Data lifecycles analysis: towards intelligent cycle

M. EL ARASS [(1)(3)], I. TIKITO [(2)]
(1) Institut National des Sciences Appliquées Lyon, France
mohammed.elarass@gmail.com
(2) Mohammed V University in Rabat
EMI-SIWEB Team Rabat, Morocco
tikito.iman@gmail.com

SOUISSI N.[(2)(3)]
(3) Ecole Nationale Supérieure des Mines de Rabat
Computer Science Department Rabat, Morocco
souissi@enim.ac.ma

***Abstract*—As companies generate and handle increasingly large amounts of data along with the Big Data era, several model of data lifecycle have been proposed to deal with this situation. The analysis, the management and the use of data becomes more complicated or almost impossible in some cases for the companies. To transform these data to a knowledge, the choice of the adequate lifecycle that matches with the company expectations becomes essential.**

For this goal, this paper aims to be a guide to assist companies to choose a lifecycle that fits their data management vision. For this, we identify the relevant criteria of selection cycles and defined a rating system to each of these criteria. In this paper, we study the available lifecycles of data in the literature that we consider relevant. As a result of this study, we classify these cycles following two types: first analysis oriented phases and the second based on relevant criteria.

Keywords—Data lifecycle, Data management, Big Data, Smart Data, Benchmarking.

I. INTRODUCTION

Data lifecycle is defined in [1] by a set of steps (or phases) by which the data cross from they enter a set of system until they leave. DataONE in [2] consider that the data management tasks depend on data lifecycle. However, lifecycle management of data or DLM[1] was defined in [3] as the set of processes implemented to manage the data of the company for their definition until withdrawn.

USGS[2] in [4] describe that the data must be handled and managed once decision making to collect or use data until they become obsolete or no longer needed. The phases crossed by a data from creation to deletion, or external storage, is initially related to customer needs, context and type of data. The importance to manage data lifecycle model is explained in [5]; it provide a structure for considering the many operations that will need to be performed on a data record throughout its life. Many conservation actions can be much easier if they have been prepared in advance on the creation of the data. Each failure during the cycle can present a big risk (data loss, non-compliance with rules, data inconsistency, and disclosure of privacy ...), therefore good management of this cycle is required to have a good performances. If cycle management of all types of data is not new, needs to systemize and automate emerged more strongly in recent years especially with the advent of Big Data era which was been defined in [6] as a process of extracting the relevant information from a data set. This data set is characterized by Volume, Variety, Velocity, Veracity and the Value as specified in [7].

Indeed, the problem that begins to impose with Big Data is the word **Big**, because too much data kills data. Thus, the analysis, management and use of data becomes very complicated tasks and even impossible in some cases. Our solution to correct this situation is to **make data intelligent and have a good quality of content**. This solution lies in two words: **Smart Data** described in [8] as the evolution of the mass initially unstructured data in intelligent processing of data and its transformation into knowledge. However, a lifecycle model that takes part in this intelligence will extract the relevant value and useful outcome of the big mass of received data that will improve the operation of the value added.

Our contribution through this paper is to make a detailed analysis of the different lifecycles of data that has been proposed in the literature to help companies choose the most relevant and appropriate approach for their context. Thus, a lifecycle that makes raw data with no value became **Smart Data** help to manage a large number of problems present during the extraction of the added value of raw data. This approach will also resolve the huge volume of digital information that must be operated efficiently in spite of requirements [9; 10].

To do this, we will study in the second section, twelve most interesting lifecycles through literature during the last twenty years. To build a solution, we will identify, in the third section, the phases and the relevant criteria that we will use to compare and analyze the lifecycles models. Finally, in the fourth section, we will discuss the results of our two analyzes to identify the most interesting lifecycles that can help companies to better manage their data.

II. STATE OF THE ART OF DATA LIFECYCLES: LITERATURE REVIEW

In this section, we present initially a study of 12 data lifecycles that have been proposed over the past 20 years. We explain through this study all steps for the cycles in order to identify the strengths and weaknesses of each model. We do a summary of that study in a second time.

[1] Data Lifecycle Management
[2] US Geological Survey

978-1-5090-4062-9/17/$31.00 ©2017 IEEE

A. *Study of the data lifecycles*

1) Information pyramid

Jade Georis-Creuseveau refers in [11] the information pyramid of Reynolds and Busby [12] to illustrate the transformation processes from "bottom to top", in which the raw data are transformed into information and then indicators that will constitute knowledge of the company. The pyramid concept illustrates the dependence between levels to progress in the hierarchy. At the base of the pyramid, very large numbers of data are needed to produce information, and more advanced indicators, but much less quantity. In its pyramid model, it distinguishes four phases: collection, integration, analysis and publication.

