2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA) | 978-1-6654-9079-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIBA56860.2023.10165495

Research on Data Quality Improvement Program Based on Big Data Application

Xuerong Zuo
School of Marxism, Hohai University,Nanjing, Jiangsu, China
Corresponding author: 1362900168@qq.com

***Abstract*—In the rapidly developing information age, the explosive growth of data increases the difficulty of big data application. Through analyzing the challenges faced by big data applications such as data collection, data timeliness, data security, data quality and data expected results, based on the key characteristics of big data and the measurement indicators of data quality, the improvement procedures of data quality are proposed, that is, following the pre-processing stage, the quality measurement stage, the quality defect identification stage and the pre-processing work tracing stage, in addition, the effect of the data quality improvement procedure is evaluated, to improve the data quality and reduce the potential risks in the application process of big data.**

Keywords-big data application; data quality; improve procedures

I. INTRODUCTION

With the rapid development of the information age, modern society has shown the characteristics of data explosion, and various types of data and information have been steadily and continuously growing. Today, big data is widely used in all walks of life. From the perspective of daily life, correct and effective application of big data can help people get more convenient and faster life. The application of big data in social networking can help us expand our circle of life and friends; When applied to traffic control, it can coordinate and standardize people's travel safety; It can reduce the losses caused by natural disasters when applied to weather forecast; Health management can also be carried out in combination with body data information. From the perspective of business intelligence, the key for enterprises and organizations to improve their market competitiveness lies in these huge amounts of data. Big data application technology can predict consumers' psychology and behavior, provide enterprises and organizations with the basis for planning and decision-making, and improve their economic benefits. All the above prove that big data applications have penetrated into all aspects of our lives and gradually become the focus of people's research [1-2]. Therefore, how to improve people's quality of life by improving the quality of data is particularly important.

II. CHALLENGES FACED BY BIG DATA APPLICATIONS

Although big data applications can provide people with more convenient lives, generally speaking, big data applications must first overcome the technical problems faced by big data processing operations before successfully achieving the expected benefits.

A. Challenges of data collection

In real life, most enterprises and organizations will obtain users' personal data, network transactions and business management strategies through various network collection tools. In order to prevent the intrusion of competitors, most enterprises and organizations adopt a reasonable anti-collection protection mode for websites. For example, some special characters will be selected to replace conventional vocabulary, and a specific cyberspace will be planned to store core data, which has caused certain obstacles to data collection and semantic analysis, and greatly impacted the efficiency of big data collection operations.

B. Challenges of data timeliness analysis

When enough and diversified data are collected, these data should be classified and stored first, and then handed over to special data analysts for data model construction, so as to develop appropriate and effective analysis methods. Data analysis is time-sensitive. In order not to affect the subsequent processing time and the laughter of big data analysis, data analysts must complete data analysis within the specified time. At present, the lack of experienced data analysts and data scientists in the field of big data application has also become another challenge for data processing.

C. Challenges of data security

In the process of continuous growth of data volume and value, driven by economic interests, problems related to data security frequently occur. Similarly, in the process of big data application, data collection and analysis often generate disputes about data security. Therefore, we must pay attention to data security issues, and avoid violating personal privacy information and protecting personal unpublished data when conducting data collection and analysis.

D. Challenges of data quality

Data quality has a decisive impact on big data applications. In the era of big data, a lot of information has not been verified, and special attention should be paid to the analysis and processing of data. If the collected data is wrong or untrue, there will be no analysis value, and on the contrary, it will waste the human resources and financial resources generated in the process of data collection and management. A large number of redundant and useless data will only lead to the inaccuracy of prediction and the error of decision-making, resulting in unexpected risk losses.

E. Challenges of expected results of data

If enterprises and organizations can't deeply understand their expected goals, they can't know the required data resources, and they can't determine the region, time period and content of data collection, which makes big data applications unable to achieve the expected goals. Therefore, the application of big data must overcome the specific confirmation of data resources and data requirements in order to promote the effective development of big data applications.

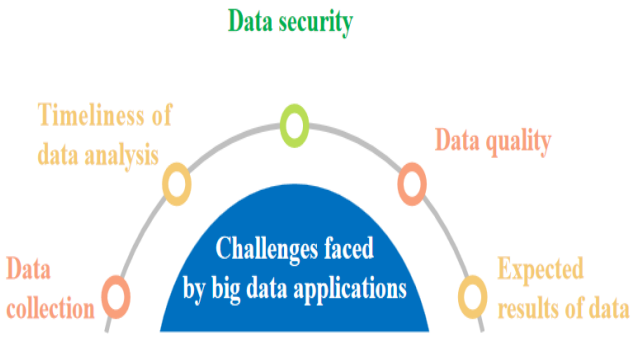

Figure 1. Challenges faced by big data applications

III. Big Data Application and Data Quality

A. The key characteristics of big data

A large amount of data collected from the network environment has the following four attributes, which are closely related to the application of big data and need our attention.

