Improving Data Quality using Big Data Framework: A Proposed Approach

Shikha Soni¹, Amritpal Singh²

¹Research Scholar, Department of CSE, Lovely Professional University, India
²Assistant Professor, Department of CSE, Lovely Professional University, India

shikhasoni355@gmail.com
amritpal.17673@lpu.co.in

Abstract. The field of Big Data and Big Data Analytics is one of the most emerging fields in today’s technology. Data being the most important aspect of Big Data, it is important to select it in such a way that it provides maximum efficiency when analysed. Data quality is a concept which is taken in account while measuring the quality of the dataset. Data quality imposes certain defined characteristics on data which needs to be fulfilled up to a certain extent for the data to be considered reliable for analysis. In this paper, we have proposed an approach which introduces a ‘believability factor’ to measure the reliability or believability of the dataset by taking a sample of the unstructured dataset. This paper along with believability factor proposes the methodology to calculate execution time of the sampled dataset and Mean Absolute Error between the believability of the sampled and unstructured dataset.

Keywords. Data quality, Big Data, Spark, Believability, Unstructured data

1. Introduction
There have been a lot of changes and improvements in the field of database systems since the last decade. With the phrase “Data is the new oil”, it is pretty clear that data has become one of the most important aspect in technology and business field. In the early 2000s, the data which was stored and used everywhere was majorly relational i.e. in a tabular format consisting of rows and columns. But since then there have been a lot of changes resulting in majority of the data being unstructured i.e. not able to use in a tabular format [15]. Since the boom of social media applications and usage of internet in day-to-day life, it has become extremely difficult to keep the data structured and ready to analyze. With over 80% of the data being unstructured, more and more data-based and technology related organizations are focusing on cleaning the data. This makes analyzing the data easier and results in higher accuracy and efficiency. Another fields where data is an important aspect to yield better results is Business Analytics and Business intelligence (BI) [10].

Data quality is the way to observe or setting up standard set of factors in a given data to measure the quality of data. These data quality factors are taken in account while deciding the usage of data for analysis by business or technological organizations. These factors include terms such as accuracy, timeliness, completeness, reliability and consistency of data. Dirty data can lead to a lot of issues which are faced by organizations and lead to economic loss. For example, if an e-commerce organization delivers its products to a wrong customer or wrong shipping address due to inaccurate customer data stored, it can lead to bad reputation or value loss for the organization. Hence, data cleaning and data quality are important aspects that need to be taken in account [16] [18].
Big Data is one of the fields where data quality plays an important role. With most of the data generated in Big Data being in unstructured format, it has become important to maintain the data quality to make the analysis of data easier. The factor which decreases the efficiency of the analysis drastically is the speed at which the data is being generated [17]. There are various tools which analyze the datasets in batch and streaming mode but the data quality factor is not taken in account. There are various data cleaning tools available which clean the data to make the analysis and decisions more efficient but time is a factor which needs to be taken in account. The characteristics of data quality is often challenged by the characteristics of Big Data: volume, variety, velocity, veracity and value [11].

Considering the challenges of data quality in Big Data, we propose an approach which can be used for applying data quality measures in the unstructured datasets in a way to reduce the space being used for computation leading to the analysis of the dataset. This is achieved through sampling the dataset and introducing a “believability factor” which takes in account the sample of the dataset as well as the original dataset. We propose an approach which is able to: (i) calculate believability factor of the given unstructured dataset, (ii) calculate the time evaluation of the sampled dataset, and (iii) calculate Mean Absolute Error between the accuracy measure of sampled dataset and original unstructured dataset.

2. Literature review
Maryam Ghasemaghaei and Goran Calic (2019) [1] proposed a structured model by surveying various organizations to observe how big data affects an organization’s decision. For this, the authors took information of 130 different organizations and used “Organizational Learning Theory and Data Quality Factor (DQF)” for studying the effects of big data. The authors have tried to explain the data quality factors and how it affects the decisions of an organization along with big data. There are four major hypothesis which were formulated to show “big data utilization”. In the proposed structural representation of the research, the hypothesis consists of four major segments on which they work: (1) “big data utilization”, (2) “data quality”, (3) “data diagnosticity”, and (4) “decision quality”. It was concluded that the standard of the decision taken by an organization does not depend on big data application but on data quality and characteristics.