To understand the different behaviours of each phase, we refer to [12]:

- **Collection:** consist in receiving the raw data of various natures.
- **Integration:** is to set rules and policies to integrate the distributed data because the methods of collection are different.

In this pyramid model comes a first view of the lifecycle of the data. It is a basic model whose later work has been based to develop and adapt it according to their needs.

2) CRUD lifecycle

Yu, Xiaojun and Wen propose in [13] a lifecycle model called **CRUD** (Create, Read, Update and Delete) in five phases. In this model, in relation to the pyramid of information, phases have changed names, others have been subdivided into sub-phases. The phases here have not been explained explicitly to be able to make a comparison with the pyramid model but starting from [14] and the semantics of each phase we were able to draw the following points:

- **Create** corresponds to the data collection phase which is located at the bottom of the pyramid of the information or the data are in the raw state and without workable value.
- In **Store**, the data is stored either for use or publication or for archiving.
- **Destruct** is the last phase of the data lifecycle if it becomes useless and without added value.
- **Use and Share** correspond to the exploitation of the data and the publication of the results after their analysis.
- **Archive** consists in storing the data for a long term.

This lifecycle model highlights the main classical phases of the data lifecycle in the literature. The advantage of this model lies in the fact that the data do not all follow the same lifecycle. However, Create, Store, and Destruct phases are mandatory, while Use and Share and Archive phases are optional.

3) Lifecycle for Big Data

Yuri Demchenko defined a new lifecycle for Big Data in [15] given the specificity of this type of data which has raised the interest of several works in the literature as in [3; 16; 17; 18; 19; 20]. Big Data requires the adoption of new scientific discovery methods that include improving a new iterative model and improving collect data process and reusing them with an improved model. Thus, a filter and enrichment phase was added after collect to reduce the mass of data initially collected. The other phases have retained the same meaning and role. The characteristic of this model lies in the storage phase where the data are retained during all the stages of the lifecycle, allowing the re-use of the data and their reformatting. However, this is possible only if the identification of the data is complete, their references are crossed and their links are implemented in the Big Data Infrastructure (BDI). Data integrity, access control and accountability must be supported throughout the lifecycle. Verifying the reliability of data content is an important component of this cycle and it must also be done in a secure and trustworthy manner. However, storage has not been further detailed in order to conclude whether this model facilitates the complexity of Big Data or contributes to their redundancy, since the data are stored several times.

This lifecycle adapted to Big Data is not very different from other traditional data cycles. The new phase of filtering and enriching the data after their collection seems interesting to us, however the storage of the data throughout the lifecycle appears to us 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.

4) IBM lifecycle

IBM considers in [21] that management tasks are part of the data lifecycle. Thus IBM adds layers of management over the traditional lifecycle. It defines three essential elements for managing the lifecycle of the data during the different phases of the existence of the data:

- **Test Data Management:** in the process of developing new data sources, test technicians must automate the creation of realistic data sources of the same size to reflect the real behaviors of existing production databases.
- **Data masking:** the technique of data masking consists not only of using fictitious data to protect privacy, but also of preserving the actual production data of the company.
- **Archiving:** effective management of the cycle includes intelligence not only to archive data, but also the policy to archive should be based on specific parameters or business rules, such as the age of data.

IBM considers a new concept of data lifecycle, by adding layers of management and management policy over the classical data lifecycle.

5) DataOne lifecycle

DataOne has adopted in [2] a lifecycle model specific to the field of scientific research. It focuses on "data". This cycle is useful because it makes possible to identify data flows and work processes for scientists. DataONE defines eight phases in the lifecycle of scientific data: **Plan, Collect, Assure, Describe, Preserve, Discover, Integrate and Analyze**.

- **Plan:** lifecycle starts when scientists make a plan to undertake their researches. In this phase, a description of the data which will be exploited is carried out, as well as how these data will be managed and made accessible throughout their lifecycle.

- **Collect:** observations are made by hand or with sensors or other tools and data is placed under digital format.
- **Assure:** controls and inspections are carried out to ensure the quality of the data
- **Describe:** metadata standards are used accurately and correctly to describe the data.
- **Preserve:** Data is stored in a suitable long-term archive. At this stage, the data are discoverable and can be consulted by other scientists.
- **Integrate:** data from different sources are combined to have a homogeneous set of data.
- **Analyze:** data is exploited and analyzed to draw conclusions and interpretations of decision support.

