*Int. J. Advance Soft Compu. Appl, Vol. 7, No. 3, November 2015*
*ISSN 2074-8523*

# Data Quality in Big Data: A Review

**Noraini Abdullah[1], Saiful Adli Ismail[1], Siti Sophiayati[1], and Suriani Mohd Sam[1]**

[1]Advanced Informatics School, Universiti Teknologi Malaysia
Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
e-mail: norainiabdullah@gmail.com, saifuladli@utm.my,
sophia@utm.my, suriani.kl@utm.my

**Abstract**

***The Data Warehousing Institute (TDWI) estimates that data quality problems cost U.S. businesses more than $600 billion a year. The problem with data is that its quality quickly degenerates over time. Experts say 2 percent of records in a customer file become obsolete in one month because customers die, divorce, marry, and move. In addition, data entry errors, system migrations, and changes in source systems, among other things, generate bucket loads of errors. More complex, as organizations fragment into different divisions and units, interpretations of data elements change to meet the local business needs. However, there are several ways that the Company should concern, such as to treat data as a strategic corporate resource; develop a program for managing data quality with a commitment from the top; and hire, train, or outsource experienced data quality professionals to oversee and carry out the program. The Organizations can sustain a commitment to managing data quality over time and adjust monitoring and cleansing processes to changes in the business and underlying systems by using the Commercial data quality tools. Data is a vital resource. Companies that invest proportionally to manage this resource will stand a stronger chance of succeeding in today's competitive global economy than those that squander this critical resource by neglecting to ensure adequate levels of quality. This paper reviews the characteristics of big data quality and the managing processes that are involved in it.***

**Keywords**: *Big data, five v's, Data Quality, Big Value, Data Quality Attributes, Data Quality Methodology.*

# 1 Introduction

Data quality is part of the big data concern. Big data is a term applied to a new generation of software, applications, and system and storage architecture, all designed to derive business value from unstructured data. Advanced tools, software, and systems are required to capture, store, manage, and analyze the data sets, all in a timeframe that preserves the intrinsic value of the data.

Big data may be defined as "techniques and technologies that make handling data at extreme scale affordable" [1]. Big data is not only a technology, but also involves people with the appropriate analysis skills, and makes dealing with extreme scale affordable. It was originated as a tag for a class of technology with roots in high-performance computing, as pioneered by Google in the early 2000.

This paper firstly defines what big data are, then introduces the characteristics of quality data, and, then reviews the processes to manage the quality of data.

# 2 Overview of Big Data

Big data has been defined in terms of the five v's [1]: volume, velocity, variety, veracity and value. The detail explanation as shown in Table 1.

Table 1: Explanation on 5vs of Big Data

| No | 5V's | Description |
|---|---|---|
| 1. | Volume | The quantity of data relative to the ability to store and manage it |
| 2. | Velocity | The speed of calculation needed to query the data relative to the rate of change of the data[2] |
| 3. | Variety | A measure of the number of different formats the data exist in (e.g. text, audio, video, logs etc.) |
| 4. | Veracity | Refers to the messiness or the trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable (posts with hashtags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of the content) but big data and analytics technology now allows us to work with these type of data. The volumes often make up for the lack of quality or accuracy. |
| 5. | Value | There is another v to take into account when looking at Big Data: Value! Having access to big data is no good unless we can turn it into value. Companies are starting to generate amazing value from their big data. |

Most definitions of big data focus on the size of data in storage. However there are other important attributes of big data, namely data variety and data velocity[3]. The three v's of big data (volume, velocity and variety) bust the myth that big data is only about data volume. In addition, each of the three v's has its own

ramifications for analytics. Now, there are 2 additional attributes exist i.e. veracity[4] and value[5] as shown in Figure 1.