Tadhg Nagle et. al. (2020) [2] proposed and conducted a study among various organizations to observe the data quality of said organizations and reached the conclusions based on the results. The authors collected about 75 different datasets in 2 years of time from various organizations and businesses. The authors started the explanation with an example of an Irish helicopter crash which resulted due to errors in the database of the system. This example explains how a bad database can have adverse effects on something. The authors have mentioned five protocols that are needed for acquiring the data quality: (1) easy to recognize, (2) simple to carry out, (3) comparatively fast with low cost, (4) productive in providing results, and (5) can be applied among different dominions.

The authors used a specific protocol called FAM for assessment of the data. After the assessment, it was observed that the data quality was 71% as well as 29% of the recorded databases were of poor quality. It was also observed that the organizations lose millions of dollars due to these inconsistencies in database. The authors took information from five different industries to check the levels and accuracy
of the data quality and found out that the median and mean value ranges from 50-60%. It was observed that about 50% of the database consists of at least one critical error.

The authors concluded that the organizations tend to show an unhealthy forbearance for poor data and accentuate the changes an organization has to make. The authors also defined a topology to set actions for the common occurring data errors.

Yoram Timmerman and Antoon Bronselaer (2019) [3] proposed a framework to analyze the “data quality in information system” using the “rule-based measurement principle”. The suggested framework can handle the problems in dataset that arises due to incorrect implementation, explanation of data collection and its authorization. The framework also deals with the precariousness of data. The authors took a “survey data” to study the quality measurements and a practical example to show the application. The authors concluded that the quality of data depends upon various factors and thresholds.

Manel Souibgui et. al. (2019) [4] proposed an outline of data quality viewpoint using Extract, Transform and Load (ETL) process. The authors have also used different tools for ETL process and provided an observation of how every tool examines the characteristics of data quality. The analysis of ETL process has been done using two open source tools namely “Talend Data Integration” and “Talend Data Quality”. The demonstration has been carried out in six consecutive phases. The first three phases include the instigation, categorization of data quality issues with ETL process, and traditional data quality approaches in ETL process. The next three phases consist of comparison between tools, observations using ETL explanations, and shortcomings of the tools.

By this study, the authors have also showed that there are two ways in which data quality issue in ETL can be approached: (i) “Process centered” and (ii) “Data centered”. They have also listed the shortcomings using the said open source tools. There are a lot of issues like diverse datasets which makes data quality unresolved; this research combines the ETL process with data quality issue. The authors successfully showed that data quality characteristics help in improving the constituents of ETL mechanism.

Shixia Liu et. al. (2018) [5] proposed to traverse different research topics to extract knowledge about data quality. The research can be summarized in three steps: (i) taking different types of data to encapsulate common ways of making the data quality better at various stages, (ii) propose a “visual analytics” foundation, and (iii) issues and shortcomings taking in data and humans as factors. The authors have reflected the importance of data quality by explaining how ambiguous data leads to discarding of data in later stages and making the data unfit.

The backdrop of this study has taken in account various types of dataset. The datasets have been categorized as tabular and non-tabular data. There are existing tools for tabular data already but non-tabular data takes efforts due to diversification of the data. One of the examples taken to describe this point is the usage of detection of words, copied content, wrong grammar usage, etc. in datasets of text format. The authors have mentioned different types of data (text format, media-based, graphical format, etc.) and how each data type deals with cleaning data and improves data quality in certain ways. The authors proposed a framework for data quality using visual analytics to assist a user in figuring out the issues of data that needs to be analyzed.

Naresh Sundar Rajan et. al. (2018) [6] proposed a data quality archive for analysis of data quality. The authors have taken in account the various data quality tools and frameworks which are used to analyze the quality of various types of data. The backdrop of this study consists of taking in account various articles for review. Using this, the authors took into consideration the description, notion, concepts and associations. The factors were then taken to create a “meta-model” and executed in a knowledge depository. The authors successfully created a meta depository and further broadened this framework for analyzing various data models in existence.

