Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 173 (2020) 364–371

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on *Smart Sustainable Intelligent Computing and Applications* under ICITETM2020

Data Life Cycle Management in Big Data Analytics

Kumar Rahul*[1] Rohitash Kumar Banyal[2]

[1, 2] Department of Computer Science and Engineering, Rajasthan Technical University, Kota, Rajasthan 324010, India
[1]Basic and Applied Science, NIFTEM, Sonepat, Haryana 131028, India
Email: kumarrahul.niftem@gmail.com, rkbayal@gmail.com

Abstract

Data life cycle management is very much useful for any enterprise or application where data is being used and processed for producing results. Data's appearance for a certain period time ensures accessibility and usability in the system. Data generated through different sources and it is available in various forms for accessibility. A big data-based application such as the healthcare sector generates lots of data through sensors and other electronic devices which can be further classified into a model for report generations and predictions for various purposes for the benefits of patients and hospitals as well. The data life cycle presents the entire data process in the system. The lifecycle of data starts from creation, store, usability, sharing, and archive and destroy in the system and applications. It defines the data flow in an organization. For the successful implementation of the model, there is a need to maintain the life cycle of data under a secured system of data management. This paper deals with the data life cycle with different steps and various works are done for data management in different sectors and benefits of the data life cycle for industrial and healthcare applications including challenges, conclusions, and future scope.

*Keywords:*Data life cycle, Data creation, Data usability, Healthcare, Big data

1. Introduction

Data is considered a valuable object in industries in today's scenario. Big data is an emerging subject under management information systems, computer science, and social science as well. The impact of big data reflects on social media, sensors system used in various industries (including healthcare sectors), mobile devices, marketing, education, smart city, production, and e-commerce and so on.

* Corresponding author. Tel.: +91-7206070670; fax: +91-130-2281020.
*E-mail address:*kumarrahul.niftem@gmail.com

10.1016/j.procs.2020.06.042

Big data is also widely suitable for IoT (Internet of things) based products and applications. Big data analytics is used for an application where voluminous data generated which are highly scalable, stored at fault-tolerant distributed database systems and processes effectively and efficiently. Big data is a term used for processing heterogeneous data (i.e. unstructured and structured data) for bringing to informative data form. To make the data effective for an application, it passes through data extraction, transformation, and visualization. Big data is used for collection, stored, process and transforming the huge amount of data to meaningful data and information through strategic tools and techniques. Big data is characterized by volume, variety, value, velocity, veracity, variability, visualization and so on [1].

Big data analytics is a means to analyze and interpretation of any digital information and data. Technical and analytical advancement in big data analytics (BDA) determines the functional aspects of digital products and services [2]. As with the fastest and growing data generations, it is important to streamline and understand the process and mechanism that how big data analytics can add value addition to the industries in several ways. Data life cycle management is important in big data analytics where it encompasses several steps. Big data analytics associated with different technical aspects comprises searching, mining, analysis, and usability which is suitable and important for enhancement for any applications. Now a day's data life cycle management under BDA is so important to recognize and streamline to understand marketing strategies to popularize their product in comparison to other industries.

Under data life cycle management, huge data sets exist where it matters to examine which data suite is important for what so ever conditions. It reflects conditions of access and represents big data investment for the production of business values which are underexplored. Data life cycle management under bigdata engineering or big data analytics is believed by the corporate sector executive to be the new milestone to nurture business opportunities. Big data analytics is suitable for cost reduction, better and faster decision making and to develop new products and services. During data acquisition, degree of data consistency, degree of data completeness, benefits of external data usages, benefits of eternal data usages support in the acquisition of intention of data analytics [3]. Data life cycle under big data analytics follows capture, process, analyze and visualize data for accessibility. Big data analytics is suitable for business intelligence, real-time streaming analysis, predictive analysis and enterprise data warehousing. Data inconsistencies, incompleteness are two important issues under data life cycle management.

