Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2021, 8(1):89-96

Research Article ISSN: 2394-658X

Ensuring Data Quality in Business Intelligence

Pranay Mungara

ABSTRACT

Many businesses have moved their operations to the cloud in an effort to fortify their data protection and enhance the standard of their business dealings. In modern operations, data quality is paramount. Data used to generate and gather information represents events and occurrences as they happen in real time. Customer happiness, the organization's decision-making strategy, and the organization's plan of execution are all negatively affected by low-quality data. The data quality has a major influence on how well machine learning and deep learning accomplish their duties in terms of accuracy, complexity, and efficiency. Several techniques and instruments are available for assessing the quality of data in order to guarantee a seamless absorption into the construction of models. The vast majority of data quality technologies only allow for the evaluation of data sources at a particular instant in time; as a result, the user is responsible for the organization and automation of the process. The collection of data from a variety of sources and the monitoring of data quality are two of the many processes involved in assuring the quality of the data that is automatically generated. Any issues that arise with the data quality must be addressed in an appropriate manner. A gap existed in the existing body of literature due to the fact that no previous attempts had been made to compile all of the advancements that had been made in various aspects of automatic data quality. The purpose of this restricted narrative evaluation of the current literature was to attempt to fill this gap by establishing a correlation between the many phases and improvements that are associated with information quality systems that are automatic.

Key words: Data Quality, Business Intelligence, Data protection

INTRODUCTION

Maintaining high standards of data quality has grown in importance whenever it pertains to an organization's data management. Better decision-making procedures, improved corporate strategy plans, and the discovery of outstanding patterns for improved problem-solving would be available to organizations that possess high data quality. Dissatisfied customers, excessive operating costs, and improper judgments as a result of faulty data are some of the problems that have arisen as a result of enterprises' inability to supply high-quality data [1]. The definition of data measures, dimensions, procedures, quality, evaluation, and improvement models has been the subject of various studies in the field of data quality, which have ultimately led to the discovery of numerous breakthroughs. Data that is fit for use and meets all of the constraints imposed by users, considering the field of application in which it is being used, is good data quality, according to scientists and researchers who have reviewed the long history of data quality.

The reliability of the data is a major issue for many different types of applications. The reliability of the model's prediction is greatly affected by the use of high-quality data throughout the model's execution [2]. Data quality assessment relies heavily on it, and it is essential for determining whether or not the results of the Software Process and empirical software are meaningful. Over the years, researchers in many fields have studied data quality, including multimedia, the internet of things, smart cities, artificial intelligence, medicine, big data analytics and management, drug databases, and ecological systems.

Research on some parts of data quality has appeared in the literature, but nothing that covers the whole scope of an automated data quality system altogether. This void in the literature exists because the area of automated data quality systems is currently evolving [3]. This study aims to fill the gap by collecting all the important components of data quality management that affect the development of automated data quality management systems. This qualitative study employed a narrative evaluation of the relevant literature as its technique. The review was conducted primarily from the point of view of practitioners who are employed in the process of establishing automated data quality systems.

A. Data Quality Toolkit

Previous studies have shown that a major component in deciding how well machine learning and deep learning models work is the quality of the data used for training and testing these models [4]. Although there have been persistent attempts to improve the quality of models, there has been very little documented work towards improving the data quality.

