IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-9, Sep 2018 E-ISSN: 2347-2693

Semantic Web Approach of Integrating Big Data- A Review

Jeelani Ahmed[1*], Muqeem Ahmed[2]

[1, 2] Department of CS&IT, Maulana Azad National Urdu University, Hyderabad, India

[*]*Corresponding Author: Jeelani.jk@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 25/Sept/2018, Published: 30/Sept/2018

***Abstract*—** Semantic Web is spread by World Wide Web Consortium (W3C) and international standardization body of the web. It is an extended form of the current web which provides an easier way to search, reuse, combine and share the information. Therefore Semantic Web is subsequently viewed as an integrator crosswise over various content, data applications, and frameworks. Today's big data is usually pronounced as consisting of 3 V's: volume, variety, and velocity. Variety of data discusses to deal with different formats of data and a large number of various data sources. Thus the problems of big data variety are important for solving many real-world difficulties. Semantic Web is used as an integrator to incorporate data from various kinds of sources like web services, relational databases, and spreadsheets etc. and in different formats. Due to the presence of data heterogeneity, this work presents various difficulties that may not be totally settled with the existing system. This paper is an attempt to focus on the various challenges that involved in integrating data from different types of sources and how different semantic web technologies and tools are used for the integration of disparate data.

***Keywords*—**Semantic Web, Big Data, Disparate Data

I. INTRODUCTION

Big data is defined as consisting of main 3v's namely volume, variety, and velocity. Volume alludes to the issue of how to manage vast informational collections, which normally requires execution in a conveyed cloud-based foundation. Velocity alludes to managing continuous gushing information, for example, video sustains, where it might be difficult to store all information for later preparing. Variety alludes to managing diverse sorts of sources, distinctive arrangements of the information, and expansive quantities of sources [14]. A significant part of the work on enormous information has concentrated on volume and speed; however, the issues of variety are similarly vital in taking care of numerous true issues [1].

The variety related with huge information prompts challenges in data integration. Huge information originates from a variety of spots online networking streams, the worker made records, email frameworks, enterprise applications and so forth. Combining every one of that information and accommodating it with the goal that it can be utilized to make reports can be extraordinarily troublesome. Sellers offer an assortment of ETL and data integration tools intended to make the procedure less demanding, yet numerous ventures say that they have not tackled the data integration issue yet [2].

The process of allowing users to gain access to deliver and utilize data across whole organizations while preserving its integrity and quality is known as data integration. It additionally empowers changes made to information put away in one source to be reflected in different sources continuously [2]. Big Data can be gathered from multiple types of resources such as spreadsheets, web services relational databases and more and in broadly different forms including both logical and hierarchical data [1].

Big data is made conceivable by data integration [10]. By empowering access to information put away in divergent information stockrooms, mapping changes starting with one endeavor application then onto the next, and conveying real-time data to proposed clients, data integration empowers enterprises to gather and clean huge information originating from different frameworks for analysis.

Beside huge information examination, information joining likewise empowers different business advantages, for example, having a 360-degree perspective of datasets, quicker coordinated effort crosswise over whole organizations, and more noteworthy adaptability in choosing endeavor applications and frameworks and streamlining and mechanizing business functions [3].

This paper is an attempt to revisit and explores the challenges that may arise in the integration of big data. Section II

explores the different challenges and Section III contains the introduction of the semantic web and semantic solution to address the challenges and section IV concludes the work with the future research directions.

II. CHALLENEGS OF INTEGRATING BIGDATA

Big Data is a wide term for far-reaching and complex data sets where regular data planning applications are inadequate. The integration of this gigantic informational index is very intricate. There are a few difficulties one can look amid this reconciliation, for example, information generation, analysis, catch, search, sharing, representation, data storage, and privacy.

A. Data Management

Data Management is one of the most important tasks of integrating big data. When data is in huge volume then management of this data becomes difficult. To avoid these difficulties NoSQL [5][14] databases can be used. These Databases are contemporary to the traditional relational databases and provide great execution of different big data applications. These Databases works on the concept of the key-value pair [5] [14] so that it can deal with a huge amount of information with faster response.

B. Bad Data

In any data integration system, data quality is the biggest problem to worry about. Inheritance data must be cleaned up going before the change and joining, or organizations will almost certainly face honestly to goodness data issues later. Legacy information pollutions have an intensifying impact; by nature, they tend to focus on high volume information clients. On the off chance that this data is degenerate, along these lines, as well, will be the choices made by it. It isn't irregular for unfamiliar information quality issues to develop during the time spent cleaning data for use by the integrated framework [7].

C. Lack of Skills

Due to the coming of new technologies in a day to day market customers are moving from old traditional relational databases to new information processing system like NoSQL Databases, in-memory analytics, and Hadoop etc. Actually, in the market, there is the absence of required skills for big data innovations [2]. There is an average number of masters are available in the markets to work on these systems that process the huge amount of information.