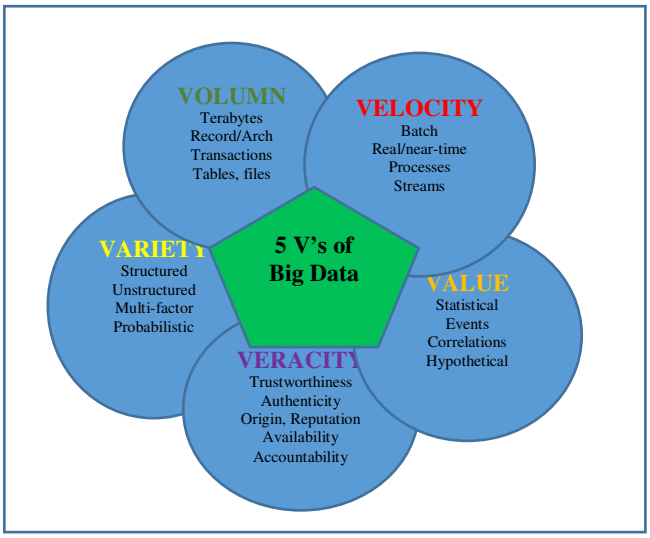


Fig. 1: The Five V's of big data

New technologies like Hadoop, NoSQL and MPP databases have emerged to address Big Data challenges and to enable new types of products and services to be delivered by the business. New technologies enable the analysis of big data based on each component.

Big Data Components [6] consist of the following: Hadoop: Provides storage capability through a distributed, shared-nothing file system, and analysis capability through MapReduce.[7], and, NoSQL: Provides the capability to capture, read, update, in real time, the large influx of unstructured data and data without schemas; examples include click streams, social media, log files, event data, mobility trends and sensor and machine data.

## 2.1 Creating Big Value from Big Data

A three-step approach below can help to determine how to create Big Value [8] from Big Data.

i. Start with the Right Big Data Store

Matching the business problem or opportunity with the right technology is an important first step. Big Data stores fit into one of several categories: Hadoop (which is a software framework that includes a Big Table clone called HBase)[1], NoSQL (which is subdivided into several more categories, including Key Value Stores, Document Databases, Graph Databases, Big Table Structures, and Caching Data Stores), Analytical Databases (e.g. Infobright, VectorWise, Vertica, Netezza, etc.).

ii. Add Deep Domain Knowledge

Domain knowledge is the human intelligence that accumulates within a certain practice or process. A "domain" in this sense could be a functional application area (like CRM or Supply-Chain), a vertical industry (like financial services, pharmaceuticals, or energy/utilities), or a specific process (like after-sales support). Domain expertise is necessary to genuinely know which data, from all the possible sources, are valuable and which are not. Domain knowledge is the primary reason the Big Data opportunity requires business unit personnel to lead rather than follow more than ever before.

iii. Apply the Right Reporting & Analysis Tool

Choosing the right reporting and analysis tool that enables the right overall big data approach (or architecture) is perhaps the most important step.

# 3 Big Data Quality

Data quality is not necessarily data that is devoid of errors. Incorrect data is only one part of the data quality equation. Most experts take a broader perspective. Larry English says data quality involves "consistently meeting knowledge worker and end-customer expectations." Others say data quality is the fitness or suitability of data to meet business requirements. In any case, most cite several attributes that collectively characterize the quality of data [9].

In order for the analyst to determine the scope of the underlying root causes and to plan the ways that tools[10] can be used to address data quality issues, it is valuable to understand these common data quality attributes[11].

## 3.1 Characteristics of Data Quality

Figure 2 shown the first five attributes (i.e. Accuracy, Integrity, Consistency, Completeness and Validity) generally pertain to the content and structure of data, and cover a multitude of sins that we most commonly associate with poor quality data: data entry errors, misapplied business rules, duplicate records, and missing or incorrect data values. But defect-free data is worthless if knowledge workers cannot understand or access the data in a timely manner[12]. The last two attributes (Timeliness and Accessibility) above address usability and usefulness, and they are best evaluated by interviewing and surveying business users of the data.

Data quality is an essential characteristic that determines the reliability of data for making decisions. High data quality[13] is:

- Complete: All relevant data such as accounts, addresses and relationships for a given customer is linked.

- Accurate: Common data problems like misspellings, typos, and random abbreviations have been cleaned up.
- Available: Required data are accessible on demand; users do not need to search manually for the information. [14]
- Timely: Up-to-date information is readily available to support decisions.[15]

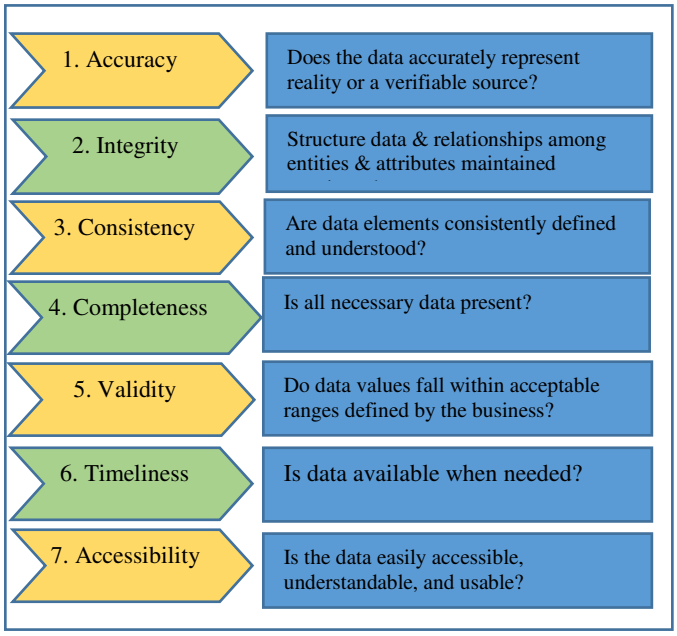


Fig. 2: Data quality attributes

Amongst others, there are several conditions that contributed to the data quality problem such as [16] lack of validation routines, data valid, but not correct [17], mismatched syntax, formats, and structures[18], unexpected changes in source system, spiderweb of interfaces, lack of referential integrity checks, poor system design and data conversion errors.

According to TDWI's Data Quality Survey [1], almost half of companies (40%) have suffered losses, problems, or costs due to poor quality data and 43% have yet to study the issue. The two most common problems caused by poor quality data are extra time required to reconcile data and loss of credibility in the system or application. Fig. 3 represents the output of 286 respondents who could select multiple answers on the data quality problem.

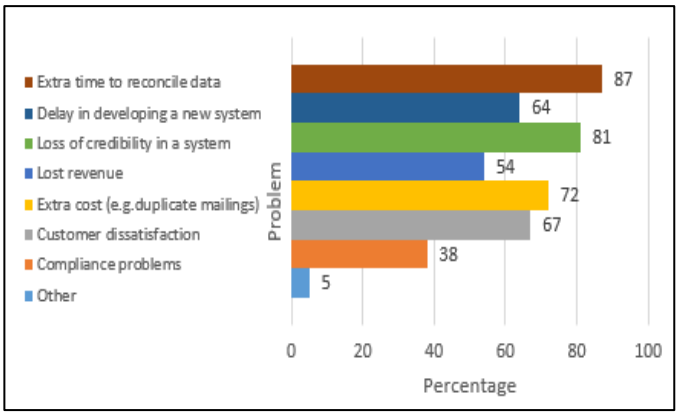


Fig. 3: Problem due to poor data quality

Defective data cause a litany of problems. Without good data, organizations are running blind. They unable to make good decisions because they have no accurate understanding of what is happening within their company or the marketplace. They rely on intuition, which is dangerous in a fast moving market with nimble competitors and finicky customers.

These two problems are related since an inability to reconcile data between the data warehouse and the source systems causes end users to lose confidence in the data warehouse. This is true even if the data in the warehouse is more accurate. Without Meta data to track the origin and transformation of data in the warehouse, users typically trust source systems before a data warehouse.

Companies also cite extra costs due to duplicate mailings, excess inventory, inaccurate billing and lost discounts as well as customer dissatisfaction, delays in deploying new systems, and lost revenue. Several survey respondents also noted an extremely serious problem that Impact on strategic planning and programs. Poor data quality has undermined strategic plans or projects.

Companies that have invested in managing and improving data quality can cite tangible and intangible benefits, often the inverse of the problems mentioned above. For example, a data quality project at a medium-sized financial institution is generating cost-savings of $130,000 annually on an outlay of $70,000 ($40,000 for software and $30,000 for data cleansing services). This project's internal rate of return is 188 percent and the net present value is $278,000 with a several month payback. Almost half of our respondents (47%) said their companies have derived benefits from better quality data. Based on fig. 4 below, topping the list were customer satisfaction, creating a single version of the truth and greater confidence in analytical systems and so on.

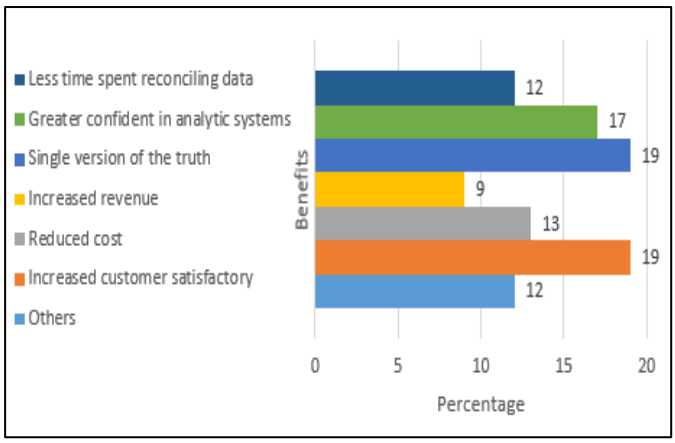


Fig. 4: Benefit of high quality data

## 3.2 Managing Data Quality

Managing data quality is a never ending process. Even if a company gets all the pieces in place to handle today's data quality problems, there will be new and different challenges tomorrow. That's because business processes, customer expectations, source systems, and business rules all change continuously. To ensure high quality data, companies need to gain broad commitment to data quality management principles and develop processes and programs that reduce data defects over time. To lay the foundation for high quality data, companies need to adhere to a methodology depicted in fig. 5.

The detail of data quality methodology[19] are described as follows:

i. Launch a Data Quality Program.

   The first step to delivering high quality data is to get top managers to admit there is a problem and take responsibility for it i.e. getting executives on board.

ii. Develop a Project Plan

   The next step is to develop a data quality project plan, or series of plans. A project plan should define the scope of activity, set goals, estimate ROI, perform a gap analysis, identify actions, and measure and monitor success.

iii. Build a Data Quality Team Positions.

   To implement a data quality plan, organizations must assign or hire individuals to create the plan, perform an initial assessment, scrub the data, and set up monitoring systems to maintain adequate levels of data quality. Management should establish the Data Quality Team[20] with regards to the following functions as per table 2.

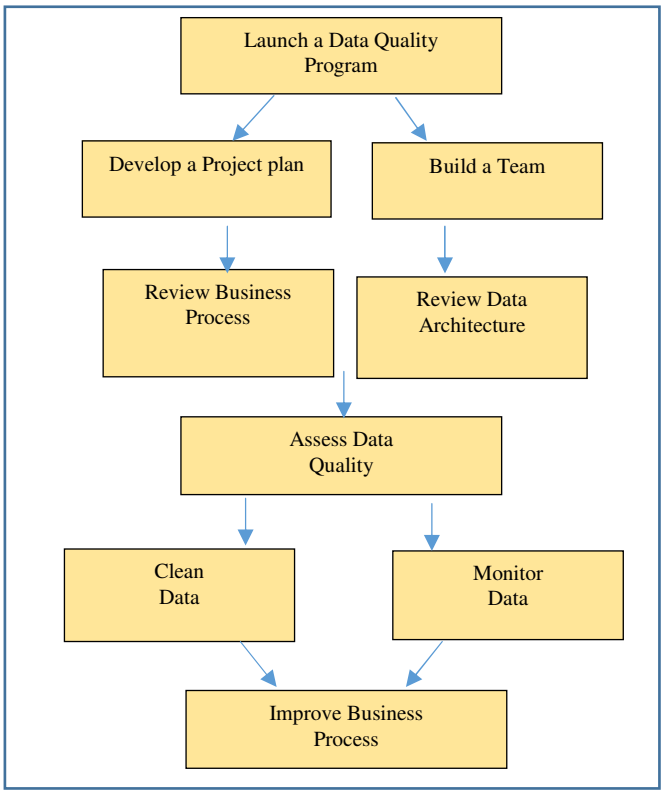


Fig. 5: An eight step methodology for maintaining data quality

Table 2: The Functions of Data Quality Team

| **Data Quality Team** | **Functions** |
|---|---|
| Chief Quality Officer | A business executive who oversees the organization's data stewardship, data administration, and data quality programs. |
| Data Steward | A business person who is accountable for the quality of data in a given subject area. |
| Subject Matter Expert | A business analyst whose knowledge of the business and systems is critical to understand data, define rules, identify errors, and set thresholds for acceptable levels of data quality. |
| Data Quality Leader | Oversees a data quality program that involves building awareness, developing assessments, establishing service level agreements, cleaning and monitoring data, and training technical staff. |
| Data Quality Analyst | Responsible for auditing, monitoring, and measuring data quality on a daily basis, and recommending actions for correcting and preventing errors and defects. |

| | |
|---|---|
| Tools Specialists | Individuals who understand either ETL or data quality tools or both and can translate business requirements into rules that these systems implement. |
| Process Improvement Facilitator | Coordinates efforts, to analyze and reengineer business processes to streamline data collection, exchange, and management, and improve data quality. |
| Data Quality Trainer | Develops and delivers data quality education, training, and awareness programs. |

iv. Review Business Processes and Data Architecture

Once there is corporate backing for a data quality plan, the stewardship committee or a representative group of senior managers throughout the firm needs to review the company's business processes for collecting, recording, and using data in the subject areas defined by the scope document. With help from outside consultants, the team also needs to evaluate the underlying systems architecture that supports the business practices and information flows.

v. Assess Data Quality, Data Auditing.

After reviewing information processes and architectures, an organization needs to undertake a thorough assessment of data quality in key subject areas. This process is also known as data auditing or profiling. The purpose of the assessment is to (1) identify common data defects (2) create metrics to detect defects as they enter the data warehouse or other systems, and (3) create rules or recommend actions for fixing the data.

vi. Clean the Data

Once the audit is complete, the job of cleaning the data begins[21]. A fundamental principle of quality management is to detect and fix defects as close as possible to the source to minimize costs. There are four basic methods[22] for data cleaning as described in table 3 below.

Table 3: Four Basic Methods of Data Cleaning

| **Cleaning Data Method** | **Description** |
|---|---|
| Correct | Most cleansing operations involve fixing both defective data elements and records. |
| Filter | Filtering involves deleting duplicate, missing or nonsensical data elements, such as when an ETL process loads the wrong file or the source system corrupts a field. Caution must be taken when filtering data because it may create data integrity problems. |
| Detect and Report | Want to change defective data because it is not cost-effective or possible to do so. |

| | |
|---|---|
| Prevent. | Prevention involves educating data entry people, changing or applying new validations to operational systems, updating outdated codes, redesigning systems and models, or changing business rules and processes. |

vii. Monitor Data

It is time consuming to prepare data files when loading a database for the first time. But organizations can quickly lose the benefits of these data preparation efforts if they fail to monitor data quality continuously. To monitor data quality, companies need to build a program that audits, data at regular intervals, or just before or after data is loaded into another system such as a data warehouse. Companies then use the audit reports to measure their progress in achieving data quality goals and complying with service level agreements negotiated with business groups

# 4 Conclusion

Companies that manage their data as a strategic resource and invest in its quality are already pulling ahead in terms of reputation and profitability of those that fail to do so. The quality of a company's data generates both tangible and intangible costs and benefits. Clearly, the further we move into the information economy, the more important it will be for companies to invest in maintaining good quality data.

**ACKNOWLEDGEMENTS**.
This research is a part of the Potential Academic Research Scheme (Vot Number 01K63). The authors would like to thanks Research Management Centre (RMC), Universiti Teknologi Malaysia (UTM) for the support in R & D.

# 5 References

[1] P. Russom. "Big Data Analytics." TDWI Best Practices Report, Fourth Quarter 2011. TDWI Research. 2011.

[2] H. Chen, R. H. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact." MIS Q., vol. 36, no. 4, pp. 1165–1188, 2012.

[3] J. Gong, L. L. Wang, and Q. Xu, "A three-step dealiasing method for Doppler velocity data quality control," J. Atmospheric Ocean. Technol., vol. 20, no. 12, pp. 1738–1748, 2003.

[4] P. Hitzler and K. Janowicz, "Linked data, big data, and the 4th paradigm," Semantic Web, vol. 4, no. 3, pp. 233–235, 2013.

[5] R. N. Bolton and J. H. Drew, "A multistage model of customers' assessments of service quality and value," J. Consum. Res., pp. 375–384, 1991.

[6] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, “Data mining with big data,” Knowl. Data Eng. IEEE Trans. On, vol. 26, no. 1, pp. 97–107, 2014.

[7] “Cuzzocrea et al. - 2011 - Analytics over large-scale multidimensional data .pdf.” .

[8] B. Brown, M. Chui, and J. Manyika, “Are you ready for the era of ‘big data,’” McKinsey Q., vol. 4, pp. 24–35, 2011.

[9] W. W. Eckerson, “Data quality and the bottom line,” TDWI Rep. Data Wareh. Inst., 2002.

[10]J. Barateiro and H. Galhardas, “A Survey of Data Quality Tools.,” Datenbank-Spektrum, vol. 14, no. 15–21, p. 48, 2005.

[11]Y.-Y. R. Wang, R. Y. Wang, M. Ziad, and Y. W. Lee, Data Quality. Springer Science & Business Media, 2001.

[12]Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations,” Commun. ACM, vol. 39, no. 11, pp. 86–95, 1996.

[13]R. Jugulum, Competing with High Quality Data: Concepts, Tools, and Techniques for Building a Successful Approach to Data Quality. John Wiley & Sons, 2014.

[14]D. Boyd and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” Inf. Commun. Soc., vol. 15, no. 5, pp. 662–679, 2012.

[15]L. Sebastian-Coleman, Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Newnes, 2012.

[16]R. Y. Wang and D. M. Strong, “Beyond accuracy: What data quality means to data consumers,” J. Manag. Inf. Syst., pp. 5–33, 1996.

[17]D. McGilvray, Executing Data Quality Projects: Ten Steps to Quality Data and Trusted InformationTM. Morgan Kaufmann, 2010.

[18]Y. W. Lee, “Crafting rules: Context-reflective data quality problem solving,” J. Manag. Inf. Syst., vol. 20, no. 3, pp. 93–119, 2003.

[19]Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, “AIMQ: a methodology for information quality assessment,” Inf. Manage., vol. 40, no. 2, pp. 133–146, 2002.

[20]R. Madrigal, J. James, and others, “Team quality and the home advantage.,” J. Sport Behav., vol. 22, no. 3, pp. 381–398, 1999.

[21]K. Adu-ManuSarpong and J. Kingsley Arthur, “Analysis of Data Cleansing Approaches regarding Dirty Data A Comparative Study,” Int. J. Comput. Appl., vol. 76, no. 7, pp. 14–18, Aug. 2013.

[22]L. Zhao, S. S. Yuan, S. Peng, and L. T. Wang, “A New Efficient Data Cleansing Method,” in Database and Expert Systems Applications, A. Hameurlain, R. Cicchetti, and R. Traunmüller, Eds. Springer Berlin Heidelberg, 2002, pp. 484–493.