Information Quality Assessment within the Huge Information

Chandrashekar H N and Dr. G Mahadevan

***Abstract*—**High-quality knowledge area unit the precondition for analyzing and exploitation massive knowledge and for guaranteeing the worth of the info. Currently, comprehensive analysis and analysis of quality standards and quality assessment strategies for giant knowledge area unit lacking. First, this paper summarizes reviews of information quality analysis. Second, this paper analyzes the data characteristics of the massive data atmosphere, presents quality challenges sweet-faced by massive knowledge, and formulates a hierarchal knowledge quality framework from the attitude of information users. This framework consists of huge knowledge quality dimensions, quality characteristics, and quality indexes. Finally, on the premise of this framework, this paper constructs a dynamic assessment method for knowledge quality. This method has sensible expansibility and adaptableness and might meet the wants of huge knowledge quality assessment. The analysis results enrich the theoretical scope of huge knowledge associate lay a solid foundation for the longer term by establishing an assessment model and finding out analysis algorithms.

***Keywords*—** Big data; Data quality; Quality assessment; Data science

I. INTRODUCTION

Many vital technological changes have occurred within the data technology trade since the start of the twenty first century, like cloud computing, the net of Things, and social networking. the event of those technologies has created the number of knowledge increase endlessly associated accumulate at an unprecedented speed. All the higher than mentioned technologies announce the approaching of huge information (Meng&Ci, 2013). Currently, the number of world information is growing exponentially. the info unit isn't any longer the GB and TB, however the metal (1PB = 210TB), EB (1EB = 210PB), and ZB (1ZB = 210EB). per IDC's "Digital Universe" forecasts (Gantz&Reinsel, 2012), forty ZB of knowledge are going to be generated by 2020.

The emergence of associate era of huge information attracts the eye of trade, academics, and government. By quickly exploit and analyzing huge information from numerous sources and with numerous uses, researchers and decision-makers have step by step completed that this huge quantity of knowledge has advantages for understanding client desires, rising service quality, and predicting and preventing risks. However, the utilization and analysis of huge information should be supported correct and high-quality information that may be a necessary condition for generating worth from huge information. Therefore, we have a tendency to analyze the challenges long-faced by huge information and planned a top-quality assessment framework and assessment method for it.

Chandrashekar H N, Research Scholar of Ph.D, Department of Computer Science, Rayalaseema University, Kurnool, India

Dr. G Mahadevan, Principal, Annai College of Engineering, Kumbakonam India.

II. THE CHALLENGES OF DATA QUALITY IN BIG DATA

A. Features of big data

Because massive knowledge presents new options, its knowledge quality additionally faces several challenges. The characteristics of big knowledge comes back all the way down to the 4Vs: Volume, Velocity, Variety, and price (Katal, Wazid, &Goudar, 2013). Volume refers to the tremendous volume of the info. we tend to sometimes use TB or on top of magnitudes to live this knowledge volume. Velocity means knowledge are being formed at associate new speed and should be restrained during a timely manner. Selection indicates that massive knowledge has all kinds of information types, and this diversity divides the info into structured knowledge and unstructured knowledge. These multi typed knowledge want higher processing capabilities.

Finally, price represents low-value density. price density is reciprocally proportional to total knowledge size, the greater the massive knowledge scale, the less comparatively valuable the info.

B. The challenges of data quality

Because huge information has the 4V characteristics, once enterprises use and method huge information, extracting high-quality and real information from the huge, variable, and sophisticated information sets becomes pressing issue. At present, big information quality faces the subsequent challenges:

• The diversity of information sources brings abundant information sorts and sophisticated data structures and increases the difficulty of knowledge integration.

In the past, enterprises solely used the information generated from their own business systems, such as sales and inventory information. But now, information collected and analyzed by enterprises have surpassed this scope. huge information sources square measure terribly wide, including: 1) information sets from internet and mobile internet (Li & Liu, 2013); 2) information from the web of Things; 3) information collected by varied industries;

4) Scientific experimental and empiric information (Demchenko, Grosso&Laat, 2013), such as high-energy physics experimental information, biological information, and

area observation information. These sources produce wealthy information sorts. One information kind is unstructured information, as an example, documents, video, audio, etc. The second kind is semi-structured information, including: software system packages/modules, spreadsheets, and monetary reports. The third is structured information. the number of unstructured information occupies more than eightieth of the overall quantity of knowledge breathing.

