Smart Data LifeCycle as a process cartography

Mohammed EL ARASS, Nissrine SOUISSI
EMI-SIWEB Team, Mohammed V University in Rabat, Morocco
mohammed.elarass@gmail.com nissrine.souissi@gmail.com

***Abstract*—Data management is becoming increasingly complex, especially with the emergence of the Big Data era. The best way to manage this data is to dispose of a data lifecycle from creation to destruction. This paper proposes a new Data LifeCycle (DLC) named Smart DLC that helps to make from raw and worthless data to Smart Data in a Big Data context. In order to do this, we have followed a method which consists firstly in identifying and analyzing the lifecycles from a literature review, and then in defining the phases of our cycle and finally in modeling it. The cycle is modeled as a process cartography resulting from the ISO 9001: 2015 standard to facilitate its implementation within companies. Smart DLC is qualified as a set of management, realization and support processes that could be addressed by an Information System urbanization approach. The advantage of modeling the phases such as processes is to be concerned not only with the technical activities but also with management, which is a major player for the success of the technique.**

***Index Terms*—Big Data, Data lifecycle, Data management, Smart Data, Smart DLC, Cartography, Process.**

I. Introduction

The majority of Big Data projects contain dozens of powerful servers, nested in complex architecture and with many dependencies. However, these solutions have not been able to solve the problem of Big Data because the size of this data is beyond the capacity of conventional database software tools to collect, store, manage and analyze data [1]. Also, these data are too large to be manipulated and parsed by traditional database protocols such as SQL [2].

Indeed, the survey conducted by Capgemini in November 2014 and revealed at the beginning of 2015: only 27% of IT managers interviewed described their Big Data project as a success. This situation sums up a rather catastrophic situation.

A Big Data system must guarantee compliance with the defined Service Level Agreements (SLA), and can only confirm its relevance if it is able to investigate very quickly source problems to repair them.

This situation forces us to advocate smart management of these data called Big through an adequate lifecycle. The lifecycles that have been analyzed in [2] are not all adequate for the Big Data context but based on the ranking found in [2], the lifecycle *Hindawi* [3] has been recommended for companies especially those working in a Big Data context. But this model presents some weaknesses, notably the absence of the *Planning, Enrichment and Destruction* phases and does not deal with sensitive aspects of the Big Data context like *Quality Control and Management*.

In [2], we have identified relevant phases in a Big Data context. In this paper, we propose to design a new data lifecycle in a Big Data context. The objective here is to propose a cycle to answer all the aspects of Big Data and to make data smart. This is in a way *SMART DATA* which was described in [4] as the evolution of the mass of initially unstructured data to the smart processing of data and its transformation into knowledge. The proposed lifecycle will participate in this intelligence and will make it possible to extract from the gigantic mass of the received data, the relevant and useful value. This cycle will solve the limitations on the enormous volume of digital information that must be efficiently exploited in spite of the requirements [5], [6].

We found in [2] that the phases which constitute the data lifecycle in a Big Data context are very complex. Each phase is considered as one or more complex, operational and independent processes, but these processes are linked to one another and to make data management more flexible and smart. Through this article, we try to explain the methodology we used to construct this lifecycle in the form of process cartography from the standard ISO 9001: 2015.

To do this, we identify and define the phases of our data lifecycle model that we have named Smart Data Lifecycle (*Smart DLC*) in the second section. Then, we identify, in the third section flows among Smart DLC processes and we model it. In the fourth section, we position our cycle against existing cycles that have selected in the literature review. In the fifth section, we apply the proposed process-oriented architecture to the 4 V's of Big Data to verify its adaptation to this context in order to show our contribution and added value. Finally, we conclude in the sixth section.

II. Smart Data Lifecycle: Phases

In this section, we present all the phases that made up our DLC.

We noted in [2] that the majority of phases are shared by most data lifecycles, although their nomenclature is sometimes different. Phases are cut into sub-phases for certain lifecycles,

others are grouped together to form one phase. For example, the collection phase includes the following phases: data reception, data creation, filtering, data integration and anonymity. Some lifecycles introduce the visualization phase in the analysis phase; others detach it, to have it as one phase wholly.

Following the analysis of data lifecycles presented in [2], we have retained 14 phases: *Planning, Management, Collection, Integration, Filtering, Enrichment, Analysis, Visualization, Access, Storage, Destruction, Archiving, Quality and Security.*

The nomenclature of the phases and their semantics differ from one model to another. To this end, we have found it useful to resolve this ambiguity by defining each phase in order to clarify its role in our lifecycle.

Planning

This phase is part of the lifecycles of DDI [7], DataONE [8] and USGS [12]. This phase is part of the management processes. Planning involves all other phases of the cycle and gives a preliminary view of what will happen in the medium and long term. This phase is monitored by the project team, which will determine all the necessary human and material resources for the good management of the company's data. To do this, a team dedicated to this task is essential to control the planning and correct any errors. It also requires decision-making on aspects related to data management (data lifetime, data security, data archiving, etc.) in order to define a data management plan for the cycle. During this phase, the planning team provides a detailed description of data that will be used and how they will be managed and made accessible throughout their lifecycle.

Management

The phase of Management is part of the planning processes and is presented in a transversal way. This phase was defined by the IBM lifecycle. IBM considers in [9] that management tasks are part of the data lifecycle. Thus, IBM adds layers of management over the traditional lifecycle. However, IBM considers management only at the level of data testing, masking and archiving. For us, Management concerns all the operational phases that directly manipulate the data. It is a phase that manages the whole lifecycle from end to end and makes communication between all phases effective. It helps identify and capitalize on good practices and manage internal cycle management. Also, it allows measuring, the internal satisfaction of the service rendered by the lifecycle in question. This process is managed by a dedicated team that takes care of everything concerning the lifecycle; in this case, processes of realization and also of support in case of need.