Data life cycle required to identify the impact of data on voluminous and expensive data. Data sensing can be achieved through processing, analyzing and sharing. During the identification of meaningful data, a data merger would be possible through different sources. Apart from big data analytics, now a day's augmented analysis is also in a trend where mixing IoT and machine learning(ML) used for creation, development and sharing analytics. Big data analytics areas include several applications of customer service development, enterprise content management, smart cities development, conversational analytics and natural language programming (NLP). Data life cycle management can be understood as:

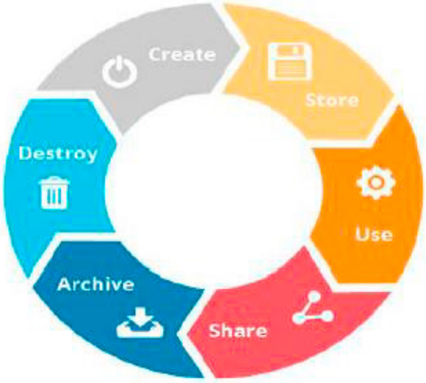

Fig. 1: Data Life Cycle Management (Source: [4])

1.1. Data creation

Data creation is the first phase of data life cycle management. Under data creation, various sources can be considered for data generations including data acquisition point, data entry point, signal capturing and processing, devices, and sensors installed in industries and so on. Big data analytics uses these data sources for analyzing and visualization for the benefit of industries. Data presentations play an important role in data creation. Data acquisition point, where data generated by industries outside the enterprises. Data entry is an electronic point that generates lots of data for the enterprise. Sensors generated data is suitable for the analytics system to filter and evaluate relevant data and remove missing and irrelevant data. The internal data source is used for decision making and business operations. Data creation can be executed through different ways including first-party data collections, third party data collection, and repurposing and surveillance data.

1.2. Datastore

Datastore is another phase of data life cycle management which is more important for storage purposes. The Data store process requires movement, integration, enrichment, and ETL (extract, transform and load). Data stored in basic backbone support for data processing. In this case, data is important for various purposes for industries to meet objectives and operations. However, data cannot reproduce or recycle it, somewhere it needs to be re-using the same for another purpose. After data creation or generation, it goes through several phases for usability and accessibility of applications. Datastore phases consist of structured, semi-structured and unstructured types of data. Datastore starts with data conceptualization and collection but it never ends [5]. It is also executed in some form of retaining for usability purposes. Various machine learning algorithms (MLA) used to allocate data sets into cluster forms. K-means is one of the extensively used techniques for data cluster analysis. It identifies Kno. of centroid and allocates the nearest cluster to every data point. K–means clustering intention is to partition ℕ objects into K cluster where every object belongs to the cluster with the nearest mean. In K-means finds the centroid through iterations process. It is of two types such as a hierarchical and central cluster. The datastore concept follows both ways to store data point value for large datasets. K-means defined as:

$$Objectivefunction\text{: } 0 = \sum_{j=1}^{p} \sum_{i=1}^{m} \left| |x_i^{(j)} - c_j| \right|^2 \tag{1}$$

Here, O = Objective function, p=Number of the cluster,m=Number of cases, x_i= case i$\left| |x_i^{(j)} - c_j| \right|^2$ is Euclidean distance function and c_j is centroid for cluster j.

1.3. Data usability

Data usability indicates how data are usable for industries through technologies and applications. The goal of data life cycle management is initiation, planning, execution, and closure. During data usability phases, critical and generated data is reviewed, analyze, process and modify. Security control and duplicate removal of data applicable through defines mechanisms. The data usability graph shows how effective the data generation system is suitable for enterprises. The growth of data in the healthcare sector explained in [6].