Danilo Ardagna et. al. (2018) [7] have proposed a strategy to provide data quality assessment depending on the user’s requirement. The dataset used for conducting this study consists of information taken from a smart city scenario-based study. The main goal of this proposed strategy is to input sample data in such a way that the data quality is satisfactory. The backdrop of this strategy consists of the concept of unstructured format in Big Data and how the continuous flow of huge amount of data affects
the data quality. To solve this issue, this study proposed appropriate sampling of data in such a way that the analysis of data is easier because of less quantity of data. The authors have proposed a “confidence value” to show how reliable the data is. The authors have also formulated evaluation for time and budget to later calculate the accuracy of the sample dataset. The implementation of the study is explained through a comparison of real and predicted data.

### Table 1. Literature review table.

| Year | Author | Description | Findings |
|------|--------|-------------|----------|
| 2018 | “Danilo Ardagna” | Proposed a strategy to provide data quality assessment depending on the user’s requirement. | Formulated evaluation for time and budget to later calculate the accuracy of the sample dataset. |
| 2018 | “Naresh Sundar Rajan” | Proposed a data quality archive for analysis of data quality. | Created a meta depository and further broadened this framework for analyzing various data models in existence. |
| 2018 | “Shixia Liu” | Proposed to traverse different research topics to extract knowledge about data quality. | Visual analytics framework |
| 2019 | “Manel Souibgui” | Proposed an outline of data quality viewpoint using ETL process. | The authors successfully showed that data quality characteristics help in improving the constituents of ETL mechanism. |
| 2019 | “Yoram Timmerman” and “Antoon Bronselaer” | Proposed a framework to analyze the “data quality in information system”. | Framework that can handle the problems in dataset that arises due to incorrect implementation and data collection. |
2019

“Maryam Ghasemaghaei” and “Goran Calic”

Proposed a structured model by surveying various organizations to observe how big data affects an organization’s decision.

It was concluded that the standard of the decision taken by an organization does not depend on big data application but on data quality and characteristics.

2020

“Tadhg Nagle”

Proposed and conducted a study among various organizations to observe the data quality of said organizations and reached the conclusions based on the results.

The authors concluded that the organizations tend to show an unhealthy forbearance for poor data and accentuate the changes an organization has to make. The authors also defined a topology to set actions for the common occurring data errors.

3. Data Quality

Data Quality, as the name refers, is the quality of a piece of information. Data Quality can be in both qualitative and quantitative form. If data follows certain characteristics and can be used for successfully analysing and generating results, then it can be said that it is a data of quality [14].

3.1 Data quality characteristics

Data quality is defined by following characteristics:

3.1.1 Accuracy. Accuracy is one of the most important factors among data quality characteristics as it is directly proportional to the quality of analyzed data. Data accuracy can be defined as a data which is precise.

3.1.2 Completeness. Completeness can be defined as the lack of missing or incomplete data. This helps the organizations to make sure if a data is usable, since incomplete data produces inaccurate results and cannot be used.

3.1.3 Reliability. Reliability can be defined as data which is dependable enough to use. For example, a data is not reliable enough if it consists of two different entries under same name in two different locations.

3.1.4 Timeliness. Timeliness can be defined as how up-to-date the given data is. For example, a dataset created in 1980 will not be as useful as one created in 2019 due to it being comparatively older.
3.1.5 Relevance. Relevance can be defined as how pertinent the given information in a data is. For example, to be able to vote in elections, the age of a citizen is relevant information whereas his religion or gender is not [19] [8] [12].

4. Data Quality Service Approach
Data Quality is a concept which can lead to better efficiency while analyzing unstructured dataset(s). With the data generated being at an extremely fast pace in Big Data, the timeliness of the data is very short. This makes time management an important issue in data quality management in Big Data. Along with the usual data quality characteristics, there are other factors which need to be taken in account as well [20]. One such factor of data quality is ‘believability’. Believability can be defined as the trustworthiness of a data. It is an important factor while analyzing a dataset since the dataset being used has to be from a reliable source.

