iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VII **Month of publication:** July 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue VII, July 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Study on Metadata Management and Quality Evaluation in Big Data Management

Anuja Kulkarni
Computer Engineering Department, Bharati Vidyapeeth Demeed University-Pune

Abstract*— In today's world huge amount of data is being continuously generated through all streams of life using different devices and systems. Such data is called as Big Data. It is important to maintain "4V" of data (i.e., volume, velocity, variety, and veracity) while capturing, storing and analysing it in a systematic manner. The processing of metadata highly influences the processing of big data, but still within the scope of big data project, metadata is often taken for granted. The main aim of this paper is to highlight why metadata is vital for big data processing and to understand how metadata is managed, and quality of data is evaluated using quality attributes and quality policies.

***Keywords*— Big Data, Metadata, Quality, Metadata Management, Storage Infrastructure, Standards**

I. INTRODUCTION

The term Big Data becomes apparent, as the amount of data is enormous, it means it is difficult to process, a massive volume of structured and unstructured data using traditional database and software techniques. While processing the data using traditional approaches, metadata is not considered as a vital priority. But with the increasing commencement of big data, the value of metadata is considered as a critical priority for the success of big data. [1]

The platforms of big data, such as Hadoop are 'schema-less' i.e.; there is no any accurate description of what this data truly 'is'. While launching any new data project, it is necessary to identify, metadata of big data, i.e., what exactly the big data is truly about with an accurate and descriptive understanding. It will cause a wide range of potential issues and become somewhat challenging if metadata is not identified at project startup. For developing and maturing big data processing services, it is necessary to establish a comprehensive enterprise metadata management program. And hence the importance of metadata processing for big data cannot be understated.

According to the report published by IDC and sponsored by EMC, enterprise data management has various sub-segments, out of which metadata is one of the fastest growing sub-segment. The problem encountered in this report is that, while metadata is growing, it is not keeping rapidity with the rapid increase of big data projects being currently initiated by firms. This problem is referred as 'big data gap' by IDC. Big data management process for collecting, integrating and analyzing the data can be significantly rationalized and enhanced by the use of metadata.

The quality and trustworthiness of the data are crucial issues in big data management because a lot of freely accessible data originates from indeterminate sources and it the data collected from these sources is commonly unstructured or semi-structured.

The quality evaluation of data can be done in one or more data processing phases of big data architecture, i.e., data extraction, data processing, data analysis and decision-making phase. It is highly risky for company's business to use unreliable data such as inaccurate or incomplete data, corrupted data as it may lead to poor or incorrect business decisions.

The purpose of this paper is to highlight the importance of metadata while processing the big data and understanding the concept of quality metadata management to ensure the quality and trustworthiness of data using quality attributes and quality policies. This paper is organized according to following sections-Section II defines what is metadata and metadata standards, the relation between big data and metadata, metadata management, quality metadata and how quality metadata can be created in data extraction phase of big data architecture. Section III provides future scope and Section IV concludes the work.

II. BACKGROUND

A. Metadata and Metadata Standards

Metadata can be simply defined as 'information about data within any data environment' which is used to describe the properties of data such as provenance, quality, and technical details. It can be used by end users to ensure the quality of data and to identify its value for business usage. Classification of metadata can be done in three categories: descriptive, structural and administrative metadata. The goal of descriptive metadata is to describe and identify information resources. Structural metadata provides the facility for navigation and presentation of electronic resources. Administrative metadata gives administrative privileges like

providing the information such as, when was a resource created, how it was created and by whom a resource was created. According to data-user metadata taxonomy, there are four classes of metadata: data quality metadata, definitional metadata, navigational metadata and lineage metadata. Data quality metadata is useful for describing the quality of data when it is used for specific purpose. To understand the meaning of data from a business viewpoint definitional metadata can be used. Users can find the desired data for a specific purpose using navigational metadata. To find out the source of data and the actions of the data lineage metadata can be used.