Although the DataONE model gives further details for data lifecycle by having eight phases, it remains a very particular model in a scientific research context where most processing is done manually and the volume of data is not very large. This type of model is therefore not suitable for Big Data.

6) Information lifecycle

In [22], the data lifecycle consists of seven phases: **Data Generation, Data Transmission, Data Storage, Data Access, Data Reuse, Data Archiving, and Data Disposal**. This cycle corresponds to a detailed Cloud environment in [23; 24].

- **Data Generation:** the cloud service receives requests from users and creates their data while specifying the access control policy.
- **Data Transmission:** the cloud service establishes a secure transmission channel to verify the reliability of user data and uses methods for encrypting data between servers and users using a digital certificate mechanism.
- **Data Storage:** The role of the cloud service is to ensure that data is placed in the right places authorized by agreements or regulations.
- **Data Access:** The Cloud Service must ensure the validity of the users' identity to protect them from spoofing and verify the proper execution of the data access policy.
- **Data Reuse:** This can lead to leakage of sensitive and personal data, which is why this service should not be in a Cloud environment. But with the Big Data era, data sharing has made this phase quite primordial. This phase corresponds to the traditional integration phase in some lifecycles.
- **Data Archiving:** Three main operations are required in this phase, band encryption, long-range storage and data retrieval.
- **Data Disposal:** The primary goal in this stage is to effectively locate data in the cloud and completely remove unnecessary data.

This data lifecycle provides services compatible with Cloud Computing. It focuses more on data security throughout its phases. The most interesting point in this model for Big Data is the intelligent management of data during the archiving and disposal phases. Another advantage of this model is the data security layer, which is ubiquitous throughout all phases.

7) CIGREF

CIGREF[3], network of large enterprises defines a link between data and knowledge: "the **data** becomes **information**, then well shared within the company, it transforms into **knowledge** and constitutes its **being**". In [25], they describe the precious value of data in a company considering the impact which it can have on its success. As a result, several companies are focusing their efforts on controlling the lifecycle of the data. Throughout the chain, information must be collected, monitored, protected, shared and analyzed to contribute to the development of the organization.

- **Data Quality Control:** we note the transition from one phase to another through "Quality Control". This concept requires a definition of the quality requirements, the qualification of the level of precision required, and the establishment of controls to measure the quality of data.
- **Acquisition:** collecting "raw" data from the "source" through business processes.
- **Storage:** storage of data in the system.
- **Transformation:** this phase is essential given the variety of typologies and supports as well as a large part of the collected data are unstructured. This phase plays the role, in some sort, of the classical integration phase.
- **Analysis:** it relies on the expertise and intelligence of decision-makers in order to generate **value for the company.**

Regarding data **Value**, one of the key axes is to identify how the data to be valued will contributes to the realization of one of the strategic axes of the company.

This cycle makes it possible to optimize overall the data processing circuit, which can be very expensive in the Big Data context. This cycle proposes to integrate the quality controls at the time of collection. But, it is necessary to find a balance between speed of access to information and quality requirements.

8) DDI lifecycle

In [26], 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.

- **Concept:** allows to define 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.

[3] CIGREF: an association gathering French large companies

- **Archiving:** allows archiving of data by moving obsolete and unused data.
- **Distribution:** in [26], the operation of this phase has not been explained, but we still understand that this is the phase where the data are ready to be diffused in various systems.
- **Discovery:** allows to describe the data by essential metadata for the purpose of discovery.
- **Analysis:** we cannot find in [26] information about the operation of this phase. But we believe that it makes it possible to examine and determine the information.
- **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.

We note in this cycle three very 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.

9) USGS lifecycle

According to [4], USGS[4] 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.

- **Plan:** The project team should consider the necessary approaches, resources, and planned outputs for each stage of the cycle. A data management plan is the recommended output of this phase.
- **Acquire:** represents the activities by which new or existing data is collected, generated, or considered and evaluated for reuse. It involves the collection or addition of data banks.
- **Process:** refers to actions or measurements taken on the data to verify, organize, transform, integrate and extract data in an appropriate output form for future use.
- **Analyze:** includes actions and methods performed on data that help to describe the facts, detect trends, develop explanations and test assumptions. This includes assurance of quality data, statistical analysis of data, modeling and interpretation of test results.
- **Preserve:** involves actions and procedures for storing data for a certain period of time and/or setting up side data for future use, and includes archiving data and/or submitting data to a reference data.
- **Publish/Share:** prepare and publish, or disseminate, data with good quality to the public and other organizations. We need to make sure that the data is shared, but with controls to protect the data ownership and pre-decision and data integrity.
- **Describe (metadata, documentation):** throughout the cycle, documents must be updated to reflect measurements taken on the data.
- **Manage quality:** protocols and methods must be used to ensure that data are properly collected, managed, processed, used and maintained at all stages of the scientific data lifecycle.
- **Backup and secure:** take security measures against accidental data loss, corruption and unauthorized access.

