Towards a Comprehensive Data LifeCycle Model for Big Data Environments

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

ABSTRACT

A huge amount of data is constantly being produced in the world. Data coming from the IoT, from scientific simulations, or from any other field of the eScience, are accumulated over historical data sets and set up the seed for future Big Data processing, with the final goal to generate added value and discover knowledge. In such computing processes, data are the main resource; however, organizing and managing data during their entire life cycle becomes a complex research topic. As part of this, Data LifeCycle (DLC) models have been proposed to efficiently organize large and complex data sets, from creation to consumption, in any field, and any scale, for an effective data usage and big data exploitation.

Several DLC frameworks can be found in the literature, each one defined for specific environments and scenarios. However, we realized that there is no global and comprehensive DLC model to be easily adapted to different scientific areas. For this reason, in this paper we describe the Comprehensive Scenario Agnostic Data LifeCycle (COSA-DLC) model, a DLC model which: *i*) is proved to be comprehensive as it addresses the 6Vs challenges (namely Value, Volume, Variety, Velocity, Variability and Veracity; and *ii*), it can be easily adapted to any particular scenario and, therefore, fit the requirements of a specific scientific field. In this paper we also include two use cases to illustrate the ease of the adaptation in different scenarios. We conclude that the comprehensive scenario agnostic DLC model provides several advantages, such as facilitating global data management, organization and integration, easing the adaptation to any kind of scenario, guaranteeing good data quality levels and, therefore, saving design time and efforts for the scientific and industrial communities.

CCS Concepts

• Information systems→Data management systems • Computer systems organization→Architectures→Other architectures→ Data flow architectures • Software and its engineering→ Software system structures→ Data flow architectures

Keywords

Big Data, Data LifeCycle, Data Management, Data Organization, Data Complexity, Vs Challenges.

1. INTRODUCTION AND MOTIVATION

A huge amount of data is constantly being produced in the world, turning Big Data as one of the hottest research topics currently. Data are being generated from multiple scientific sources, including Smart Cities, the IoT, scientific modeling, or different big data simulations [1-3]; but also from users' social, professional or everyday activities. These daily fresh data are accumulated over other historical repositories, setting up the vast and complex universe of digital data. Data can then be used in different forms during big data processing (reading, writing, transforming or removing), and then be reused in following processes, therefore drawing the life cycle of data.

An appropriate management and organization of diverse and sophisticated data sets during their entire life cycle, including data generation, data acquisition, data preservation, or data processing, becomes a complex and challenging task [4, 5]. The main objective of data management is to provide easy, efficient and safe access to data sources and repositories, in order to be able to extract any form of value through complex computing and analytical processes over big data sources. For this reason, efficient data management and organization systems are a key topic for an effective data to value generation.

Data LifeCycle (DLC) models have recently been proposed as an effective data management solution that facilitates the organization of data and the extraction of knowledge in complex data systems [4-8]. DLC models define the sequence of phases in the data life, specify the management policies for each phase, and describe the relationship among phases [6]. Furthermore, a DLC model is designed specifically for a particular field and scenario, addressing its private requirements and challenges [5, 8]. The benefits of designing and implementing a DLC model are the following:

- Easing for planning and handling complexity of data management throughout all data life stages [6, 7];
- Preparing data products for end-users access, achieving the expected constraints and efficiency requirements [5-7];
- Providing high data quality level, removing any kind of waste and noise [6];
- Identifying the appropriate sequence of essential activities related to data life [6]; and
- Helping system designers to create sustainable and efficient software [9, 10].