D. Transmission of Data into Big Data Structure

There are various people who have raised wants thinking about separating monstrous data accumulations for a noteworthy data platform. They moreover may not think about the versatile quality behind the entrance, transmission, and movement of data and information from a broad assortment of advantages and after that stacking this data in a major information stage. The marvelous parts of data transmission, get to and stacking is simply bits of the test.

E. Extracting Information

The most convenient use cases for huge data incorporate the availability of data, expanding existing accumulating of data and furthermore empowering access to end-client using business knowledge instruments with the true objective of the revelation of data. It transforms into a test in enormous data coordination to ensure the right-time data availability to the data clients.

F. Synchronization Data Sources

At the point when information is import into enormous information stages, we may in like manner comprehend that information duplicated moved from a broad assortment of sources on different rates and timetables can rapidly get away from the synchronization with the starting structure. This gathers the data starting from one source isn't obsolete when stood out from the data beginning from another source. It moreover infers the mutual quality of data definitions, thoughts, and metadata. The ordinary data administration and data circulation focus, the gathering of data change, extraction and movements all develop the situation in which there are perils for data to end up unsynchronized.

G. Other Challenges

The huge amount of data, arrangement cost, reliability, the correctness of information and rate of change of data these are the other difficulties could ascend during data integration. Actually saying it is not less than a test to practice the huge amount of data with the practical quickness with an important objective of provides the data to the necessary customer at the time of requirement. The endorsement of data accumulation is moreover fulfilled while trading data beginning with one source then onto the following or to purchasers as well.

III. SEMANTIC WEB

Data frameworks today need to determine the heterogeneity among information living in different self-ruling information sources. Specifically, the utilization of the World Wide Web as an all-inclusive medium for trading data has fundamentally changed our vision of information access and control [11]. Following the viewpoint of the semantic web, we trust that the unequivocal introduction of information semantics will encourage information interoperation in a variety of information control tasks.

Semantic web goes for specifically giving machine-understandable information on the Web. It consists of two remarkable ingredients: web ontologies [7]; data annotation. In spite of the fact that it has a long approach for the acknowledgment of the semantic web including the development of an ontology, inference engine and different segments, the rule of the semantic web advises us that information semantics is to set up and keep up the correspondence from information to the subject is planned.

A. Semantic Data Integration

Semantic information integration is the way toward joining information from divergent sources and merging it into important and significant data, however, the utilization of semantic technology. As organizations grow up an estimate, so does their information. Without the correct information management system, intradepartmental or application specific information storehouses rapidly emerge and impede efficiency and participation.

Semantic Data Integration provides an answer that goes past the standard endeavor application integration arrangements. It utilizes information-driven engineering based upon an institutionalized model for information distributing and exchange, specifically the Resource Description Framework (RDF) [9]. In this system, the heterogeneous information of an organization such as structured, semi-organized and unstructured is communicated [1], put away and got to the similarly. As the information structure is communicated through the links inside the information itself, it isn't compelled to a structure formed by the database and does not wind up out of date with the advancement of the information. Following are some important steps to perform the semantic data integration.

1. Creating an Application Profile (RDF Shape) that depicts the coveted type of the last dataset;
2. Reusing existing ontologies and building new ontologies as required;
3. Leveraging completely the accessible Linked Open Datasets in the area;
4. Designing a basic, sensible and practical URL methodology;
5. Using the assortment of accessible transformation and ETL tools to play out the integration.

To go easily through a full semantic information integration lifecycle, organizations require an arrangement of simple to utilize semantic integration devices. Using semantic integration tools, clients can rapidly plan information handling jobs and integrate a gigantic volume of information.

IV. CONCLUSION

Today's Big Data is very diverse in nature that generated from disparate sources that comprise structured, semi-structured and unstructured. Many organizations is interested to integrate this diverse data in order to perform important operational and analytical functions on this data But one may need to remember that this integration may face many challenges during the whole process. This review paper highlights some of those challenges and discusses the role of the semantic web to address those issues using semantic data integration. Future work of this paper will be proposing a new model to integrate the disparate big data using semantic web technologies.

REFERENCES

[1] C. A. Knoblock and P. Szekely, "Exploiting semantics for big data integration", AI Magazine, Vol.6, Issue.1, pp. 25-38, 2015.

[2] L. Ragusa, "Data Integration Tools for Overcoming Integration Challenges in 2017", https://www.liaison.com

[3] C. Harvey, "Big Data Challenges", https://www.datamation.com/bigdata

[4] J. Bhogal, I. Choksi, "Handling Big Data using NoSQL", In the Proceeding of the 2015 IEEE Conference on Advanced Information Networking and Applications Workshop (WAINA), pp. 393-398,2015.

