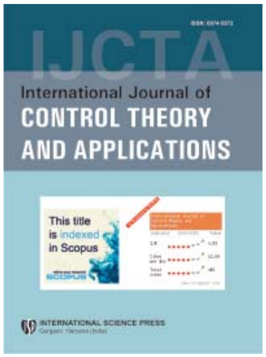

International Journal of Control Theory and Applications

ISSN : 0974–5572

© International Science Press

Volume 10 • Number 23 • 2017

Data in Big Data – A Quality Introspection

K.Vijay Kumar[a] and K. Shyamala[b]

[a] *Research Scholar, Vels University Centre for Development of Advanced Computing, Chennai, India.*

E-mail: vijaydharshan@gmail.com

[b] *Associate Professor, Dr. Ambedkar Govt. College, Chennai, India.*

E-mail: shyamalakannan2000@gmail.com

Abstract: Driven by the shift in technology paradigm, the big data applications revolutionizing IT industry is plagued with poor quality of the collected data. This survey is to consolidate the best practices. The data collected from different sources per se characterized with volume, variety and velocity is not worthy enough to do processing directly. Propelled on these prime variables, we attempted to investigate the subtleties impacting the life cycle of the nature of data administration that needs to experience pre-processing to fuel the analytical engine. Data Governance Model spotlights on data structures, data-dimensions, data measures, data investigation, data control and administration. The comprehensive data management programs can be implemented to improve the data quality solutions. Some of the ideas include using Hadoop functions with no limitations - like ETL process, cleansing, matching, de-duping, merging/purging, tokenizing, standardizing, etc. Pure data quality of YARN is achieved by replacing map-reduce typical programming complexity. Build data lakes on Hadoop repository to streamline the data management in a centralized manner. By stratifying the dimensions and matrices of data, we can achieve qualitative data. Adding to these best practices, we require defining data facts and validity rules to control quality of data from its business applications. An automated robust data quality system can be generated to accomplish the entire data quality processes. The system can be evaluated and visualized using analytical algorithms.

Keywords— Hadoop; YARN; Big Data; Data Quality; Data Management

1. INTRODUCTION

In the Big-Data era, the data quality and its administration has become the prime importance in attaining optimization and efficiency. Loading the relevant data is the fundamental step for getting insights and formulating the predictions through the outcome. So, the data should be of quality for the rest of processing strands. Ensuring the quality of your data also ensures you the quality towards perfection about the information after analytics. Data Quality is very crucial in decision-making and planning. Since the data is accumulated from a number of sources/devices, it is very important to standardize and cleanse before moving to analysis where a number of clustering algorithms can be applied [1]. The state of Data-Quality can be illustrated by the terms like consistency, efficiency; accuracy and accessibility, across different sources of data to deliver reliable data with appropriate representation in the limited time to bring completeness (update status). As [2] states, one of the biggest problems with Big Data is the tendency for errors to snowball. The sanctity of entering, storing and managing the data, does provide its reflection in the state of data quality. Sometimes, quality is also affected by the database being used for its management like relational database, NoSQL database etc. There are different ways of handling variant databases in the pre-processing phase viz., cleansing and application of other analytical algorithms. In Log Analysis, the webserver logs so collected can be optimally analyzed using MongoDB and Java at preprocessing stage [3]. Within an organization, data quality acceptability is important for operational and transactional processes and for the reliability of Business Analytics (BA) / Business Intelligence (BI) reporting. Data Quality Assurance (DQA) is the process of verifying the reliability and effectiveness of data. Challenge to possess Quality data plagued with segmented outlook, missing data with lacunas, obsolete data and inaccurate data requires Data Quality Management. In varied standards and methods managing the Data-Quality is a bottleneck for bringing DQA and is classified as statistical/quantitative vs logical/constraint based.

2. OVERVIEW

2.1. Data Quality Management Cycle

The on-going process of managing the quality of the data is in itself an administration that incorporates the role of the establishment, deployment, policies and practices with regard to the acquisition, maintenance and framing, monitoring, disposition, reporting and checksum for erroneous [4]. The entire procedure refines the data quality and brings it to the environment where it is ready for highlighting the big-info spheres. To achieve such an outcome, a strong bonding between data sources and technical groups is needed. Virtuous data quality cycle incorporates fundamental practices for implementation of core services [5]. The process cycle involves assessment of data quality on the level of impedance in attaining those business objectives. We require bridging the lacunas by framing data quality rules to attain the business performance benchmark. This involves iterative approach of evaluating the design quality to its optimization by continuous inspection till it attains quality check clearance for production. The Quality Management Life Cycle so depicted in the figure 1 requires being in place in every data incubator of the organization.