B. Big Data Metadata

The data repositories like data lake, data warehouse or a normalized relational database are structured. The data representation in these repositories is in rows and column form, and the metadata model is 'native' of the structure. Such a 'native' metadata model is not available for big data because metadata from new external data sources is essential to unlocking new meaning. For constructing the beginnings of these new metadata definitions, big data will require processing through precise analytics. These new metadata definitions are then correlated with metadata defined from other structured data sources to provide comprehensive metadata model for the entire enterprise.

Metadata allow to associate similar data assets and disassociate different data assets of various big data sources. This potential of metadata can be used to dismiss irrelevant information during the search process, which is beneficial for searching algorithms to create high confidence results. Metadata can be used by big data and analytics users to locate the right information quickly despite the large amount of content residing in these repositories. Metadata also creates and maintains data consistency. Using metadata, organizations can define a consistent definition or business rule for particular data attribute. These rules are further applied across the enterprise data level, for both structured and unstructured data stores.

C. Metadata Management

Metadata management is an important capability that enables managing the change in data while delivering trusted, secure data in the complex big data environment. Metadata management handles two core functionalities, i.e., metadata management and quality management. These management functions are performed using one dataset called as metadata. Metadata dataset acts as a storage unit for metadata, and it is also used for organizing and managing the metadata. Metadata management allows extraction of metadata and provides access to metadata. The task of quality management is to assign values to quality attributes. These values are based on the properties of data sets and associated metadata.

D. Quality Attributes, Metrics, And Policies

The quality attribute can be considered as a single construct of quality. Properties of the quality attribute can be measured using quality metric. Data quality attributes have different dimensions that are important to data consumers like intrinsic dimension, representational dimension, accessibility dimension and contextual dimension.

Quality metric can be thought of as a measure of certain properties of the quality attribute to evaluate the degree of presence of that quality attribute. It can be categorized as content-based metrics, context-based metrics, and rating-based metrics.

The quality policy is required to find out relevant quality attributes in the context of the particular task and to provide the quality metric that should be used for the evaluations of defined quality attributes.

E. Quality Metadata

Quality metadata describes the quality of data using different quality attributes and metrics for each quality attribute that can be used for a specific purpose. Table 1 describes the quality attributes.

TABLE I
QUALITY ATTRIBUTES

Quality Attribute	Description
Accuracy	Accuracy defines the degree of correctness and precision.
Believability	Trustworthiness
Completeness	Completeness ensures the complete availability of information in terms of breadth, depth and scope.
Consistency	Consistency validates the conflict between two or more values with each other.

Validity	Validity is used to check whether the information is valid in certain situation or not.
Relevance	It is used to check whether the collected information is relevant or applicable to the given task.
Timeliness	The Timestamp is important for extracted, processed and analysed data sets to check the freshness of data.
Verifiability	It can be used to check the correctness of data.

F. Quality Metadata Management

Quality metadata managed by defining the rules to identify which quality attributes and metrics can be used and when. These rules can follow the simple if-then-else structure or can use some rule language. The rules are defined according to company's quality policy(organizational policy and decision-making policy).

G. Data Quality Evaluation

End-users can evaluate the quality of data at different phases of pipelined architecture of big-data.(Data extraction, data processing, data analysis, decision making). Quality evaluation can be done using following steps:

1) Understanding the purpose of data collection.
2) Applying organizational and decision-making policies.
3) Searching for data using the keyword.
4) Receiving visual results of evaluated data.
5) Making business decisions accordingly.

H. Creation of Quality Metadata in Data Extraction Phase

Quality policies always facilitate the creation of quality metadata in different phases of big data architecture. In data extraction phase, the organizational policy facilitates the quality metadata creation process by defining the data sources that are acceptable and valid quality attribute selection. It also facilitates the process by providing metrics and methods to evaluate the quality attributes. If the data source type of data set is known, then the applicable attributes can be automatically provided. After getting the acceptable quality attributes, they are evaluated using qualitative evaluation. Once the quality attributes get evaluated, then the imported data is stored in data storage. After evaluation, the quality metadata is created for extracted data set, and the evaluated values for quality attributes are inserted into the metadata.

Same steps can be followed for the creation of quality metadata in data processing, analysis, and decision-making phases. In decision-making phase, decision-making policy helps in selection of relevant data for decision-making purpose.