The main concept behind this approach is to reduce the computation time and space by taking a sample of the dataset by Probability Sampling. We then use the sample dataset to measure ‘believability factor’ and use it to calculate time taken to execute the sampled dataset. At last, we have calculated Mean Absolute Error (MAE) to measure the difference between the believability of sampled and original dataset.

The proposed approach consists of three parts:
- Believability factor
- Time Evaluation
- Mean Absolute Error (MAE)

We have proposed this Data Quality Service approach for unstructured Big Data. There are various Big Data tools for which this approach can be followed. We have taken Apache Spark as the Big Data tool for this approach as it provides both batch as well as stream processing of datasets and is one of the most established Big Data tools.

4.1 Sampling
Large amount of unstructured dataset takes a lot of computational space and time to get analysed. Taking a sample from the dataset in such a way that elements are in equal occurrence state is a way of making sure that the analysis of data is being done in a non-partial manner.

We have proposed taking a sample dataset through Simple Random Sampling for ease and making sure that every element in the dataset have equal chances of being included.
4.2 Believability factor
We have taken sampling from an unstructured dataset to reduce the computation power and storage in the system. We have proposed a believability factor ‘b’ as the ratio of selected dataset through sampling and the original dataset. The factor can be mathematically expressed as

$$b = \frac{\text{Selection of sample dataset}}{\text{Total Dataset}}$$  \hspace{1cm} (1)

The range of ‘b’ is between 0 to 1.

4.3 Execution time
With the data generated being at an extremely fast pace in Big Data, the timeliness of the data is very short. This makes time management an important issue in data quality management in Big Data. We have formulated an execution time of the sampled dataset to reduce the computation time of the system.

Execution time has been formulated by using an approach explained in [7] and [9]. These studies conducted explained how in Big Data tools like Hadoop and Spark, the execution time can be reduced using various regression models. The study conducted in [x] concluded that Support Vector Machine (SVM) is the best regression model to reduce the execution time. The execution time has been formulated using amount of data and number of cores in a system as functions.

From [9], it can be observed that the quantity of data being used for analysis is directly proportional whereas the number of cores in a system is inversely proportional to execution time. This can be mathematically represented as

$$T_b \propto \frac{Q_D}{n}$$  \hspace{1cm} (2)

We have introduced a ‘believability coefficient’ as a constant value which will be used to formulate execution time.

$$T_b = b'(\frac{Q_D}{n})$$  \hspace{1cm} (3)

where $T_b =$ execution time,

$Q_D = $ quantity of the dataset

$n = $number of computational cores in the system

$b'$ = believability coefficient

4.4 Mean Absolute Error (MAE)
We have formulated Mean Absolute Error (MAE) to calculate the error (in percentage) between the data quality dimension of sampled and original dataset. If MAE is calculated and found out to be satisfactory, we can replace original datasets with sampled datasets for further analysis of data.

Error rate of believability dimension is inversely proportional to believability factor ‘b’ and the directly proportional to the difference between believability coefficient of original and sampled datasets. This can be mathematically represented as

$$\text{MAE} = \frac{1}{\text{Cardinality}(b)} \sum_{b \epsilon b} \left( \frac{|A_b - F_b|}{A_b} \right)$$  \hspace{1cm} (4)

where $b'$ = believability coefficient

$b$=Believability factor

$A_b$= actual data quality dimensions

$F_b$= sample data quality dimensions
4.5 Working
Step 1. Decide the aim for selecting a dataset. In this approach, the aim is to select an unstructured dataset for sampling.
Step 2. Decide the data quality dimensions that needs to be fulfilled. In the proposed approach, the data quality dimension is believability.
Step 3. Formulate the evaluation baseline. In this case, the evaluation baseline is to calculate the MAE of believability between actual dataset and sampled dataset.
Step 4. Collect the required dataset from a reliable source.
Step 5. Select a small sample from the dataset. In this case, we are using Simple Random Sampling for ease and inclusion of each type of entered dataset.
Step 6. If sampling is done properly, we can carry on the approach as mentioned and if unsuccessful, we can go back to taking a sample again until successful.
Step 7. Calculate the believability factor ‘b’ using equation (1).
Step 8. For a given believability factor ‘b’, we will calculate the execution time (T_b) using equation (2).
Step 9. Calculate the Mean Absolute Error (MAE) for actual data quality dimensions and estimates due to sampling using equation (4).
Step 10. If baseline evaluation is satisfied, perform the next step as mentioned and if not, go back to calculating the believability factor again until baseline evaluation is satisfied.
Step 11. Generate report of the calculated equations and result.