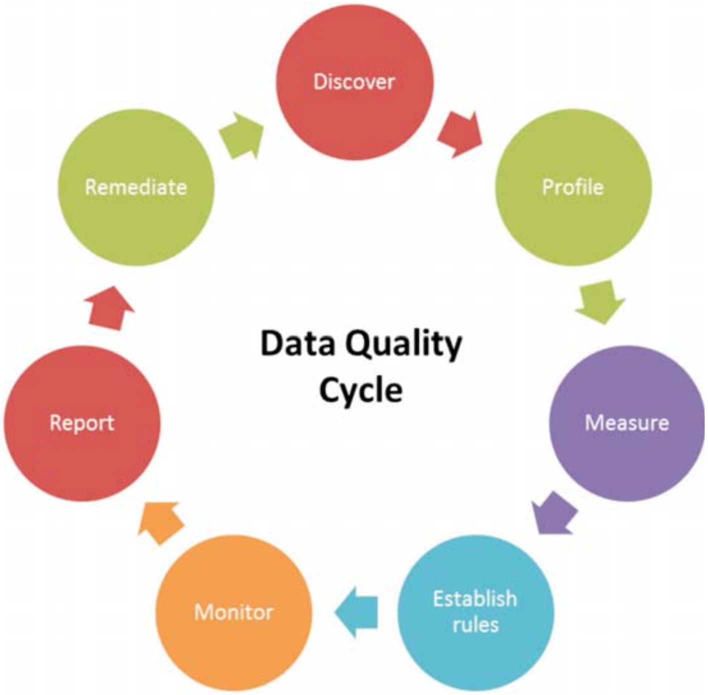

Figure 1: Data Quality Management Life Cycle

2.2. Issues in Managing Data Quality

There are several issues related to quality of data management, some of them are mentioned here [6][7][8].

1. **To identify the inaccurate data (Manual Wrong Entry):** Mainly, the data profiling process produces the inaccurate facts that identify some specific instances of wrong values and in some cases it also identifies the existence of wrong value but it fails to locate which value is wrong.
2. **Integrating data from different sources (System Consolidation) :** In bringing the data from various disparate sources, there comes the problem of common storage and cleansing. The consolidated data are the collection of heterogeneous applications and environments that's very complex to handle as it introduces a lot of errors while extraction, transformation and loading.
3. **Real -Time Interfaces :** Data is likely to be out-of-sync, when trigger procedure downstream the data volume from one database to another. The data propagation is done at a very high rate, so there is no time to verify the accuracy.
4. **Inconsistency in data :** When the data is consolidated into single database storage, the chances of data inconsistency goes up to 90%. High volume of data traffic leads to this problem.
5. **Unreliability and data-loss :** Similarly, when the data is being synchronized, some of the data may be lost and become unreliable for analytics. This results in poor data quality for later stages.
6. **Noisy Data :** Along with the possibilities of data-loss, the probability of inclusion of unwanted and non-relevant data also becomes high that will be the obstacle in finding the models for giving analytical forecasting.
7. **Data Cleansing :** Sometimes, the data cleansing also becomes an issue in terms of quality of data. As the data quality issues are intricate and interrelated, the data-cleansing steps for some part of the data may cleanse it, but may also create problem for other parts of the data.

2.3. Solutions for maintaining Data Quality

Data quality and its administration are totally relying upon the nature of the datasets that creates the assets that are definitive objectives of Business applications. Data and process quality can be improved by acknowledging the correlations between the strong or weak data dimensions. There are some fundamental dimensions viz., accuracy, consistency, totality and timeliness [9]. Keeping up high caliber of information and giving best quality of data and examination can depend on Data-Governance Model that builds up the parts and accountabilities of data overseeing staff, required in Business's customer support. Data Governance Model spotlights on data structures, data-dimensions, data measures, data investigation, data control and administration [10].

