Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Review of Data Quality Issues in Cloud Computing

B. Swetha

Research Scholar Dept. of Computer Science Sri Padmavati Mahila VisvaVidyalayam

Tirupati.

swetha.sahana@gmail.com

Dr. K. Usha Rani

Professor Dept. of Computer Science Sri Padmavati Mahila VisvaVidyalayam

Tirupati.

usharani.kuruba@gmail.com

Article Info
Page Number: 1997-2005
Publication Issue:
Vol. 72 No. 1 (2023)

Abstract—Data is important and it is part of our everyday lives. But for some organization's it is one of the most important assets as the quality of data can have a huge of impact on the functionalities of these organizations. Due to this fact, there are multiple efforts and researches have gone into the understanding and improvements on the "Data Quality" (DQ). The storage of the data in such organization is happening in the Cloud, now-a-days. Therefore, the DQ can be a challenging task due to the variety of dimensions and diversity of available data in the cloud computing environments for the organization's effective functionalities. This study provides a clear review on the current state of research in cloud related to quality of data lead towards future direction of research

Keywords— Data Quality, Cloud Computing, Quality Dimensions, Machine Learning, Data Quality Assessment, Missing Data, Data Transformation

Article History
Article Received: 15 April 2023
Revised: 24 May 2023
Accepted: 18 June 2023
Publication: 06 July 2023

I. INTRODUCTION

Data can be considered as information which is translated into the form for processing, and the users of it have the ability of storing, retrieving and elaborating through appropriate software processes, which later can be communicated via Internet to remote requirements, if any. The impact of the quality considerations on the data in any given organization can be short-term like affecting the customer-employee relation or long-term affecting the decision making in the organisation. Thus, quality of the available data should be such that it promotes further enhancements and growth of the organization in its all dimensions.

If the data quality cannot be maintained properly (poor data quality), there would be a lots of missing opportunities in business as the data being used does not depict the clear picture of the situation. In such a scenario, it becomes absolutely necessary to pay attention to the quality of data being used for making decisions. Poor DQ can cause various social and financial issues. Though companies do make proper attempts for improving DQ with various applications and tools, their improvement efforts leads to pivot narrowly on accuracy. When

the quality of data is good, it can lead to better decision making along with the added trust of customers on the employees and organisations [1]. In addition, in these days, all organizations have mechanisms to store and use the data assets in Cloud computing environments. Cloud is one of the platforms used to store data in which to improve accessibility by reduce costs. Cloud computing represents a shift away from generic computing where everything is offered as a service that is delivered over the Internet to customers. This feature adds more complexity in the context of data quality due to the fact that the cloud is an out-sourced facility or service as far as the organizations are concerned.

High quality data has been defined by different authors in different ways. Some people define it as the "degree to which a set of characteristics of data fulfil the requirements" or data that is "fit for use by its consumers". The typical characteristics of data that can be used to analyse the quality of data are completeness, accuracy, correctness etc., Giving a rough idea about data quality without actually going into the depth of analysing is an easy task but to really find out the problems in data quality and understand the problems that can result from it is a complicated task.

One of the important issues and damage caused due to poor quality data is, huge amount of capital – both in terms of technical and monitory - could be wasted apart from having additional workload on employees and processing the data [1]. These kinds of wastages lead poor throughput of the organisation. The DQ of the organization stored in the cloud which normally is owned and managed by a third party, from the utility point of views.

This paper is organized as follows: Section II describes the reasons for poor DQ. Section III describes about dimensions of DQ. Section IV describes on the data quality in cloud computing environment. Section V describes the points related to Data Quality Assessment (DQA). Section VI mentions the open research issues in this domain and Section VII describes the role of Machine Learning (ML) in managing DQ. Conclusion is given in section VIII.

II. REASONS FOR POOR DATA QUALITY

The DQ is integral to its usability. . When data is not of good quality, it cannot be used for the intended purpose. Understanding the root causes of poor data quality is the first step in resolving and preventing the issue. As per research studies [1, 2] there are five reasons why data quality deteriorates.

- Human error: "People make mistakes, and when operators are hand-keying data, errors are expected, such as missing details, typos or putting the data in the wrong fields".
- Inconsistent data-entry standards: A company's data quality will undoubtedly deteriorate if there are no guidelines for how data should be entered or recorded.
- Poor data integration: It's be aware of that DQ vary seriously over all the systems they are drawing data from. It is noted that redundancy of data is an important provocation to face as these redundancies will skew any survey.

- Lack of authoritative data source: An authoritative data source must include just one version of the truth. "The reporting won't be accurate on the other end if there are issues with the organization's data but fixes aren't taking place at the source. For example, in the case of the state data issue mentioned above, if the reporting was fixed but no one went back to make sure the data was cleaned up at the source, the business would suffer.
- Not keeping data up to date: “Data gets abandoned and forgotten”. We must ensure that it doesn't. Making sure there isn't any duplicate, erroneous, or out-of-date information in these places will help to alleviate this problem by keeping data and other relevant information up to date.