In a previous work [11], we surveyed most DLC found in the literature, and evaluated qualitatively each of them with respect to the 6 Vs challenges, namely Value, Volume, Variety, Velocity, Variability and Veracity. We concluded that although each DLC model had been designed to successfully address their required challenges, no DLC model could be considered comprehensive, in

DOI: 10.1145/3006299.3006311

the sense that no DLC model was addressing all 6Vs challenges completely. In addition, we realized that each DLC model was designed specifically for a particular scenario, and therefore no DLC model was general enough to be easily adapted to a different or new scenario. For this reason, in this paper we describe the Comprehensive Scenario Agnostic Data LifeCycle (COSA-DLC) model, a DLC model designed to address all 6Vs challenges and that can easily be adapted to any scientific scenario for big data management. The proposed comprehensive DLC model can be understood as an abstract model that can easily address the requirements and challenges of any specific area. In [12], we demonstrated the completeness of the model by evaluating it with respect to the aforementioned 6Vs challenges. In this paper, we illustrate the ease of the adaption of the model by including two use cases, where the comprehensive COSA-DLC is adapted to two different scenarios: a smart city context and a digital library scenario.

This rest of this paper is organized as follows. Section 2 reviews the main DLC models found in the literature and highlights their limitations. Section 3 describes our model proposal for data complexity management, i.e., the comprehensive, scenario agnostic, DLC (COSA-DLC) model. Section 4 presents two use cases to illustrate the ease of adaption and utilization of the proposed model. And finally, Section 5 highlights the contributions of this model and concludes the paper.

2. RELATED WORK

Several traditional technologies for data management, such as Relational Database Management Systems (RDBMS) and the recent Extract-Transform-Load (ETL) process, have been proposed for modeling the data life cycles in the context of data warehousing environments [3, 13, 14]. In addition, the advent of big data imposes additional challenges to the traditional data management and organization systems, by including non-structured and heterogeneous data sets [3, 15]. The DLC models go beyond these technologies to provide a more global framework for data management and organization, from data creation to data consumption, which is not related to any particular hardware or software technology or system.

Many DLC models can be found in the literature to tailor specific scenarios and use cases, and which addresses a particular set of requirements and challenges. One of the first formal definition of DLC was provided by Levitin and Redman in [6]. In following works, Rose in [16], defined the concept of Data Life Management (DLM) which represents a policy based approach to depict the data flow among all phases of data life cycles, and describes the implementation of each phase. More recently, research in intensive data sciences, like eScience, are getting interested in organizing and managing massive amount of scientific data with the definition of Scientific Data LifeCycle Management (SDLM) [4]. Eventually, there are some strong efforts in the field of big data among researchers in academia and industries to organize and handle the difficultly and complexity of vast and massive amount of data in big data environments, as part of big data management [3, 15].

Our initial research interests have been oriented towards proposed DLC models in any field and any environment [11]. We surveyed most DLC models found in the literature and realized that each only model addressed some specific data stages, such as data discovery [5], data quality [6, 9], data sharing [17] and data security [18]. In addition, we also found that each DLC model was proposed to address the particular requirements for a particular scientific field, for example, library and information science in [19], geomorphology in [1920 ecology in [7, 8], or to address specific scenarios, such as Smart Cities in [21, 22], or support of sciences in [4, 10]. In addition, we assessed the completeness of all existed DLC models with respect to the 6Vs challenges as a benchmark test. The conclusion is that it seems that each DLC model is a perfect contribution for its specific scenario of application, which has been tailored to cover some subset of the 6Vs challenges, but we highlighted that there is not any model that completely address all 6Vs challenges in global.

As a summary, we can conclude all previously proposed DLC models have the following barriers or limitations:

- No model can be adjusted to a new scenario or environment easily and quickly, because each proposed model has been drawn for a special scenario and environment;
- No model covers all stages of the data life cycles, because each proposed model has been designed for a specific set of data requirements;
- And, no model can effectively and globally address all 6Vs challenges completely.

For these reasons, in this paper we propose a comprehensive, scenario agonistic, DLC model which is able to overcome the aforementioned barriers and challenges and, eventually, can be tailored to any scenario and science quickly and easily and, furthermore, overcome all 6Vs challenges.