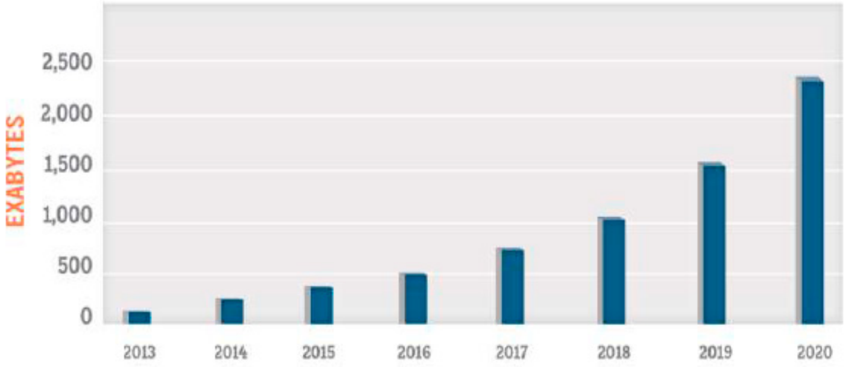

Fig. 2: Growth in healthcare data (Source [6])

Multiple linear regressions are used to show the relationship among different values in large data set through equations:

$$X_1 = b_0 + b_1X_2 + b_2X_3 + . \ldots \ldots \ldots \ldots \ldots \ldots . + b_nX_{n+1} \tag{2}$$

1.4. Data sharing

Data sharing is important in terms of accessibility. All components and modules of the system use data-sharing technologies. The purposes of data life cycle management are to re-use the data for various purposes and it should be shared among modules in the application [4]. Therefore, every component in the system monitors different data (structured, unstructured and semi-structured data) [4]. It passes through different locations, systems, operating atmosphere and access by the different platforms. It helps to support to provide meaningful data at the right time [4]. Since industries' core resources are data, therefore, it should possess the infrastructure to maintain functionalities associated with data [7]. Data has become a central core for enterprises in today's scenario [8]. Pearson correlation coefficient helps to find out the relationships among different variables such as accuracy, sensitivity, specificity, precision, etc. It will be calculated as:

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{[x \sum_x^2 - (\sum x)^2][y \sum_y^2 - (\sum y)^2]}} \tag{3}$$

It provides the strength and relationship degree of association among variables.

1.5. Data Archive

Since data are required again and again in the system, therefore it transferred to the new location for future purposes in the data life cycle management system for an application. In case of need, it can be brought back to an active module of applications for execution. The data archive is a process to transfer and store under data life cycle management.

1.6. Data destroy or reuse

Data destroy is the phase, where data is no longer required for an application. It is of no use in the system. However, data generate again through different sources and it may repeat also. In a real application, data destroy is the removal of such data having no sense of access or re-use in any module. Data cannot reproduce itself [8]. The author in [9] described at the end, data either destroy or recycle/reuse when their usability is exhausted.

2. Related Work

Data life cycle management helps to accelerate data delivery and accessibility for an organization. According to a survey performed by Accenture, 83% of world reputed organizations have shown interest to use bigdata and somehow, they have started using it for benefit of the organization. Some of the organizations adopted, some of them are in pipeline to adopt and so [10]. An author in [9] described several reasons to destroy data including of selection of new data in place of the old one, retention law, etc. An author in[11] described data life cycle management and shown scientific data that need to be collected for usability purposes and preserve it for the long-term for the future requirement for industries.

The authors defined a few challenges of data quality such as high volume, heterogeneity, data change, and data security [12]. The authors discussed some of the challenges of data security comprises of confidentiality, integrity, availability, privacy and so on [12].

The authors described data protection laws with the salient feature of some countries [13]. Privacy and security issued discussed in big data life cycle management in [13]. It was discussed the big data technologies which are used for protecting and securing data in the healthcare sector that includes authentication, encryption, data masking, and access control [13]. The adoption of cloud computing in the healthcare sector can solve various issues in industries [14].