Collection

Data collection is generally the first step in every data lifecycle. This phase consists of receiving the raw data of different natures and making the conversions and modifications necessary to organize them. Cleaning of data received in real time saves calculation time and memory space. Data quality must be carried out at this level because it makes it possible to optimize the data processing circuit overall, which can prove to be very costly, possibly in the context of Big data. A balance must be struck between the speed of access to information and quality requirements [10].

Integration

The purpose of the integration phase is to provide a coherent pattern of data from multiple independent, distributed and heterogeneous sources of information so as to facilitate users accessing and querying such data as if they were accessing only one data source. This phase involves putting in place rules and policies defined by the planning team to integrate the distributed data because the collection methods are different. Data from different sources is combined to form a homogeneous set of data that can be easily analyzed. Because there isn't any agreement on data standards, stakeholders tend to use different methods of data collection and management, which complicate their integration. This phase exists in the cycles proposed in [3], [8], [10]–[13]. For the rest of the cycles, this phase is included in the collection phase.

Filtering

This phase was introduced by the best lifecycle Hindawi following the analysis in [2]. Indeed, it consists in restricting the large data flow. It is necessary to distinguish between Restriction that exists in the Collection and this phase. The restriction concerns noisy data and errors. On the other hand, filtering concerns data in good quality, which have not been blocked by the Restriction rules but the planning team has considered it useful to filter them because they don't add value for the company. This filtering can have a positive impact on data analysis in order to make correct decisions. It also makes it possible to reduce the calculation time and the memory space occupied by data to optimize the phases that come after this phase.

Enrichment

This phase was defined in the Big Data lifecycle in [14]. Data enrichment involves making structural or hierarchical changes to received data. It allows adding information on collected data to improve their quality. This phase assumes having ready and standard data (repositories) to enrich the newly collected data. The enrichment data are updated continuously and automatically to contribute to their quality. The planning team defines the enrichment rules and ensures their application. It always controls them because the rules can change according to several parameters (time, place, data themselves, decision making...).

Analysis

It is the most important phase and has been introduced in almost all lifecycles [3], [7]–[10], [12]–[17]. In this phase, the data are exploited and analyzed to draw conclusions and interpretations of decision-making. It makes it possible to draw information

and knowledge from received raw data. This knowledge provides a basis for decision-making for policymakers. The planning team sets and defines the methods for analyzing data according to the objectives set by the strategic decision-makers. The chosen methods must respect certain criteria that are sensitive to the company. Indeed, data mining methods are multiple and choosing a specific method requires a considerable effort by the planning team to make this method practical, communicable and objective. It must always be aware of the fundamental questions that have not disappeared.

Access

The access within the Big Data Application Provider is focused on the communication/interaction with the Data Consumer. Similar to the collection, the access may be a generic service such as a web server or application server that is configured by the Data manager to handle specific requests from the Data Consumer. This activity would interface with the visualization and analytic phases to respond to requests from the Data Consumer (who may be a person) and uses the processing and platform frameworks to retrieve data to respond to Data Consumer requests.

Visualization

The Hindawi and Big Data DLCs have introduced this phase respectively in [3], [14]. The other cycles introduce this phase in the data analysis phase. It consists of displaying the results of the analysis in a clever and intelligent way so that decision-makers can easily understand these results and then make decisions. There are several ways to view the data. Opting for any kind of display requires considerable effort because a badly chosen type of display can distort the analysis and thus mislead the decision-makers in their decisions. The planning team should pay attention to this phase and predict the predefined visualization types or design specific graphic representations to their use case. The results of the visualization must be checked before publishing them to the decision makers and also must keep privacy anonymity.

Storage

This phase is part of all data lifecycles. It falls within the support processes and must be transversal in the cycle. The storage concerns all the other phases of the cycle and makes it possible to store the data throughout its lifecycle in order to have continuous traceability of data in each phase of the cycle and to know its state of progress. Storage must be managed with reliability, availability, and accessibility. Storage infrastructures must provide reliable space and a robust access interface, which analyzes large amounts of data and also stores, manages and determines data. Thus, the storage capacity must take into account the large increase in the volume of data. However, the enormous amount of data received obliges the planning team to recommend intelligent management of this data because storage is a fundamental and sensitive element of the company information system.

Destruction

This phase is to delete the data when it is successfully used and will become useless and without added value. There are few data lifecycles that introduce the data destruction phase, including cycles of CRUD [18], PII [19], Information lifecycle [20] and Enterprise lifecycle [11] because they consider that data can still be used despite the fact that the needs are not visible at the present moment. However, we believe that from a moment and despite the fact that enough space is available for archiving data, we will probably find ourselves in a situation where we will have to choose between removing obsolete data or invest more to continually increase storage and archiving capacity. This causes additional costs for the memory capacity and the processing of data volume increasingly gigantic. We have chosen to destroy the data once it has reached its end of the cycle because our cycle is a hybrid: linear and cyclical at the same time. The destruction of data must be done intelligently so that it only concerns unnecessary data. To do this, the planning team defines rules and policies for the destruction of data in consultation with the company's decision-makers.

Archiving

This phase consists of long-term storage of the data for possible use. In [9], effective data lifecycle management includes intelligence not only at the archive data level but also the policy of archiving based on specific parameters or business rules, such as the age of the data or the last date of their use. It can also help the planning team develop a hierarchical, automated storage strategy to archive dormant data in a data warehouse, thereby improving its overall performance.

Security

This phase refers to the support processes. It must be present throughout the cycle and is a transversal process. Security has been present in the lifecycles of IBM [13] and Hindawi [6]. This stage of the data lifecycle describes the implementation of data security and its means as well as the roles in data management to make them confidential [21], [22]. This phase concerns three essential security parameters that we explained in [2] namely: data integrity, access control, and privacy. Indeed, these parameters must be checked throughout the cycle.

Quality

In the analysis in [2], only two DLCs include this phase namely USGS [12] and CIGREF [16] cycles.

