Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 5, 1032 – 1039

# Towards Standardization of Data – Focusing on Data Quality as a Service

**[1]Amit Khan, [2]Anirban Bhowmick, [3]Abhijit Das, [4]Dipankar Majumdar**

*[1]RCC Institute of Information Technology, Beliaghata, Kolkata, amit.it.khan@gmail.com*
*[2]RCC Institute of Information Technology, Beliaghata, Kolkata, cityofjoy.ami@gmail.com*
*[3]RCC Institute of Information Technology, Beliaghata, Kolkata, ayideep@yahoo.co.in*
*[4]RCC Institute of Information Technology, Beliaghata, Kolkata, dipankar.majumdar@gmail.com*

## Abstract

In today's world Data Quality is the fundamental to the health of any organizations. This organization can be a private player in the business world or even it can be any national body managed and controlled by the government. Poor data quality leads to myriad problems and it is a monumental effort to standardize them manually. Data quality issues trace back their source to the early days of computing. A wide range of area specific practices to measure and improve the quality of data exist in the works. These solutions mainly target data which exist in relational databases and data warehouses. The recent advent of big data analytics and resurgence in machine learning demands evaluating the suitability relational database-centric approaches to data quality. In this paper, we plan to target data quality issues related to the Address World in the context of big data and machine learning, and devise a systematic and planned data governance-framework to improve the data quality of the Address as a whole, finally describe the approach to its implementation.

**Keywords**: Data Quality, Machine Learning, Big Data, Data Governance, Standardization.

## 1. INTRODUCTION

Data quality plays a critical role in computing applications overall and data-concentrated applications in particular. Data acquisition and authentication are among the major tasks in data-intensive applications. High-quality data brings business value in the form of more knowledgeable and quicker decisions, increased returns and reduced costs, increased ability to meet legal and regulatory compliance, among others. Data quality depends on the task and is often defined as the degree of data fitness for a given purpose. It indicates the degree to which the data is complete, consistent, free from duplication, accurate and timely for a given purpose. The application of relevant practices and controls to improve data quality is referred to as data quality management. Defining and assessing data quality is a difficult task as data is captured in one context and used in totally different contexts. Additionally, the data quality assessment is industry-specific, less objective, and requires noteworthy human participation.

The quality of data is defined by different factors such as the accuracy, the completeness, the consistency, validity, uniqueness and the timeliness as shown in the figure1. That quality is necessary to fulfill the needs of an organization in terms of operations, planning and decision-making.


Figure 1: *Data Quality Factors*

Today, the majority of a company's operations and strategic decisions are based on data, thus quality is even more important. High quality data has lots of advantages as shown in figure 2. A Low-quality data is, in fact, the greatest cause of failure for advanced data and technology initiatives, costing American firms $9.7 million each year (not counting businesses in every other country of the world). Low-quality data, in general, can have a negative influence on productivity, profitability, and overall return on investment (ROI).


Figure 2: *Advantages of Data Quality*

Data quality management (DQM) is a set of techniques aimed at ensuring that data is of good quality. DQM encompasses all aspects of data management, from data collection to advanced data processing to data delivery. It also necessitates a managerial oversight of the data you've gathered. Effective DQM is considered as critical to any consistent data analysis, since data quality is critical to draw actionable and, more importantly, correct insights from your data.

We'll get into some of the consequences of poor-quality data in a moment. However, let us not fall into the "quality trap," because the ultimate purpose of DQM is to maximize return on investment (ROI) for those business sectors that rely on data, not to establish subjective views of what "high-quality" data is.

The benefits of good DQM can have a ripple effect on an organization's performance, from customer relationship management through supply chain management to enterprise resource planning. Organizations can create data warehouses with high-quality data to examine trends and develop future-oriented plans. The favorable ROI on quality data is broadly acknowledged across the industry. According to recent Accenture big data surveys, 92 percent of executives who use big data to manage are happy with the results, and 89 percent consider data to be "very" or "extremely" significant because it will "revolutionize operations in the same way the internet did." The rest of the article is organized as follows. Section 2 presents related works, section 3 discuss traditional vs. machine learning data quality management, section 4 presets our proposed method, followed by result and discussion in section 5 and conclusion and future scope in section 6.