III.FUTURE SCOPE AND CONCLUSIONS

A. Implementation of New Types of Metadata

In future, big data utilization is going to increase. Therefore new kinds of metadata must be defined to meet the special requirement of different and growing market segments provisioning big data.

B. Implementation of Several Quality Policies

Currently, organizational quality policies are associated with very few attributes and DataSourceType. It is the duty of Quality Policy Manager to initialise organizational quality policies. In the future, more organizational quality policies could be defined for different DataSourceType.

C. Evaluation of More Quality Attributes

Currently, there are very few quality attributes are used for quality evaluation. New quality attributes could be defined by developing algorithms to improve the utility of solution. This paper concludes a study of the importance of the most critical component of any robust data governance practice, i.e., metadata management and quality evaluation at different phases of big data architecture using quality metadata concept and quality attributes and policies.

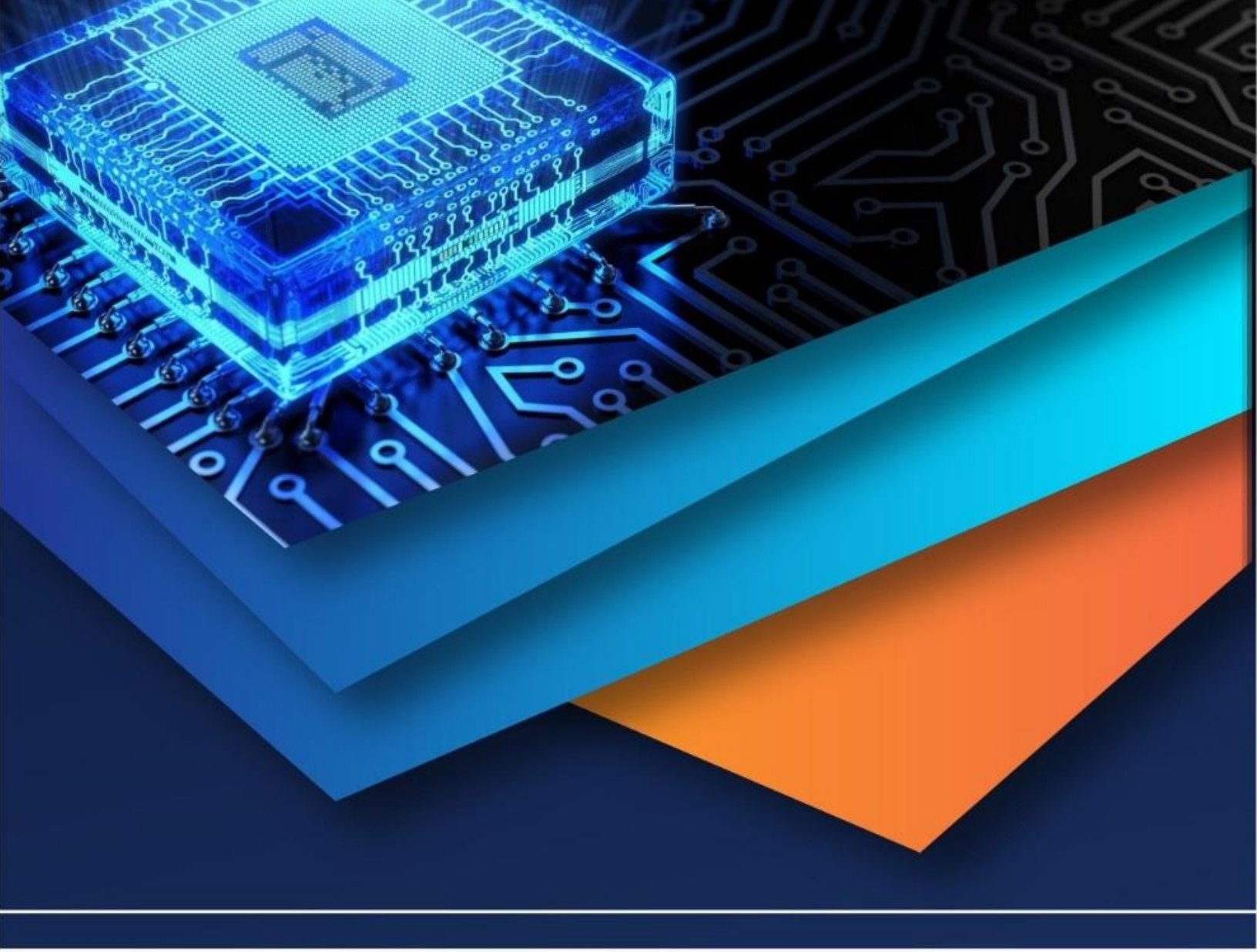
Stewarding metadata is necessary for the implementation of enterprise data governance practice because it provides the value and