In [10], we notice that Quality control is provided during transitions from one phase to another. This is achieved through a definition of the quality requirements, the qualification of the level of precision required and then the implementation of controls to measure the satisfaction of the data quality. However, the USGS lifecycle sees this phase differently in [12]. It introduces into this phase the protocols and methods that must

be used to ensure that data are properly collected, managed, processed, used and maintained at all

Table 1 describes for each Smart DLC process its objective, inputs, outputs, characteristic, rules, and actors.

Table 1. Smart DLC processes details

	Objective	Inputs	Outputs	Characteristic	Actors
Planning	• Define plans for each process	• Company requirements	• Plan for each process • Data description	• Transversal	• Planner • Planning team
Management	• Manage all realization and support processes • Validate all realization and support processes	• Management plan • All processes indicators	• Rules for each process • Orders for each realization process	• Transversal	• Smart DLC manager • Management team
Collection	• Collect all company data • Collect external data • Collect archived data	• Company data • External data • Collection rules • Collection plan • Security means • Quality means	• Raw data • Indicators of collection rules application	• Single	• Collection supervisor
Integration	• Provide a coherent pattern of data from multiple independent, distributed and heterogeneous sources • facilitate users accessing and querying	• Row data • Integration plan • Integration rules • Security means • Quality means	• Integrated data • Indicators of integration rules application	• Single	• Integration supervisor
Filtering	• Restrict the large data flow	• Integrated data • Filtering plan • Filtering rules • Security means • Quality means	• Filtered data • Indicators of filtering rules application	• Single	• Filtering supervisor
Enrichment	• Add information on collected data to improve their quality	• Filtered data • Enrichment data • Knowledge data • Enrichment plan • Enrichment rules • Security means • Quality means	• Information • Enrichment data • Archived data	• Single	• Enrichment supervisor
Analysis	• Exploit and analyze data to draw conclusions and interpretations of decision-making	• Information • Implementation of anonymity • Analysis plan • Analysis rules • Security means • Quality means	• Knowledge • Indicators of analysis rules application	• Single	• Analysis supervisor
Access	• Provide interface to data consumer	• Knowledge • Access plan • Access • Security means • Quality means	• Interface • Indicators of access rules application	• Single	• Access supervisor

Visualization	• Display knowledge with a smart manner	• Knowledge • Visualization plan • Visualization rules • Security means • Quality means	• Dashboards • Decisions • Indicators of visualization rules application	• Single	• Visualization supervisor
Storage	• Store data throughout its lifecycle in order to have continuous traceability of data in each process and to know its state of progress	• Raw data • Integrated data • Filtered data • Information • Knowledge • Archived data • Requests for storage • Storage plan • Storage rules	• Raw data • Integrated data • Filtered data • Information • Knowledge • Archived data • Indicators of storage rules application	• Transversal	• Storage supervisor
Archiving	• Provide long-term storage of data for possible use	• Knowledge • Archiving plan • Archiving rules • Security means • Quality means	• Archived data • Indicators of archiving rules application	• Single	• Archiving supervisor
Destruction	• Delete the data when it is successfully used and will become useless and without added value.	• Destruction plan • Destruction rules • Security means • Quality means	• Indicators of destruction rules application	• Single	• Destruction supervisor
Quality	• Measure and control data quality throughout the cycle • Provide to all realization processes quality means	• Request for security means • Quality plan • Quality rules	• Implementation of quality • Indicators of quality rules application	• Transversal	• Support supervisor
Security	• Measure and control data security throughout the cycle • Provide to all realization processes security means to make data confidential.	• Request for quality means • Security plan • Security rules	• Implementation of security • Indicators of security rules application	• Transversal	• Support supervisor

III. SMART DATA LIFECYCLE: MODEL

The phases we have chosen do not manipulate all the data directly. There are phases that collaborate and support the operational phases that have a direct impact on the data received. For example, the planning and management phases do not process data but determine how other operational phases should work. Since the selected phases represent processes in their own right and do not have the same roles as we have underlined, we opt for process cartography from the ISO 9001 Version 2015 standard [23] and the CIGREF framework [24]. Each phase is a process and all processes constitute our data lifecycle. The advantage of this process cartography is to ensure a better organization of the activities carried out within each phase, at the managerial and operational levels. [25]

In order to develop process cartography, it is necessary to identify the different flows that circulate between all processes. Every process has inbound and outbound flows. The realization processes are cyclic; they pass the processed data to the following realization process only after having been validated by the planning process, which continuously monitors them via indicators. Figure 2 illustrates the input and output streams for an operational process.

We used the process approach because designing activities as interrelated processes that make up a system helps achieve more coherent and predictable outcomes. People, teams and processes do not work in silos, and the efficiency will be much better if everyone knows the activities of the body and knows how they relate to each other [23].

We propose a data lifecycle as process cartography. To do this, we have identified processes that belong to three types: *Management* processes, *Realization* processes and *Support* processes [26]–[28].

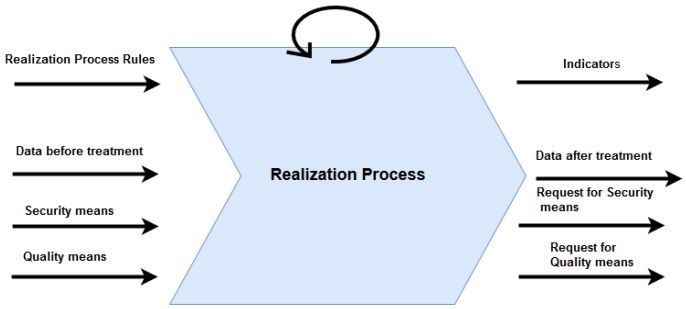

Fig. 2 Inputs and outputs flows for a realization process

The data, after being processed by the various business processes, is supplemented by an additional layer that improves them further and increases lifecycle intelligence. Thus, the data is initially raw and then integrated, filtered, hidden, enriched and finally becomes information and knowledge for the company after the analysis process. For management processes, entries are requirement rules of data Management Company and how strategic decision makers want to exploit their data; and also the results of the application of these rules by the realization processes in order to verify them. Outputs are planning parameters derived from requirement rules that apply to all production processes. In terms of support processes, inputs are requests for means of security and quality according to the policies derived from the requirements of the company to enable the realization processes to fulfill their mission.