The authors also discussed the role of cloud computing technologies with advantages and disadvantages in the healthcare sector [14]. The authors explained the remote therapy concept for easier accessibility, receiving, sharing and storing patient information through cloud computing technologies [14]. Impact and effect of big data technologies in the healthcare sector discussed in [15]. Patient-centric healthcare systems developed where various parties including insurers, service providers, patients, practitioners, and pharmaceuticals are involved for functioning and execution [15].

The authors explained data integration tools, searching and processing tools, machine learning tools, real time and stream data processing tools, etc. [15]. Sources of big data (which is important for data creation under data life cycle management), nature of analytics and analytical techniques (through various references) discussed in the paper [15]. There is a concept of energy-efficient during the entire data lifecycle management system for a big data-based project discussed in the paper listed at [16]. Data life cycle management deals through a strategy of the find, get and use, deliver and maintain tools and techniques in the system.

The authors discussed data in the context of a data-centric system that is of under different categories of descriptive, homogenous, closed or open, centralized and life-cycle type, etc [17]. Big data analysis techniques perceived by the industries; however, it needs certain instructions and guidelines as well [18]. The authors identified big data analysis opportunities with the data life cycle for AEC (architecture, engineering, and construction) industries [18]. Data analysis with conditional probability can be understood through Bayesian statistics from the below equation as:

$$p\left(\frac{A}{B}\right) = p\left(\frac{B}{A}\right) X p(A)/p(B) \tag{4}$$

It is suitable in least data set case as well. Naïve Bayes classifiers are used to classify data point through Bernoulli Naïve Bayes and Gaussian Naïve Bayes. Gaussian Naïve Bayes is suitable to find continuous value generation from data sources [19].

The authors identified risk evaluation models for the whole life cycle for a wind power project [20]. Similarly, the authors defined the importance of the data life cycle in clinical data analysis under the healthcare sector [21]. The authors defined a few dimensions of data quality under clinical data such as completeness, correctness, concordance, plausibility, and currency. Data generation through EHR (electronic health record), administrative data, registries and so on and it passes to data transformation, data reuse and finally, post-reuse data quality reporting executes [21]. Different challenges encounter in the clinical data life cycle including of loss of information and context, loss of constructive, misuse of clinical data, lack of best practices of quality data, etc. [21].

The application of data life cycle analysis and concept discussed in identifying the e-commerce system recommendations system for BDA [22]. When the data life cycle used in the healthcare sector, there is a chance to use the intelligent community system. Such type of community system developed with the support of IoT (Internet of things). In this concept of healthcare services, various data sources, processed data, and stored data used to provide services to the public to access medical facilities [23]. The authors show a grey relation estimation method suitable for an online e-commerce system [24]. The authors analyzed and discussed the technology diffusion system for agriculture sectors in [25]. The authors also described data management through cloud computing [26]. The data life cycle has been used in discrete simulation modeling in any application system [27]. According to the author [28], it was designed and developed a model for secured preserving data publishing. For any data modeling and data life cycle process under business software or applications, information and computing theory needed to assemble technology, text, signal, code, etc. [29]. The author defined traditional data and big data which is required for the data life cycle initially to differentiate among data creations and storage [30]. The authors also explained traditional data used in GB (Gigabyte) whereas big data measured in TB (Terabytes). The authors discussed different aspects of the data life cycle starting from the find, capture, normalize, aggregate, report, understand and act upon data [31]. The authors discussed product life cycle management including quality system [32]. The below diagram shows the adoption status of big data technology in industries.