The number is extremely large: as long as the key words are retrieved on the Internet, a very large amount of data can be obtained in an instant, from trillions of bytes to exabytes. Therefore, it is necessary to plan a set of equipment to properly manage data and effectively help the subsequent work.

Rapid analysis and processing: in the process of online transactions, social activities and social networking, data is continuously and rapidly generated. Data scientists and data analysts need to build a set of rapid analysis and processing models according to the needs of data, so as to effectively grasp the timeliness of huge data resources, and then help enterprises and organizations gain market competitive advantages.

Diversified storage format: the current data format of big data shows a highly diversified feature. In addition to structured text and numbers, it can also be unstructured new multimedia data, such as video, film, sound, etc. Not only do they need different interpretation methods, but data in different formats cannot be stored in a structured database.

Data authenticity: this feature is an important feature recently imported. The authenticity of data is directly related to the analysis results of big data. In addition, the authenticity of the data can also make up for the lack of diversity, large volume and rapid development of big data characteristics, and can improve the practicability and effectiveness of big data.

B. Measurement index of data quality

Only by improving efficiency in the process of collecting and analyzing big data can we obtain higher substantive benefits. At present, all kinds of online transaction behaviors and activity records are the content to be collected after graduation of big data application. The larger the amount of data collected and the more diverse the format, the greater the difficulty of big data analysis and processing. Therefore, in order to complete the analysis of big data within the specified time and scope, display the concrete analysis results, and help the subsequent decision-making and processing work, big data must have efficient analysis tools and methods, and cooperate with the requirements of enterprises and organizations to achieve the ultimate goal of big data application.

In the era of big data, in order to achieve the practicability and effectiveness of subsequent big data applications, it is necessary to analyze massive data, screen and eliminate abnormal data in advance, so as to improve the quality of data and improve the efficiency of data analysis. Before the application of big data can obtain the expected benefits, it is necessary to check the problem of data quality and eliminate the improper information in advance to improve the efficiency and application results of bureau analysis[3-4].

Measurement Indicator 1: basic quality of data. In order to ensure the practicability and effectiveness of data, the data of the same data source should have basic quality characteristics such as data integrity, data consistency and data accuracy.

Measurement indicator 2: data availability. When massive data information is collected, we must invest corresponding time and resources to manage and store the data. Therefore, it is necessary to confirm whether the data is available before the next step of data analysis to avoid excessive use of human resources. The availability of data can be confirmed from the following three aspects: the source of data has high credibility, the time of data generation can be confirmed, and the content of data expression can be authenticated.

Measurement indicator 3: data manageability. Generally speaking, there is no fixed or uniform format for data collected on the Internet, and there are differences in the interpretation and processing methods of data. If the effective information contained in the data cannot be extracted, then the data will lose its existence value, and it is also impossible to analyze the data. Therefore, data needs to be manageable, able to transform data format independently, easy to access and compare, and easy to conduct semantic and statistical analysis.

Measurement indicator 4: maintainability of data. After classifying, managing and storing the data collected from different websites, organizations or groups, it is necessary to properly integrate and summarize the data to provide a basis for subsequent data modeling and analysis. The maintainability of data is reflected in its scalability, associativity and non-repeatability.

Table 1. Measurement indicators of data quality

Quality characteristics	Contents
Basic quality of data	Data integrity
	Data consistency
	Data accuracy
Data availability	The data source has high credibility
	The time of data generation can be confirmed
	The content of data expression can be confirmed
Data manageability	Data can be converted into format
	Data can be accessed and compared
	Data can be used for semantic and statistical analysis
Data maintainability	Data is extensible
	Data can be combined
	Data has no repeatability

C. *Process flow of big data application*

When we want to extract the necessary key information from such a complex data system, we must carry out at least five key steps of big data application and processing[5-6］.

In the data collection step, the data quality of the storage space should be classified and managed according to the data needs, and data related to users' browsing records and trading activities should be collected comprehensively and quickly from social groups, communications or commercial websites with the help of various network collection tools.

In the process of data management and storage, because the data collected from the network are diverse in form and content, it is necessary to use unstructured databases to store these diverse data in a timely manner. Of course, in order to properly manage such a huge amount of data, it is necessary to clearly classify the data first, and then transfer it to the cloud service to deal with subsequent matters.