## 2. RELATED WORKS

Data quality is a great matter of interest in many application areas[1]. Let us consider the software engineering area. The usefulness of an estimate models in the empirical software engineering critically dependent on the quality of the data used in building the prototypes [2]. Data quality analysis plays an important role in appraising the practicality of data composed from the Software Process frameworks [3] and empirical software engineering research [4]. Cases Inconsistency Level is a metric for investigating conflicts in software engineering datasets [5]. Data quality is studied in numerous other domains including cyber-physical systems [6], assisted living systems

[7], citizen science [8], ERP systems [9], accounting information systems [10], drug databases [11], smart cities [12], sensor data streams [13], linked data [14], data integration [15], [16], multimedia data [17], scientific workflows [18], and customer databases [19]. Big data management [20], Internet of Things (IoT) [21], and machine learning [22] domains are generating renewed interest in data quality research. A extensive variety of area specific practices to measure and enrich the class of data exist in the works [23], [24]. Authors in [25] proposed a Firefly Update Enabled Rider Optimization Algorithm (FU-ROA), which is the hybridization of the Rider Optimization Algorithm (ROA) and Firefly (FF). The impact of data quality management on supply chain presented by authors in [26]. In [27] authors presented an extensive survey on different data cleaning techniques for web information system to improve the quality of data. Authors in [28] showed how different determinants impact specific information quality (IQ) dimensions of shared demand-related information in dyadic supply chain relationships. In [29] authors proposed an approach for the analysis of variance and distribution of datasets for modeling product quality prediction. These characteristics have to be analyzed to interpret the results correctly. In [30] authors explored outcomes that arise from a data quality improvement process implementation in an operations management environment. Over a three-year period, they were conducted a longitudinal single case study at an organization that maintains a large fleet of aircraft, collecting and analyzing qualitative interviews and observations.

## 3. TRADITIONAL VS MACHINE LEARNING DATA QUALITY MANAGEMENT

Current view: Traditional Data Quality Management relies on manual data correction by data stewards. Determining how to prevent the Data Quality exceptions in the future requires additional effort in terms of revisiting the Data Quality rules. This is having high cost associated with it. Also the scalability is another issue because increase in data volume may impact efficiency.

Future view: Leveraging Machine Learning to understand new data patterns improves the client's data quality and causes data stewardship efforts to be limited to validation rather than correction. This to be a solution which will not only be scalable but also will adaptive with little efforts can be easily applied to any data patterns or specific structured data use case.

The architecture of both traditional and machine learning data quality management are shown in the figure 3 and figure 4 respectively.