• Information volume is tremendous, and it's difficult to judge information quality within a reasonable amount of time.

After the economic revolution, the number of data dominated by characters doubled each 10years. After 1970, the number of data doubled each 3 years. Today, the worldwide quantity of information may be doubled each 2 years. In 2011, the number of worldwide information created and derived reached 1.8 ZB. it's troublesome to gather, clean, integrate, and at last get the mandatory high-quality data among an affordable time-frame. as a result of the proportion of unstructured information in huge information is incredibly high, it'll take loads of your time to remodel unstructured sorts into structured sorts and additional method the data. this can be an excellent challenge to the prevailing techniques of knowledge process quality.

• Information modification very fast and the "timeliness" of information is extremely short, that necessitates higher requirements for process technology.

Due to the fast changes in massive information, the "timeliness" of some information is extremely short. If firms can't collect the specified information in real time or subsume the information desires very long time, then they may get superannuated and invalid data. process and analysis supported this information will turn out useless or dishonest conclusions, eventually resulting in decision-making mistakes by governments or enterprises. At present, real-time operation and analysis software system for large information is still in development or improvement phases; very effective business merchandise is few.

III. QUALITY CRITERIA OF BIG DATA

Big information could be a new conception, and academe hasn't created an identical definition of its information quality and quality criteria.

The literature differs on a definition of knowledge quality, however one issue is certain: information quality depends not only on its own options however additionally on the business atmosphere mistreatment the information, as well as business processes and business users. solely the information that adjust to the relevant uses and meet needs are often thought of qualified (or sensible quality) information. Usually, information quality standards area unit developed from the attitude of data producers. Within the past, information customers were either direct or indirect information producers that ensured the data quality. However, within the age of huge information, with the range of knowledge sources, information users don't seem to be essentially data producers. Thus, it's terribly tough to live information quality. Therefore, we have a tendency to propose a gradable information quality commonplace from the attitude of the users, as shown in Figure 1.

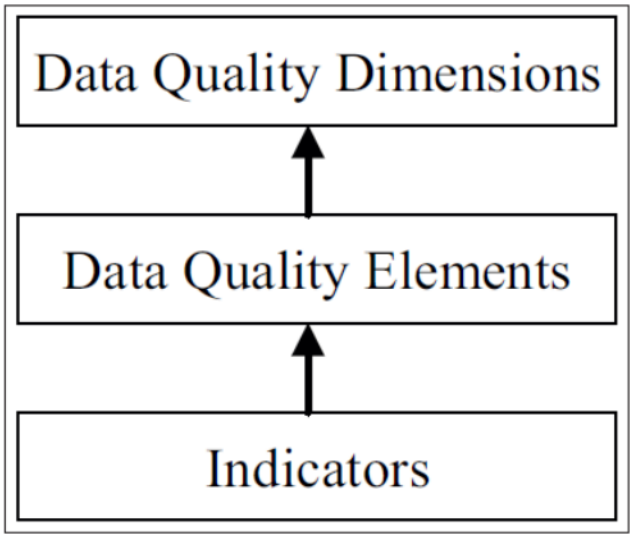

Fig. 1: Data quality framework.

We selected information quality dimensions ordinarily accepted and wide used as huge information quality standards and redefined their basic ideas supported actual business desires. At constant time, every dimension was divided into several typical parts related to it, and every part has its own corresponding quality indicators. During this method, gradable quality standards for giant information were used for analysis. Figure2 shows a universal two-layer information quality commonplace.

