DATA CONVERSION LAYER FOR DATA WAREHOUSE OR DATA LAKE USING NLP BASED TECHNIQUES

Yohan Hemal De Silva

A dissertation submitted in partial fulfilment of the requirement for Bachelor of Engineering (BEng) Honours in Software Engineering

Department of Computing
Informatics Institute of Technology, Sri Lanka in collaboration with University of Westminster, UK

2020

ABSTRACT

In the current context, data and information play a vital role where data is called the new oil. As the complexity of the world is progressing, the value given to the intelligence over data is also becoming more significant. In the present climate, business organizations, professionals thrive to interpret data and information as it's proven to facilitate more realistic insights for decision making. As the requirements of stakeholders of data and also the nature of the data is progressing to be more sophisticated, new architectural patterns should be created to tackle the intensity of data. To completely utilize the benefit of the concept of big data, associations need to have adaptable data models and ready to extricate the most extreme incentive from their data environment.

With the dynamic and complex nature of the world, it is proven that the definition of the analyzable data is changed when data analysis is not just based on structured data but it involves different types of data which facilitate more realistic insight generation. With the mentioned changes in the paradigm of data, new architecture call, 'Data Lake' has been established (Pradeep Menon Jul 6, 2017). Data Lake functions as a platform which store data of purest way. It can be structured or unstructured. The concept takes business analysis to a sophisticated level where it provides data users such as data scientist and business analysts to explore data to create a more realistic hypothesis and facilitate proactive insight generation for decision making.

However, structured data is traditionally easy to capture and understand compared to unstructured data. Exclusively, it's been proven that capturing and analyzing structured data provides a restrictive set of insight looking at the statistics of data availability in the world (20%). Therefore, the rest of the 80% of unstructured data needs to be considered for decision-making processes of organizational (Pradeep Menon Jul 6, 2017), economical, socialization which is proven to generate more realistic insight. Hence it justifies the importance of a firm solution for unstructured data processing.