Data is always changing. Hence, regular phases of data collection, transformation and updation are necessary. However, each of these stages in the data lifecycle presents opportunities for poor practice or errors to cause a degradation of the DQ [3]. Hence, the fight against bad data quality is never-ending. Nonetheless, by being aware of these probable causes and utilizing the appropriate technological assistance, one may believe to have a solid foundation for high-quality data.

III. DIMENSIONS OF DATA QUALITY

DQ has often been addressed with various aspects of the quality parameters of data. In this context, dimensions are commonly used to define and evaluate the quality of data. As anyone can guess, there is no single quality of data definition for DQ. Data quality is defined by Wang and Strong [3, 4] as data that are suitable for use by data consumers. A criteria of data quality set includes accuracy, completeness, consistency, and timeliness, are commonly referred to as DQ. The various dimensions on which data quality can be represented depend to some extent on one or more of these attributes. The effectiveness of the corresponding data applications might be severely impacted by poor (or low) data quality [3, 4].

The following six dimensions for DQ is given by the International Data Management Association (IDMA) provides a complete list of the data quality dimensions as represented in Figure 1. IDMA defines the six data quality dimensions [5] in the following way:

Fig. 1. Data Quality Dimensions [9]

- Completeness: The amount of data stored against the total of “100% complete” data;
- Uniqueness: Based upon the identification of the thing no data will be stored more than once.
- Timeliness: How much amount of data is represented with respective to the required time.
- Validity: If the entered data is according to syntax (format, type, range) then only it is valid of its definition;
- Accuracy: How well the data represents the real world data.
- Consistency: The lack of variation between two or more representations of a thing against a definition.Yet another popular studies [5] and later standardized by ISO/IEC 25012 [6] the following are the ones as group of dimensions:
- Accessibility: The extent to which information may be accessible in a certain context of use, including appropriateness of depiction.
- Accuracy: The extent by which properties of correctly represents the true value of the required thing.
- Completeness: For all properties expected and instance of entity will be given to every entity.
- Consistency: Representation of the data must be in constant manner so that everywhere the data will be same

The above explanation show clearly how data quality is achieved and measured- understood for various applications where data are more important. We observe that though there are little drifts in making and defining the dimensions majority of the dimensions are being used [8], the definitions found in ISO/IEC 25012 are also quite generic, and requires to be updated considering the context and application domains.

IV. DATA QUALITY ISSUES IN CLOUD COMPUTING

With the latest technology, cloud computing has been used by numerous companies worldwide. Using cloud computing in an organization yields several benefits while it also has some downsides [7]. One such challenge is that data quality. Data quality arises when data, as well as the data applications, move around in the cloud and in between cloud and on-premise due to requirements of application execution. Certain investigations as detailed below are the main causes for affecting data quality in Cloud Computing:

- Movement of data - There is a risk in formatting issues or data loss when data are moved for a specific reason, whether within the cloud or even between the cloud and infrastructure. In addition, it might have a reliable timestamp and other minor problems that could potentially degrade the data's quality.

- Huge Quantity – Data stored in the cloud has the advantage of being quickly available. It is simple to store vast volumes of data and information on the cloud because it is also scalable. And because there is so much data, there is a chance that the quality would suffer or even be hampered by it, making it more difficult for the cloud service provider to uphold the standard of cloud computing.

- Update automatically -Cloud services are on-demand and interactive processes [8, 9]. Due to this set of reasons, they are getting updated automatically by means of appropriate services. The cloud-based tools might not notify the auto-updates and associated issues, if any arise, to the customers of it. Therefore, whenever data is getting modified in the auto-update process it can create issues in DQ

V. DATA QUALITY ASSESSMENT

When evaluating whether data satisfies a company's quality standard, scientists and statisticians do a data quality assessment, or DQA. The standard could specify a particular amount, kind, or data format be used for a particular project or operation. It may also include a collection of rules and strategies used to gather, purify, and apply data.

The goal of doing a DQA is to use company data to reveal process inefficiencies and problems. But erroneous defaults, mismatched structures, and empty data fields are often simple to spot. DQA seeks to identify the root causes of more complicated prob issues.

A DQA focuses on using the criteria or dimensions for data quality and assessing the systems and methods for gathering data to ascertain if they are likely to deliver high quality data over time or not. In other words, it is likely that good quality data will be produced provided the requirements for data quality are met and the data collection approach is well thought out [9].