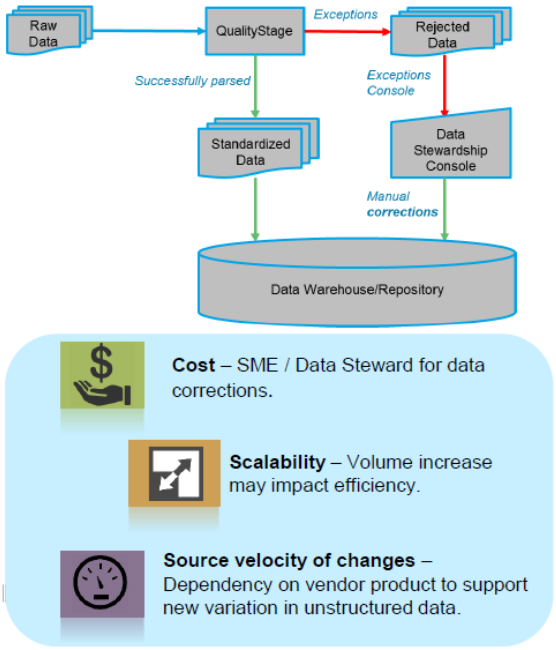


Figure 3 : *Current View*

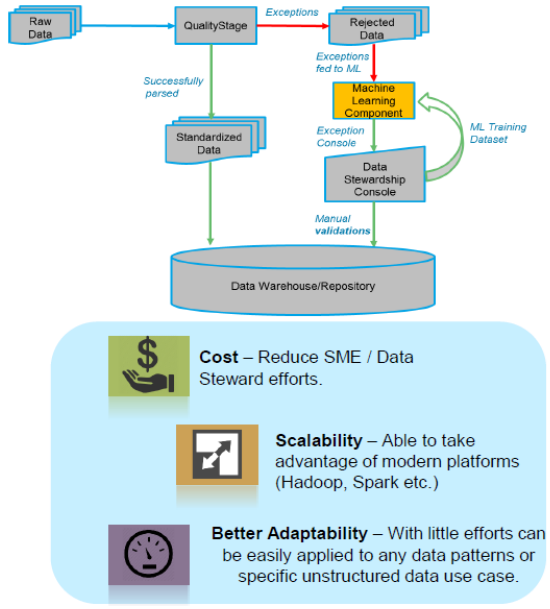


Figure 4 : *Future View*

## 4. PROPOSED METHOD:

Our proposed work addresses the problem s of traditional approaches of data quality management. Here we proposed a machine learning based data standardization technique to improve the quality of data so that industry can directly use the standardized data for quick processing. Overall flow chart of our proposed solution is shown in figure 5.

### 4.1 Flow chart of our Proposed Solution:

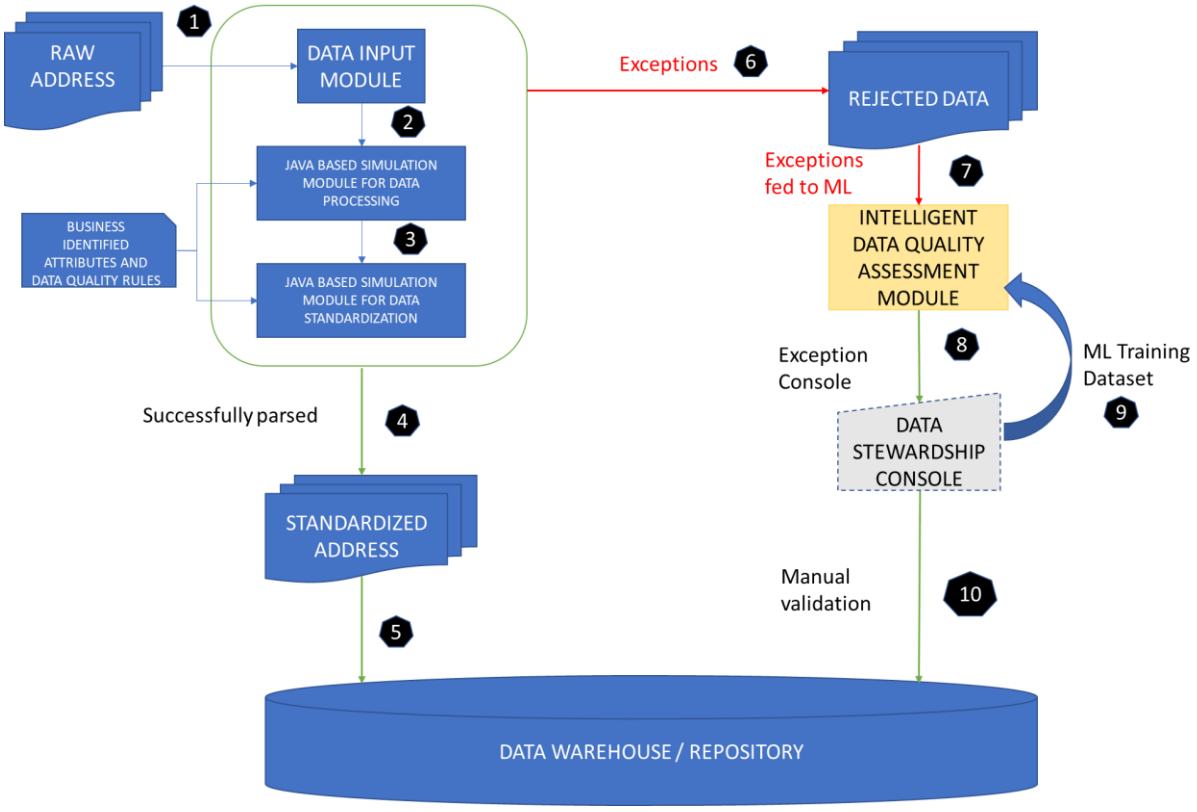


Figure 5: *Flow Chart*

### 4.2 Algorithm of our Proposed Solution

1. Unstructured address, i.e., RAW address will be the input to the proposed model. This data set can be in the form of any standard file format. e.g., .csv, pipe delimited .txt file etc.