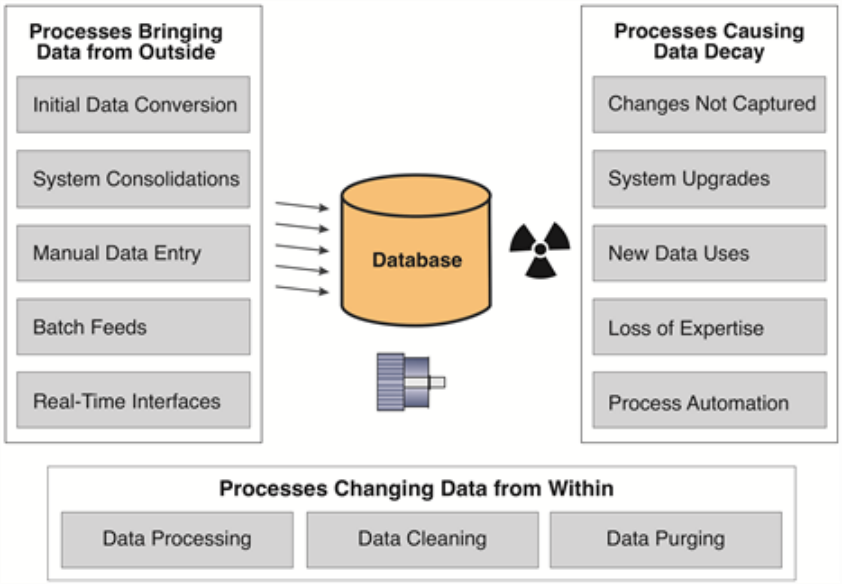

Figure 1: Processes affecting data quality

A variety of processes can impede data, and the majority of these activities have an impact on the quality of the data to some degree. Depending on the circumstances under which they are discovered, the quality of the data deteriorates in a variety of ways. There are thirteen distinct types of processes that could cause problems with data. The three distinct advanced categories shown in Figure 1 represent some of the potential causes of data difficulties.

The steps required to manually or automatically import data from other sources, as well as to use data integration methods and different interfaces, are shown on the left side of Figure 1.

Errors will be introduced throughout the processes of data loading, transformation, and extraction as a result of the entry of new data; these issues will be further exacerbated by the traffic created by high data volumes. The procedures that are responsible for the alteration of the data are located on the right side. The process of data manipulation involves a number of activities, some of which are routine, while others are the result of database reorganization, periodic mass data updates, and many other ad hoc acts. On the other hand, during the implementation stage, some of these processes do not have the resources, time, and trustworthy metadata that are required to attain the desired effects on data quality.

Table 1: Classification and issues with data quality.

Data Quality Problem	Type	Description
Single-source Problem	Instance Level	Misspelling
		Inconsistent values
		Data entry errors
		Redundancy Replicas
	Schema Level	Referential Reliability
		Poor schema designer
		Lack of integrity checks
		Uniqueness checks
Multi-source Problem	Instance Level	Overlapping, challenging and erratic timing, data, and aggregating
	Schema Level	Naming Clashes
		Mixed data models and schema design

The process chain that leads to the gradual corruption of otherwise accurate data over time in the absence of any external physical change is illustrated in the bottom section. It only indicates that the data values remain unchanged, although with reduced precision. This happens when the data-gathering procedures fail to notice a change to a real-time item, rendering the previously collected data erroneous and irrelevant. Additional categorization of data quality issues into single-source and multi-source difficulties was provided in [5].

According to previous studies, there are four distinct types of data quality, as shown in Table 1. About 1,250 results are returned when the search string "automated data quality" is entered into Google Scholar. The outcomes, however, are all talks about plans to build or install sector-or industry-specific automated data quality systems. Just two results were returned when the query "automated data quality systems" was entered [6]. Rather than focusing on the system itself, both of these articles discussed how automated data quality systems can be applied to certain tasks. As a result, we broadened our scope to include data quality assurance, monitoring, and all other related aspects.

DATA QUALITY IN HEALTH RESEARCH

The efforts that governments all over the world are making to develop infrastructure and technology with the intention of expanding their capacity to make use of generated data are proof of the significance and performance of digital media in the field of health. It must be emphasized that technology cannot transform raw data into actionable intelligence; rather, human intervention is crucial for deriving knowledge from raw data. Insights, beliefs, contexts, and experiences all make up knowledge, which can provide a basis for assessment and choice making [7]. This is achieved through studies that enhance the alignment of effective policies and optimize health interventions. One of the main factors that undermine health-related research and decision-making is the inadequacy of health data in terms of quality, availability, and integration.

Also, it is important to point out that there are a great number of databases that are completely separate from one another and can only be accessed in a specific setting. These variables lead to issues with the quality of the data, which in turn leads to the loss of information. Due to the information's continued dispersion, it is unable to inform choices. This presents difficulties in terms of both coordination and evaluation, despite the fact that the volume is particularly high.