3. THE COMPREHENSIVE SCENARIO AGNOSTIC DATA LIFECYCLE MODEL

The comprehensive scenario agnostic DLC (COSA-DLC) model considers all phases of data management and organization, from data acquisition to data preservation and processing, but also includes other fundamental aspects related to data quality and data security, among others. The COSA-DLC model can be easily customized and fitted to any scenario to guarantee the specific requirements while providing high level of data quality and, in addition, has been proved to be comprehensive according to the 6Vs challenges.

Some potential advantages of the 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.

The main organization of the COSA-DLC model is defined in three main blocks, named Data Acquisition, Data Processing, and Data Preservation. Each block, in turn, is further described into a set of more detailed data phases, covering all cycles involved in the data life. 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.

3.1 Main blocks in the COSA-DLC model

The COSA-DLC model is defined as a modular three blocks structure (see Figure 1), with the Data Acquisition, the Data Processing and the Data Preservation main blocks. These blocks are responsible for gathering, storing and organizing data for processing or any other use purposes, while guaranteeing high standards of data quality. This modularity eases the process of adaption to specific scenarios by tailoring this model to the specific scenario requirements.

The Data Acquisition block is the responsible for collecting data into the system, gathering data from different sources, assessing data quality, and tagging this data with any additional description required in the business model. Collected data can then be stored, through the Data Preservation block, or processed, through the Data Processing block. The Data Processing block is the responsible for performing the main big data processing, extracting knowledge or generating additional value, through sophisticated data analysis techniques. The results of the processed data (higher value data) can be delivered to the end users, or stored for future additional data reuse or reprocessing. The Data Preservation block is the responsible for data storage, performing any eventual action related to data curation or data classification. This data is ready for future publication or dissemination, or for further processing.

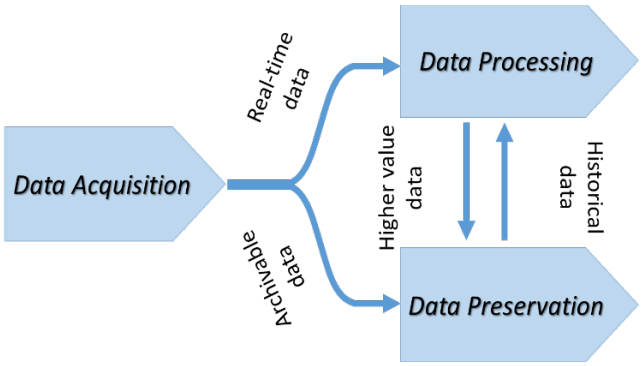

Figure 1. Blocks in the DLC models.

The data flow is the following. When data is created it is collected through the Data Acquisition block (this block is also responsible for data discovery). If data is immediately requested and processed, this is considered real-time data; otherwise if it is preserved, this is considered archivable data. Note that either 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 from the Data Preservation block is required and used for processing, this is considered as some sort of 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 stored back through the Data Preservation block: this data is considered higher value data (with respect to the original data before processing).

3.2 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 Figure 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.