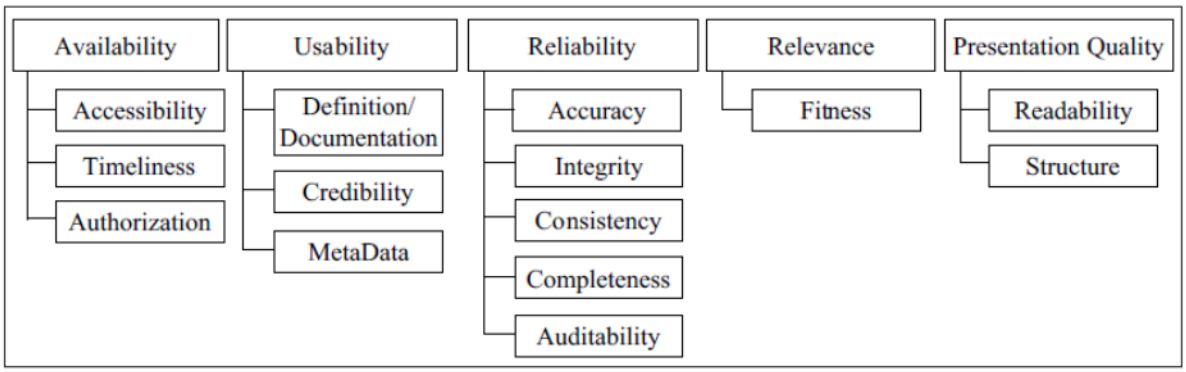

Fig. 2: A universal, two-layer big data quality standard for assessment.

In Figure 2, the information quality commonplace consists of 5 dimensions of knowledge quality - convenience, usability, reliability, relevance, and presentation quality. for every dimension, we have a tendency to known 1–5 parts with good practices. The primary four quality dimensions area unit considered indispensable, inherent options of knowledge quality, and also the final dimension is extra properties that improve client satisfaction. Convenience is defined because the degree of convenience for users to get information and connected info, that is split into the 3 parts of accessibility, authorization, and timeliness. The conception of usability means that whether or not the data area unit helpful and meet users' desires, as well as information definition/documentation, dependableness, and information.

Reliability refers as to if we will trust the data; this consists of accuracy, consistency, completeness, adequacy, and auditability parts. Connection is employed to explain the degree of correlation between information content and users' expectations or demands; ability is its quality part (Cappiello, Francalanci, &Pernici, 2004). Presentation quality refers to a sound description technique for the information that permits users to fully perceive the information. Its dimensions area unit readability and structure. Descriptions of the information quality elements area unit given below.

Accessibility

Accessibility refers to the issue level for users to get information. Accessibility is closely coupled with information openness, the upper the information openness degree, the lot of information sorts obtained, and also the higher the degree of accessibility.

Timeliness

Timeliness is outlined because the times delay from information generation and acquisition to utilization (McGivray, 2010). Information ought to be accessible among this delay to permit for important analysis. In the age of massive information, information content changes quickly thus timeliness is extremely vital.

Authorization

Authorization refers as to whether a private or organization has the correct to use the information.

Credibility

Credibility is employed to judge non-numeric information. It refers to the target and subjective elements of the credibleness of a supply or message. Quality of knowledge has 3 key factors: dependableness of knowledge sources, information social control, and also the time once the information area unit created.

• Definition/Documentation

Definition/document consists of knowledge specification, which has information name, definition, ranges of valid values, commonplace formats, business rules, etc. Normative information definition improves the degree of knowledge usage.

•Metadata

With the rise data sources and data sorts, as a result of information customers distort the that means of common word and ideas of knowledge, victimization information could bring risks. Therefore, information producers have to be compelled to give data describing completely different aspects of the datasets to cut back the issues caused by misunderstanding or inconsistencies.

•Accuracy

To ascertain the accuracy of a given knowledge price, it's compared to a familiar reference price. In some things, accuracy will be simply measured, like gender, that has solely 2 definite values: male and female. however, in alternative cases, there's no familiar reference price, creating it tough to live accuracy. as a result of accuracy is related to with context to some extent, knowledge accuracy ought to be determined by the appliance scenario.