2. Any industry standard data quality tool, in this case a Java based simulation module to perform the data processing followed by data standardization on the raw address data based on predefined attributes.

a. Output can either be standardized address which will directly go to the data repository, else

b. Partially formatted address to be fed to the ML based Intelligent Data Quality Assessment component to further standardize the unhandled data into identified tokens

c. This will be exposed as REST API service for integration platforms.

3. ML based Intelligent Data Quality Assessment in this case to be simulated by a Java module for error rectification and structure the address fully.

4. Intelligent platform offerings will also be deployed as a REST API for further integrations.

5. Any exceptions at this layer will be used as the fixes to the ML training data set for the learning of the ML component for either real-time or near real-time feed.

4.3 High Level Architecture of the Proposed System

High level architecture of our proposed work is shown in figure 6. The given architecture works based on the aforesaid algorithm.

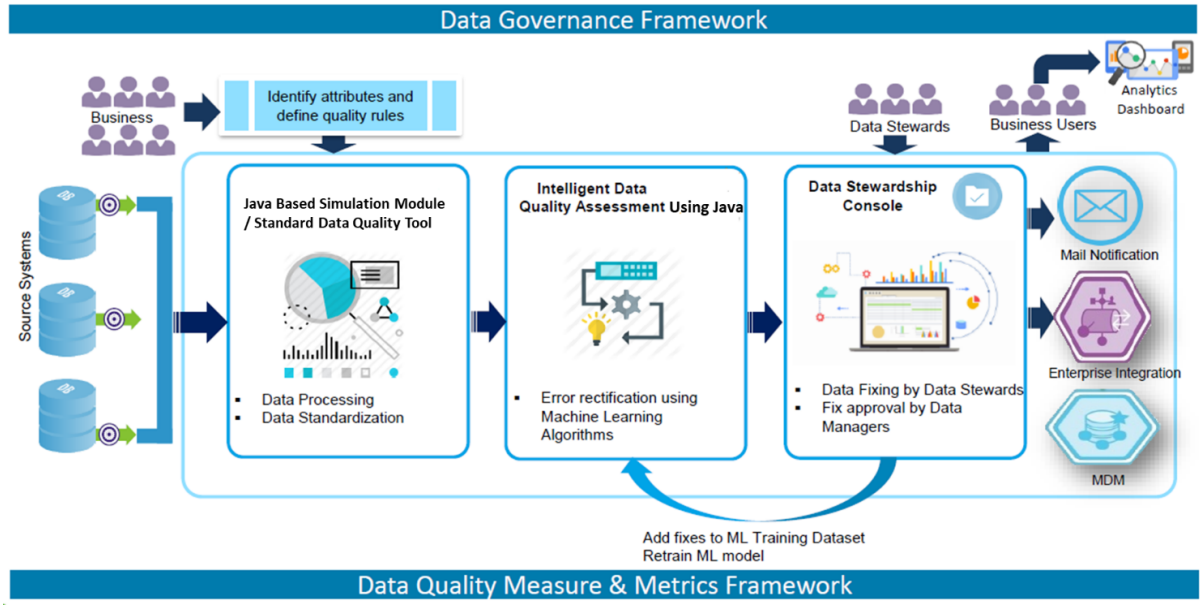


Figure 6: *High Level Architecture of our proposed model*

## 5. RESULTS AND DISCUSSION

From the analysis it is clear that manual checking time increases exponentially as the data volume increases but in case of automatic checking using machine learning approach time increase gradually as volume of data increases. Figure 7 shows the performance the performance improvement of machine learning approach over traditional approach.