4.6 Pseudocode of Data Quality Service
BEGIN
1. Select goals (believability of a sampled dataset)
2. Select data quality dimensions (believability)
3. Formulate evaluation baseline
   i. Believability factor
   ii. Execution time
   iii. Mean Absolute Error
4. Collect dataset
5. Perform Simple Random Sampling
6. IF
   Simple Random Sampling= Successful
   Go to previous step
   ELSE IF
   Simple Random Sampling= Unsuccessful
   Go to next step
7. Calculate the believability factor
8. Calculate the execution time
9. Calculate the Mean absolute error (MAE)
10. IF
    Satisfy evaluation baseline= Yes
    Go to next step
    ELSE IF
    Satisfy evaluation baseline= No
    Go to step 7 and repeat
11. Generate report
END
4.7 *Flowchart*

![Flowchart of Data Quality Service](image)

Figure 3. Flowchart of Data Quality Service
5 Conclusion
In this paper, we have provided an approach to analyse datasets in Big Data by taking a small sample using simple random sampling and using it to formulate believability factor, time execution and MAE between data quality dimension of sampled and original dataset. The goal of this paper is to provide a way of reducing the execution time and computational storage of datasets.

With data becoming one of the most important aspects of technology in various fields, it has become increasingly important to make sure that the data being used is efficient to provide good results using less resources. The constant developments in the field of Big Data and Big Data Analytics have improved many aspects of how we analyse data efficiently. However, data cleaning and data quality still remain big issues which are hard to tackle due to huge amount of data generated in a very short time. Data quality is a technique which if use in Big Data can make significant change in how we use our resources in analysing the datasets. Through this paper we tried to provide our idea through an approach of combining the two fields together for efficient usage of resources while analysing data.

7. References
[1] Ghasemaghaei, M., & Calic, G. (2019). Can big data improve firm decision quality? The role of data quality and data diagnosticity. Decision Support Systems, 120(December 2018), 38–49. https://doi.org/10.1016/j.dss.2019.03.008
[2] Nagle, T., Redman, T., & Sammon, D. (2020). Assessing data quality: A managerial call to action. Business Horizons, 63(3), 325–337. https://doi.org/10.1016/j.bushor.2020.01.006
[3] Timmerman, Y., & Bronselaer, A. (2019). Measuring data quality in information systems research. Decision Support Systems, 126(February), 113138. https://doi.org/10.1016/j.dss.2019.113138
[4] Souibgui, M., Attigui, F., Zammalı, S., Cherfi, S., & Yahia, S. Ben. (2019). Data quality in ETL process: A preliminary study. Procedia Computer Science, 159, 676–687. https://doi.org/10.1016/j.procs.2019.09.223
[5] Liu, S., Andrienko, G., Wu, Y., Cao, N., Jiang, L., Shi, C., … Hong, S. (2018). Steering data quality with visual analytics: The complexity challenge. Visual Informatics, 2(4), 191–197. https://doi.org/10.1016/j.visinf.2018.03.002
[6] Rajan, N. S., Gouripeddi, R., Mo, P., Madsen, R. K., & Facelli, J. C. (2019). Towards a content agnostic computable knowledge repository for data quality assessment. Computer Methods and Programs in Biomedicine, 177, 193–201. https://doi.org/10.1016/j.cmpb.2019.05.017
[7] Ardagna, D., Cappiello, C., Samà, W., & Vitali, M. (2018). Context-aware data quality assessment for big data. Future Generation Computer Systems, 89, 548–562. https://doi.org/10.1016/j.future.2018.07.014
[8] Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data Science Journal, 14, 1–10. https://doi.org/10.5334/dsj-2015-002
[9] Ataie, E., Gianniti, E., Ardagna, D., & Movaghar, A. (2017). A combined analytical modeling machine learning approach for performance prediction of MapReduce jobs in cloud environment. Proceedings - 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2016, 431–439. https://doi.org/10.1109/SYNASC.2016.072
[10] Taleb, I., Serhani, M. A., & Dssouli, R. (2019). Big Data Quality Assessment Model for Unstructured Data. Proceedings of the 2018 13th International Conference on Innovations in Information Technology, IIT 2018, 69–74. https://doi.org/10.1109/INNOVATIONS.2018.8605945
[11] Taleb, I., Kassabi, H. T. E., Serhani, M. A., Dssouli, R., & Bouhadioui, C. (2017). Big Data Quality: A Quality Dimensions Evaluation. Proceedings - 13th IEEE International Conference on Ubiquitous Intelligence and Computing, 13th IEEE International Conference on Advanced and Trusted Computing, 16th IEEE International Conference on Scalable Computing and Communications, IEEE International Conference on Cloud and Big Data Computing, IEEE International Conference on Internet of People and IEEE Smart World Congress and Workshops, UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld 2016, 759–765. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122
[12] Cichy, C., & Rass, S. (2019). *An overview of data quality frameworks*. IEEE Access, 7, 24634–24648. [https://doi.org/10.1109/ACCESS.2019.2899751](https://doi.org/10.1109/ACCESS.2019.2899751)