As an illustration, the utilization of data of low quality in the process of developing models for artificial intelligence might result in decision-making procedures that include the formation of incorrect conclusions. Labeled collections of large amounts of data have been utilized by artificial intelligence systems in order to construct their models. These systems are increasingly being utilized to assist in decision-making. The collection and labeling of data is frequently carried out by algorithms that are not compensated, and research is progressively demonstrating the disadvantages associated with this strategy. During the process of training and testing their models, algorithms may exhibit biases in their assessments of a person's profession, nationality, or character, as well as fundamental flaws that are concealed inside the data that they employ. Therefore, it is possible to conceal predictions, which makes it difficult to differentiate between models that are correct and those that are incorrect [8].

Information gathered from health services and research is various and complex, which is why there is an inseparable relationship between the heterogeneity of data in this field and them. The computerization of health systems and research is inevitable, and it is the foundation for computerizing and promoting knowledge generation and administration [9]. Reasons for this foundation include the ever-changing and extremely diverse medical terminology, the massive amounts of data produced by process automation and emerging technologies, and the absolute necessity of processing and analyzing this data in order to make decisions.

The term "data quality" can refer to a variety of distinct things. There are a number of characteristics that must be present in order for data to be considered of high quality, as stated by the World Health Organization. These characteristics include the following: accuracy and validity, reliability, completeness, readability, timeliness and punctuality, accessibility, meaning or usefulness, confidentiality, and security. It is possible for the quality of the data to be affected at various stages, including the method of data collection, coding, and the absence of term standardization. Interference can be caused by a variety of factors, including environmental, behavioral, organizational, and technical factors.

Even in situations where data are available, there are some factors that render its utilization by researchers, managers, and health professionals impossible. Health information systems suffer from a number of issues, such as insufficient computerization of procedures, data duplication, heterogeneity, and errors that arise during data collection and processing [10]. Decisions and strategies to enhance service delivery also necessitate reliable health data in order to offer consistent proof of health status. Data quality control is thus an essential procedure for guaranteeing accurate results. Departments at various healthcare facilities are required by action protocols to launch programs aimed at enhancing service quality while decreasing costs. Consequently, the healthcare industry has adopted a plethora of strategies aimed at raising the bar for service quality. Similarly, new strategies for improving research quality through making it more repeatable and giving groups of researchers and patients tools to securely share data while still meeting privacy regulations are constantly being discussed in scientific communities.

Through the process of data standardization, databases that originate from a variety of sources are converted into a standard format that has shared criteria and architecture. In addition, it facilitates the sharing of digital resources amongst users at several universities, which in turn might encourage the consolidation of data from different sources and the establishment of federated research networks [11]. There are two things you need to do to make

this happen: 2) the standardization of the database structure through the use of a CMD, which explains the location and storage of data values in the database; and 3) the standardization of individual data components, which entails adherence to terminology guidelines [12]. There has been an improvement in the functioning and coding structures of the electronic collection program, which reportedly reduced the mistake rate. You should also be familiar with the platform used to perform the study and have access to secondary data sources that could be useful. Thus, the necessary interdisciplinary collaboration could be enhanced through transparent systemic data quality dissemination, good communication, well defined procedures, and appropriate tools [13]. A number of data governance areas were also enhanced as a result of organizational-level awareness initiatives on the topic. According to their findings, the best way to avoid mistakes is for professionals to keep learning and for data collectors to receive constant training throughout the duration of their investigations. In order to improve the reliability and correctness of records and to facilitate their frequent auditing, in-service training should promote the proper use of names generated by organized systems. There was an improvement in the reliability of health systems' data when they were offered financial incentives to do research [14].

BUSINESS INTELLIGENCE AND ANALYTICS (BIA) USAGE IN THE BANKING INDUSTRY SECTOR

An organization can gain a competitive advantage, improve operations and product development, and strengthen relationships with customers through the use of Business Intelligence and Analytics (BIA), which is commonly recognized as one of the most important technologies, systems, practices, and applications [15]. BIA is also thought of as a crucial system, practice, and technology. Experts and managers in the banking sector greatly benefit from BIA since it helps them make better, more accurate, faster, and relevant judgments. The bank is doing this to increase its efficiency and profitability while simultaneously meeting all of the environmental and regulatory requirements that are specific to this industry.

