-time data (in critical applications) is requested in relatively small sizes, otherwise (if large size) its management would not be that fast. 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.

Fig. 3 shows a graph that illustrates the basic data life cycles, including the aforementioned considerations. When data is created and collected, it can be used immediately (real-time data) for processing, or archived for a later use. When archived data is accessed for processing, it is considered (relatively) historical data. Finally, data after being processed can also be archived for a later use. In this case, this data can be considered to be either higher value data, more mature data, or more processed data.

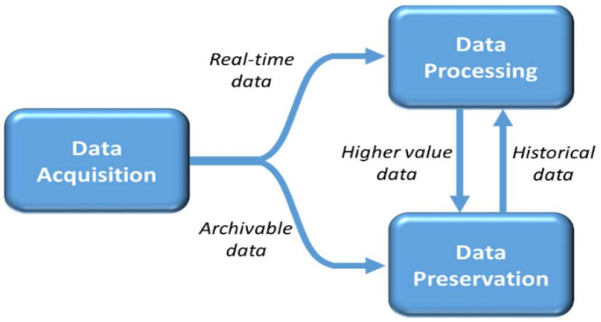

Figure 3. Basic data life cycle.

IV. The COSA-DLC Model For Smart Cities

As explained before, we already proposed in [12] a Comprehensive Scenario Agnostic DLC model (COSA-DLC), as an efficient and global data management model to be easily tailored for any scenario. In this section we tailor the COSA-DLC model to a Smart City scenario, turning into the Smart City Comprehensive DLC (SSC-DLC) model.

The original COSA-DLC model consists of three main blocks representing the main data cycles (as seen in Fig. 3), each built upon a set of basic phases, as shown in Fig. 4. The COSA-DLC model is scenario agnostic, meaning that it has been designed for completeness and it is not specific for any particular scenario. Thus, adapting the COSA-DLC model to a Smart City scenario means choosing the appropriate subset of phases required to provide the desired city services.

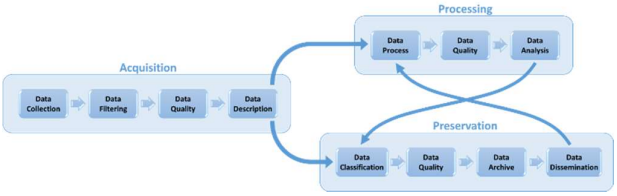

Figure 4. The COSA-DLC model.

In a Smart City, one of the most important tasks is Data Acquisition, because the more information collected from the city, the more sophisticated services can be provided (as long as this data is verified and with quality enough). For this reason, all phases of the Data Acquisition block are critical in the model adaptation. Similarly, Data Preservation and Data Processing are also important tasks in Smart Cities, because many smart applications and services depend on historical data (or accumulated data) obtained in the city and their corresponding processing.

Fig. 5 shows the Smart City Comprehensive DLC (SCC-DLC) model, the COSA-DLC model adaptation to our Smart City scenario, according to the aforementioned considerations. In this case, the adaption has been made by just removing the Data Quality phases from the Data Processing and Data Preservation blocks. The reason is that, in this scenario, all data entering the system comes from the Data Acquisition block where Data Quality has already been checked and assured. For this reason, both the SCC-DLC and the original COSA-DLC are practically the same, meaning that data life cycles in a Smart City are complex and comprehensive. As in the COSA-DLC, the SCC-DLC consists also of three main blocks, as described below.

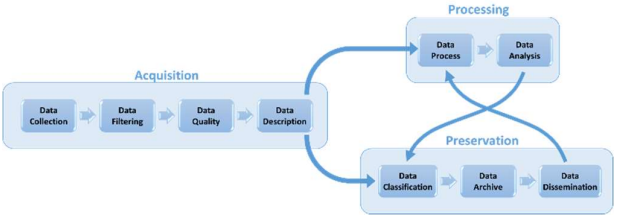

Figure 5. The SCC-DLC model, a COSA-DLC adaptated to a Smart City.

A. Data Acquisition

The Data Acquisition block contains all phases defined in the original comprehensive COSA-DLC model. Their management is described as follows:

Data Collection, responsible for:

- Collecting data directly from physical devices spread along the city, such as sensors, surveillance cameras, users' smart phones and vehicles, and so on.
- Collecting data indirectly from other city sources, for instance, data created in city's local business or public institutions, and offered to the city as open data for smart services.
- Exploring and discovering new data sources that may extend the available data scopes at the city.

Data Filtering, responsible for:

- Applying some methods for data optimization, such as data filtering, data aggregation, data compression, data polishing, and so on. They are intended to optimize the volume of data managed in the system.
- Classifying or sorting data in order to provide enhanced performance. The actual classification will depend on the city's business model.

Data Quality, responsible for:

- Checking the data quality level (namely Quality Control) according to different techniques and algorithms. The particular quality methods required will depend on the city requirements.
- Discarding or repairing low quality data, according to the city's requirements and policies. In case of continuous failure, the data source could be blocked.
- Monitoring the quality of data flows and, in case of continuous failures, proceed according to the provided policies (namely Quality Assurance).

Data Description, responsible for:

- Tagging data with additional description for optimized future retrievals.
- Any metadata considered in the business model can be used, such as timing information (creation, collection, modification, etc.), location positioning (city, country, GPS coordinates), authoring, privacy, and so on.

A user interface for accessing just collected data (i.e., real-time data) should be considered at the end of any of these phases. If *the most recently possible* real-time data was preferred, lacking any quality control, then the interface would be in the Data Collection phase. Alternatively, if quality data is important for the city services, even real-time and critical, then the interface could be in the Data Quality or Data Description phases.

B. Data Preservation

The Data Preservation block contains all phases defined in the original comprehensive COSA-DLC model except the Data Quality phase. The reason is that in the context of this Smart City, all stored data come from the Data Acquisition block and, therefore, its quality is granted. The phases' management is described as follows:

Data Classification, responsible for:

- Classifying and organizing data before storing, according to the city's business model.
- Adding some additional metadata regarding storage, such as expiry time, usage and reuse capabilities, security level, and so on.
- And eventually, implementing the corresponding management techniques in order to implement any data versioning, data lineage or data provenance.

Data Archive, responsible for:

- Storing (large sets of) data collected and processed in the city. Data will be stored in temporal sites, distributed along the city, and a selection of data (aggregated) will be permanently stored in the cloud.
- This phase is responsible for the long term preservation, but also responsible for some additional tasks, such as data cleaning according to the corresponding expiry time, or implementing other business related policies.

Data Dissemination, responsible for:

- Providing a user interface for safe private or public access to stored data, and managing data sharing according to the access permissions policies.
- Implementing the protection, privacy and security policies according to the business requirements.

C. Data Processing

The Data Processing block contains all phases defined in the original comprehensive COSA-DLC model except the Data Quality phase. As with the Data Preservation block, the data quality checking is not necessary. Their management is described as follows:

Data Process, responsible for:

- Performing all data processing required in the application or service to convert raw data into some more sophisticated, higher level information, which provide smartness to the service. These processes could include one or several internal steps, such as pre-processing or post-processing, depending on the particular applications requirements.

Data Analysis, responsible for:

- Performing all deep data analysis and data analytics algorithms for extracting knowledge and discovering new insights. Again, the analysis or analytics processes tightly depend on the users' application or service.
- This phase also provides a user interface for accessing the results of data processing of an application or service. Alternatively, processed data can also be considered for archiving and stored.

Note that processed data can be either consumed directly by the end-user, or stored back to the system to allow data re-using and data re-processing.

V. THE SCC-DLC MODEL FOR A F2C SMART CITY

The Smart City F2C resources management architecture, including the SCC-DLC model, is shown in Fig. 6. The architecture is illustrated with three layers, representing Layer-1 Fog Nodes, Layer-2 Fog Nodes (any other number of upper level layers could be considered), and the Cloud.

Fog Nodes in Layer-1 are composed by a set of devices in the edge clustered according to their physical location, including the sensors in that area. Therefore, Fog Nodes in Layer-1 are the main responsible for data collection. They have very limited preservation and processing possibilities, according to their capabilities, but they provide the lowest latencies, so they are supposed to be the best option for real-time applications. Data collected at this level is periodically sent to higher levels, after applying some filtering and aggregation (yet to be decided).

Fog Nodes in Layer-2, or upper, receive aggregated data from Layer-1. They have higher preservation and processing capabilities, although yet limited. This layer is appropriate for deeper processing over a broader data set. Finally, the Cloud is the highest layer in the F2C architecture. It provides the highest preservation and processing capabilities, although latency becomes much higher.

A. Data production and storage

Data is generated from multiple devices in Layer-1. Some data (the most recent) can be stored at this level as cached data (temporary data preservation). The cached data size will depend on the storage capacities of the actual Fog Nodes. Data in Layer-1 is periodically sent to upper levels, while still keeping some cached copies for fast access.

Similarly, data in Layer-2 and upper levels, constitute a hierarchy of higher level data caches, storing temporal data copies of their corresponding broader area, and according to their respective capacities. Data from the highest Fog Node layers are finally sent to the Cloud, where the long term data preservation of historical data can be performed.

B. Data consumption

Applications and services are launched in the scope of Fog Nodes. If processes are real-time, which usually require the newest data from nearby areas, and computation is limited (otherwise they would not be real-time), then they can be executed in the same Layer-1 Fog Node. If processes require data not stored (or cached) in the current Fog Node, then they are moved upwards to higher layers –owning less recent data but from broader areas–, until the node scope covers the required data. The Cloud, at the highest level of the hierarchy, contains all accumulated data from the whole city, so this would be the last stair of the chain.

CONCLUSIONS

In this paper we have presented a tentative architecture for a comprehensive data management model particularly tailored to a Smart City with Fog-to-Cloud (F2C) resources management. The main advantage of F2C management is

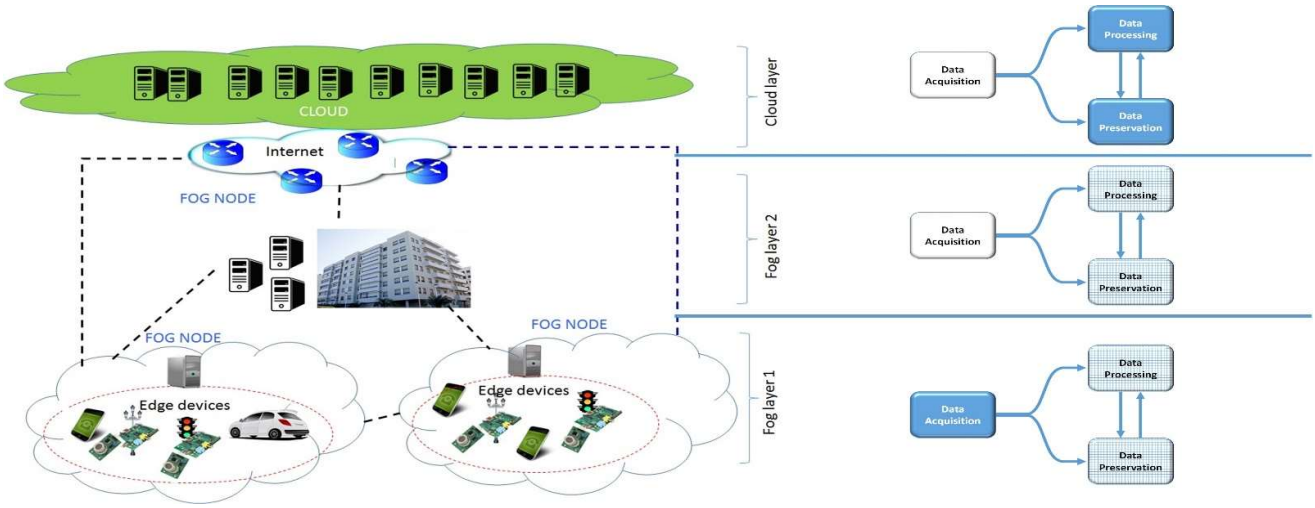

Figure 6. Illustration of the SCC-DLC model, a COSA-DLC adaptated to a Smart City.

that it combines both, the low latency and enhanced data privacy of fog technology, with the computing capacity of cloud technology. The data management model, referred to as the SCC-DLC, is an adaption of the Comprehensive Scenario Agnostic DLC model, which has been proved to be complete according to the 6Vs challenges (Value, Volume, Variety, Velocity, Variability and Veracity), easily adaptable to any particular scenario, as well as correctly suitable to address some additional data issues, such as data quality and data security. The model considers data during their whole data life cycles, from production to consumption and cleaning, including storage and processing.

The contributions of the presented research are diverse. By the one side, it is the first data model architecture for Smart Cities with explicit and global management of all data life cycles. By the other side, it is also the first proposal with explicit focus on data that combines the advantages of centralized and distributed data management strategies in the context of a Smart City, by using F2C resources management. This research is in a preliminary stage. The next steps are to start the real implementation and solve the multiple design decisions yet open.

ACKNOWLEDGMENTS

Work 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 grant 2015FI_B100186.

REFERENCES

[1] B. Tang, Z. Chen, G. Hefferman, T. Wei, H. He, and Q. Yang, "A hierarchical distributed fog computing architecture for big data analysis in smart cities," in *The Fifth ASE International Conference on Big Data*, 2015.

[2] A. Sinaeepourfard, J. Garcia, X. Masip-Bruin, E. Marin-Tordera, J. Cirera, G. Grau, F. Casaus, "Estimating Smart City sensors data generation: current and future data in the city of Barcelona," in *The 15th IFIP Annual Mediterranean Ad Hoc Networking Workshop*, 2016.

[3] X. Masip-Bruin, E. Marín-Tordera, 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," in *IEEE Wireless Communications Magazine, October* 2016.

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

[5] 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, 2013.

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

[7] 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, 2016.

[8] P. A. Pena, D. Sarkar, and P. Maheshwari, "A Big-Data Centric Framework for Smart Systems in the World of Internet of Everything," in *The 2015 International Conference on Computational Science and Computational Intelligence (CSCI'15)*, 2015.

[9] 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.

[10] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the Suitability of Fog Computing in the Context of Internet of Things," *Journal of IEEE Transactions on Cloud Computing,* 2015.

[11] A. Sinaeepourfard, X. Masip-Bruin, J. Garcia, and E. Marín-Tordera, "A Survey on Data Lifecycle Models: Discussions toward the 6Vs Challenges," Technical Report (UPC-DAC-RR-2015-18).

[12] A. Sinaeepourfard, J. Garcia, X. Masip-Bruin, and E. Marín-Tordera, "A comprehensive scenario agnostic Data LifeCycle model for an efficient data complexity management," in The IEEE 12th International Conference on eScience (IEEE eScience 2016), 2016.

[13] 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, 2015.