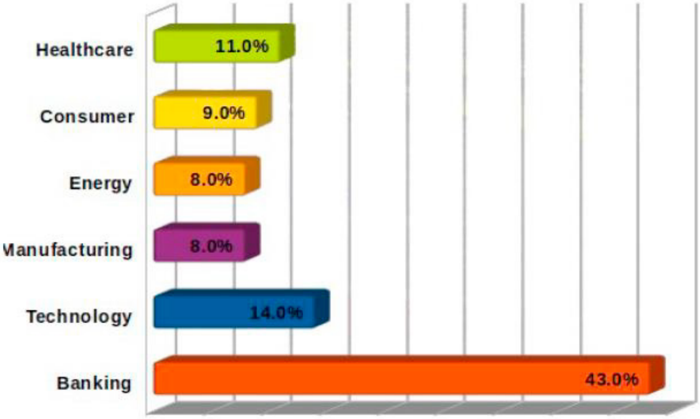

Fig. 3: Big data used in industries (Source: Peer Research-Big data analytical survey)

3. Benefits for Data Lifecycle Management

Big data lifecycle management adheres to various challenges of capturing storing, processing, analyzing and cleaning large volumes of data. By implementing data life cycle management, the industry can do data ingestion through different sources and store in form of HADOOP. Any applications of big data can be implemented in MATLAB as well to show the progress and life cycle of data in a different form. MATLAB tool including simulation used to represent data sources and generations and analysis. The analysis report indicates the performances of data usability and accessibility. With data life cycle management, analysis and evaluation are possible by the integration of BDA (Big data analytics) with the mainframe system. With data life cycle management, creating streaming databases can be provided to software like Kafka to evaluate real-time analysis with the support of data generated through different sources. With data life cycle management, analysis and evaluation are possible by the integration of BDA (Big data analytics) with the mainframe system. In this way, the mainframe system can be off-loaded and enable Apache Kafka and BDA system with the support of the HADOOP system.

Data life cycle management optimizes the data warehousing system by using HADOOP technology to reduce the workload that helps to reduce cost and provides high performances. Data life cycle management reduces and optimized enterprises with intelligence.

Data life cycle management support to adopt policies for data access and data usages in an organization. It supports to use data for a longer period using defined policies within organizations. Defined policies are required because of data storage and access to life within an organization.

The benefits of data life cycle also indicate to reduce data risk and help in data quality improvement. It is used to map business tasks or processes with data maps to meet goals of reduce process and data bottleneck, maintain redundancy, and improve consistency and so on.

4. Conclusion and Future Work

Data life cycle concepts used in various applications where it can be effectively defined and efficiently used. Industries need to manage the data lifecycle for better utilization of the data life cycle. In the healthcare sector, it includes different aspects of challenges such as data collection, security, and privacy, EHR (Electronic Health Record) integration, data sharing, unstructured data, data authentication, real-time availability of data, etc. The data life cycle of an application leads to the big data life cycle, where data collection, data cleaning, data aggregation, data representation, data modeling, and analysis and data delivery executed sequentially

References

[1] H. Khaloufi, K. Abouelmehdi, and A. Beni-hssane. (2018) “Science Direct ScienceDirect Security model for Big Healthcare Data Lifecycle,” *Procedia Comput. Sci*., vol. **141**, pp. 294–301.

[2] C. Loebbecke and A. Picot. (2015) “*Journal of Strategic Information Systems Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda*,” J. Strateg. Inf.Syst.

[3] O. Kwon, N. Lee, and B. Shin. (2014) “*International Journal of Information Management Data quality management, data usage experience and acquisition intention of big data analytics*,” Int. J. Inf. Manage., vol. **34,** no. **3**, pp. 387–394

[4] “Data Lifecycle Management (DLM) _ Spirion.”.

[5] N. R. C. S. T. S. Panel and W. Dc.(2011) “for the Data Life Cycle Working
premise,”.

[6] “Bytes to Bucks_ The Valuation of Data - HealthCare Appraisers.”.

[7] P. Mikalef, I. O. Pappas, and J. Krogstie,(2018) “review and research agenda,” *Inf. Syst. E-bus. Manag*., vol. **16,** no. **3**, pp. 547–578.

[8] “7 phases of a data life cycle _ Bloomberg Professional Services.”.

[9] “Data Management Life Cycle Final report.”

