FRAMEWORK FOR A SEMANTIC DATA TRANSFORMATION IN SOLVING DATA QUALITY ISSUES IN BIG DATA

[1]Grace Onyeabor & [2]Azman Ta'a

[1]Universiti Utara Malaysia, Federal University of Technology, Minna, Nigeria

[2]Universiti Utara Malaysia

grameenah@gmail.com

grace_amina@ahsgs.uum.edu.my

azman@uum.edu.my

ASTRACT

Purpose - Today organizations and companies are generating a tremendous amount of data. At the same time, an enormous amount of data is being received and acquired from various resources and being stored which brings us to the era of Big Data (BD). BD is a term used to describe massive datasets that are of diverse format created at a very high speed, the management of which is near impossible by using traditional database management systems (Kanchi et al., 2015). With the dawn of BD, Data Quality (DQ) has become very imperative. Volume, velocity and variety – the initial 3Vs characteristics of BD are usually used to describe the main properties of BD. But for extraction of value (which is another V property) and make BD effective and efficient for organizational decision making, the significance of another V of BD, veracity, is gradually coming to light. Veracity straightly denotes inconsistency and DQ issues. Today, veracity in data analysis is the biggest challenge when compared to other aspects such as volume and velocity. Trusting the data acquired goes a long way in implementing decisions from an automated decision making system and veracity helps to validate the data acquired (Agarwal, Ravikumar, & Saha, 2016). DQ represents an important issue in every business. To be successful, companies need high-quality data on inventory, supplies, customers, vendors and other vital enterprise information in order to run efficiently their data analysis applications (e.g. decision support systems, data mining, customer relationship management) and produce accurate results (McAfee & Brynjolfsson, 2012). During the transformation of huge volume of data, there might exist data mismatch, miscalculation and/or loss of useful data that leads to an unsuccessful data transformation (Tesfagiorgish, & JunYi, 2015) which will in turn leads to poor data quality. In addition of external data, particularly RDF data, increase some challenges for data transformation when compared with the traditional transformation process. For example, the drawbacks of using BD in the business analysis process is that the data is almost schema less, and RDF data contains poor or complex schema. Traditional data transformation tools are not able to process such inconsistent and heterogeneous data because they do not support semantic-aware data, they are entirely schema-dependent and they do not focus on expressive semantic relationships to integrate data from different sources. Thus, BD requires more powerful tools to transform data semantically. While the research on this area so far offer different frameworks, to the best of the researchers knowledge, not much research has been done in relation to transformation of DQ

in BD. The much that has been done has not gone beyond cleansing incoming data generally (Merino et al., 2016). The proposed framework presents the method for the analysis of DQ using BD from various domains and applying semantic technologies in the ETL transformation stage to create a semantic model for the enablement of quality in the data.

Methodology - The proposed framework offers task of producing semantic data in the format of RDF triples from the source data base on the semantics encoded in the semantic data web TBox. A Knowledge Base (KB) usually consists of two mechanisms. That is, TBox and ABox. The TBox presents the vocabulary of a domain. And the ABox composed of the set of declarations representative of instances. The ABox declarations always follow the TBox. The assumption in this paper, is that the mechanisms of a KB are given the description by a set of RDF triples, that is, a KB is an RDF graph without distinctiveness between classes and instances. Furthermore, discussion on the components of the framework were made which can be implemented on any specific domain of interest. The framework is divided into three layers: namely Definition Layer, ETL Layer, and Data Warehouse Layer. The definition of the Semantic Data Web (SDW) schema, data sources, and the mappings between the data sources and the target are given in the first layer – **Definition Layer**. The Definition segment of the semantic data web TBox intends to define a TBox giving the description of the appropriate data and the SDW schema built on the requirements. Definition of mappings between a data source TBox and the target TBox can be done by the user making use of the segment of Define Mapping. ETL process are intended in the **ETL Layer**. The extraction of data from various sources are intended in the extraction segment while the various task of cleansing and to conversion of the format such as data normalization, integrity violation checks, filtering, sorting and grouping are envisioned in the traditional transformation segment after the data have been extracted which will then be stored in a **Staging Area** which is the storage that is used to store the intermediary ETL sub process results. Still on the transformation, but this time the Semantic Transformation which does the conversion of the data into RDF triples based on the target TBox. And lastly, the Load segment will either straightly load the produced RDF triples by the Semantic Transformation segment or will load from the RDF dump file to the triple store. The **Data Warehouse Layer** is the layer for storage of the transformed semantic data. SPARQL queries will be used to analyze the stored data.

Findings - It will be proven in the evaluation that will be carried out the degree to which the created triples are consistent and complete (which are the data quality dimensions used in the measurement of quality in data) and linked semantically comparing the results with previous studies. The future work is the implementation of the framework by using BD set from a chosen domain. Likewise, the proposed framework will be validated by the future work.

Keywords: Semantic Transformation, Data Quality, Big Data, ETL

CONCLUSIONS

Transforming BD which is composed of large volume with various formats running at a high speed demands a semantic approach. The framework inclines towards overcoming the limitations

associated with obtaining quality data from vast amount of heterogeneous data. Also, the proposed framework pointed out the significance of using semantic technologies in transforming BD. Furthermore, discussion on the components of the framework were made which can be implemented on any specific domain of interest. The future work is the implementation of the framework by using BD set from a chosen domain. Likewise, the proposed framework will be validated by the future work.

REFERENCES

Agarwal, B., Ravikumar, A., & Saha, S. (2016). A novel approach to big data veracity using crowdsourcing techniques and bayesian predictors. *Proceedings of the 9th Annual ACM India Conference* ACM. pp. 153-160, (October),

Kanchi, S, Sandilya, S., Ramkrishna, S., Manjrekar, S & Akshata, A. (2015). Challenges and solutions in big data management - An overview 3rd International Conference on Future Internet of Things and Cloud.

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 60-68.

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for big data. *Future Generation Computer Systems*, *63*, 123-130.

Tesfagiorgish, D. G., & JunYi, L. (2015). Big data transformation testing based on data reverse engineering. *Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom),* pp. 649-652. IEEE.