These days, business intelligence analysis (BIA) is a topic that is currently trending and a necessary precondition for developing an exceptional corporate image. This is in line with the implementation of a successful plan for the broad use of technology. Research and development (R&D) requires tremendous investment, but this pays off in the end by bolstering company decisions and giving them an edge in today's fast-paced market. Information is the future since it can be analyzed and used efficiently to back dangerous events and decisions that can have a significant impact on company performance.

Figure 2: Business Intelligence in banking

The banking industry is reaping enormous benefits as a result of changes brought about by digital technology. In addition to increasing the number of access points to bank accounts, it also provides a method for data warehousing by keeping data in the branches of the bank. Through the use of transactions that are completed online, automated teller machines (ATMs), cash and check deposit machines, and electronic wire transfers, the banking system becomes more stable in terms of both its technical and customer-oriented aspects. All of the transactions and the data that is associated with them have been saved. Consequently, in today's world, banks keep enormous electronic data warehouses as their electronic storage. The size and dimensionality of data are always expanding, and this process is ongoing. The financial institutions will be able to transform their massive amounts of data into the most valuable asset they own by employing the techniques of big data analytics. These data include the occurrence of fascinating patterns as well as knowledge that is useful. As a consequence of this, the banking industry has a significant opportunity to implement data mining techniques in order to recognize similar patterns and knowledge on an individual basis, which can assist with crucial decision-making processes such as risk management, marketing, the identification of money laundering, and the detection of fraud [16].

According to [17] Structures, databases, apps, tools, and procedures all fall under the BI umbrella, which is used to describe a wide range of solutions for business data analysis. To achieve this goal, data is transformed into actionable insights that can back up managerial decision-making. Several business ideas, analytics, technology, and tools can be applied very well in the banking industry, particularly in the following areas: branch performance, sales, risk assessment, electronic banking, client segmentation, and retention. Decision support systems (DSS), data warehousing, and data mining (DM) are a few instances of such fields. In order for the banking sector to thrive in today's business environment, the top brass must always keep their sights set on overcoming obstacles and capitalizing on opportunities. This in turn calls for analytics, decision support, and business intelligence systems since managers will have to rely on computers to help them make decisions. Technological advancements in data integration, analytics, and mining have led to the creation of business intelligence systems (BIS).

The purpose of these solutions is to supply stakeholders at different levels with important information that enables them to make decisions that are both effective and successful. In this regard, data analyses have the potential to contribute to the development and resolution of banking issues, as well as to the achievement of the most favorable outcomes for decision making. Managers are unable to recognize the connection between the different factors in corporate data because the data volume is huge and constantly increasing. There are extra steps that managers must take in order to determine the client's wants and needs and the pattern of behavior. Consequently, managers and product managers can benefit from business intelligence through data analyses in the following ways: identifying client categories; creating products or services that meet customer needs; defining pricing strategy and competition; improving revenue management; increasing sales; and expanding the customer segment. Gaining insight into your ideal customers, keeping them as clients, and attracting new ones all necessitate a great deal of extra effort.

According to researchers, business intelligence is the capacity of companies to think, plan, forecast, solve problems, comprehend, and develop novel approaches to enhance decision-making and business processes, facilitate effective actions, and contribute to the establishment and attainment of organizational objectives. Consequently, it is believed that the data, databases, applications, procedures, technologies, scorecards, dashboards, and online analytical processing (OLAP) all contribute to bolstering the abilities that define business intelligence [18].

Not only that, but it also involves comparing statements on an annual basis in order to discover fraudulent activity. In addition, the fraudulent financial statements were said to have a greater number of activation languages, a lower utilization of lexical variety, and a lower utilization of group reference in comparison to the non-fraudulent fiscal statements. Banks are able to identify fraudulent financial statements by utilizing that facts, and decisions are being made in accordance with those findings.

First, let's examine their primary methods and find out if they are adequate.