• Consistency

Data consistency refers as to whether the logical relationship between related to knowledge is correct and complete. Within the field of databases (Silberschatz, Korth, &Sudarshan, 2006), it always implies that an equivalent knowledge that placed in numerous storage areas ought to be thought of to be equivalent. Equivalency implies that the information has equal price and also the same which means or are basically an equivalent. Knowledge synchronization is that the method of creating knowledge equal.

• Integrity

The term knowledge integrity is broad in scope and will have wide completely different meanings counting on the precise context. in a very info, knowledge with "integrity" are same to own an entire structure. Knowledge values are standardized in line with data model and/or data sort. All characteristics of the information should be correct – together with business rules, relations, dates, definitions, etc. In info security, knowledge integrity suggests that maintaining and reassuring the accuracy and consistency of knowledge over its entire life-cycle. This implies that knowledge can't be changed in associate unauthorized or undiscovered manner.

• Completeness

If a information has multiple elements, we will describe the standard with completeness. Completeness implies that the values of all elements of one information valid for instance, for image color, RGB will be accustomed describe red, green, and blue, and RGB represents all components of the colour knowledge. If the colour price of an explicit element is missing, the image cannot show the important color and its completeness is destroyed (Wang & construction, 1995).

•Auditability

From the angle of audit application, data life cycle includes 3 phases: data generation, data assortment, and information use (Wang & Zhu, 2007). however here audit ability means auditors can fairly appraise information accuracy and integrity inside rational time and workforce limits throughout the data use section.

• Fitness

Fitness has two-level requirements: 1) the quantity of accessed information employed by users and 2) the degree to that the information created matches users' desires within the aspects of indicator definition ,elements, classification, etc.

• Readability

Readability is outlined because the ability of knowledge content to be properly explained consistent with far-famed or well outlined terms, attributes, units, codes, abbreviations, or alternative info.

• Structure

More than eightieth of all information is unstructured; therefore, structure refers to the amount of problem in transforming semi-structured or unstructured information to structured information through technology.

We gift an enormous information quality assessment framework in Table one, that lists the common quality parts and their associated indicators. Generally, a top-quality part has its own multi-indicators.

IV. QUALITY ASSESSMENT PROCESS FOR BIG DATA

An acceptable quality assessment methodology for large knowledge is critical to draw valid conclusions. during this paper, we propose a good knowledge quality assessment method with a dynamic feedback mechanism supported huge data's own characteristics, shown in Figure three.

Determining the goals of knowledge assortment is that the start of the total assessment method. huge knowledge users rationally select the information to be used in step with their strategic objectives or business needs, such as operations, higher cognitive process, and designing. the information sources, types, volume, quality needs, assessment criteria, and specifications in addition because the expected goals got to be determined ahead.

In different business environments, the choice of knowledge quality parts can take issue. For instance, for social media knowledge, timeliness and accuracy square measure 2 vital quality options. However, as a result of it's troublesome to directly choose accuracy (Shankaranarayanan, Ziad, & Wang, 2012), some extra data is required to judge the information, and alternative knowledge sources function supplements or proof. Therefore, credibleness has become a very important quality dimension. However, social media knowledge square measure sometimes unstructured, and their consistency and integrity aren't appropriate for analysis. the sector of biology is a very important supply of massive data. However, because of the shortage of uniform standards, knowledge storage code and knowledge formats vary wide. Thus, it's troublesome to take consistency as a top-quality dimension, and also the wants of relating to timeliness and completeness as knowledge quality dimensions aren't high.

In order to more quality assessment, we want to settle on specific assessment indicators for each dimension. These need the information to adjust to specific conditions or options. The formulation of assessment indicators conjointly depends on the particular business atmosphere.

Each quality dimension wants completely different activity tools, techniques, and processes that end up indifferences in assessment times, costs, and human resources. In an exceedingly clear understanding of the work needed to assess every dimension, selecting those dimensions that meet the wants will well outline a project's scope. The preliminary assessment results of knowledge quality dimensions confirm the baseline whereas the remaining assessment as a locality of the business method is employed for continuous detection and knowledge improvement.