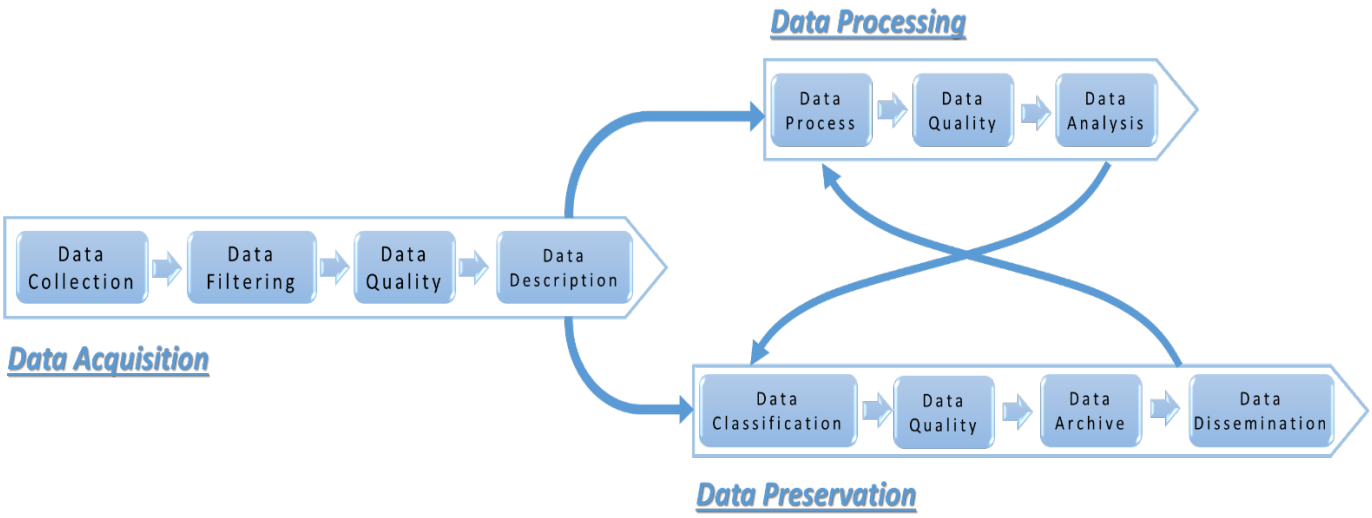

Figure 2. The proposed DLC model.

This phase is the last step in the procedure of value generation, and it is usually the natural interface with 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:

- 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.

3.3 The 6Vs Challenges

Several authors [23-25], propose a list of problems and challenges that should be considered in large and complex data management systems, often related to big data, known as the Vs challenges. The main challenges in big data have originally been described by Gartner through the 3Vs challenges, named Volume, Variety and Velocity, as referenced in [26]. This set has later been extended to the 5Vs challenges, mainly summarized into Volume (huge volume of data), Variety (various data formats), Velocity (rapid generation of data), Value (huge value but very low density), and Veracity (quality and security of data) [27], although it may be considered as 4 +1, since the latter challenge differs depending on the reference (it may be either Variability or Veracity). Recently, there is some effort to show that the challenges can assume 7Vs, including both Variability and Veracity, and adding Visualization as a new challenge. And some other authors propose Volatility, Viscosity, or Virality as additional challenges, although this is perhaps not mature enough to be considered.

After reviewing all definitions about the Vs challenges in big data, in [11], we proposed a 6Vs challenges model, which includes Value, Volume, Variety, Velocity, Variability and Veracity, as the appropriate model to evaluate the comprehensiveness of the different DLC models. Later, in [12], we evaluated the completeness of the COSA-DLC with respect to the 6Vs challenges and concluded this model is global and comprehensive.

4. USE CASES

In this section, we present two different use cases to illustrate how easy is the customization and adaption of the proposed COSA-DLC model to any kind of science or scenario. The first use case adapts the COSA-DLC model for data management in a Smart City scenario. The second use case adapts the COSA-DLC model into a library, which represents a scientific sample.

4.1 A DLC model for a Smart City

In the first example we use the Smart City of Barcelona as a use case to illustrate the COSA-DLC model adaption. Data in the Barcelona Smart City is currently managed through the Sentilo platform [28], a framework that collects data from different sources (mainly sensors, but also other information sources from the city),

organizes and stores it, and provides a public interface to access the datasets, either real-time or historical data1. Figure 3 shows the Sentilo architecture, a middleware providing a unified access to the public data.