Figure 1 illustrates our proposed DLC named *Smart DLC* in the following process cartography.

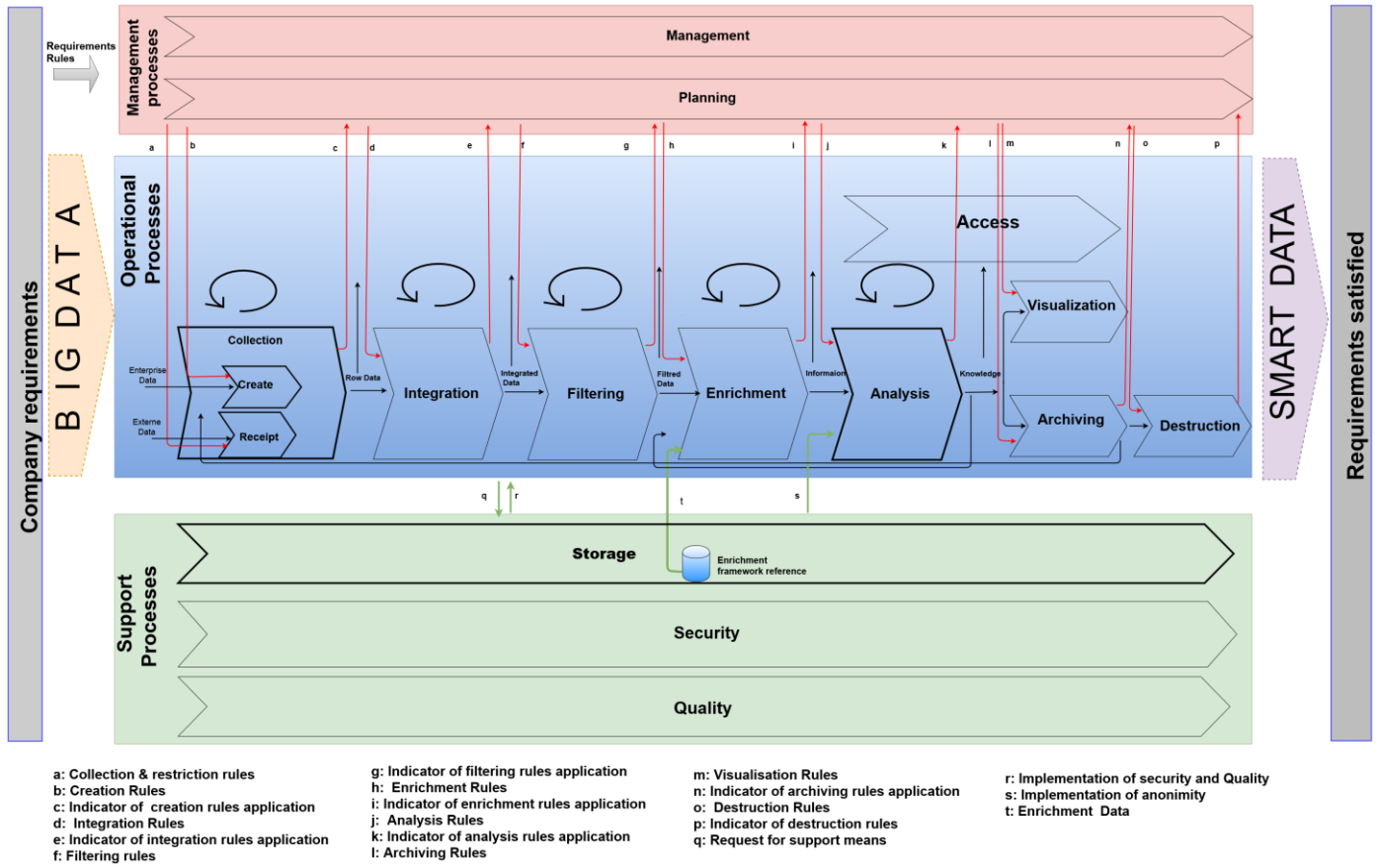

Fig. 1 Smart Data LifeCycle

IV RELATED WORKS

It should be noted that there is no cycle in the literature that presents all phases of Smart DLC. As well, the transversality of the phases has been little discussed in the literature. The sequence of Smart DLC phases has been revised compared to other cycles.

In order to situate our DLC with other lifecycles of the literature, we present in this section among the twelve lifecycles studied in [2] the top six cycles following the final ranking found in [2]: Hindawi DLC [3], Information DLC [20], Big Data DLC [14], USGS DLC [12], IBM DLC [9] and DDI DLC [7]. And then we illustrate our contribution in relation to these relevant cycles carried out in this paper.

A. *Hindawi Lifecycle*

Most companies view their data as a valuable asset. In this sense, they provide considerable effort for the development and optimal use of these data. A data lifecycle is modeled to obtain a consistent description between data and processes. According to [3], Hindawi DLC consists of the following phases: *collection, filtering & classification, data analysis, storing, sharing & publishing, and data retrieval & discovery*. They are presented in figure 3.

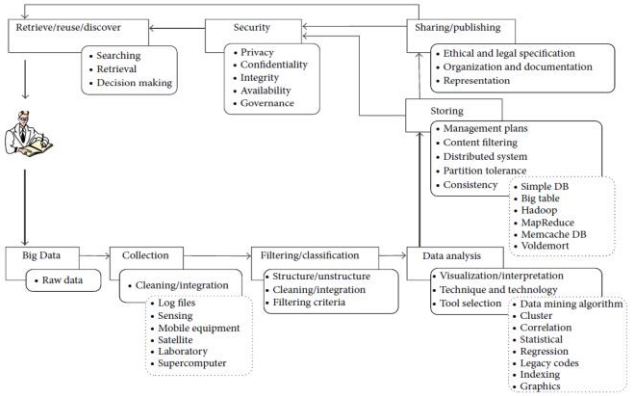

Fig. 3 Hindawi DLC [3]

It's the first in the literature using Big Data terminologies and technologies. We note a very important phase that is filtering. This phase, which comes before the analysis makes it possible to restrict the large data flow.