Record linkage and integrity constraints are just two examples of the many strategies that have been suggested to analyse and enhance the quality of data. Due to the complexity and diversity of these strategies, the academic and practitioner groups in data management have made an effort to develop methodologies that aid in the selection, personalization, and use of DQA and improvement techniques. A set of rules and methods known as a data quality methodology establishes a logical procedure for evaluating and enhancing the quality of data starting with input information characterizing a particular application context. We observe that all DQ approaches must include evaluation of DQ, which evaluates the caliber of data collections along pertinent quality dimensions. DQA typically consists of the following steps [10]:

- **Data analysis**: This looks at data structures and conducts interviews to fully comprehend data and any underlying architectural and management principles.

- **Data quality requirements analysis**: which solicits feedback from administrators and consumers of the data in order to pinpoint quality problems and establish new quality goals.

- **Identification of critical areas:** which chooses the most pertinent databases and data flows to be evaluated
- **Process modelling**: This gives a representation of the procedures used to create or update data.
- **Quality Measurement**: Based upon quantitative metrics measurements will be objective, by users and data administrators qualitative evaluation will be subjective

VI. OPEN RESEARCH ISSUES IN DQ

The advancement in DQ is still evolving in various application domains in the real world. Since we are in the world of cloud computing, the issues related to DQ are still more complex. One of the key problems in this situation is important to note: data quality characteristics are still connected to the usage environment. Research on the various aspects of data quality is still done in the context of well-structured data being kept in a cloud environment. Reviewing some of the data quality aspects revealed that completeness, consistency, accuracy, and timeliness of the data were very important in obtaining high DQ. In the context of big data and cloud computing technology, the measurement and assessment methods of these critical dimensions could be different and require more technology oriented solutions. In our research we aim to attempt to use machine learning approach to address one or two issues of DQ in cloud computing environment.

VII. MACHINE LEARNING APPROACHES FOR DQ

Automated processes for evaluating data quality, cleaning it up, and other related tasks have developed to the point where they now use Artificial Intelligence (AI)-based approaches like Machine Learning (ML) techniques, with the expectation of a vast array of improvements for the enhancement of organizational business processes. By utilizing distributed improved computations, ML methodologies hope to cut down on the time spent on managing data quality as well [11].

In contrast to conventional computer software, which is intended to perform in a certain precise manner, ML techniques learn from the data they are fed. In essence, ML picks up knowledge the same way people do—through mistakes and a variety of experiences

A generic understanding on ML is depicted in figure-2.

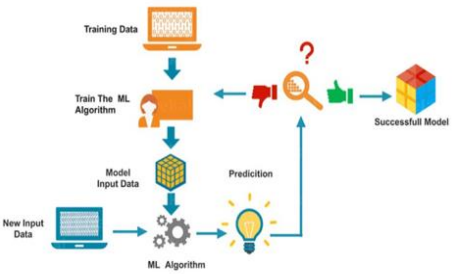

Figure -2 A generic machine learning model – approach [12]

Classifications and predictions are well trained by ML algorithms. Decision making in applications are based on key metrics of growth which was showed in the above figure-2.

Following are the tasks related to DQ that could be benefited over the use of ML approaches:

- Automating Data Capture - Much of the poor quality capturing related problems can be resolved using ML as a solution model.
- Using ML techniques, automated data entry and ingestion can raise the quality of the data.
- Error reduction - When people enter or amend data, they run the risk of making mistakes. However, ML methods essentially eradicate these mistakes.
- Detecting Data Errors - A data set's overall quality and usability can be impacted by even the tiniest inaccuracy. Data mistakes can sometimes be found using ML methods
- Finding Duplicate Records – Machine learning is also good at finding duplicate data or records. When data is collected from many sources, duplicate data might be a problem.
- Filling in Missing Data – In automation systems data cleaned by the programming rules explicitly, to filling the missing data gaps is impossible without intervention in adding data resources. ML can do assessments for data missing based upon knowing the current situation.
- Validating Data – Validating the data in the system for accuracy by comparing it to existing data sources – this also can be automated with the use of ML
- Supplementing Existing Data – DQ can be improved by adding original data in certain situations and later by knowing the original data sets based upon the expansion of additional data. ML can be used to detect the patterns and connections between the data points.
- Accessing Relevance – ML approaches can also be used to recommend supplementary data that is pertinent to the original data set and to identify data within the original data set that is no longer valuable or relevant. By identifying the meaningless data points, we may efficiently clean the data.
- Scaling DQ Operations can also be performed using ML as an automated mechanism in the context of cloud computing and big data.

VIII. CONCLUSION

In this review details related to reasons for poor data quality, critical data quality dimensions and data quality assessment methods have been successfully explained. As new technology become available, data in organization is no longer stored in the database. In this context, cloud computing environments are alternate large storage can be considered and the data

quality issues in this component has also been explained. Later this review suggests future consideration for data quality research in organizations could be based on ML approaches.