The comprehensive data management programs can be implemented to improve the data quality solutions [11]. Some of the ideas in these programs can include

1. **Specific Hadoop Functions :** By using some functions in Hadoop with no limitations - like ETL process, cleansing, matching, de-duping, merging/purging, tokenizing, standardizing, etc.
2. **YARN data qualities :** Can be used in place of map-reduce typical programming complexity like redpoint [12] which is a pure-YARN data quality.
3. **No migration of Data :** Rather to migrate the data in different environment, manage data in traditional Hadoop Repository that will build a data-lake or centralized data storage.
4. **Stratify the data dimensions and data matrices :** According to the data that complies with completeness, timeliness, consistency, and uniqueness. The results of this process will be the collection of measures that collectively contributes to qualitative data.
5. **Defining Data Facts and validity rules** are directly integrated to the Business applications to control and verify that the data quality expectations are up to the mark or not in the information flow.

3. FUTURE SCOPE

In this survey paper, we tried to deliberate the industry standard prevailing issues and solutions in data incubators. An Automated Robust Data Quality System can be generated that can solemnly accomplish the entire data quality processes including data-profiling, data-conversion, data-cleansing, data-purging etc. and the overall result can be analyzed using analytical algorithms and visualized through the same system [13]. The Data Quality System will encapsulate provision for evaluation of the qualitative data using analytical algorithms on dynamic selection of the criteria to build domain specific policy providing weightage to Usability, Accessibility and Timeliness of data life cycle [14][15].

REFERENCES

[1] T. Sajana, C. M. Sheela Rani, K. V. Narayana, “A Survey on Clustering Techniques for Big Data Mining,” *Indian Journal of Science and Technology*. vol 9(3). Jan 2016.

[2] Barna Saha, Divesh Srivastava, “Data Quality: The other Face of Big Data,” *IEEE 30th International Conference on Data Engineering (ICDE)*, 2014.

[3] P. Parthiban, S. Selvakumar , “Big Data Architecture for Capturing, Storing, Analyzing and Visualizing of Web Server Logs,” *Indian Journal of Science and Technology*. vol 9(4). Jan 2016.

[4] (2016) Building successful data quality management program, [Online], Available: http://knowledgent.com/whitepaper/building-successful-data-quality-management-program.

[5] David Loshin, “White Paper: Data Quality & Data Integration,” *Five Fundamental Data Quality Practices*, *Pitney Bowes Business Insight*. pp. 1-12.

[6] (2016) Thirteen causes of enterprise data quality problems, [Online], Available: http://searchdatamanagement.techtarget.com/ feature/Thirteen-causes-of-enterprise-data-quality-problems.

[7] (2016) How to maintain data quality and provide high quality information management and analysis, [Online], Available: http://searchdatamanagement.techtarget.com/answer/How-to-maintain-data-quality-and-provide-high-quality-information-management-and-analysis.

[8] (2016) Data Quality Management, [Online], Available: https://www.talend.com/resource/data-quality-management.html

[9] Payam Hassany Shariat Panahy, Fatimah Sidi, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, Aida Mustapha, “A Framework to Construct Data Quality Dimensions Relationships,” *Indian Journal of Science and Technology*, vol 6(5), May 2013.

[10] Fan W, Geerts F. “Foundations of data quality management,” *Synthesis Lectures on Data Management*, vol 4(5). pp 1–217. 2012.

[11] Jonathan G. Geiger, “Data Quality Management, The Most Critical Initiative You Can Implement,” Intelligent Solutions, Inc., Boulder, CO, Paper 098- 29.

[12] (2016) Big Data Quality Integration, [Online], Available: http://www.redpoint.net/data-marketing-solutions/enterprise-big-data/big-data-quality-integration.

[13] Sang Gi Lee, Byeonghee Lee, Hanjo Jeong, “A Study on the Problem Analysis and Improvement Plan of the Data Quality Management System of National R&D Data,” *Indian Journal of Science and Technology*. vol 8(23). Sep 2015.

[14] Guma Abdulkhader Lakshen, Sanja Vraneš, “Big Data and Quality: A Literature Review,” *IEEE, 24th Telecommunications forum (TELFOR)*. pp 22-23. Nov 2016.

[15] Ikbal Taleb et al., “Big Data Quality: A Quality Dimensions Evaluation,” International IEEE Conferences on Ubiquitous Intelligence & Computing. July 2016.