Hindawi DLC was the best cycle according to the analysis conducted in [2]. It is a lifecycle adapted to Big Data context which has several advantages [3]: the data sources are from Big Data. This cycle can handle various formats of structured and unstructured data; in the collection phase, the data are cleaned to allow for better treatment thereafter; before analyzing the large mass of data, a data filtering phase is carried out in order to focus only on the company needs; the storage is managed intelligently because a management plan is implemented to conduct them with reliability, availability and accessibility; and finally, data security is ensured by a dedicated phase. It encompasses policies and procedures to protect legitimate privacy, confidentiality and intellectual property.

However, Hindawi DLC does not introduce the removal of data that become obsolete. We believe that data destruction is very important in a Big Data context. Also, it does not check the quality of the data throughout the cycle. Security is not transverse throughout all phases. Smart DLC encompasses all phases of the Hindawi cycle and positions them more intelligently. The security phase is ensured at any phase of the cycle, unlike the Hindawi cycle, which presents it as a phase limited in time. Similarly, our cycle destroys, at a given moment, the data that becomes useless.

B. *Information Lifecycle*

This DLC corresponds to a cloud environment. It consists of seven phases: *Data Generation, Data Transmission, Data Storage, Data Access, Data Reuse, Data Archiving, and Data Disposal* [20].
The advantage of this DLC in the Big Data context is the smart data management during the archiving and disposal phases.

The cycle of *Information lifecycle* solves the disadvantage of the Hindawi cycle concerning the limited presence of security in its cycle and the lack of data deletion phase. Information lifecycle introduces security as a cross-cutting phase that concerns all other phases. The strongest point for the Big Data context for this cycle is the intelligent management of the destruction and archiving phases. However, this cycle is not very interested in the phases of data collection, planning, management and quality control, which are key phases for us to make the entire cycle intelligent. *Smart DLC* participates in the intelligence of data collection by introducing data restriction and filtering processes for data not desired by the company and also their integration. In addition, *Smart DLC* presents the results of data analysis in an intelligent way to enable decision-makers to make efficient and timely decisions.

C. *USGS Lifecycle*

USGS DLC cannot justify or allow the acquisition of useless data. Data must be acquired and maintained to meet a scientific need. For this, the idea of data management throughout a lifecycle becomes more relevant. This cycle focused on all issues of documentation, storage, quality assurance and ownership [12].

USGS DLC consists of the following phases as mentioned in figure 4: *Plan, Acquire, Process, Analyze, Preserve, Publish/Share, Describe, Manage quality and Backup & secure.*

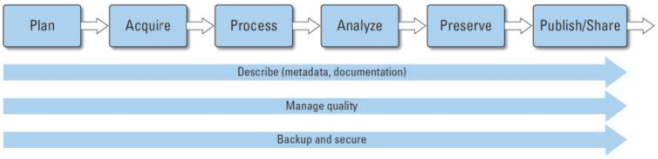

Fig. 4 USGS DLC **[12]**

This DLC introduces new phases that haven't a direct impact to data like planning, metadata description, quality and security. Another advantage of this DLC is the cross-sectional presence of data description, quality and security.

The major advantage of the **USGS** cycle is the cross-sectional presence of the quality and security phases [12]. For this cycle, planning is the first phase to be carried out, so it is a phase that does not concern the other phases in a transversal way. There is

no destruction of data. It is obvious that *Smart DLC* solves all the disadvantages found in the USGS cycle. Indeed, planning is a cross-sectional phase and data destruction is present at the end of the cycle.

D. Big Data Lifecycle

This DLC adapted to Big Data [14] is not very different from other traditional data lifecycles. The new phase of filtering and enriching the data after their collection seems interesting, however, the storage of the data throughout the lifecycle appears incompatible with the Big Data since it does not solve the concern of Volume relating to Big Data but on the contrary, it makes their management more complicated by their redundancy. Big Data DLC phases are illustrated in figure 5.

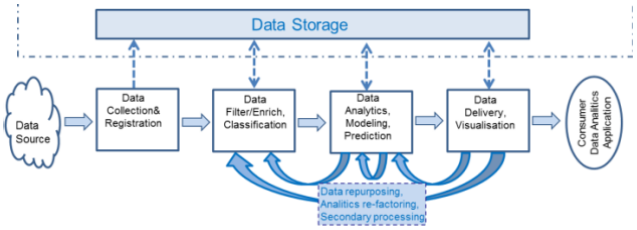

Fig. 5 Big Data DLC **[14]**

The *Big Data* cycle initially introduced by Yuri Demchenko in [14] and included in [1], [18], [29]–[32] has adopted new methods for improving the collection process. Thus, a filtering and enrichment phase was added after collection to reduce the mass of data initially collected. The particularity of this model lies in the storage phase where the data is retained during all the stages of the lifecycle, which will allow the re-use of the data and their reformatting. Although this data lifecycle has been introduced in a Big Data context, its intelligence is not reached; as no planning nor security nor quality phases are present. The data cannot be deleted.

Smart DLC enjoys the benefits of this lifecycle and solves its shortcomings to make the end-to-end cycle smarter.

E. IBM Lifecycle

In [9], IBM considers that management tasks are part of the data lifecycle. IBM DLC adds over the traditional lifecycle layers of management. It defines three essential elements for managing data lifecycle during the different phases of the data existence as presented in figure 6.

Fig. 6 IBM DLC [9]

The Smart DLC has the advantage of IBM DLC that consists of adding management, masking layers and adding layers of quality and security that make the data lifecycle more intelligent.