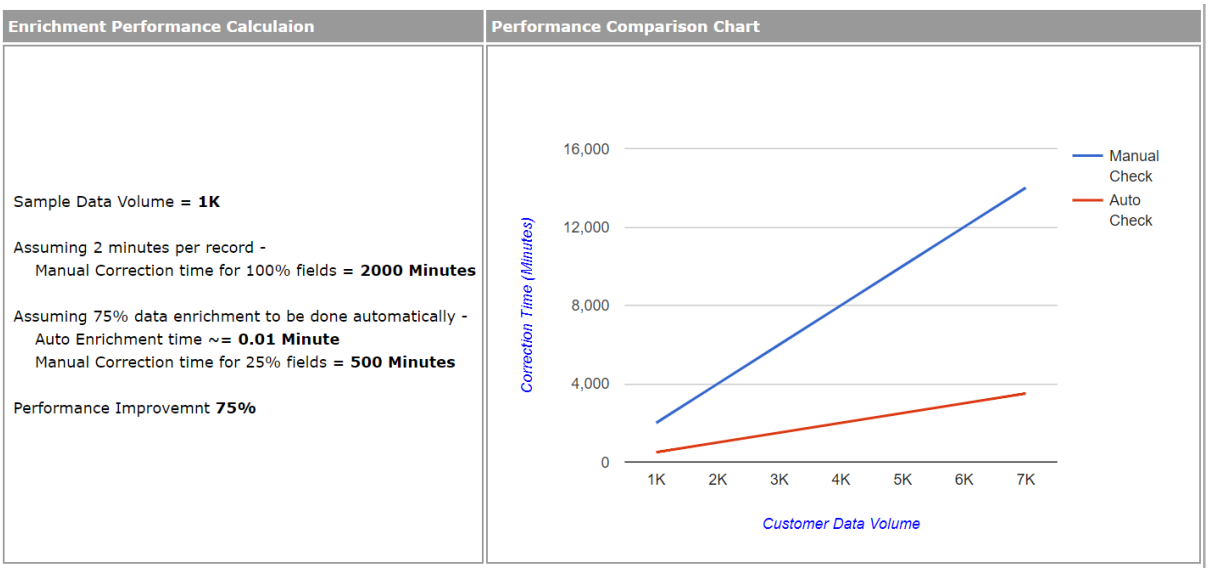


Figure 7: *Performance improvement over traditional approach*

## 6. CONCLUSION AND FUTURE SCOPE

Our proposed machine learning based data quality improvement technique provides lot of benefits in terms of scalability, variability of data change, effort reduction, variety of data and seamless integration. Proposed technique reduced data correction time considerably by ~50-60% over traditional data correction technique. It reduces data processing cost and time. But one limitation of approach is that due to the lack of large volume training data the algorithm may produce inefficient result. In near feature we want to tune our proposed approach using different deep learning paradigms to achieve more refined result.

## References

[1] V. Gudivada, D. Rao, and W. Grosky, "Data quality centric application framework for big data," in Proceedings of the Second International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA 2016). Lisboa, Portugal: IARIA, Feb. 2016, pp. 24 – 32.

[2] M. F. Bosu and S. G. MacDonell, "Data quality in empirical software engineering: A targeted review," in Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering. New York, NY: ACM, 2013, pp. 171–176.

[3] Y. Shirai, W. Nichols, and M. Kasunic, "Initial evaluation of data quality in a tsp software engineering project data repository," in Proceedings of the 2014 International Conference on Software and System Process. New York, NY: ACM, 2014, pp. 25–29.

[4] M. Shepperd, "Data quality: Cinderella at the software metrics ball?" in Proceedings of the 2nd International Workshop on Emerging Trends in Software Metrics. New York, NY: ACM, 2011, pp. 1–4.

[5] P. Phannachitta, A. Monden, J. Keung, and K. Matsumoto, "Case consistency: A necessary data quality property for software engineering data sets," in Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering. New York, NY: ACM, 2015, pp. 19:1–19:10.

[6] K. Sha and S. Zeadally, "Data quality challenges in cyber-physical systems," J. Data and Information Quality, vol. 6, no. 2-3, pp. 8:1–8:4, Jun. 2015.

[7] J. McNaull, J. C. Augusto, M. Mulvenna, and P. McCullagh, "Data and information quality issues in ambient assisted living systems," J. Data and Information Quality, vol. 4, no. 1, pp. 4:1–4:15, Oct. 2012.