After the standard assessment preparation is completed, the method enters the info acquisition part. There are many ways to gather information (Zhu &Xiong, 2009), including: information integration, search-download, net crawlers, agent ways, carrier monitors, etc. within the age of massive information, information acquisition is comparatively straight forward, but much of the info collected isn't forever smart. We want to enhance information quality as so much as attainable underneath these conditions while not an oversized increase in acquisition price.

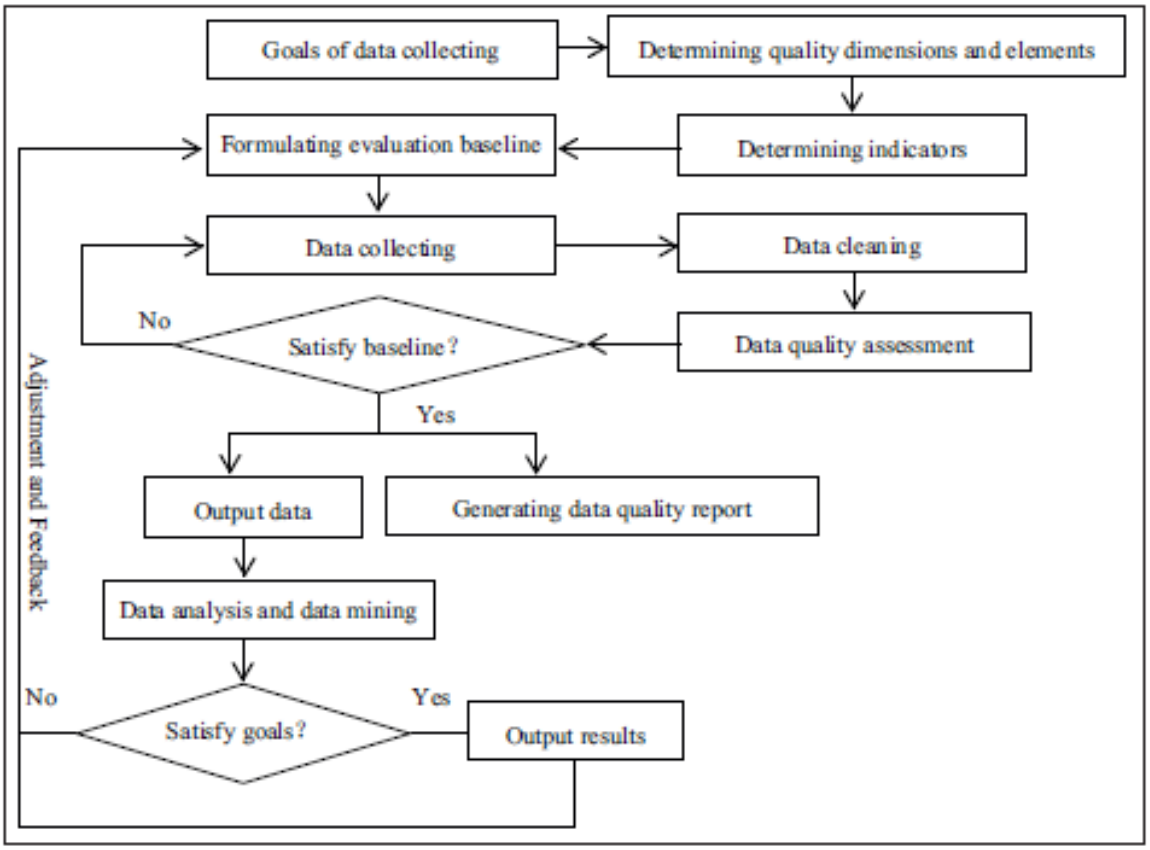

Fig. 3 : Quality assessment process for big data

Big information sources are terribly wide and information structures are advanced. the info received could have quality issues, such as information errors, missing data, inconsistencies, noise, etc. the aim of knowledge cleanup(data scrubbing) is to sight and take away errors and inconsistencies from information so as to enhance their quality. Data cleanup may be divided into four patterns supported implementation ways and scopes (Wang,Zhang, & Zhang, 2007): manual implementation, writing of special application programs, information cleanup unrelated to specific application fields, and finding the matter of a kind of specific application domain. In these four approaches, the third has smart sensible worth and might be applied with success.