F. DDI DLC

In [7], a data lifecycle is defined as a representation of the set of data management procedures. Data Documentation Initiative (DDI) presents a combined lifecycle model, which is linear but becomes circular in 8 steps as shown in figure 7:

- **Concept:** allows defining an initial and global vision of the data. The system contains a list of concepts and definitions that can be grouped into a hierarchical structure.
- **Collection:** capture information about the survey itself, the question text and the information response domain, as well as all interview instructions, including additional descriptive information and instructions visible at the time of filling out the questionnaire.
- **Processing:** data processing takes place at different points in the lifecycle. Specific areas of information captured in data processing include control operations, clean-up operations, weighting factors, data evaluation, and coding.
- **Archiving:** allows archiving of data by moving obsolete and unused data.
- **Distribution:** is the phase where the data are ready to be diffused in various systems as is explained in [2].
- **Discovery:** allows describing the data by essential metadata for the purpose of discovery.
- **Analysis:** this phase makes possible to examine and explore data to determine information and knowledge.
- **Repurposing:** This step reflects a new conceptual framework. The implications of this point of view include the need to define the relationships between the data conceived during the design process and the possibility of defining both primary and secondary data sources in the collection phase.

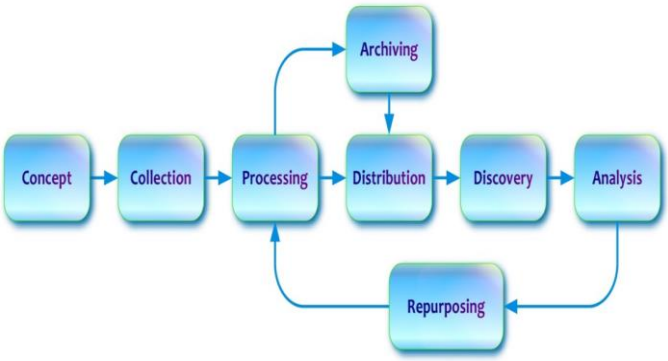

Fig. 7 DDI DLC [7]

Smart DLC has the advantages of DDI DLC that consist of three important concepts: the use of ontologies in data collection and archiving which reduces conceptual incoherence within an institution, the re-use of data generating a gain in various fields and finally archiving. Smart DLC enjoys all these advantages and add more ones such as the pervasive presence of management, security and quality control.

G. *Synthesis*

Smart DLC Smart DLC encompasses all above DLCs phases and positions them more intelligently. Its advantages can be summarized as follows:

- Manage Big Data issues by cleaning data at the beginning by the Collection phase and filtering just company needs in order to optimize their treatment. So, Smart DLC participates in the intelligence of data collection by introducing data restriction and filtering processes for data not desired by the company and also their integration.
- Their phases are organized according to a standard ISO 9001:2015 [33] and a well-defined framework CIGREF [24].
- Smart DLC is interested in both technical and managerial issues.
- Smart DLC phases are organized into three types:
  - Management processes: determine how operational phases should work.
  - Realization processes: have a direct impact on the data received.
  - The Support processes collaborate with the operational phases to achieve their assignment.
- Management and Support processes are cross-cutting which allows all realization processes to benefit from them at any time.
- Each type of processes is managed by a dedicated team that takes care of everything concerning Smart DLC.
- Planning team defines all processes plans executed by realization processes and verified by the management team.
- Data processed by one realization process only goes to the next process after being validated by management team, otherwise, they will be rerun by the same process until validation by the management team.
- Smart DLC begins with a definition of the company requirements to allow moving from Big Data to Smart Data.
- Smart DLC is generic that could interest any company that is interested in data management.
- Smart DLC offers security and quality services.
- As long as data are processed by the realization processes as long as its quality is increased and so intelligence is added.

V. Application of Smart DLC to Big Data

To validate the proposed lifecycle in the Big Data context, we illustrate, in this section, the use of this cycle in this context. We define for this purpose Big Data and its 4 Vs.

A. *Big Data definition*

The number of Vs that characterizes Big Data has evolved over time. The most cited definition characterizes Big Data in 3 Vs: *Volume, Variety* and *Velocity* [34]. This definition has since been taken up by IBM and others to include a fourth V: *Veracity* to address issues of trust and uncertainty [29], [30], [35], [36]. Since then, more and more Vs have been added to arrive at 9 Vs [37]: the 4 most popular Vs that we have just quoted plus the following five: *Variability, Value, Visualization, Volatility and Validity.*

We chose the most popular and most quoted definition of Big Data namely the definition by the 4 Vs to apply our cycle proposed to these Vs. We did not retain the Visualization because we have a process dedicated to this characteristic.

The literature review of different definitions the research and professional world in [1], [29]–[32], [36], [38]–[40] allows us to retain this definition: "*Big Data is an approach* [39] *which consists of extracting value from a gigantic data set (Volume) of various forms (Variety) in continuous movement (Velocity) and with uncertain reliability (Veracity).*"

This definition characterizes Big Data by the 4 Vs:

- *Volume (Data in rest)*: the amount of data generated and managed by the company must be constantly increasing.
- *Velocity (Data in motion)*: it is to be able to use the data as they are collected. This requires computational power and analytical tools cut to measure. If data are not processed at the right time, usually in real time, they will have no added value.
- *Variety (Data in many forms)*: data are of several structured or unstructured forms and of several types: document, image, video, log files ... hence, the treatment tools must be adequate for this diversity.
- *Veracity (Data in doubt)*: measures the accuracy and reliability of data. These criteria will be understood if the data source is uncertain or inaccurate or from the poorly-known origin.

B. *Smart DLC applied to the 4 Vs of Big Data*

By mapping the characteristics of Big Data Vs to each proposed cartography process, we get questions and problems that a Data Manager or a Big Data project manager can ask when faced with such a project [41]. The answers to these

questions will help to make decisions for the infrastructure and to guide the data management activities of a Big Data project.

Table 2 illustrates the application of the 4 Vs of Big Data selected for the proposed cycle **Smart DLC**.