[8] S. A. Sheppard and L. Terveen, "Quality is a verb: The operationalization of data quality in a citizen science community," in Proceedings of the 7th International Symposium on Wikis and Open

Collaboration. New York, NY: ACM, 2011, pp. 29–38.
[9] L. Cao and H. Zhu, "Normal accidents: Data quality problems in ERP-enabled manufacturing," J. Data and Information Quality, vol. 4, no. 3, pp. 11:1–11:26, May 2013.
[10] H. Xu, "What are the most important factors for accounting information quality and their impact on ais data quality outcomes?" J. Data and Information Quality, vol. 5, no. 4, pp. 14:1–14:22, Mar. 2015.
[11] O. Curé, "Improving the data quality of drug databases using conditional dependencies and ontologies," J. Data and Information Quality, vol. 4, no. 1, pp. 3:1–3:21, Oct. 2012.
[12] P. Barnaghi, M. Bermudez-Edo, and R. Tönjes, "Challenges for quality of data in smart cities," J. Data and Information Quality, vol. 6, no. 2-3, pp. 6:1–6:4, Jun. 2015.
[13] A. Klein, "Incorporating quality aspects in sensor data streams," in Proceedings of the ACM First Ph.D. Workshop in CIKM. New York, NY, USA: ACM, 2007, pp. 77–84.
[14] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri, "Test-driven evaluation of linked data quality," in Proceedings of the 23rd International Conference on World Wide Web. New York, NY: ACM, 2014, pp. 747–758.
[15] S. K. Bansal and S. Kagemann, "Integrating big data: A semantic extract-transform-load framework," Computer, vol. 48, no. 3, pp. 42–50, 2015.
[16] N. Martin, A. Poulovassilis, and J. Wang, "A methodology and architecture embedding quality assessment in data integration," J. Data and Information Quality, vol. 4, no. 4, pp. 17:1–17:40, May 2014.
[17] K.-S. Na, D.-K. Baik, and P.-K. Kim, "A practical approach for modeling the quality of multimedia data," in Proceedings of the Ninth ACM International Conference on Multimedia. New York, NY: ACM, 2001, pp. 516–518.
[18] A. Na'im, D. Crawl, M. Indrawan, I. Altintas, and S. Sun, "Monitoring data quality in kepler," in Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. New York, NY: ACM, 2010, pp. 560–564.
[19] H. M. Sneed and R. Majnar, "A process for assessing data quality," in Proceedings of the 8th International Workshop on Software Quality. New York, NY: ACM, 2011, pp. 50–57.
[20] V. Gudivada, R. Baeza-Yates, and V. Raghavan, "Big data: Promises and problems," IEEE Computer, vol. 48, no. 3, pp. 20–23, Mar. 2015.
[21] Y. Qin, Q. Z. Sheng, N. J. Falkner, S. Dustdar, H. Wang, and A. V. Vasilakos, "When things matter: A survey on data-centric internet of things," Journal of Network and Computer Applications, vol. 64, pp. 137 – 153, 2016.
[22] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, Learning From Data. AMLBook, 2012.
[23] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," ACM Comput. Surv., vol. 41, no. 3, pp. 16:1–16:52, Jul. 2009.
[24] V. Ganti and A. D. Sarma, Data Cleaning: A Practical Perspective, ser. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2013
[25] K. Rahul, R.K. Banyal , "Detection and Correction of Abnormal Data with Optimized Dirty Data: A New Data Cleaning Model" International Journal of Information Technology & Decision Making ,Vol. 20, No. 02, pp. 809-841 , 2021
[26] K. Phasha, S. Mankazana, S. Mukwakungu, N. Sukdeo, "The Impact of Quality Management on Supply Chain " , 30th Annual Conference of the International Association for Management of Technology (IAMOT 2021) Sep. 2021, Cairo, Egypt.
[27] Wang, J., Wang, X., Yang, Y., Zhang, H., Fang, B. , "A Review of Data Cleaning Methods for Web Information System", Computers, Materials & Continua, vol.62, no.3, pp.1053-1075, 2020
[28] Myrelid, P., Jonsson, P., "Determinants of information quality in dyadic supply chain relationships ", International Journal of Logistics Management, Vol. 30, no.1,pp. 356-380, 2019
[29] Weiß, I., Vogel-Heuser, B., "Assessment of variance & distribution in data for

effective use of statistical methods for product quality prediction ", at - Automatisierungstechnik, April, 2018

[30] Hazen, Benjamin T. & Weigel, Fred K. & Ezell, Jeremy D. & Boehmke, Bradley C. & Bradley, Randy V., "Toward understanding outcomes associated with data quality improvement," International Journal of Production Economics, Elsevier, vol. 193(C), pages 737-747, 2017