Then, the method enters the info quality assessment and observation phases. The core of knowledge quality assessment is the way to value every dimension. The present technique has 2 categories: qualitative and quantitative ways. The qualitative analysis technique relies on bound analysis criteria and needs, according to assessment functions and user demands, from the attitude of analysis to describe and assess information resources. Analysis ought to be performed by subject consultants or professionals. The quantitative technique could be a formal, objective, and systematic methods during which numerical information are utilized to get data. Therefore, sound judgment, generalizability, and numbers are options typically associated with this technique, whose analysis results are a lot of intuitive and concrete.

After assessment, the info may be compared with the baseline for the info quality assessment established above. If the info quality accords with the baseline commonplace, a follow-up information analysis part may be entered and a knowledge quality report is going to be generated. Otherwise, if the info quality fails to satisfy the baseline common place, it is necessary to accumulate new information.

V. Conclusion

The arrival of the large knowledge era makes knowledge in numerous industries and fields gift explosive growth. How to ensure massive knowledge quality and the way to research and mine info and information hidden behind the info become major problems for business and domain. Poor knowledge quality can result in low knowledge utilization potency and even bring serious decision-making mistakes. We tend to analyzed the challenges faced by massive knowledge quality and proposed the institution and data structure of an information quality framework. Then, we tend to develop a dynamic massive knowledge quality assessment method with a feedback mechanism that has set a decent foundation for more study of the assessment model. Following stage of analysis can involve the development of a giant data quality assessment model and formation of a weight constant for every assessment indicator. At the same time, the analysis team can develop an rule accustomed build a sensible assessment of the large knowledge equality in an exceedingly specific field.

REFERENCES

[1] Cao, J. J., Diao, X. C., Wang, T., et al. (2010) Research on Some Basic Problems in Data Quality Control. Micro-computer Information 09, pp 12–14.

[2] Cappiello, C., Francalanci, C., &Pernici, B. (2004) Data quality assessment from user‘s perspective. Procedures of the 2004 International Workshop on Information Quality in Information Systems, New York: ACM, pp 78–73. https://doi.org/10.1145/1012453.1012465

[3] Demchenko, Y., Grosso, P., de Laat, C., et al. (2013) Addressing Big Data Issues in Scientific Data Infrastructure. Procedures of the 2013 International Conference on Collaboration Technologies and Systems, Califor-nia: ACM, pp 48–55. https://doi.org/10.1109/CTS.2013.6567203

[4] Feng, Z. Y., Guo, X. H., Zeng, D. J., et al. (2013) On the research frontiers of business management in the con-text of Big Data. Journal of Management Sciences in China 16(01), pp 1–9. https://doi.org/10.1016/S1001-0742(12)60053-9

[5] Gantz, J., &Reinsel, D. (2012) THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Retrieved February, 2013 from the World Wide Web: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-western-europe.pdf

[6] Meng, X. F., &Ci, X. (2013) Big Data Management: Concepts, Techniques and Challenges. Journal of Computer Research and Development 50(1), pp 146–169.

[7] Nature (2008) Big Data. Retrieved November 5, 2013 from the World Wide Web: http://www.nature.com/news/specials/bigdata/index.html

[8] Science (2011) Special online collection: Dealing with data. Retrieved November 5, 2013 from the WorldWideWeb: http://www.sciencemag.org/site/special/data/

[9] Shankaranarayanan, G., Ziad, M., & Wang, R. Y. (2012) Preliminary Study on Data Quality Assessment for Socialized Media. China Science and Technology Resources 44(2), pp 72–79.

[10] Wang, J. L., Li, H., & Wang, Q. (2010) Research on ISO 8000 Series Standards for Data Quality. Standard Science 12, pp 44–46.

[11] Zhu, Y. Y., &Xiong, Y. (2009) Datology and Data Science, Shanghai: Fudan University Press.

[12] Zong, W., & Wu, F.(2013) The Challenge of Data Quality in the Big Data Age. Journal of Xi’an Jiaotong University (Social Sciences) 33(5), pp 38–43.