In the data analysis and processing step, the big data application should rely on the clear data requirements put forward by the user, and the data scientists and data analysts should conduct deeper processing of the big data that has been managed and stored, so as to establish the data model and data statistics to complete the big data analysis.

In the data visualization step, the results obtained from the big data analysis need to be presented through the visualization charts and formulas, so that the superior decision-makers of the user unit can confirm the analysis content in more detail and easily, and then promote the subsequent work.

In the data decision-making, prediction and planning step, the data requirements formulated by the superior decision-maker in advance must be clear and complete, so that the data scientists and data analysts can analyze and process the specific data content, and then help the user unit to achieve the goals and tasks of decision-making, prediction and planning through the concrete charts.

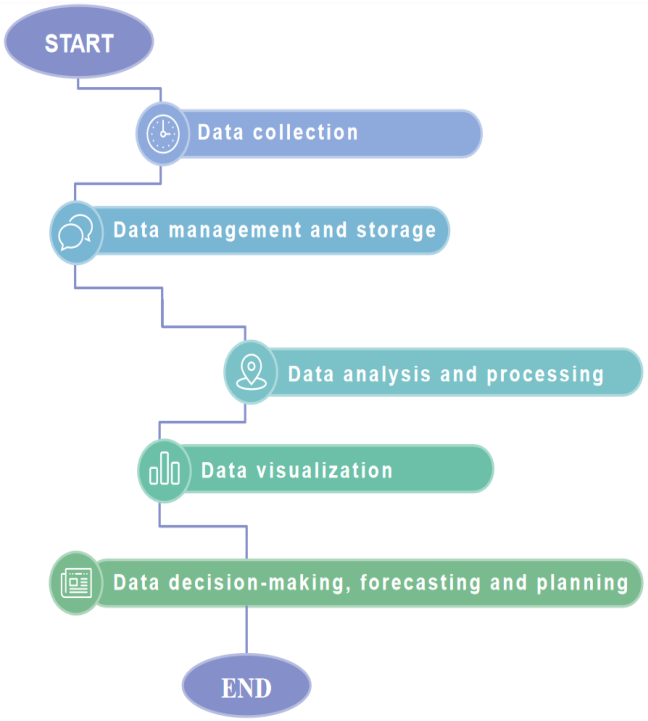

Figure 2. Big data application processing flow chart

IV. DATA QUALITY IMPROVEMENT PROCEDURES

Data quality errors or inaccurate analysis will affect the efficiency and quality of big data applications, so it is necessary to improve data quality. The collection and analysis of big data must ensure its efficiency if it wants to achieve the expected practical benefits. In the current era of big data, the workload of data collection is huge and the data format is diversified, which makes the analysis task of big data more difficult. The significance of data preprocessing is to improve the quality of collected data. However, data preprocessing cannot ensure that the improved data quality can meet the needs of big data applications. Therefore, it is also necessary to identify and reconstruct the lack of data partial quality. The data quality improvement program includes four stages: data pre-processing, data quality measurement, data quality defect identification, and tracing stage of pre-processing work.

Data pre-processing stage: the work content of data pre-processing stage mainly includes data cleaning, data transposition, data integration and data reduction [7-8]. Specifically, as the core task of data pre-processing, data cleaning first needs to identify data missing, fill, modify, check or delete abnormal data according to the types of data missing or other data problems, and adjust incomplete or inconsistent data. On this basis, we can ensure the data quality and better promote the application of big data. Data transposition means that it takes a lot of time and resources to collect data from different fields and extensive environments, and it often encounters difficulties in data processing and analysis in different formats. Therefore, it is necessary to transform the collected data into a suitable cluster in a unified format. Data integration refers to the sorting and merging of data in multiple data stores, databases or documents, which is an important link to increase the efficiency of data analysis. High-quality data integration can minimize or avoid the repeatability and inconsistency of data collection and storage, and can greatly improve the accuracy and efficiency of subsequent data analysis. Data reduction refers to the deletion

of similar or repeated contents collected. Duplicate or similar data identification tools can be adopted to identify, merge or delete duplicate or similar contents, so as to reduce the amount of data and the time of data analysis, and pay attention to avoid deleting data that should not be deleted. At the same time, each sub-work item is accompanied by detailed work tasks to be completed. The basic premise for the pre-processing work to confirm the data quality is to complete the assigned work.

Data quality measurement stage: in the data quality measurement stage, in order to ensure the quality of data, it is necessary to carefully inspect before the completion of pre-processing work. In terms of quality characteristics, data inspection activities mainly identify the lack of data quality by collecting quantitative quality factors. Most of the measurement methods of quality characteristics are based on the combination of some basic measurement data values[9-10].