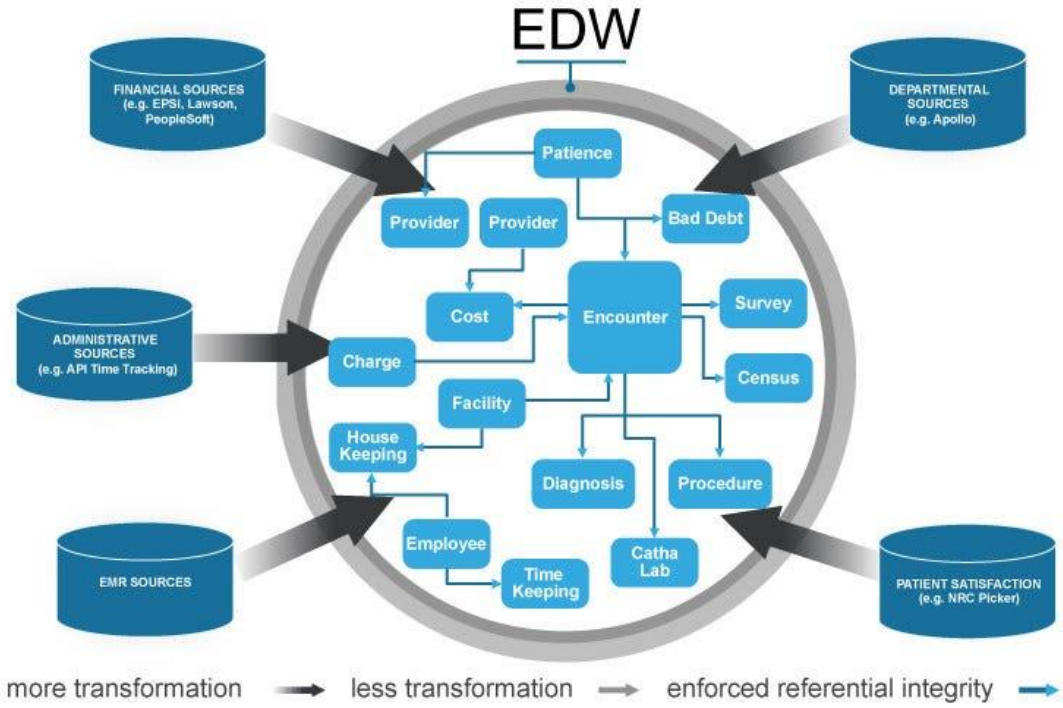

Figure 3: Banks usage of Data Warehousing

Data warehousing is a technique that banks employ to properly organize and manage their data in order to use this data in a variety of different industries. It is often believed that data-driven decision-making cannot be achieved without the methods and resources of business intelligence analysis. Furthermore, they offer the structure and backing that financial institutions need to make judgments that are true and grounded in facts, and to operate in a successful and unique manner [19].

A. Business Intelligence and Analytics (BIA)

In the 1990s, the phrase "business intelligence" (BI) began to gain traction. It refers to a broad category of tools and methods used to gather, evaluate, and share information with the goal of bettering decision making. Methodologies, databases, applications, tools, and infrastructures are all part of the software and processes in question. Many different ideas and methods for enhancing company decision making with the use of fact-based support systems have been grouped under the umbrella term "business intelligence" (BI). One of the main goals of business intelligence is to make it easier to create valuable and relevant information that can help analysts and managers make better decisions for their companies. Another goal is to make it easier to process and transform data so that users can easily and interactively access a lot of different types of data. One example of a software solution or technology that is technically involved in business intelligence (BI) is an ETL tool. Other examples include data warehouses, OLAP technology, data extraction, reporting applications, and an interface that allows users and web access. By utilizing ETL technologies, the data will be extracted and stored in the data warehouse (DW). Reports will be generated with the help of OLAP and data mining as part of the Business Intelligence Analysis (BIA) process. Users are able to access more reporting tools using this user interface. Data extraction, cleansing, and loading into the data warehouse are all tasks that fall within the purview of ETL packages throughout the extract (ETL) process. These packages also ensure that no duplicate or inaccurate data is sent, and that users receive data that is specifically tailored to their needs. Data from several different databases in different departments can be consolidated in the data warehouse. Organizational decision-making is aided by subsequent data aggregation and evaluation. Data mining will eventually supply sales, budgeting, and forecasting reports for businesses, based on the OLAP technology needed by these organizations. These approaches incorporate relational databases and report writing within their multidimensional structure. Business intelligence analysis (BIA) relies on data mining (DM), a core technology that uses a quantitative data analysis tool to find rules and patterns in data resources, as well as logical relationships that summarize data in a new, understandable, and useful way to support organizational management decisions and business intelligence. The capacity to predict a particular action or result by means of data models is among data mining's most significant advantages in the field of business intelligence and analytics. This type of analytics is referred to as predictive analytics, and it provides the most likely conclusion, which ultimately leads to improved management decisions and future planning.