The key feature of this lifecycle is the cross-sectional presence of quality control, qualification of metadata also documentation, and finally the backup and security phase which allows a backup to be made during accidental loss of data.

10) PII lifecycle

The data lifecycle in [27] cover PII data (Personal Identifiable Information) from creation to storage. The obsolete data is subsequently deleted.

- **Collection:** collect of a user's personal data following a collection plan. The information collected is protected, and the user is aware of how his or her data is used.
- **Processing:** includes the use and access to information of users, as well as the possible modification of this one. During this phase, all information (which, where, when, what and how) related to this processing must also be saved.
- **Storage:** in this step information security is present, but also the user can manage or delete information about him/her.
- **Transfer:** PII transfer results in internal sharing and external distribution/publication.
- **Maintenance:** the user can manage or delete information about him/her.

Although this lifecycle is closed, but it contains the step of destroying obsolete data. In each step of this cycle we notice that there is additional information added to maintain the traceability of the information.

11) Entreprise data lifecycle

According to [28], the lifecycle of the data can be visualized as follows: the data created is treated as assets for a period defined by the company where it is used. After the data enters a semi-active phase where there are no defined usage cases for the data-consuming element, they are archived and then deleted.

- **Creation/Receipt:** information is created when data is received at their point of origin in the business process.
- **Distribution:** the data is created, and distributed to the concerned people who exploit them in the management of the basic operational processes.
- **Consumption:** there are two types of consumers: final consumers of data where the mode of use is mainly oriented analysis and reporting as well as commercial functions and consumer services of data in the context of their business processes and transaction needs. Here, the data are used to make a decision.
- **Disposition/Archival:** each company needs to define the useful period for which a data item must be preserved in the registration system or other storage devices.

[4] US Geological Survey

- **Destruction/Retire:** data that has reached the end of its lifecycle and thus classified as inactive will be destroyed/retired.

In this closed lifecycle, we note an important notion that is the destruction of data. The proper use of this concept with archiving will considerably reduce the amount of data stored.

12) Hindawi lifecycle

According to [29], the proposed lifecycle consists of the following steps: collection, filtering & classification, data analysis, storing, sharing & publishing, and data retrieval & discovery.

- **Big Data:** allows to enhance the raw data collected by researchers and organizations. The data are transformed from its initial state and are stored in a state of added value.
- **Collection:** large amounts of data are created and others come from sensors, mobile equipment, satellites, laboratories, supercomputers, research forms, tchat sites, messages on internet forums and microblog messages.
- **Filtering/classification:** the data collected are of low density and high value. This phase allows the classification of the data in structured/ unstructured, as well as a filtering according to certain criteria.
- **Data analysis:** allows an organization to process abundant information that can affect the company and accurately predict future observations.
- **Storing:** Storing, managing and determining large amounts of data.
- **Sharing/publishing:** Enables the public, tribal governments, academics, researchers, scientific partners, federal agencies and other stakeholders to benefit from the information being processed.
- **Security:** describes data security and roles in data management to protect legitimate privacy, confidentiality and intellectual property.
- **Retrieve/reuse/discover:** Data recovery ensures data quality, adding value and retaining data by reusing existing data to discover new and valuable information.

We note a very important notion that is filtering. This phase which comes before the analysis makes it possible to restrict the large data flow.

B. Synthesis

The review of these different lifecycles made us note that phases are common to all of these models although the terminology differs in some cases. Other phases are specific to a model which itself is linked to a well-defined field. For example, the addition of the **Plan** phase before the collection phase corresponds to the scientific research domain as for the models described in [2; 4]. Before beginning our analysis, we considered it will be useful to illustrate the progression of the data lifecycles over the last twenty years. Figure 1 shows that 2014 was a landmark year in the proposal for a set of data lifecycle models. This is due, from our point of view, to the new era of Big Data where data became so important and critical that it has to be controlled from creation to destruction. For this, a model that accurately describes the lifecycle of the data proved to be essential. But until now, no model was retained to define a standard which justifies that these models did not yet reached the level of maturity necessary and sufficient to be the object of a standard.