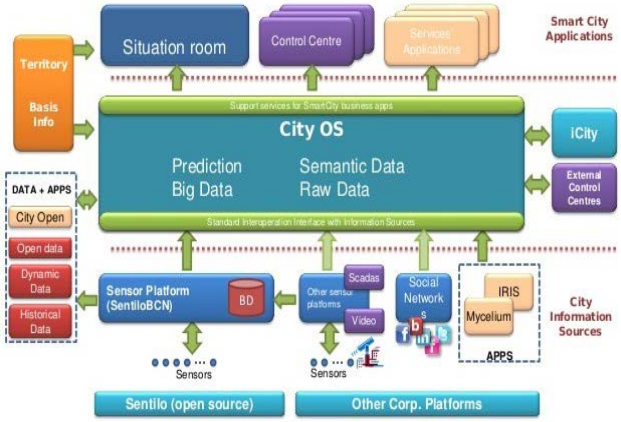

Figure 3. The Barcelona Smart City IT architecture.

In order to adapt the COSA-DLC model to Sentilo, data acquisition, data processing and data preservation should be considered. Our proposal is illustrated in Figure 4. The Data Acquisition block includes Data Collection, Data Filtering and Data Description phases. Note that no quality control is performed in Sentilo as they provide the raw data as they get it, with just some additional descriptive data about dates and positioning. All collected data is archived in the Sentilo databases, and developers can access both real-time and historical data. The Data Preservation block includes Data Classification, Data Archive and Data Dissemination phases. The classification phase organizes information to be stored according to its type and format. And the stored data can then be retrieved through the interface managed in the data dissemination phase. Finally, some services are offered providing more processed data, although these services will be very customer dependent and then, no generalization can be easily made. For this reason, we include a Data Processing phase to allow some basic processing in the DLC model.

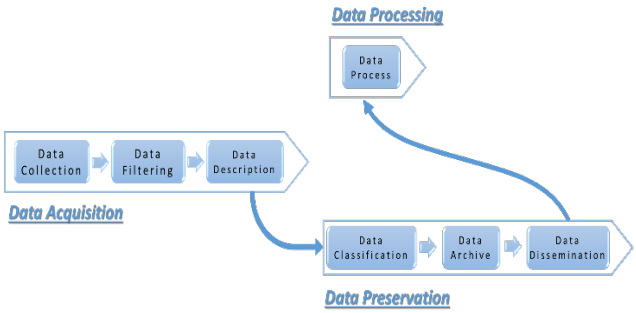

Figure 4. COSA-DLC model proposal for Sentilo.

4.2 A DLC model for a Scientific Library

In the second example we will use the Library of the UPC BarcelonaTech as a use case to illustrate the COSA-DLC model adaption in an eScience field. The library is connected to different university campuses to collect, aggregate and share all digital resources among internal libraries and departments. Several types and formats of information are collected by the library procedures, including books, journals, digital video, doctoral theses, examination records, etc., and all the information is available for online access, under registration and according to any eventual copyright statement [29]. The data in the UPC BarcelonaTech Library is currently managed through the framework shown in Figure 5.

The COSA-DLC model can easily adapt this framework by considering data acquisition and data preservation, as illustrated in Figure 6. The Data Acquisition block includes Data Collection, Data Filtering, and Data Description phases. Note that in a Scientific Library many metadata is required to facilitate the catalog retrieval. The Data Preservation block includes Data Classification, Data Archive and Data Dissemination phases. In this case, some analysis is performed to check if different entries are referencing the same physical document in the quality phase. The dissemination phase provide a basic interface to do advanced searching.

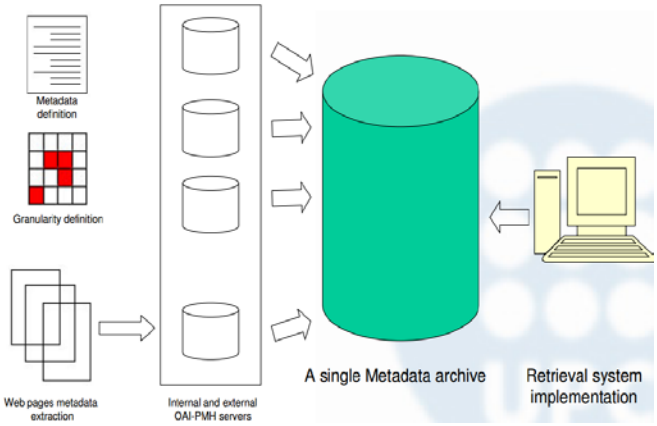

Figure 5. The UPC BarcelonaTech Library architecture.