B. Business Intelligence Application in the Banking Industry Sector

Businesses in the banking sector have historically been early adopters of cutting-edge software, hardware, and networking solutions that can improve internal processes, increase output, boost revenue, and differentiate the company from rivals. Aside from offering more accurate and exact data analysis, business intelligence and analytics have the ability to help the bank get far more and better insights than the typical report technologies that are currently available. They can really increase sales and profitability by enhancing data underrating and analysis at the operational and managerial levels, which is made feasible by business intelligence analysis [20].

Figure 4: Applications of BI in Banking

However, a key differentiator and success element for banks is their management and usage of business analytics and intelligence. Not only must business intelligence allow the bank to use all data sources and business intelligence applications, but it must also integrate client information throughout the whole bank in order to enhance efficiency and customer service. There is a strong and crucial relationship between business intelligence and analytics and the success of banks in several areas, including product marketing, client acquisition and retention, risk management, and overall performance. In conclusion, the literature study on the strategic effects of BI done by demonstrates the critical importance of BI and analytics. More study is required, though, to fill our gaps in knowledge regarding the advantages and use of these technologies in the banking industry.

CONCLUSION

The ability to manage massive volumes of varied data, whether structured or unstructured, is a critical capability of business intelligence and analytics. This enables banks to gain significant competitive advantages and operate more efficiently. Every single bank needs to prioritize the installation of a huge business-intelligence system in order to stay up with the rapid tech advancements in light of the current ever-changing environment. To that end, we set out to investigate what variables impact financial institutions' use of BI and analytics. We have looked at a variety of elements, including the technical, organizational, and environmental aspects. The study's findings validated the significance of these elements on the uptake and utilization of business intelligence and analytics, as well as the immense benefit that can be produced by this technology for the banking industry. When it comes to banking, it's crucial to grasp the dynamics of these aspects for better business intelligence usage and tech planning and implementation. Numerous problems concerning data quality management, evaluations, dimensions, and types were addressed in this research. Because of technological advancements, data in enterprises is no longer restricted to databases but can now span different domains. Online communities and social media platforms, among others, are rapidly becoming vital data sources for businesses, academics, and customer relationship builders. Various new technological topics were included in this assessment, which focused on improving data quality. The report also highlighted six crucial data quality aspects, all of which are extremely important for assessing and controlling data quality. When dealing with data quality, researchers and organizations alike should implement data quality management strategies. Integrating this management model and approaches might necessitate improvements to accommodate various data sources in unlabeled data. In addition, data quality evaluation is an essential instrument for effective data quality management. Important aspects of data quality evaluation that contribute to reliable data were discussed in this analysis. Researchers have observed that more studies focusing on unstructured data could improve the quality of unstructured data.

REFERENCES

[1]. Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S. and Munigala, V. (2020) Overview and Importance of Data Quality for Machine Learning Tasks. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, 6-10 July 2020, 3561-3562. https://doi.org/10.1145/3394486.3406477

[2]. Bandyopadhyay, B., Bandyopadhyay, S., Bedathur, S., Gupta, N., Mehta, S., Mujumdar, S. and Patel, H. (2021) 1st International Workshop on Data Assessment and Readiness for AI. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Delhi, 11 May 2021, 117-120. https://doi.org/10.1007/978-3-030-75015-2_12

[3]. Gupta, N., Patel, H., Afzal, S., Panwar, N., Mittal, R.S., Guttula, S. and Saha, D. (2021) Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. arXiv preprint arXiv:2108.05935.