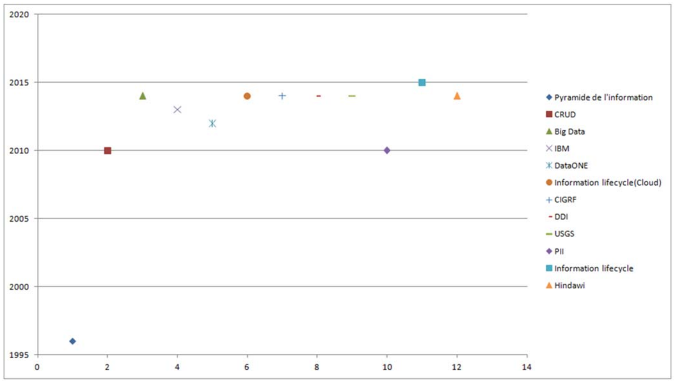

Figure 1 : Lifecycles for 20 years

III. BENCHMARKING OF DATA LIFECYCLES

In this section, we analyze the lifecycles studied in the previous section. Our analysis will be carried out in two stages: a first analysis oriented **Phases** and a second analysis oriented **relevant Criteria**.

A. Analysis oriented phases

For the purposes of this analysis, we have selected 11 phases: **Planning, Creation/Reception, Integration, Filtering, Anonymity, Enrichment, Analysis, Visualization, Storage, Destruction and Archiving**. The choice of these phases is the result of the synthesis of the state of the art which enabled us to identify the most relevant phases of each cycle studied. In order for our comparison to be objective and meaningful, we have put in place a table with the different phases selected, we assign the value 1 if the phase in question exists in the lifecycle and the value 0 if not. The answer to each section of the table makes it possible to evaluate the phase itself and the whole cycle following a given model. When filling the table, phases for certain cycles do not have the same nomenclature as our phases; for this purpose, we have based on the semantics of these phases to find those which correspond to our phases, for example, the phase acquisition in the cycle CIGREF described in [25] corresponds to the phase **Create/Receive**. Table 1 summarizes the correspondence of the phases of some cycles studied with our selected phases.

Table 1 : Correspondence between the phases of the studied cycles and those retained

Retained phases	**Pyramid of Information**	**CRUD**	**Big Data**	**IBM**	**DataOne**
Plan					
Create /Receive	Collect	Create	Data collection and registration	Create	Collect
Integration	Integration			Repurposing	Integrate
Filtering			Data filter		
Anonymity					

Enrichment			Enrichment, and classification		
Analyze	Analyze, publishing	Use and Share	Data Analytics, Modeling, Prediction	Use, Share	Analyze, discover
Visualization			Data delivery, Visualization		
Storage		Store	Data storage	Store /Retain	Preserve
Destruction		Destruct			
Archiving		Archive		Archive, dispose	

Afterward, we assigned a score for each lifecycle based on the scores found, which allows us to evaluate the models. We have assigned the value 1 if the phase in question exists in the cycle and the value 0 otherwise. The results of our first phase oriented analysis illustrated in Table 2 enabled us to conclude a ranking of the data lifecycles illustrated in Figure 2.

Table 2 : Lifecycle score according to the retained phases

	Information lifecycle	**Hindawi**	**DataONE**	**USGS**	**Big Data**	**IBM**	**DDI**	**CIGREF**	**CRUD**	**Enterprise**	**Pyramid**	**PII**
Plan	0	0	1	1	0	0	1	0	0	0	0	0
Create/Receive	1	1	1	1	1	1	1	1	1	1	1	1
Integration	1	1	1	1	0	0	0	1	0	0	1	0
Filtering	0	1	0	1	1	0	0	0	0	0	0	0
Anonymity	0	1	0	0	0	1	0	0	0	0	0	1
Enrichment	0	0	0	0	1	0	0	0	0	0	0	0
Analyze	0	1	1	1	1	1	1	1	1	0	1	0
Visualization	0	1	0	0	1	0	0	0	0	0	0	0
Storage	1	1	0	1	1	1	0	1	1	1	0	0
Destruction	1	0	0	0	0	0	0	0	1	1	0	1
Archiving	1	1	1	1	0	1	1	0	1	1	0	0
Total	5	8	5	7	6	5	4	4	5	4	3	3