[13] Immonen, A., Pääkkönen, P., & Ovaska, E. (2015). *Evaluating the Quality of Social Media Data in Big Data Architecture*. IEEE Access, 3, 2028–2043. [https://doi.org/10.1109/ACCESS.2015.2490723](https://doi.org/10.1109/ACCESS.2015.2490723)

[14] Christopher Tozzi. (2020). *How to Measure Data Quality – 7 Metrics to Assess the Quality of Your Data*. Retrieved from https://www.precisely.com/ website: [https://www.precisely.com/blog/data-quality/how-to-measure-data-quality-7-metrics](https://www.precisely.com/blog/data-quality/how-to-measure-data-quality-7-metrics)

[15] https://profisee.com. (2020). *DATA QUALITY – WHAT, WHY, HOW, 10 BEST PRACTICES & MORE!* Retrieved from https://profisee.com/ website: [https://profisee.com/data-quality-what-why-how-who/](https://profisee.com/data-quality-what-why-how-who/)

[16] Mikhailouskaya, I. (2018). *Your guide to data quality management*. Retrieved from https://www.scnsoft.com website: [https://www.scnsoft.com/blog/guide-to-data-quality-management](https://www.scnsoft.com/blog/guide-to-data-quality-management)

[17] Talha, M., El Kalam, A. A., & Elmarzouqi, N. (2019). *Big data: Trade-off between data quality and data security*. Procedia Computer Science, 151, 916–922. [https://doi.org/10.1016/j.procs.2019.04.127](https://doi.org/10.1016/j.procs.2019.04.127)

[18] Schuh, G., Rebentisch, E., Riesener, M., Ipers, T., Tönnes, C., & Jank, M. H. (2020). *Data quality program management for digital shadows of products*. Procedia CIRP, 86, 43–48. [https://doi.org/10.1016/j.procir.2020.01.027](https://doi.org/10.1016/j.procir.2020.01.027)

[19] Günther, L. C., Colangelo, E., Wiendahl, H. H., & Bauer, C. (2019). *Data quality assessment for improved decision-making: A methodology for small and medium-sized enterprises*. Procedia Manufacturing, 29, 583–591. [https://doi.org/10.1016/j.promfg.2019.02.114](https://doi.org/10.1016/j.promfg.2019.02.114)

[20] Tian, Q., Liu, M., Min, L., An, J., Lu, X., & Duan, H. (2019). *An automated data verification approach for improving data quality in a clinical registry*. Computer Methods and Programs in Biomedicine, 181. [https://doi.org/10.1016/j.cmpb.2019.01.012](https://doi.org/10.1016/j.cmpb.2019.01.012)