

Ensuring Data Integrity and Reliability in Big Data*

Michal Zrutta

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
xzrutta@stuba.sk

5. November 2023

Abstract

We live in the era of Big Data. These days data comes too fast and in an enormous sizes, this requires better understanding of data and processing. Data generation increased by 95 zettabytes between 2010 and 2022.[4] However, the upward trajectory poses challenges in data quality. Consequently, the demand for realistic data sets, accurate and consistent information, and overall data quality has become a pressing concern.[4] This study delves into the challenges of data quality assurance in the realm of Big Data analytic, exploring methodologies to validate diverse and extensive data-sets. The research is shown on a context of gas consumption in various car models.[5]. The paper later shows potential but already used integration of Artificial Intelligence. AI algorithms are supposed to identify and rectify inaccuracies, while also learning from the data to continuously enhance data quality.

1 Introduction

1.1 Background

In recent times, the term "Big Data" has become widespread, describing a massive amount of information flooding businesses, organizations, and research institutions daily. Big Data includes data-sets that are too large, complex, or fast-moving for regular data processing tools. Data comes from various sources like social media, sensors, online transactions, and multimedia. What makes Big Data unique isn't just its enormous size but also its speed, diversity, and reliability.[4] The rapid growth and widespread use of Big Data technologies have transformed industries and research fields globally. With the increasing digitization of information and advancements in data processing capabilities, organizations now have the ability to collect, store, and analyze massive volumes of data in real-time. This growth has been driven by the need to extract valuable insights, improve decision-making processes, enhance customer experiences, and

*Semestrálny projekt v predmete Metódy inžinierskej práce, ak. rok 2023/24, vedenie: Mirwais Ahmadzai

gain competitive advantages. Industries ranging from business and healthcare to science and social sciences have embraced Big Data technologies, leading to their everywhere use in diverse applications.[2] This requires accurate and precise data analysis to extract valuable insights, ensuring its usefulness across a wide range of industries. Trustworthiness in dealing with Big Data is really important. It means the information used is reliable and accurate, preventing mistakes. It also builds people's confidence and protects their privacy in our data-driven world.[2]

1.2 Problem Statement

Big data analysis can transform raw data collected from users in a big data program into useful information.[8] Wrong data and incorrect inputs can significantly alter the overall results, which are crucial in many cases.[1] There are many works out there on data reliability. For example, consider buying a new laptop. You can rely on the specifications provided for the product. However, you also seek user experiences by reading comments. If there are many positive comments, you may feel confident about the purchase. But this system lacks credibility in verifying the trustworthiness of the person who wrote the comment.[8] The question still remains: How can we ensure data quality?

- Data Validity: Is the data fulfilling its intended purpose?
 - Apple is a fruit, but it is also a company
- Data Completeness/Accuracy: Is the data reliable for organizational use?
 - Does all processed data necessary for decision-making exist?
 - Can we trust the data used for calculations?
- Data Consistency: Is the data consistent in its behavior?
 - Apple is a fruit, but it is also a company

[5]

1.3 Research Objectives

This paper aims to explore strategies ensuring data quality within the realm of Big Data. By addressing the challenges associated with incorrect data and not verified data it proposes methodologies to validate data quality. Through these strategies, the paper seek to help sort and clean data for reliable, accurate, and trustworthy data analysis in the era of Big Data.

2 Previous Studies on Data Quality in Big Data

2.1 Data Accuracy and Reliability Evaluation for Big Data

In the proposed data quality assessment process for big data, as outlined by Li Cai from Yunnan University, the initial step involves defining data collection goals, aligning with strategic objectives like operations and decision-making. Selection of quality elements varies based on different business environments. For example, social media data prioritize timeliness and credibility due to their unstructured nature. Assessment indicators

are tailored for each dimension, considering specific conditions and business contexts. Data cleaning, crucial for error detection and removal, follows, enhancing data quality. The evaluation involves qualitative and quantitative methods. Qualitative analysis,

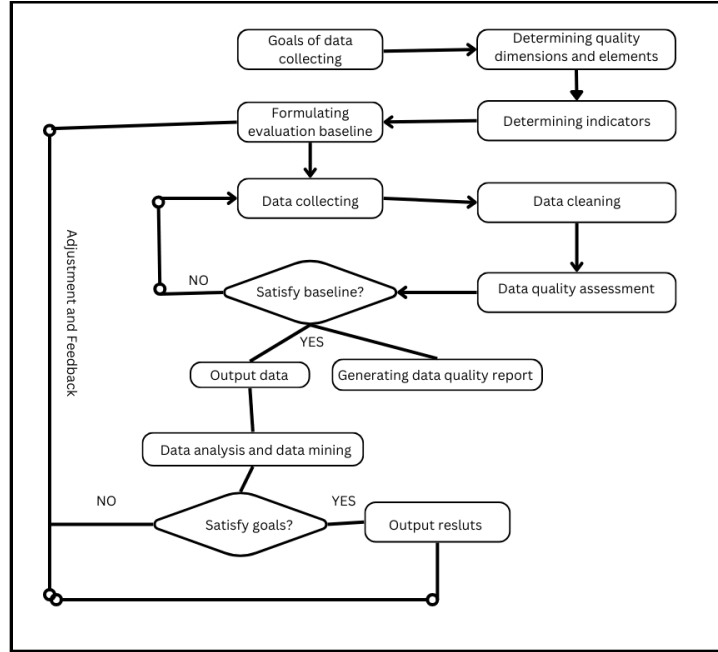


Figure 1: Quality assessment for big data

performed by experts, assesses data resources based on specific criteria. Quantitative evaluation, relying on numerical data, offers objective insights. After assessment, data quality is compared with the established baseline. If it meets standards, data analysis proceeds, generating quality reports. If not, adjustments are made, and new data acquisition might be necessary. Data analysis and mining, while not directly part of quality assessment, aid in discovering valuable insights within big data. Successful analyses are fed back into the quality assessment system, providing support for subsequent assessments. Adjustments are made if results do not align with goals, ensuring continuous improvement.[3]