Data quality defect identification stage: mainly based on the principle of data quality, identify the data quality defect with the help of quantitative quality factors.

Tracing stage of pre-processing work: because the data pre-processing work can not be completed successfully, the data quality is lost. In order to improve the data quality loss, it is necessary to trace back to the data pre-processing and re-complete this work.

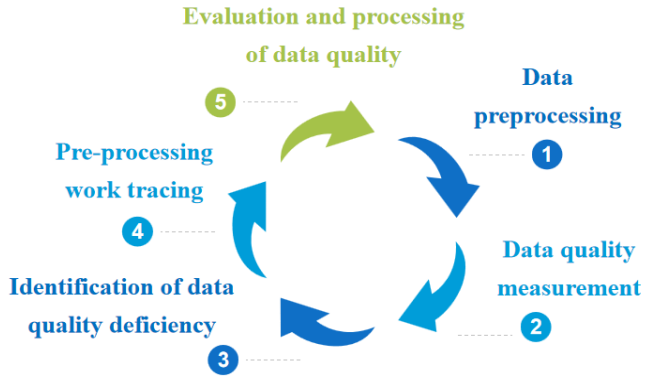

Figure 3. Flow chart of data quality improvement process

V. EFFECT EVALUATION OF DATA QUALITY IMPROVEMENT PROGRAM

As an extremely important step in big data application, after the data pre-processing stage, how to examine and improve the data quality is a problem that must be faced by the current big data application quality and practicability. The data quality improvement program proposed in this paper can redo the previously imperfect pre-processing work items to improve the data quality. From the perspective of application software development, the quality of application software is continuously adjusted and improved on the basis of periodic review, in order to improve the quality and efficiency of subsequent development work. To sum up, the quality of big data application is affected by four factors: improving accurate prediction or correct decision, improving data classification and storage effect, improving data attribute recognition and improving data statistical analysis efficiency. The effect of the data quality improvement program can be evaluated through inspection, as shown in the table.

Table 2. Effect evaluation of data quality improvement procedure

Big data prepro-cessing / Big data applica-tion work items	Preprocessing with data quality improvement program	Adopt the unimproved data preprocessing	Unused data preprocessing
Improve accurate prediction or correct decision	****	****	*
Improve data classification and storage	****	***	**
Improve data attribute recognition	****	***	**
Improve the efficiency of data statistics and analysis	****	***	*

Remark：****Excellent, ***Good, **Ordinary, *Bad

VI. CONCLUSION

In the modern society with fierce market competition, big data application is particularly important. It can not only help enterprises and organizations improve their market competitiveness, but also help improve people's quality of life. Therefore, the improvement of data quality has become a prerequisite for big data application. When the data has high quality, it can well promote the analysis and processing of big data, and then lay the foundation for big data application. The data quality improvement program can improve the data quality, further improve the efficiency and quality of big data application, and effectively reduce the potential risks in the process of big data application.

REFERENCES

[1] C.L. Philip Chen, C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Information Sciences 275, 2014, pp. 314–347.

[2] Nada Elgendy, Ahmed Elragal, "Big Data Analytics: A Literature Review Paper," Lecture Notes in Computer Science, pp.214-227, 2014.

[3] B. Saha and D. Srivastava, "Data quality: The other face of Big Data," in 2014 IEEE 30th International Conference on Data Engineering (ICDE), 2014, pp. 1294–1297.

[4] .J. Liu, J. Li, W. Li, J. Wu, "Rethinking big data: A review on the data quality and usage issues," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 115, 2016, pp. 134-142.

[5] Dave Wagner, "The importance of big data analytics in business", October, 2014 ,World of tech http://http://www.techradar.com/news/world-of-tech/the-impor-tance-of-big-data-analytics-in-business-1267606/2

[6] L. Cai, and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," Data Science Journal, 14: 2, pp. 1-10, 2015.

[7] Ikbal Taleb, Rachida Dssouli, Mohamed Adel Serhani "Big Data Pre-processing: A Quality Framework", 2015 IEEE International Congress on, 2015, pp.191-198.

[8] Bala Deshpande, "5 situations which drive data pre-processing before data mining," 2013, http://www.simafore.com/blog/bid/116618/5-situations-which-drive-data-pre-processing-before-data-mining.

[9] N. E. Fenton, Software Metrics - A Rigorous Approach, Chapman & Hall, 1991.

[10] Daniel Galin, Software Quality Assurance, Addison-Wesley, 2004.