[5] J. Pokorny, "NoSQL databases: a step to database scalability in web environment", International Journal of Web Information Systems, Vol. 9 Issue: 1,pp.69-82, 2013

[6] S. Siwoon, G. Myeong-Seon, and Y.S. Moon, "Anomaly detection for big log data using a Hadoop ecosystem", In the Proceeding of the 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), pp.377-380, 2017

[7] S. Kumar, V. Singh, and B. Saini, "A survey on ontology matching techniques", In the Proceeding of 2014 International Conference on Computer and Communication Technology (ICCCT) , pp. 13-15, 2014.

[8] R. Hammami, H. Bellaaj, and A.H. Kacem, "Interoperability of healthcare information systems", In the Proceeding of the 2014 International Symposium on Networks, Computers and Communications, pp. 1-5, 2014

[9] A. Cuadra, M. M. Cutanda, D. Fuentes-Lorenzo and L. Sánchez, "A semantic web-based integration framework," In the Proceeding of the 2011 International Conference on Next Generation Web Services Practices, pp. 93-98, 2011.

[10] Y. Shi, X. Liu, Y. Xu and ZhenyanJi, "Semantic-based data integration model applied to heterogeneous medical information system", In the Proceedign of the 2010 International Conference on Computer and Automation Engineering (ICCAE), pp. 624-628, 2010.

[11] S.K. Bansal and S. Kagemann, "Integrating Big Data: A Semantic Extract-Transform-Load Framework" IEEE Computer Society,Volume 48, Issue 3, 42-50, 2015.

[12] I. Merelli, H. Pérez-Sánchez, S. Gesing and D. D'Agostino, "Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives", BioMed Research International, Volume 2014, pp.1-13, 2014.

[13] D. Ostrowski, N. Rychtyckyj, P. MacNeille and M. Kim,"Integration of Big Data Using Semantic Web Technologies" In the Proceeding of IEEE International Conference on Semantic Computing, pp.382-385, 2016.

[14] J Ahmed, R Gulmeher "Nosql databases: New trend of databases,emerging reasons, classification and security issues" International journal of engineering sciences & research technology, Volume 4, Special issue 6, June 2015.

[15] Oluigbo Ikenna V., Nwokonkwo Obi C., Ezeh Gloria N., Ndukwe Ngoziobasi G., "Revolutionizing the Healthcare Industry in Nigeria: The Role of Internet of Things and Big Data Analytics", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.6, pp.1-12, 2017

[16] Anitya Kumar Gupta, Srishti Gupta, "Security Issues in Big Data with Cloud Computing", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.6, pp.27-32, 2017

[17] Loukia Karanikola, Isambo Karali and Sally McClean (2014), "Uncertainty reasoning for the Big Data Semantic Web", In the Proceeding of IEEE 15th International Conference on Information Reuse and Integration, 147 – 154, 2014

[18] J.F. Sánchez-Rada, M. Torres, C.A. Iglesias, R. Maestre, and E. Peinado, "A Linked Data approach to sentiment and emotion analysis of twitter in the financial domain", In the Proceedings of the Second International Workshop on Finance and Economics on the Semantic Web (FEOSW 2014), Anissaras, Crete, Greece. 26th May 2014, pp.51-62, 2014

[19] Zhang, J., & Huang, M. L. "5Ws model for big data analysis and visualization", In the Proceeding of IEEE International Conference Computational Science and Engineering, pp. 1021-1028, 2013

[20] D. Jeon, and W. Kim, "Development of semantic decision tree," In the Proceedings of 3rd International Conference on Data Mining and Intelligent Information Technology Applications (ICMiA), pp. 28-34, 2011

[21] R.T. Bedeley, and L.S. Iyer, "Big Data opportunities and challenges: the case of banking industry", In the Proceedings Southern Association for Information Systems Conference (SAIS), Macon, pp. 1-6, 2014

Authors Profile

Jeelani Ahmed pursued Bachelor of Engineering and Master of Technonology in Computer Science from Visvesvaraya Technical University Belagavi, India in 2012 and 2015. He is currently pursuing Ph.D from Maulana Azad National Urdu University, Hyderabad, India since 2017. His main research work focuses on Big Data Analytics, Semantic Web, Network Security and Cloud Security. He has 4 years of teaching experience and 2 years of Research Experience.

Dr. Mugeem Ahmed working as an Assistant Professor at the Department of Computer Science and Information Technology Hyderabad (India). He received his doctoral degree in computer Science from Jamia Millia Islamia New Delhi India. His professional experience spans over more than 10 years of teaching, research, and project supervision. He has supervised various students for interdisciplinary research and industrial projects. Over the years, he has published many research papers with national and international journals of repute. In addition to these, he is also in the Editorial Boards and Reviewers' Panels of various journals. His primary area of research focuses on semantic web applications,Distributed Database Machine learning and Big data Analytics.

