A Comprehensive Scenario Agnostic Data LifeCycle Model for an Efficient Data Complexity Management

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

Abstract— **There is a vast amount of data being generated every day in the world, coming from a variety of sources, with different formats, quality levels, etc. This new data, together with the archived historical data, constitute the seed for future knowledge discovery and value generation in several fields of eScience. Discovering value from data is a complex computing process where data is the key resource, not only during its processing, but also during its entire life cycle. However, there is still a huge concern about how to organize and manage this data in all fields, and at all scales, for efficient usage and exploitation during all data life cycles. Although several specific Data LifeCycle (DLC) models have been recently defined for particular scenarios, we argue that there is no global and comprehensive DLC framework to be widely used in different fields. For this reason, in this paper we present and describe a comprehensive scenario agnostic Data LifeCycle (COSA-DLC) model successfully addressing all challenges included in the 6Vs, namely Value, Volume, Variety, Velocity, Variability and Veracity, not tailored to any specific environment, but easy to be adapted to fit the requirements of any particular field. We conclude that a comprehensive scenario agnostic DLC model provides several advantages, such as facilitating global data organization and integration, easing the adaptation to any kind of scenario, guaranteeing good quality data levels, and helping save design time and efforts for the research and industrial communities.**

Keywords — Data LifeCycle, Data Management, Data Complexity, Vs Challenges

I. Introduction and Motivation

Hundreds of terabytes are being generated every second all over the world. Data is created from multiple, different and distributed sources and devices (Smart Cities, IoT, scientific modeling and simulations [1], users' social, professional and everyday activities, etc.), with distinct data types and formats. Such huge flow of heterogeneous and often noisy data is accumulated over the historical data sets, constituting the complex universe of digital data. This data can then be used in different forms (reading, writing, transforming or removing), therefore drawing the life cycle of data. Managing data during its life cycle, i.e., from data creation and collection, to data storing and processing, becomes a complex and challenging issue [2]. The ultimate goal of managing data is to obtain some kind of value, by generating knowledge or discovering new insights. To that end several complex computing and analytical processes are required, where data is the key resource during all its entire life cycle. For this reason, efficient data management and organization is a key issue for an effective extraction of value from data.

Data LifeCycle (DLC) models provide an effective solution for data management. They provide a high level framework to plan, organize and manage all aspects of data during their life stages, from data generation to knowledge extraction [2-5]. DLC models clarify all phases of data life, from production to consumption, define the policies for each phase and the relationship between phases [3]. Normally, a DLC model is usually targeted to a particular scenario, addressing its specific requirements and challenges. Consequently one effective DLC model can be defined for any science, scenario or use case [2, 5]. The benefits of designing a DLC model are i) easing for planning and handling complexity of data management in all data life stages [3, 4]; ii) preparing data products ready for end-users, matching the expected constraints and efficiency [2-4]; iii) showing elucidate the quality level of data, removing any kind of waste and noise [3]; iv) illustrating a sequence of any essential activities related to data life [3]; and v) helping designers to create sustainable software [6, 7].

In a previous work [8], we surveyed most related DLC models and evaluated qualitatively each of them in terms of a set of 6Vs challenges, namely Value, Volume, Variety, Velocity, Variability and Veracity. Although each DLC model successfully addresses its specific set of requirements and challenges, and each DLC model is targeted to its particular field of application, we concluded that: i) there is no DLC model completely addressing all 6Vs challenges, and ii) there is no DLC model to be considered comprehensive, in the sense that could be useful to manage data in any scenario or scientific field. For this reason, in this paper we present the design of a comprehensive DLC model that successfully addresses the 6Vs challenges, and that can be easily adapted to manage any scenario and science. The proposed comprehensive DLC model can be understood as an abstract model, i.e. it has been designed agnostic to any particular context and scenario. But because of its completeness, it can easily address the requirements and challenges of any specific scenario.

This rest of this paper is organized as follows. Section 2 reviews the main existing DLC models and highlights their limitations. Section 3 describes our comprehensive model proposal for managing data complexity, i.e., the comprehensive, scenario agnostic, DLC (COSA-DLC) model.

In Section 4, the comprehensive DLC model is evaluated with respect to the 6Vs challenges as benchmark test. And finally, Section 5 highlights the contributions of this model and concludes the paper.

II. Related Work

Research on data management and analysis has traditionally been oriented to the context of Relational Databases Management Systems (RDBMS) and the recent Extract-Transform-Load (ETL) process for modeling the typical data life cycles in data warehousing [9-11]. However, the advent of Big Data imposes additional difficulties to the traditional data management and analysis systems [11, 12]. DLC models are conceived to manage data beyond these systems, and consider diverse, heterogeneous, and large volumes of data, applicable to any scenario or science, throughout all their data life stages.

Several DLC models have been designed targeted to particular scenarios and addressing specific requirements and challenges. One of the first authors to define DLC were Levitin and Redman in [3]. Later, authors in [13] introduced the Data LifeCycle Management (DLM) as a policy based approach to clarify and organize the flow of data throughout all phases of the data life cycle. New generation intensive data sciences, such as eScience, are interested to manage large and heterogeneous amounts of data through Scientific Data LifeCycle Management (SDLM) [14]. And more recent, some research in the field of Big Data recognize the importance of DLC models in data complexity organization [11, 12].

In a recent survey [8], we evaluated a number of public DLC models. We observed that some models concentrated only on a limited number of data stages, such as data discovery [2], data sharing [15], data security [16], data quality [3, 6]. We also realized that each DLC model was specifically designed to address a specific science, such as ecology in [4, 5], library and information science in [17], or geomorphology in [18], or to address special scenarios, such as Smart Cities in [19, 20], support of sciences in [7, 14], etc. Finally, we evaluated the completeness of each DLC model by analyzing to what extent the model was addressing the 6Vs challenges that we found convenient for DLC models. Although all DLC models were able to address most challenges, we found that none of them was addressing all 6Vs challenges.

As a summary, we can observe that all existing DLC models have the following limitations:

- As they have been designed for a particular scenario, they cannot be easily adaptable to any new scenario;
- As they have been designed according to specific requirements, most of them do not always include all stages of the data life cycle; and finally,
- There is not any DLC model ready to address completely all challenges included in the 6Vs.

For this reason, in this paper we present a comprehensive, scenario agnostic, DLC model that overcomes all these limitations and, therefore, can be easily adaptable to any scenario and science and, in addition, is able to address successfully the set of 6Vs challenges.

III. Proposing a Comprehensive Scenario Agnostic Data LifeCycle Model

A comprehensive scenario agnostic DLC (COSA-DLC) model provides all aspects of data management and organization, including data collection, data process, data storage, and other issues regarding data quality, data veracity, and data security, among others. In addition, a COSA-DLC model can be customized and fitted to any science and scenario easily and quickly to achieve specific requirements and guaranteeing high level of data quality.

Some potential advantages of a COSA-DLC model are: i) managing and organizing global datasets for any future data discovery, integration, and processing; ii) providing easy customization and adoption to any science or scenario; iii) improving data quality levels in any specific context, and; iv) eliminating any additional waste and effort for designers, including data, software and system designers, to design their appropriate and efficient architecture.

Our proposed COSA-DLC model is organized in three main blocks, each one further developed into a set of more detailed phases, covering main actions related to the data management and organization for the whole data life cycles. In addition, each phase is specified in terms of the Data Lifecycle Management (DLM), which defines the phase's policies and actions, and the interrelation among phases.

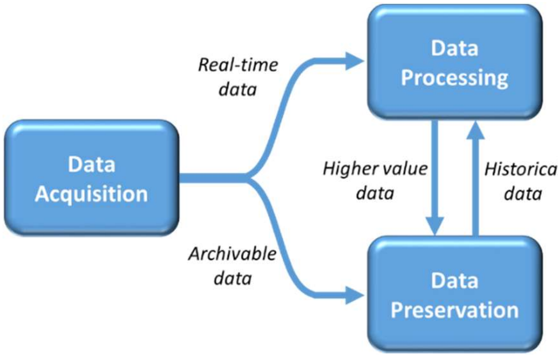
A. Main blocks in the COSA-DLC model

We propose a modular design for the COSA-DLC model structured into three blocks (see Fig. 1), Data Acquisition, Data Processing and Data Preservation. These blocks are responsible for collecting and organizing high level data quality for further processing and storing purposes. This modularity eases the process of adaption to specific scenarios by tailoring the architecture to the specific scenario demands.

Data is gathered into the system through the Data Acquisition block, which collects data from different sources, assesses quality, and tags it with any additional description required in the model. Data can then be processed or stored. The Data Processing block is responsible for performing any data to information/knowledge/value transformation, through complex analysis and/or analytical techniques. Processed data can be used by end users or stored for future reuse. Finally, the Data Preservation block is responsible for data archiving, storing high quality data (curated in the acquisition and/or processing blocks). This data can then be prepared for publication or dissemination, and used by end users or in further processing steps.

The data flow is as follows. Data is created and collected through the Data Acquisition block. If data is immediately processed, this is assumed real-time data; otherwise, if stored, it is considered archivable data. Note that all or part of the processed data can also be preserved, and vice versa, i.e. these two data sets are not exclusive. When archived data in the Data Preservation block is used for processing, this is considered historical data. So the Data Processing block is able to use both real-time and historical data for processing. Finally, the results of data processing can be stored back through the Data Preservation block: this data is considered higher value data.

Fig. 1. Main structure of the comprehensive DLC model.



B. *The COSA-DLC model*

The proposed COSA-DLC model is organized in three main blocks, as described in the previous section. This set of blocks includes a sophisticated set of phases implementing all required tasks to make comprehension, agnosticism and adaption true. Thus, as shown in Fig. 2, the Data Acquisition block is developed in four phases, the Data Processing block is developed in three phases, and the Data Preservation block is developed in four phases. The description of all functionalities and activities of each phase, together with the relationship among phases, is called the Data LifeCycle Management (DLM), and is presented next.

The Data Acquisition block is made by the following four phases, namely Data Collection, Data Filtering, Data Quality and Data Description:

- The Data Collection phase aims to collect data from all sources and devices, according to the business requirements and scientific demands. Specifically, it is responsible of:
  - Collecting data, directly and indirectly, from any valid source, such as basic or complex devices (sensors, smart devices), databases, web-generated data, third party applications, etc.
  - Managing the ranges of valid and trusted sources for data collection.
  - Exploring and discovering new sources for data collection.
- The Data Filtering phase is responsible for performing some basic data transformations in order to optimize the volume of data flowing from the collection to the quality phases. Particular data transformations are specific of the context and business requirements. However, filtering, aggregation, curation, sorting, classification, or compression, are some data transformations that could be considered as well.
- The Data Quality phase aims to appraise the quality level of collected data. It is responsible for guaranteeing both, Quality Control (QC) and Quality Assurance (QA), in particular:
  - Checking the quality level of data and discarding or repairing low quality data, according to the provided policies (QC).
  - Monitoring the quality of data flows and, in case of continuous failures, proceeding according to the provided policies (QA).
- The Data Description phase aims to tag data with some additional information for an optimal future usage. Any available metadata considered in the business model can be used, such as timing (creation, collection, modification, etc.), location or origin (city, country, coordinates), authoring, and so on.

  Once the data has been described appropriately, it can be used for either processing on real-time, or for archiving for future queries over historical data.

The Data Processing block consists of the following three phases, namely Data Process, Data Quality and Data Analysis:

- The Data Process phase provides a set of processes to transform (raw) data into more sophisticated data/information. These processes could include one or several internal steps, such as pre-processing or post-processing, depending on the particular business requirements.

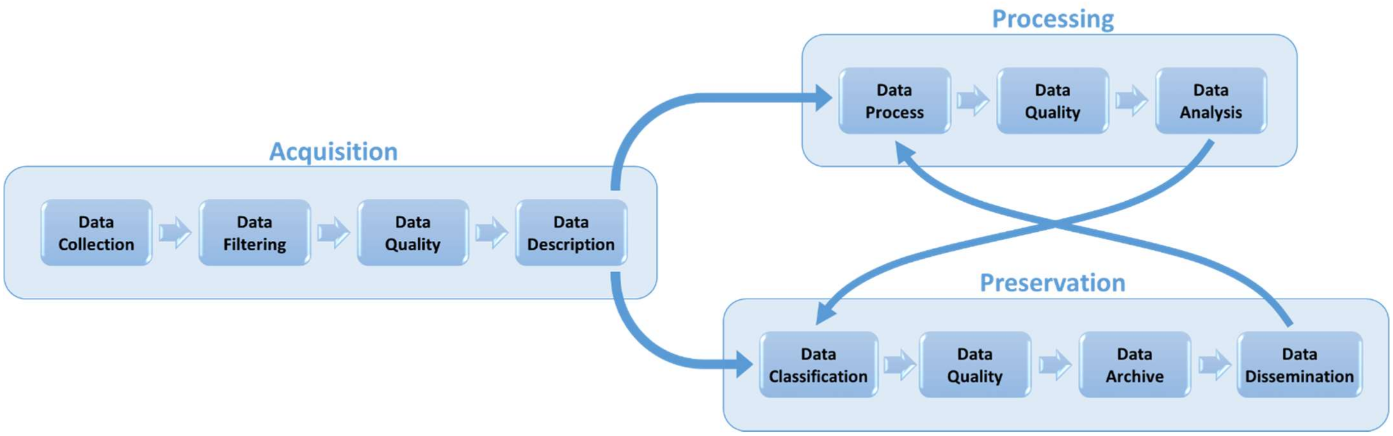
  Data considered for processing can be either real-time, just generated, data (from the Data Acquisition block), or historical archived data (from the Data Preservation block). The output of this phase is considered higher value data, meaning that this data is more mature than the original (raw) input data.
- The Data Quality phase aims to appraise the quality level of processed data. It can check both QC to the output of the processing and QA to the processing procedure.

  This phase could seem redundant or repetitive with respect to the Data Quality phase in the Data Collection block; however, they perform specific checking targeted to the specific life cycles. In addition, in order to provide completeness and guarantee a maximum level of quality, any additional quality appraising is always useful.
- The Data Analysis phase is responsible for developing all data analysis and data analytics for extracting knowledge and discovering new insights.

  This phase is the last step in the procedure of value generation, and it is usually the natural interface with the end-user. Alternatively, this data can also be considered for storing, as part of the Data Preservation block, thus allowing future data re-processing.

The Data Preservation block consists of the following four phases, namely Data Classification, Data Quality, Data Archive and Data Dissemination:

Fig. 2. The proposed DLC model.



- The Data Classification phase aims to organize and prepare data for efficient storage, by applying some optimization, such as classification, arrangement, compression, etc.

  Furthermore, some additional descriptive information could be also attached to this data related to the archiving policies, such as access permissions, privacy, expiry time, or sharing, use and reuse capabilities. In this phase, data provenance or data versioning could be considered.

- The Data Quality phase aims to appraise the quality level of classified data, before storing. It can check both QC to the output of the classification and QA to the preservation procedure.

  Again, note that this phase is specific for the data preservation block, and it is aimed at guaranteeing a maximum level of quality.

- The Data Archive phase aims to store a large set of high quality data in the available permanent or temporary storage resources. This phase must be able to perform long-term preservation over large amounts of data. It is also responsible of some additional tasks, such as data cleaning according to the corresponding expiry time or other business policies.

- The Data Dissemination phase aims to prepare archived data for private or public end-users' access. Any sharing procedures could be managed in this phase to guarantee access permissions, privacy, expiry time, or any other sharing capabilities.

  This phase is the natural interface with the end-user for stored data. Additionally, this data can also be considered for processing, as part of the Data Processing block.

IV. Comprehensive DLC Model Evaluation

Several authors [21-23], propose a list of problems and challenges that should be considered in large and complex data management, often related to Big Data. These contributions have already identified several challenges, such as Value, Volume, Variety, Velocity, Variability, Veracity, and some others (note they all start with V). This set of challenges is known as the Vs challenges. Thus, the Vs challenges depict some strong barriers, difficulties and complexities for data management in different scenarios. However, existing contributions propose to work with different number of challenges, including 3Vs, 4Vs, 5Vs, 6Vs, or 7Vs challenges –perhaps more in the future–. In a previous work [8], we analyzed the appropriate challenges to be addressed in DLC models and considered the aforementioned 6Vs challenges. We also revisited most DLC models and evaluated them with respect to the 6Vs challenges as benchmark test. We concluded that although each model is adequate for its particular purpose, there is no any comprehensive model that addresses the 6Vs challenges completely. In this section we evaluate the proposed COSA-DLC model in order to demonstrate it is certainly comprehensive according to the 6Vs challenges.

1. **Value**: The value challenge refers to the valuable information that can be extracted from (a huge volume of) data, after some processing and/or analysis steps.

   The proposed COSA-DLC model has several merits to address the value challenge. Firstly, the Data Process and Data Analysis phases are precisely included for extracting value from data. Secondly, all Data Quality phases included in the model guarantee a high level of data quality and, therefore, data is more valuable. In fact, any DLC model, just because of its nature, induces to be designed for obtaining any kind of goal, or benefit. So this challenge can be assumed for any DLC model.

2. **Volume**: The volume challenge refers to the huge volumes of data, in any format, that must be considered for management.

   The proposed COSA-DLC model addresses this challenge in the Data Collection phase, as it is prepared to collect data from multiple sources, and in the Data Archive phase, as it should be designed to store large amounts of data. Again, any DLC model is able to address this challenge by definition.

3. **Variety**: The variety challenge refers to the diverse types and formats of the data to be considered, mainly because they provide from different sources with different types.

   The proposed COSA-DLC model addresses this challenge in the Data Collection phase by collecting data, directly and indirectly, from any source, such as basic or complex devices (sensors, smart devices), databases, web-generated data, third party applications, etc. In this phase new sources for data collection are also explored, therefore expanding the candidate sources for data collection. In addition, other phases, such as Data Filtering, Data Description or Data Classification have been included precisely for supporting data organization and classification with high variety of formats.

4. **Velocity**: The velocity challenge refers to the speed rate of data stream generation and the subsequent capability to process it efficiently. This challenge is closely related to performance.

   The proposed COSA-DLC model has been designed for achieving high performance, both during data stream collection and during data processing. For this reason, a specific phase is proposed to manage each of these tasks, namely Data Collection and Data Process. Of course, the final performance will depend on the particular resources deployment, but by considering specific phases, the design helps allocating specific resources in these steps.

5. **Variability**: The variability challenge refers to the possibility that historical data varies its semantic meaning over time.

   The proposed COSA-DLC mode includes a Data Description phase to tag data for future usage and a Data Classification phase where additional tagging could be done, including expiring date. The Data Archive phase also offers the option to implement some data cleaning policies. Finally, in the Data Analysis phase, some data analytics processes could be implemented to analyze and predict eventual context changes.

6. **Veracity**: The veracity challenge can also be understood from two perspectives, according to different authors' interpretations: data quality or data security. Data quality concepts include quality of control (QC) and quality of assurance (QA). And data security prevents datasets from any kind of modification from unsecured and unauthorized sources and devices.

   The proposed COSA-DLC model includes a Data Quality phase in all blocks (Data Acquisition, Data Processing and Data Preservation), guaranteeing both QC and QA. First, all data is checked and if quality is too low (according to the business model), this can be discarded (QC). If a low quality level is reported continuously, the whole process can then be checked, in order to improve procedures for better quality and performance (QA).

   Furthermore, the COSA-DLC model is able to address the data security challenge in different phases. Initially, by guaranteeing sources to be secure and trusted during data collection. Some additional metadata can also be included during the Data Description phase to implement some eventual encryption mechanisms. In addition, during Data Dissemination, different access policies can be defined and implemented to manage accesses, permissions, etc. And finally, a deep security analysis can be performed on data during the data quality phase.

V. Conclusions

In this paper we have presented a comprehensive scenario agnostic DLC (COSA-DLC) model, and demonstrated its completeness with respect to the 6Vs challenges. This model is abstract, as it has not been designed for any specific scenario; however, it can be easily adapted to any particular scenario or eScience, where data complexity has to be addressed. In fact, adapting the COSA-DLC model just requires selecting those phases that are relevant according to the specific scenario requirements

The advantages of the COSA-DLC model are numerous. It is an interesting reference for data engineers that must design a new DLC model for their particular environment. Instead of designing from scratch, they can use this model and easily adapt it to fit their requirements and business model, thus saving design time. In addition, eventual modifications or extensions can also be made, just keeping in mind the original COSA- DLC model. Furthermore, note that during the adaption process some additional facilities are provided, such as the facility to analyze and detect an eventual lack of data quality checking. On the other side, the COSA-DLC model has been proved to address all 6Vs challenges considered in this work. This means that any adaption of this model will also be ready to address these challenges, as long as the particular business model requires such feature.

As part of our future work, we are adapting the COSA-DLC model to a real Smart City scenario with complex and flexible data management requirements. We will implement the different phases of this model and test the performance of the new model.

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] R. Darby, S. Lambert, B. Matthews, M. Wilson, K. Gitmans, S. Dallmeier-Tiessen*, et al.*, "Enabling scientific data sharing and re-use," in *IEEE 8th International Conference on E-Science (e-Science)*, 2012, pp. 1-8.

[2] R. Grunzke, A. Aguilera, W. E. Nagel, S. Herres-Pawlis, A. Hoffmann, J. Krüger*, et al.*, "Managing complexity in distributed Data Life Cycles enhancing scientific discovery," in *IEEE 11th International Conference on E-Science (e-Science)*, 2015, pp. 371-380.

[3] 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.

[4] 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.

[5] J. Rüegg, C. Gries, B. Bond-Lamberty, G. J. Bowen, B. S. Felzer, N. E. McIntyre*, 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.

[6] J. M. Schopf, "Treating data like software: a case for production quality data," in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, 2012, pp. 153-156.

[7] 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.

[8] 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-ANA-2015-1), 2015, submitted for publication.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] M. Rouse. (2010). *Data Life Cycle management (DLM) definition*. Available on: http://searchstorage.techtarget.com /definition/data-life-cycle-management

[14] 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.

[15] A. Burton and A. Treloar, "Publish my data: a composition of services from ANDS and ARCS," in *IEEE 5th International Conference on E-Science (e-Science)*, 2009, pp. 164-170.

[16] X. Yu and Q. Wen, "A view about cloud data security from data life cycle," in *International Conference on Computational Intelligence and Software Engineering (CiSE)*, 2010, pp. 1-4.

[17] 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.

[18] 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.

[19] 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.

[20] 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 Journal on IEEE,* vol. 1, pp. 112-121, 2014.

[21] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of “big data” on cloud computing: Review and open research issues," *Journal of Information Systems on Elsevier,* vol. 47, pp. 98-115, 2015.

[22] R. Rossi and K. Hirama, "Characterizing Big Data Management," *Journal on Issues in Informing Science and Information Technology (IISIT),* vol. 12, 2015.

[23] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Journal on Mobile Networks and Applications,* vol. 19, pp. 171-209, 2014.