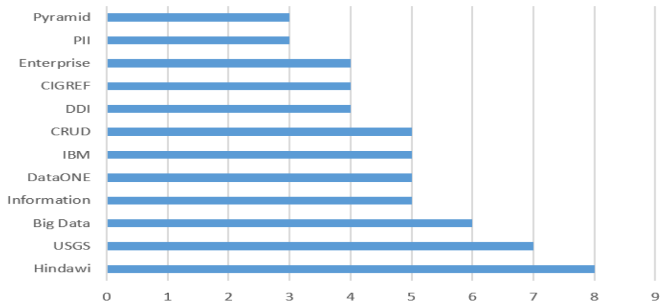

Figure 2 : Classification of the lifecycles of the data according to the analysis phases oriented

This ranking reflects the progress of data lifecycle models over time. Indeed, they are classified, with some exceptions, according to the increasing order of their dates of appearance. Thus, the latest models in our analysis are **Pyramid of Information**, **PII** described in [12; 27] and appeared respectively in 1996 and 2010. The lifecycle information model described in [28] which is the most recent (2015) is not in the first rank. The best model in our analysis is **Hindawi** data lifecycle described in [29] and appeared in 2014. It is a model that contains almost all the phases that we had retained in our analysis, adapted to Big Data, and makes it possible to add intelligence to the data in their trail. The only disadvantage we had with this cycle was the **planning, enrichment and destruction** phases which remain absent and are important in our analysis. The three missing phases in the first model do not exist at the same time in the successive cycles of USGS, Big Data and Information lifecycle. Indeed, the USGS cycle has the only planning phase, Big Data has the **enrichment** phase while the cycle Information lifecycle contains the phase **Destruction**. Thus, we recommend this lifecycle model as a basis for possible proposals that companies can use to describe their data while taking advantage of the **USGS, Big Data and Information lifecycle** models to address weaknesses in the **Hindawi** model. Although the analysis according to the **Phases** parameter allowed us to obtain a ranking, some important parameters are not highlighted in this analysis, hence the interest of our second analysis which is based on criteria that we judge relevant.

B. Analysis by relevant criteria

The state of the art of the cycles detailed in the second section, allowed us to identify certain relevant criteria but which are not fully phases. In order to distinguish the best models, we decided to make a second analysis based on these criteria. The relevant criteria we have chosen are: **Adaptation to Big Data, Security, Supervision, Management, Quality Control, Green, Intelligence level, Flexibility of the cycle**. Criteria have been splitting into sub-criteria to better evaluate them. Table 3 describes our criteria with their components:

Table 3 : Criteria and sub-criteria of the second analysis

Criteria	**Sub-criteria**
Adaptation to Big Data	Velocity
	Variety
	Veracity
	Value
	Volume
Security	Access control
	Dara Integrity
	Private life
Green	Storage
	Destruction
	Archiving
Supervision	
Management	
Quality Control	

Intelligence level	
Flexibility of the cycle	

Depending on how the lifecycle responds to a criterion, we have scored each criterion from 0 to 5. To do so, a questionnaire was used for each criterion. The answer to each question is "yes" or "no". We assign the score 1 if the answer is yes and the score 0 otherwise. Then we aggregate all the notes to have a score of the criterion between 0 and 5. A score is calculated by summing the scores for each criterion as shown in Table 4. Figure 3 illustrates the results of our second analysis oriented relevant criteria.

Table 4 : Lifecycle score according to relevant criteria

	Information lifecycle	Hindawi	DataONE	USGS	Big Data	IBM	DDI	CIGREF	CRUD	Enterprise	Pyramid	PII
Adaptation to Big Data	2	4	0	1	3	0	1	1	0	0	0	0
Security	5	4	0	0	0	5	0	0	0	0	0	0
Supervision	2	4	0	0	2	5	0	0	0	0	0	0
Management	1	1	2	5	0	3	2	0	0	0	0	0
Quality Control	0	0	0	5	0	0	0	5	0	0	0	0
Green	5	4	1	1	4	1	1	0	1	2	0	1
Intelligence level	2	5	2	2	5	0	3	1	2	0	0	0
Flexibility of the cycle	2	4	3	0	2	0	3	0	4	2	0	1
Total	19	26	8	14	16	14	10	7	7	4	0	2