2.2 Semantic Integrity Analysis for Unstructured Big Data

This study is dedicated to the evaluation of unstructured Big Data, focusing on a assessment process for unstructured datasets. The model show us several essential components that collectively contribute to the comprehensive evaluation of data quality, resulting in a detailed quality assessment report.

1. **Quality Requirements:** The process initiates with defining precise quality requirements aligned with the strategic objectives, ensuring that the assessment is tailored to meet specific business goals.
2. **Data Sources:** Unstructured data is diverse and originates from varied sources. Evaluating its quality necessitates a clear understanding of these sources, encompassing social media, sensors, multimedia, and other channels.

3 PROPOSED METHODOLOGY FOR DATA QUALITY ASSURANCE IN BIG DATA4

3. **Data Quality Repository:** A centralized repository is established to store and manage data quality-related information, facilitating seamless tracking and analysis.
4. **Sampling and Profiling:** The data undergoes sampling and profiling, enabling a deeper understanding of its characteristics and patterns, which is fundamental for quality evaluation.
5. **Preparation Process for Quality Evaluation:** Preparing the data for evaluation involves preprocessing steps to ensure consistency and completeness, laying the foundation for accurate assessment.
6. **Feature Selection and Quality Mapping:** Relevant features are meticulously selected, and quality mapping techniques are applied, allowing for a focused evaluation process and ensuring that quality assessment aligns with the specific attributes of unstructured data.
7. **Quality Assessment:** Utilizing both qualitative and quantitative methods, the data undergoes a rigorous evaluation. Qualitative analysis conducted by experts evaluates data based on specific criteria, while quantitative methods provide objective insights.

[7]

2.3 AI and Crowdsourcing Solutions

In this study, they demonstrate how AI can assist in managing Big Data lakes and handling diverse data from various sources, addressing challenges such as incomplete data, invalid entries, and uncertainties. AI and crowdsourcing can be used in this problem. AI is able to remove duplicates from database or knowledge base. And there are continuous studies in data quality.[6]

3 Proposed Methodology for Data Quality Assurance in Big Data

If we had to collect data from all specific car model in the world for particular information, for example gas consumption. Firstly we would need to specify:

1. **Where?** It is really important where the car is driven. If it is somewhere rainy, for example Britain, or in Australia where it is hot, or in Canada where it is cold. The car and its use is totally different and collected data which are not selected correctly would result in not proper product. For example electric cars like Tesla have difficulties in cold environment.
2. **Who?** Everybody drives their cars differently so in order to get good results we need drivers who have similar driving skills.
3. **When?** The time of data collection is crucial. Gas consumption can vary based on the time of day, traffic conditions, and seasons. For instance, in urban areas, gas consumption might differ during rush hours compared to non-peak times. Similarly, seasonal variations, such as summer and winter, can impact a car's fuel efficiency. Collecting data over different times and seasons can provide a comprehensive understanding of gas consumption patterns.

Secondly, I recommend implementing a user ranking system for data contributors. Because if he is not qualified as a driver who has driving style as we need. For example, if a participant drives their car in a desert, which is vastly different from the study's target environments, their data might not be suitable for the overall analysis.[7] Once the data is collected from verified users we have to clean them and sort them out. As we can see in previous study we can use AI to help us clear and clean data to enhance data accuracy. AI can automatically identify and correct mistakes. Additionally, AI possesses a deep understanding of physics, enabling it to perform calculations to validate the data accuracy and identify errors. Moreover, AI systems incorporate self-learning mechanisms, allowing them to learn from the collected data. This self-learning capability ensures continuous improvement in data quality over time by learning from past mistakes and refining its algorithms for future analyses.[6]

References

- [1] B. Agarwal, A. Ravikumar, and S. Saha. A novel approach to big data veracity using crowdsourcing techniques and bayesian predictors. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1020–1023, 2016.
- [2] M. Anisetti, C. A. Ardagna, and F. Berto. An assurance process for big data trustworthiness. *Future Generation Computer Systems*, 146:34–46, 2023.
- [3] L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, 05 2015.
- [4] V. N. Gudivada, R. Baeza-Yates, and V. V. Raghavan. Big data: Promises and problems. *Computer*, 48(03):20–23, 2015.
- [5] A. Mittal. Trustworthiness of big data. *International Journal of Computer Applications*, 80(9), 2013.
- [6] D. E. O’Leary. Embedding ai and crowdsourcing in the big data lake. *IEEE Intelligent Systems*, 29(5):70–73, 2014.
- [7] I. Taleb, M. A. Serhani, and R. Dssouli. Big data quality assessment model for unstructured data. In *2018 International Conference on Innovations in Information Technology (IIT)*, pages 69–74, 2018.
- [8] J. Yu, K. Wang, P. Li, R. Xia, S. Guo, and M. Guo. Efficient trustworthiness management for malicious user detection in big data collection. *IEEE Transactions on Big Data*, 8(1):99–112, 2022.