Table 2. Issues raised by the characteristics of Big Data applied to Smart DLC

Process	Volume	Variety	Velocity	Veracity
Planning	What is the estimate of data volume and growth rate?	How do data policies from different sources combine? What steps are being taken to address sensitive data?	Are the expected bandwidth and storage sufficient to accommodate input speeds?	What are the data sources? Who will have the derived data and the data resulting from aggregation?
Management	Are the systems with their actors ready to handle the volume of managed data?	What is the policy of the management team to handle different data formats? How does it handle a new format?	How does the management team plan for a continuous change in bandwidth and storage?	What allows Managers to trust data sources? How do Managers measure the accuracy and reliability of data?
Collection	How much data can this process manage and generate?	What are the formats (structured or unstructured) and types (document, image, video, log files) of data?	Who collects data? Do they have the tools and skills to choose the best sources available?	The collected data has not been altered or changed? How to verify that the received data have not undergone any changes?
Integration	What are the implications of volume in the preparation of data sets?	What steps are required to integrate data in different formats?	Will datasets be aggregated in series? Will meta-data apply to individual datasets, series, or both?	How to ensure the reliability of embedded data? How do we check that the integration has been successful? Are there test tools?
Filtering	What is the impact of filtering on data volume?	Filtering can be applied to different data formats?	Does Filtering apply to meta-data, individual data, series, or both?	The choice of filtering rules does not compromise the reliability of the data?
Enrichment	Is there a consistent enrichment base? Is it standardized (Repository)?	How is the enrichment database updated?	How often is the enrichment database updated?	Is the enrichment base reliable? How to check this reliability?
Analysis	Are processing and analysis means available?	Are the different analytical methods compatible with the different datasets?	Is the data analyzed when they are collected?	How accurate are the analytical methods used?
Storage	Should the raw data be preserved? What storage space is needed in the long term? Who will provide it?	What is the best form of storage (Databases, NoSQL, Cloud ...)?	What mechanism should be in place to deal with storage velocity?	When does the stored data require archiving?
Archiving	What are the archived data? When is it necessary to archive?	Are there different methods of archiving?	How are the archived data maintained for future use?	When does the data become obsolete and therefore have to be removed?
Access	How many Data consumers can access to the results of analysis and visualization?	Are there different interfaces to access?	How long do the Data Consumers have to access?	How is Data Consumers authenticity verified?
Visualization	How to display a potentially large number of results?	How can we display different results at the same time?	What is the degree of viewing latency tolerated?	How to check the relevance of the visualization to avoid distorting the representation of the results?
Destruction	What is the Data Removal Policy?	Can different types of data formats be deleted at the same time?	What is the maximum time for archived and unused data to delete it?	What is the impact of destroying obsolete data on the reliability of other processes?

				Are there ways to recover the destroyed data?
Security	Is there a secure transmission channel to collect data?	Is there multiple encryption algorithms? How do you manage security keys?	What is the latency of data encryption and decryption?	Is there an access control from different sources to ensure the reliability of these sources?
				Does masking of personal data impact analysis accuracy?
Quality	Are quality measurement tools commensurate with the volume of data?	Are the different methods of quality control compatible with the different data sets and also with all the implementation processes?	What is the impact of quality verification on other processes in terms of execution time and management?	Is there a standard or a repository for applying quality control?

The questions presented in Table 2 highlight the most important issues for each process of our Smart cycle compared to the Big Data features. The answer to each question depends on the nature of the project knowing the questions are sustainable for both simple projects or Big Data projects. To anticipate and match the needful in terms of technical and management tools of this project, the project supervisor should take the questions in Table 2.

VI. CONCLUSION

In this paper, we proposed a new vision for data lifecycle that considers data management as an Information System urbanization project. To this end, we have modelled this lifecycle as process cartography from the ISO 9001: 2015 standard and the CIGREF framework [23], [24].

The advantage of considering the data lifecycle as process cartography is to support effectively the company data management tasks and their transformations. This cartography takes into account the existing situation and makes it possible to better anticipate the internal and external evolutions or constraints impacting data lifecycle and, if necessary, relying on technological opportunities.

Representing the data lifecycle in urbanized process cartography facilitates its transformation and change. Indeed, data lifecycle changes from one company to another and it can change within a single organization that often changes strategies; which implies major structural changes and complicates interdependence. This increasing complexity has implications for costs, times and risks for data management and decision-making.

To gradually control the evolution of data with the necessary reactivity and to reduce IT costs, a response is provided by the approach of urbanization of data lifecycle processes. This method of urbanization in process cartography aims at a lifecycle capable of supporting data management in the best cost / quality / time. We have chosen cartography from ISO 9001: 2015 standard [23] and from the CIGREF [24] framework that integrates the process approach and distinguishes three types of processes within the company: *management, operational, and support processes.*

The application of our architecture to the characteristics of Big Data highlighted questions and problems that would have to be solved before starting a Big Data project.

REFERENCES

[1] J. Manyika *et al.*, ‘Big data: The next frontier for innovation, competition, and productivity’, 2011.

[2] M. El arass, I. Tikito, and N. Souissi, ‘Data lifecycles analysis: towards intelligent cycle’, in *Proceeding of The second International Conference on Intelligent Systems and Computer Vision, ISCV’2017, Fès 17-19 April*, Fez, Morocco, 2017. https://doi.org/10.1109/ISACV.2017.8054938

[3] N. Khan *et al.*, ‘Big data: survey, technologies, opportunities, and challenges’, *The Scientific World Journal*, vol. 2014, 2014.

[4] A. Lenk, L. Bonorden, A. Hellmanns, N. Roedder, and S. Jaehnichen, ‘Towards a taxonomy of standards in smart data’, in *Big Data (Big Data), 2015 IEEE International Conference on*, 2015, pp. 1749–1754.

[5] D. Farge, ‘Du Big data au smart data : retour vers un marketing de l'émotion et de la confiance’, *LesEchos.fr*, 2015.