the context for understanding the data and its components to the end users. Using metadata concept, organizations can define a consistent definition or business rule for specific data attribute. These rules are further applied across the enterprise data level, for both structured and unstructured data stores.

REFERENCES

[1] Dhawan, U., Vasilakis, N., Rubin, R., Chiricescu, S., Smith, J.M., Knight Jr, T.F., Pierce, B.C. and DeHon, A., 2014, June. Pump: a programmable unit for metadata processing. In Proceedings of the Third Workshop on Hardware and Architectural Support for Security and Privacy (p. 8). ACM.
[2] Dhawan, U., Hritcu, C., Rubin, R., Vasilakis, N., Chiricescu, S., Smith, J.M., Knight Jr, T.F., Pierce, B.C. and DeHon, A., 2015. Architectural support for software-defined metadata processing. ACM SIGARCH Computer Architecture News, 43(1), pp.487-502.
[3] Nevěřilová, Z., 2010. Metadata Processing.
[4] Wang, T., Wang, J., Yu, Y., Shen, R., Liu, J. and Chen, H., 2004, November. Metadata pro: Ontology-based metadata processing for web resources. In International Conference on Web Information Systems Engineering (pp. 34-45). Springer Berlin Heidelberg
[5] Raval, K.S., Suryawanshi, R.S., Naveenkumar, J. and Thakore, D.M., 2011. The Anatomy of a Small-Scale Document Search Engine Tool: Incorporating a new Ranking Algorithm. International Journal of Engineering Science and Technology, 1(3), pp.5802-5808.
[6] Archana, R.C., Naveenkumar, J. and Patil, S.H., 2011. Iris Image Pre-Processing and Minutiae Points Extraction. International Journal of Computer Science and Information Security, 9(6), p.171.
[7] Jayakumar, M.N., Zaeimfar, M.F., Joshi, M.M. and Joshi, S.D., 2014. INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET). Journal Impact Factor, 5(1), pp.46-51.
[8] Naveenkumar, J. and Joshi, S.D., 2015. Evaluation of Active Storage System Realized through MobilityRPC.
[9] Jayakumar, D.T. and Naveenkumar, R., 2012. SDjoshi, ". International Journal of Advanced Research in Computer Science and Software Engineering," Int. J, 2(9), pp.62-70.
[10] Jayakumar, N., Singh, S., Patil, S.H. and Joshi, S.D., Evaluation Parameters of Infrastructure Resources Required for Integrating Parallel Computing Algorithm and Distributed File System.
[11] Jayakumar, N., Bhardwaj, T., Pant, K., Joshi, S.D. and Patil, S.H., A Holistic Approach for Performance Analysis of Embedded Storage Array.
[12] Naveenkumar, J., Makwana, R., Joshi, S.D. and Thakore, D.M., 2015. OFFLOADING COMPRESSION AND DECOMPRESSION LOGIC CLOSER TO VIDEO FILES USING REMOTE PROCEDURE CALL. Journal Impact Factor, 6(3), pp.37-45.
[13] Naveenkumar, J., Makwana, R., Joshi, S.D. and Thakore, D.M., 2015. Performance Impact Analysis of Application Implemented on Active Storage Framework. International Journal, 5(2).
[14] Salunkhe, R., Kadam, A.D., Jayakumar, N. and Thakore, D., In Search of a Scalable File System State-of-the-art File Systems Review and Map view of new Scalable File system.
[15] Salunkhe, R., Kadam, A.D., Jayakumar, N. and Joshi, S., Luster A Scalable Architecture File System: A Research Implementation on Active Storage Array Framework with Luster file System.
[16] Jayakumar, N., Reducts and Discretization Concepts, tools for Predicting Student's Performance.
[17] Jayakumar, M.N., Zaeimfar, M.F., Joshi, M.M. and Joshi, S.D., 2014. INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET). Journal Impact Factor, 5(1), pp.46-51.