REFERENCES

1. A Survey on Data Quality: Classifying Poor Data by NunoLaranjeiro, SeymaNurSoydemir, and Jorge Bernardino https://ieeexplore.ieee.org/abstract/document/7371861
2. A Framework for Analysis of Data Quality Research Richard Y. Wang, Veda C. Storey, and Christopher P. Firth https://ieeexplore.ieee.org/document/404034
3. Discovering Data Quality Problems The Case of Repurposed Data by Ruojing Zhang, Marta Indulska, ShaziaSadiqhttps://link.springer.com/article/10.1007/s12599-019-00608-0
4. An Overview of Data Quality Frameworks by CORINNA CICHY1,2 AND STEFAN RASS 1 https://ieeexplore.ieee.org/document/8642813
5. Methodologies for Data Quality Assessment and Improvement https://dl.acm.org/doi/abs/10.1145/1541880.1541883
6. "Data Definition in the Cambridge English Dictionary," 2015. Available: http://dictionary.cambridge.org/dictionary/english/data
7. H. Baldwin, "Drilling Into the Value of Data."http://www.forbes.com/sites/howardbaldwin/2015/03/ 23/drilling-into-the-value-of-data/ W. Fisher and B. R. Kingma, "Criticality of data quality as exemplified in two disasters," Information & Management, vol. 39, no. 2, pp. 109–116, 2001.
8. M. Ge and M. Helfert, "A review of information quality research—develop a research agenda," in International Conference on Information Quality 2007.
9. Batini, M. Palmonari, and G. Viscusi, "Opening the closed world: A survey of information quality research in the wild," in The Philosophy of Information Quality. Springer International Publishing, 2014, pp. 43–73.
10. L. Sebastian-Coleman, Measuring data quality for ongoing improvement: a data quality assessment framework. Newnes, 2012.
11. Batini and M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques, 1st ed. Springer Publishing Company, 2010.
12. E.R. Hruschka, A.J.T. Garcia, E.R. Hruschka Jr., N.F.F. Ebecken, On the influence of imputation in classification: practical issues, Journal of Experimental and Theoretical Artificial Intelligence (2009) 43–58.
13. E.R. Hruschka, E.R. Hruschka Junior, N.F.F. Ebecken, Towards efficient imputation by nearest-neighbors: a clustering-based approach, 6th Australian Conference on Artificial Intelligence, Spring-Verlag, 2004, pp. 513–525.
14. G. Jagannathan, R.N. Wright, Privacy-preserving imputation of missing data, Data & Knowledge Engineering 65 (2008) 40–56.
15. M.K. Kerr, M. Martin, G.A. Churchill, Analysis of variance for gene expression microarray data, Journal of Computational Biology (2000) 819–837.
16. H. Kim, G.H. Golub, H. Park, Missing value estimation for DNA microarray gene expression data: local least squares imputation, Bioinformatics (2005) 187–198.

17. K.-Y. Kim, B.-J. Kim, G.-S. Yi, Reuse of imputed data in microarray analysis increases imputation efficiency, BMC Bioinformatics 5 (2004) 160.
18. R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1–2) (1997) 273–324.
19. Kononenko, I. Bratko, E. Roskar, Experiments in automatic learning of medical diagnostic rules, Tech. rep., Jozef Stefan Institute, Ljubjana, Yogoslavia, 1984.
20. S. Oba, M.-A. Sato, I. Takemasa, M. Monden, K.I. Matsubara, S. Ishii, A Bayesian missing value estimation method for gene expression profile data, Bioinformatics 16 (November 2003) 2088–2096.
21. Pyle, Data Preparation for Data Mining (The Morgan Kaufmann Series in Data Management Systems), Morgan Kaufmann, March 1999.
22. Y. Ren, G. Li, J. Zhang, W. Zhou, The efficient imputation method for neighborhood-based collaborative filtering, Proceedings of the 21st ACM international conference on Information and knowledge management, CIKM '12ACM, 2012, pp. 684–693.
23. D.B. Rubin, Formalizing subjective notion about the effects of non respondents in samples surveys, Journal of the American Statistical Association (1977) 538–543.
24. M. Saar-Tsechansky, F. Provost, Handling missing values when applying classification models, Journal of Machine Learning Research (2007) 1623–1657.
25. J. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall/CRC, 2000.
26. J.L. Schafer, J.W. Graham, Missing data: our view of the state of the art, Psychological Methods 7 (2002) 147–177.
27. R. Tibshirani, T. Hastie, D. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proceedings of the National Academy of Sciences, 2002, pp. 6567–6572.
28. O.G. Troyanskaya, M. Cantor, G. Sherlock, P.O. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 6 (2001) 520–525.
29. V. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proceedings of the National Academy of Sciences, 2001, pp. 5116–5121.