[6] F. Meleard, ‘Smart data, l'avenir du contenu’, *Les echos.fr*, 2015.

[7] X. Ma, P. Fox, E. Rozell, P. West, and S. Zednik, ‘Ontology dynamics in a data life cycle: challenges and recommendations from a Geoscience Perspective’, *Journal of Earth Science*, vol. 25, no. 2, pp. 407–412, 2014.

[8] S. Allard, ‘DataONE: Facilitating eScience through collaboration’, *Journal of eScience Librarianship*, vol. 1, no. 1, p. 3, 2012.

[9] IBM, ‘Wrangling big data: Fundamentals of data lifecycle management’, 2013.

[10] S. BOUTEILLER, *Enjeux business des données. Comment gérer les données de l'entreprise pour créer de la valeur?* CIGREF, 2014.

[11] S. Chaki, ‘The Lifecycle of Enterprise Information Management’, in *Enterprise Information Management in Practice*, Springer, 2015, pp. 7–14.

[12] J. L. Faundeen *et al.*, 'The United States Geological Survey Science Data Lifecycle Model', US Geological Survey, 2014.

[13] J. B. Jade Reynolds, *In the context of the Convention on Bilogical Diversity*. World Conservation Monitoring Centre, 1996.

[14] Y. Demchenko, C. De Laat, and P. Membrey, 'Defining architecture components of the Big Data Ecosystem', in *Collaboration Technologies and Systems (CTS), 2014 International Conference on*, 2014, pp. 104–112.

[15] C. L. Borgman, J. C. Wallis, M. S. Mayernik, and A. Pepe, 'Drowning in data: digital library architecture to support scientific use of embedded sensor networks', in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 2007, pp. 269–277.

[16] E. Deelman and A. Chervenak, 'Data management challenges of data-intensive scientific workflows', in *Cluster Computing and the Grid, 2008. CCGRID'08. 8th IEEE International Symposium on*, 2008, pp. 687–692.

[17] I. Gam, 'Ingénierie des exigences pour les systèmes d'information décisionnels: concepts, modèles et processus: la méthode CADWE', Paris 1, 2008.

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

[19] A. Michota and S. Katsikas, 'Designing a seamless privacy policy for social networks', in *Proceedings of the 19th Panhellenic Conference on Informatics*, 2015, pp. 139–143.

[20] L. Lin, T. Liu, J. Hu, and J. Zhang, 'A privacy-aware cloud service selection method toward data life-cycle', in *Parallel and Distributed Systems (ICPADS), 2014 20th IEEE International Conference on*, 2014, pp. 752–759.

[21] T. Alam and M. Benaida, 'The Role of Cloud-MANET Framework in the Internet of Things (IoT)', *Int. J. Onl. Eng.*, vol. 14, no. 12, p. 97, Dec. 2018.

[22] T. Alam and M. Benaida, 'CICS: Cloud–Internet Communication Security Framework for the Internet of Smart Devices', *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 12, no. 6, pp. 74–84, Oct. 2018.

[23] Organisation internationale de normalisation, 'Quality management systems requirements'. 2015.

[24] CIGREF, 'Les référentiels de la DSI : Etat de l'art usage s et bonnes pratiques'. 2009.

[25] M. El arass and N. Souissi, 'Data Lifecycle: From Big Data to SmartData', in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, Marrakech, 2018, pp. 80–87. https://doi.org/10.1109/CIST.2018.8596547

[26] M. El arass, I. Tikito, and N. Souissi, 'An Audit Framework for Data Lifecycles in a Big Data context', in *2018 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*, Tangier, 2018, pp. 1–5. https://doi.org/10.1109/MoWNet.2018.8428883

[27] M. El arass, K. Ouazzani Touhami, and N. Souissi, 'The System of Systems paradigm to reduce the complexity of data lifecycle management. Case of the Security Information and Event Management', *IJSSE*, 2019.

[28] M. El arass and N. Souissi, 'Smart SIEM: From Big Data logs and events to Smart Data alerts', *IJITEE*, vol. 8, no. 8, pp. 2655–2662, Jun. 2019.

[29] M. Chen, S. Mao, and Y. Liu, 'Big data: A survey', *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[30] K. Davis, *Ethics of Big Data: Balancing risk and innovation*. O'Reilly Media, Inc., 2012.

[31] K. Krishnan, *Data warehousing in the age of big data*. Newnes, 2013.

[32] A. Reeve, *Managing Data in Motion: Data Integration Best Practice Techniques and Technologies*. Newnes, 2013.

[33] ISO, 'ISO 9001 Quality management', 2015. [Online]. Available: https://www.iso.org/iso-9001-quality-management.html. [Accessed: 10-Jul-2018].

[34] L. Douglas, '3d data management: Controlling data volume, velocity and variety', *Gartner. Retrieved*, vol. 6, p. 2001, 2001.

[35] IBM, 'The four v's of big data'. .

[36] P. Zikopoulos, C. Eaton, and others, *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[37] I. TIKITO and N. SOUISSI, 'Data Collect Requirements Model', in *Proceeding of the BDCA'2017, Tetuan 28-30 March*, 2017. https://doi.org/10.1145/3090354.3090358

[38] M. Karoui, G. Davauchelle, and A. Dudezert, 'Big data. Mise en perspective et enjeux pour les entreprises.', *Ingénierie des Systèmes d'Information*, vol. 19, no. 3, pp. 73–92, 2014.

[39] M. B. Sophie Bouteiller, 'Big-Data-Vision-grandes-entreprises-Opportunites-et-enjeux-CIGREF', 2013.

[40] J.-S. Vayre, 'Les big data et la relation client', in *12ème Journées Normandes de Recherches sur la Consommation: Société et Consommation*, 2013, pp. 1–20.

[41] L. Pouchard, 'Revisiting the Data Lifecycle with Big Data Curation', *International Journal of Digital Curation*, vol. 10, no. 2, pp. 176–192, May 2016.