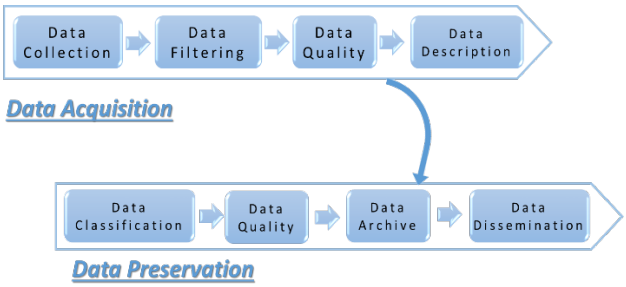

Figure 6. COSA-DLC model proposal for UPC Library.

These two use cases illustrate how easy is to adapt the COSA-DLC model into a variety of different scenarios and sciences. Indeed, adapting the model just requires selecting those phases that can be relevant according to the particular scenario requirements.

5. CONCLUSIONS

In this paper we have described the comprehensive scenario agnostic DLC (COSA-DLC) model, a data management and organization model for complex data systems. This model is abstract in the sense that it has not been designed assuming any particular scenario nor application, but can be easily adapted to address all requirements and constraints of any specific scenario in the fields of big and complex data management. For this reason, we have provided two use cases to show the ease of adaptation to different environments: the Smart City of Barcelona and the Scientific Library of the UPC BarcelonaTech.

The advantages of the COSA-DLC model are numerous. It is an interesting starting point 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 can be assumed, such as the facility to analyze and detect an eventual lack of data quality checking (as shown in the first use case). On the other side, the COSA-DLC model has been proved to address comprehensively all 6Vs challenges assumed in this work. This means that any adaption of this model will also be ready to address these challenges globally and completely, 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 Smart City scenario with hierarchical, complex and flexible data management requirements. We are defining the new data management architecture, and will implement the different phases and test performance for this new model.

6. 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.

7. REFERENCES

[1] J. Wang, Y. Tang, M. Nguyen, and I. Altintas, "A Scalable Data Science Workflow Approach for Big Data Bayesian Network Learning," in Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computing (BDC), 2014, pp. 16-25.

[2] V. Kolias, I. Anagnostopoulos, and E. Kayafas, "A Covering Classification Rule Induction Approach for Big Datasets," in Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computing, 2014, pp. 45-53.

[3] 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.

[4] 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.

[5] R. Grunzke, A. Aguilera, W. E. Nagel, 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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), 2015.

[12] A. Sinaeepourfard, J. Garcia, X. Masip, et al., "A Comprehensive Scenario Agnostic Data LifeCycle model for an efficient data complexity management," in IEEE 12th International Conference on E-Science (e-Science), Baltimore, USA, 2016.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, et al., "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.

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

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

[26] I. A. T. Hashem, I. Yaqoob, N.B. Anuar, et al., "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.

[27] Demchenko, Yuri, et al. "Big security for big data: addressing security challenges for the big data infrastructure." *Workshop on Secure Data Management*. Springer International Publishing, 2013.

[28] A. Sinaeepourfard, J. Garcia, X. Masip, et al., "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.

[29] R. Gómez-Enrich, M. López-Vivancos, M. Mestre-Vidal, J. Prats-Prat, and A. Rovira-Fernández, "Towards the integral management of library collections at the Technical University of Catalonia (UPC)," presented at the 25th IATUL Annual Conference on The International Association of Scientific and Technological University Libraries (IATUL), Krakow, Poland, 2004.