[4]. Afzal, S., Rajmohan, C., Kesarwani, M., Mehta, S., and Patel, H. (2021) Data Readiness Report. Proceedings of 2021 IEEE International Conference on Smart Data Services (SMDS), Chicago, 5-10 September 2021, 42-51. https://doi.org/10.1109/SMDS53860.2021.00016

[5]. Günther, L.C., Colangelo, E., Wiendahl, H.H. and 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

[6]. Rukat, T., Dustin, L., Sebastian, S. and Felix, B. (2020) Towards Automated Data Quality Management for Machine Learning.

[7]. Hekler E, Tiro JA, Hunter CM, Nebeker C. Precision Health: The Role of the Social and Behavioral Sciences in Advancing the Vision. Ann Behav Med. 2020 Apr 27;54(11).

[8]. Harrison K, Rahimi N, Carolina Danovaro-Holliday M. Factors limiting data quality in the expanded programme on immunization in low and middle-income countries: A scoping review. Vaccine. 2020 Jun;38(30):4652–63.

[9]. Peng C, Goswami P. Meaningful Integration of Data from Heterogeneous Health Services and Home Environment Based on Ontology. Sensors (Basel). 2019;19(8):1747.

[10]. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc. 2020 Nov 9;27(12):1999–2010.
[11]. Sarafidis M, Tarousi M, Anastasiou A, Pitoglou S, Lampoukas E, Spetsarias A, et al. Data Quality Challenges in a Learning Health System. Stud Health Technol Inform [Internet]. 2020 Jun 16; 270:143–7. Available from: https://pubmed.ncbi.nlm.nih.gov/32570363/
[12]. Scobie HM, Edelstein M, Nicol E, Morice A, Rahimi N, MacDonald NE, et al. Improving the quality and use of immunization and surveillance data: Summary report of the Working Group of the Strategic Advisory Group of Experts on Immunization. Vaccine. 2020 Oct;38(46):7183–97.
[13]. Tian Q, Liu M, Min L, An J, Lu X, Duan H. An automated data verification approach for improving data quality in a clinical registry. Comput Methods Programs Biomed. 2019 Nov; 181:104840.
[14]. Ajah, I.A.; Nweke, H.F. Big data and business analytics: Trends, platforms, success factors and applications. Big Data Cogn. Comput. 2019, 3, 32. [Google Scholar] [CrossRef] [Green Version]
[15]. Nithya, N.; Kiruthika, R. Impact of Business Intelligence Adoption on performance of banks: A conceptual framework. J. Ambient. Intell. Humaniz. Comput. 2021, 12, 3139–3150. [Google Scholar] [CrossRef]
[16]. Al-Okaily, A.; Al-Okaily, M.; Teoh, A.P. Evaluating ERP systems success: Evidence from Jordanian firms in the age of the digital business. VINE J. Inf. Knowl. Manag. Syst. 2021; ahead-of-print.
[17]. Aws, A.L.; Ping, T.A.; Al-Okaily, M. Towards business intelligence success measurement in an organization: A conceptual study. J. Syst. Manag. Sci. 2021, 11, 155–170.
[18]. Wells, D. Business Analytics—Getting the Point. BeyeNetwork. 2008. Available online: http://www.b-eyenetwork.com/view/7133.
[19]. Tassey, G. The Roles and Economic Impacts of Technology Infrastructure, Version 3. National Institute of Standards and Technology; 2008. Available online: https://www.nist.gov/system/files/documents/2017/05/09/Measurement_Infrastr_Roles_Impacts_v3.pdf.
[20]. Giovinazzo, W. BI: Only as Good as Its Data Quality. Information Management Special Reports. Available online: http://www.information-management.com/specialreports/2009_157/business_intelligence_bi_data_quality_governance_decision_making-10015888-1.html (accessed on 3 December 2021).