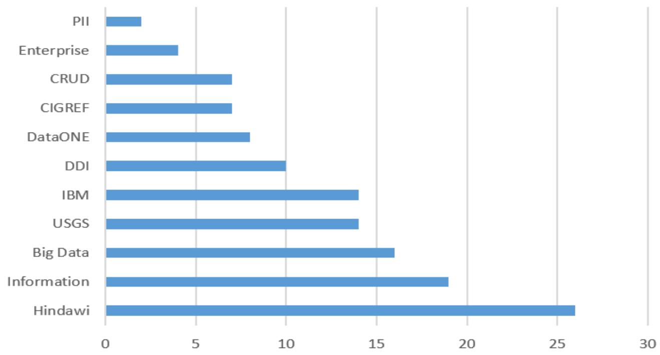

Figure 3 : Lifecycle classification according to criteria-oriented analysis

The found ranking confirms our first analysis based on the phases. Indeed, we found almost the same ranking especially for the cycles that are at the top and at the end of the ranking. The **Hindawi** model is the best cycle following this analysis and confirms its interest. On the other hand, the models **Pyramid of information and PII** remain always the last cycles following our two analyzes. The **Hindawi** model contains almost all the criteria we retained. Its only disadvantage following the second analysis lies in the absence of the criterion **Quality Control and Management** which is for us a primordial criterion in the lifecycle of the data because it makes it possible to better manage the data and to have a preliminary vision of the data way from their creation until their destruction.

IV. Discussion

In this section we will discuss the results found through the two analyzes to guide companies to identify the most relevant lifecycles that correspond to their use case.

The choice of the relevant phases and criteria was made in order to evaluate any lifecycle for any type of data. The data, whatever its field of use, will follow a cycle which contains either all the phases we have chosen or some in spite of the context, except that certain phases will be mandatory whatever the data and its field of application. The required phases are: **Creation/Reception, Analysis, Storage, Archiving**, the remainder are optional but remains important for a better control of the data. As long as we covered all the phases as long we identify the data by adding value to it and thus it become **Smart**. Indeed, a cycle that contains all the phases mentioned above is a model that answer companies' expectations whatever their field of work because it participates in the intelligence of the data from its collection to its destruction. This intelligence improves further while progressing in phases.

We found that the two analyzes converge towards the same ranking with some differences. To remove this nuance, we weighted the results of the two analyzes by summing their total scores to have a single ranking. This convergence gave us the ranking shown in Figure 4.

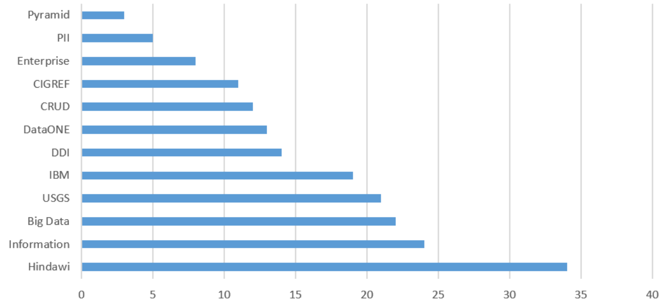

Figure 4 : Final ranking of the lifecycles following the two analyzes

The results of the fusion of the two analyzes are not far from those mentioned in the third section. They confirm our methodology for the analysis of lifecycles that we have chosen to study, analyze and compare them according to phases and criteria relevant and objectively. We noted that lifecycles that take into account the Big Data aspect are classified first. This is due to the fact that the Big Data context assumes a lifecycle which facilitates the management of the huge mass of data and makes them intelligent while adding value to transform them into information and knowledge that can help companies to make effective and reliable decisions. Based on this ranking, we recommend the **Hindawi** lifecycle for companies especially those working in a Big Data context. This model can also be the basis for possible proposals that companies can use to describe their data while taking advantage of the **USGS, Information lifecycle** and **Big Data** models to address weaknesses related to the Hindawi model including Phase

Plan, Enrichment and Destruction, also **Quality Control and Management** criteria. The relevant phases and missing criteria in the Hindawi lifecycle are present in USGS. Thus, we recommend a lifecycle inspiring from the models of **Hindawi and USGS**.

The union of these two cycles will cover most of the phases which we retained and will meet all the relevant criteria that we used to perform our analysis. We chose the USGS cycle instead of lifecycle information or big data because USGS corrects all the disadvantages related to the Hindawi cycle. On the other hand, the cycle Information lifecycle, despite being better classified than the USGS cycle, does not meet the criteria missing in the Hindawi cycle in particular the Management and Quality Control criteria.