[18] Kumar, N., Angral, S. and Sharma, R., 2014. Integrating Intrusion Detection System with Network Monitoring. International Journal of Scientific and Research Publications, 4, pp.1-4.
[19] Namdeo, J. and Jayakumar, N., 2014. Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts. International Journal, 2(2).
[20] Naveenkumar, J., Keyword Extraction through Applying Rules of Association and Threshold Values. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), ISSN, pp.2278-1021.
[21] Kakamanshadi, G., Naveenkumar, J. and Patil, S.H., 2011. A Method to Find Shortest Reliable Path by Hardware Testing and Software Implementation. International Journal of Engineering Science and Technology (IJEST), ISSN, pp.0975-5462.
[22] Naveenkumar, J. and Raval, K.S., Clouds Explained Using Use-Case Scenarios.
[23] Naveenkumar J, S.D.J., 2015. Evaluation of Active Storage System Realized Through Hadoop. International Journal of Computer Science and Mobile Computing, 4(12), pp.67–73.
[24] Rishikesh Salunkhe, N.J., 2016. Query Bound Application Offloading: Approach Towards Increase Performance of Big Data Computing. Journal of Emerging Technologies and Innovative Research, 3(6), pp.188–191.
[25] Sagar S lad s d joshi, N.J., 2015. Comparison study on Hadoop's HDFS with Lustre File System. International Journal of Scientific Engineering and Applied Science, 1(8), pp.491–494.
[26] Salunkhe, R. et al., 2015. In Search of a Scalable File System State-of-the-art File Systems Review and Map view of new Scalable File system. In nternational Conference on electrical, Electronics, and Optimization Techni ques (ICEEOT) - 2016. pp. 1–8.
[27] BVDUCOE, B.B., 2011. Iris Image Pre-Processing and Minutiae Points Extraction. International Journal of Computer Science & Information Security.
[28] P. D. S. D. J. Naveenkumar J, "Evaluation of Active Storage System Realized through MobilityRPC," Int. J. Innov. Res. Comput. Commun. Eng., vol. 3, no. 11, pp. 11329–11335, 2015
[29] N. Jayakumar, S. Singh, S. H. Patil, and S. D. Joshi, "Evaluation Parameters of Infrastructure Resources Required for Integrating Parallel Computing Algorithm and Distributed File System," IJSTE, vol. 1, no. 12, pp. 251–254, 2015.
[30] N. Jayakumar, T. Bhardwaj, K. Pant, S. D. Joshi, and S. H. Patil, "A Holistic Approach for Performance Analysis of Embedded Storage Array," Int. J. Sci. Technol. Eng., vol. 1, no. 12, pp. 247–250, 2015.
[31] J. Naveenkumar, R. Makwana, S. D. Joshi, and D. M. Thakore, "Performance Impact Analysis of Application Implemented on Active Storage Framework," Int. J., vol. 5, no. 2, 2015.

[32] N. Jayakumar, "Reducts and Discretization Concepts, tools for Predicting Student's Performance," Int. J. Eng. Sci. Innov. Technol., vol. 3, no. 2, pp. 7–15, 2014.
[33] J. Namdeo and N. Jayakumar, "Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts," Int. J. Adv. Res. Comput. Sci. Manag. Stud., vol. 2, no. 2, 2014.
[34] R. Salunkhe, A. D. Kadam, N. Jayakumar, and S. Joshi, "Luster A Scalable Architecture File System: A Research Implementation on Active Storage Array Framework with Luster file System.," in ICEEOT, 2015.
[35] Anne Immonen, Pekka Pääkkönen, And Eila Ovaska ," Evaluating the Quality of Social Media Data in Big Data Architecture" vol 3, IEEE Access, pp.2028-2043,2015



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY