ISSN:0975-9646

Enhanced Big Data Quality Frame Work

G.Mallikajruna Reddy[1], Ganesh Deshmukh[2], R.Arun kumar[3], N.Anil Babu[4]

[1,2,3,4]*Assistant Professor*
Department of CSE,VCE

Abstract: At the Present time the word Big Data content is giving boost everywhere. Big Data refers to huge amount of data such as extreme size like zeta bytes. It is prone to quality problems related to accuracy, precision, completeness, consistency and timeliness. Accuracy in big data may lead to more reliable decision making and better decisions, which can result in greater Economical efficiency, cost reduction, getting down risk and more profits. . Data quality problem is directly proportional to volume of data. For example if data volume increases 10 times then quality issues also measure 10 times more. Big Data software developers must be most sensitive to data quality issues. This paper aims to increase the data quality by taking more number characteristics of big data.

I. INTRODUCTION

As data volume increases linearly, complexity increases geometrically or exponentially. Due to the complexity, there is increase in mistakes and fault rate. Fault is an measurement that affects the data. Cost of finding out and fixing errors increases non linearly. This makes the data management critical. In this process human generated data entry errors are very less. This makes our work more efficient. Noise in the data is also treated as error. This type of error can be reduced by using analytics. Analytics here refer to finding out patterns or identifying trends. Pipelining is a process followed to reduce data quality problems. Pipelining reduces deficiencies and errors in the data. Data should be standardized and then transformed, loaded and is processed. For processing of data, Business Intelligence tools are used. Through research we can manage big data quality problem by exploring. Big data developers should also take into account data quality issues. Data Quality is important in order to have a clear analytics. Data Quality is expressed in dimensions like reliability, timeliness, consistency, dependability, availability relevance to Violating Data Quality rules and data quality issues. There are some situations like Data quality becomes worse, sometimes it becomes better, and sometimes it does not matter. Blue line refers to increase in data quality problems with increase in data .Green Zone illustrates that data quality gets better. Red Zone, as the colour depicts data quality becomes worse. White Zone refers to the situation where data quality does not matter. Data derived from automatic devices will improve data quality Errors interrelated because of which base errors will grow. In addition, cost and expense of tracking and fixing these errors will increase non-linearly. Manual errors are difficult to find and correct. Data from automated device will improve data quality. Data quality errors are consistent in Instrument-generated data. These errors are readily found and mitigated. Some Errors in big data sets can be treated as Noise and Mitigated by Analytics. Unintended Errors can be introduced by pipeline processing in Big Data. Pipeline processing is often intended to compensate for Data Quality Problems. Software industry and Research scholars provide tools for quality and management of Data.

2.ENHANCED BIG DATA QUALITY FRAME WORK

BIG	DATA
Volume **Velocity** **Variety** **Variability** **Veracity[8]** **Value** **Variability** **Venue** **Vocabulary** **Vagueness**	**Format&Structure: Structured, Un structured, Semi-structured** **Ex for Structured :Oracle tables.** **Source for Data: Machine generated, computer generated.** **Ex:Temperature sensor Data** **Ex for UN Structured Data: What'sapp messages**
QUALITY	**PROBLEM**
QBD attributes: accuracy, Timeliness Relavance Completeness, Consistency **QBD Management:** Metrics, Data Quality, Data operation management	Blue line:Data quality problems increase with data linearly Green Zone: QBD gets better Red zone:QBD gets worse White Zone: QBD is not matter

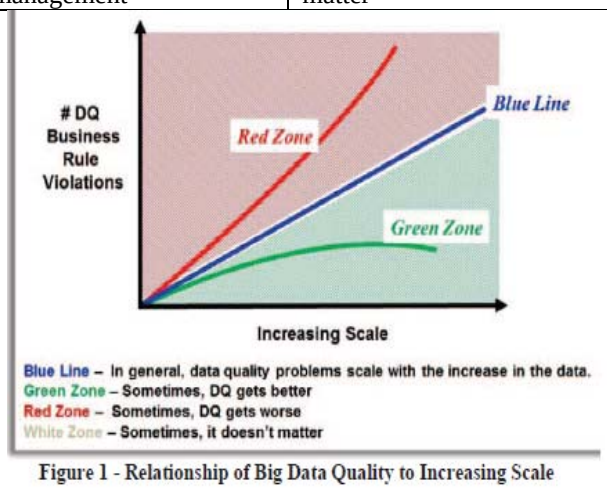

Figure 1 - Relationship of Big Data Quality to Increasing Scale

Blue line: This line depicts Data quality problems increase with data linearly
Green Zone: This line depicts QBD gets better.
Red zone: This line depicts QBD gets worse
White Zone: This line depicts QBD is not matter

3. BIG DATA CHARACTERISTICS

Volume: How much data? It measured in memory like KB,GB..etc
Velocity : Data Processing speed.
Variety: Different formats of data.
Variability: Data meaning is changing.
Veracity:[8] contains too much noise that means data needs to be clean for processing.
Value: Big data value is huge.
Venue: Data is distributed across the network.
Vocabulary: relational schema, big data models, semantics, taxonomy
Vagueness: unclear meaning in big data**. Or** it s like confusion
Visualization: Presenting the data

4. MATHEMATICAL PROOF:

Formula of standard error [10]

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where **σ** is standard deviation. This value is same for the both as data with more characteristics as well as data with less characteristics. Here n is random sample size . This n is considering as number of metrics that are applied on program. In existing paper n value is 3. In this article n value is 10. n value is substitute in the standard error formula, calculate ratio of proposed system standard fault with respect to existing and proposed system standard fault then 1.84 times better with respect to existing system. When standard fault rate is less or equal to better Quality for development, maintenance and understand ability.

5. CONCLUSION

In this Material, Enhanced Quality Frame work for big data is software maintenance and understandability. The framework involves big data characteristics volume ,variety , velocity in different forms and Quality . In the aspect level metrics, different metrics are calculated to determine the data quality of software. Then, in the software environment "If provide more characteristics for big data software then easy maintenance and understandability of Big Data". Hence enhanced big data quality frame work provides less error rate.

REFERENCES:

[1] https://en.wikipedia.org/wiki/Big_data
[2] David Becker, Bill McMullen, Trish Dunn King,The MITRE Corporation ,"BIG DATA, BIG DATA QUALITY PROBLEM" 2015 IEEE International Conference on Big Data (Big Data).
[3] Wikipedia, "Returns to scale." Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc., as modified on 3 October 2013 at 07:17, and retrieved on 5 October 2013.
[4] R.S. Pindyck, D.I. Rubinfeld, Microeconomics, 1989, pages 184-187,Macmillan Publishing Company, New York.
[5] J. Jonas., "Macro Trends in Data & Sense Making", Keynote Speech, Proceedings of 14th International Conference on Information Quality, 2009, (c) IBM Corporation.
[6] F. Chandler, I.A. Heard, M. Presley, "NASA Human Error Analysis", September 17, 2010, National Aeronautics and Space Administration.
[7] R. Panko, "Spreadsheet Research Website", Copyright 1997-2010, RayPanko, http://panko.shidler.hawaii.edu/SSR/ , retrieved on 10March 2014.
[8] http://www.iso.org/iso/big_data_report-jtc1.pdf
[9] http://www.datasciencecentral.com/profiles/blogs/top-10-list-the-v-s-of-big-data
[10] Gurland, J; Tripathi RC (1971). "A simple approximation for unbiased estimation of the standard deviation". *American Statistician* (AmericanStatistical ssociation) 25 (4): 30–32.