V. Conclusion

Several models of data lifecycles have been proposed with the appearance of the Big Data era. The added value of this paper was to guide companies to choose a lifecycle which suits their data management vision. For that, we filtered 12 most relevant models at the moment to analyze them and to detect their strengths and weaknesses. The state of the art of the literature presented in the second section allowed us to detect the most relevant phases and criteria to identify the models that makes data **Smart** and thus facilitate their management in the Big Data context. Thus, we performed two analyzes, a first oriented phases and a second oriented criteria. The results of these two complementary analyzes confirmed to us that the **Hindawi** model is the best of our selection although that it does not cover all the phases and criteria that have been raised. On the other hand, the **Pyramid of information and** PII lifecycles remain the less interesting cycles and which participate less in valuing the data and making them Smart. Companies interested in data and who want to anticipate and take into account this strategic "asset" to derive a sustainable competitive advantage from it, could be inspired by the **USGS, Big Data and Information lifecycle** models to complete the Hindawi and thus correcting its failures. We also concluded that the **Hindawi and USGS** lifecycles are complementary. Indeed, the USGS lifecycle completes and corrects the defects present in the Hindawi cycle which are mainly the absence of the criteria: **Quality Control and Management.**

References

[1] A. Simonet, “Active data : Un modele pour representer et programmer le cycle de vie des donnees distribuees,” in ComPAS’2014, 2014.

[2] S. Allard, “Dataone : Facilitating escience through collaboration,” Journal of eScience Librarianship, vol. 1, no. 1, p. 3, 2012.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, “Big data : The next frontier for innovation, competition, and productivity,” 2011.

[4] J. L. Faundeen, T. E. Burley, J. A. Carlino, D. L. Govoni, H. S. Henkel, S. L. Holl, V. B. Hutchison, E. Martín, E. T. Montgomery, C. Ladino, et al., “The united states geological survey science data lifecycle model,” tech. rep., US Geological Survey, 2014.

[5] A. Ball, “Review of data management lifecycle models,” 2012.

[6] M. B. Sophie BOUTEIL LER, “Big-data-vision-large-companies-opportunities-and-issues -cigref,” 2013.

[7] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, “Addressing big data issues in scientific data infrastructure,” in Collaboration Technologies and Systems (CTS), 2013 International Conference on, pp. 48–55, IEEE, 2013.

[8] 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, pp. 1749–1754, IEEE, 2015.

[9] F. Meleard, “Smart data, The future of content,” Les echos.fr, 2015.

[10] D. Farge, “Du big data au smart data : retour vers un marketing de l’émotion et de la confiance,” LesEchos.fr, 2015.

[11] J. G. Creuseveau, Les Infrastructures de Données Géographiques (IDG) : développement d’une méthodologie pour l’étude des usages. Le cas des acteurs côtiers et de la GIZC en France. PhD thesis, Université de Bretagne Occidentale, 2014.

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

[13] X. Yu and Q. Wen, “A view about cloud data security from data lifecycle,” in Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on, pp. 1–4, IEEE, 2010.

[14] I. Gam, Ingénierie Requirements engineering for decision-making information systems: concepts, models and processes: the CADWE method. PhD thesis, Paris 1, 2008.

[15] 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, pp. 104–112, IEEE, 2014.

[16] K. Davis, Ethics of Big Data : Balancing risk and innovation. " O’Reilly Media, Inc.", 2012.

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

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

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

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

[21] IBM, “Wrangling big data : Fundamentals of data lifecycle management,” 2013.

[22] 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, pp. 752–759, IEEE, 2014.

[23] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al., “A view of cloud computing,” Communications of the ACM, vol. 53, no. 4, pp. 50–58, 2010.

[24] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, “Cloud computing and emerging it platforms : Vision, hype, and reality for delivering computing as the 5th utility,” Future Generation computer systems, vol. 25, no. 6, pp. 599–616, 2009.

[25] S. BOUTEILLER, Business data challenges. How to manage company data to create value? CIGREF, 2014.

[26] X. Ma, P. Fox, E. Rozell, P. West, and S. Zednik, “Ontology dynamics in a data lifecycle : challenges and recommendations from a geoscience perspective,”Journal of Earth Science, vol. 25, no. 2, pp. 407–412, 2014.

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

[28] S. Chaki, “The lifecycle of enterprise information management,” in Enterprise Information Management in Practice, pp. 7–14, Springer, 2015.

[29] N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, W. K. Mahmoud Ali, M. Alam, M. Shiraz, and A. Gani, “Big data : survey, technologies, opportunities, and challenges,” The Scientific World Journal, vol. 2014, 2014.