[10] “5 Must-Have Data Engineering Skills to Land Big Data Engineer Job In 2019.” (2019)

[11] W. C. Lenhardt, S. Ahalt, B. Blanton, and L. Christopherson (2014), “Data Management Lifecycle and Software Lifecycle Management in the Context of Conducting Science,” vol. **2,** no. **1**, pp. 1–4.

[12] M. Talha, A. A. E. L. Kalam, and N. Elmarzouqi. (2019) “Science Direct ScienceDirect Big Data: Trade-off between Data Quality and Data Security Big Data: Trade-off between Data Quality and Data Security,” *Procedia Comput. Sci*., vol. **151**, pp. 916–922.

[13] K. Abouelmehdi, A. Beni-hssane, H. Khaloufi, and E. Nationale. (2017) “ScienceDirect Science Direct Big data security and privacy in healthcare: A Review Big data security and privacy in healthcare: A Review,” *Procedia Comput. Sci*., vol. 113, pp. 73–80.

[14] (2019) “International Journal of Information Management Healthcare big data processing mechanisms: The role of cloud computing,” *Int. J. Inf.Manage*., vol. **49,** no. **May**, pp. 271–289.

[15] V. Palanisamy and R. Thirunavukarasu. (2017) “Implications of big data analytics in developing healthcare frameworks – A review,” *J. KingSaud Univ. - Comput. Inf. Sci*.

[16] “20 A note on energy-efficient data, services, and memory management in Big Data Information Systems _ Elsevier Enhanced Reader.pdf.”.

[17] A. M. Cox, W. Wan, T. Tam, A. M. Cox, W. Wan, and T. Tam. (2018) “Acritical analysis of lifecycle models of the research process and research data management,”.

[18] V. Ahmed, A. Tezel, Z. Aziz, and M. Sibley. (2020) “The future of Big Data in facilities management: opportunities and challenges,” vol. **35,** no. **13**, pp. 725–745.

[19] “Naive Bayes Classifier - Towards Data Science.”.

[20] Q. Guo, X. Wang, X. Liang, and X. Wang. (2018) “Risk evaluation model of the whole life cycle of a wind power project in China: Based on the fuzzy theory,” vol. **0502**.
[21] C. Weng. (2019) “Clinical data quality: a data life cycle perspective,” *Biostat.&Epidemiology*, vol. 0, no. 0, pp. 1–9.
[22] H. Chen and H. Chen. (2018) “Personalized recommendation system of e-commerce based on big data analysis,” vol. **0502**.
[23] Y. Wang, J. He, H. Zhao, Y. Han, X. Huang, and Y. Wang. (2018) “Intelligent community medical service based on internet of things,” vol. **0502**.
[24] C. Wang and C. Wang. (2018) “A study of ranking behavior factors for online shopping by grey relation estimation,” vol. **0529**.
[25] C. Zheng and H. Huang. (2018), “Analysis of technology diffusion in agricultural industry cluster based on system dynamics and simulation model,” vol. **0529**.
[26] G. Swathi. (2018) “Secure data storage in cloud computing to avoiding some ciphertext attack,” vol. **2667.**
[27] B. A. H. Abul-huda and P. Mahnti. (2013) “Application of databases in discrete simulation data modeling,” vol. **2667.**
[28] S. Madan. (2019) “k-DDD Measure and MapReduce Based Anonymity Model for Secured Privacy-Preserving Big Data Publishing,” vol. **27**, no. **2**, pp.177–199.
[29] R. V Yampolskiy. (2013) “Efficiency Theory: A Unifying Theory for Information, *Computation, and Intelligence*,” vol. **0529**.
[30] B. K. Sarkar, “Big data for secure healthcare system: a conceptual design,” *Complex Intell. Syst*., vol. **3**, no. **2**, pp. 133–151, 2017.
[31] “7 STAGES.”
[32] A. Oracle and W. Paper. (2008) “Product Lifecycle Management in the Medical Device Industry Product Lifecycle Management in the Medical Device Industry,” no. January.