ISSN 2347 - 3983

Volume 8. No. 4, April 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter85842020.pdf

https://doi.org/10.30534/ijeter/2020/85842020

From Single Architectural Design to a Reference Conceptual Meta-Model: An Intelligent Data Lake for New Data Insights

Jabrane Kachaoui[1], Abdessamad Belangour[2]

[1]Hassan II University, Morocco, jabrane2005@gmail.com

[2]Hassan II University, Morocco, belangour@gmail.com

ABSTRACT

Recently, a new concept has appeared in the world of new technologies dealing with Big Data. This concept is called Data Lake (DL) and it is becoming the most suitable way to administer and put up Big Data new generation systems. Today's Big Data storage systems suffer from many problems related to data structure, accessibility and data quality. In order to resolve these issues, DL systems offer referential without schema for unprocessed data with a common access interface. Consequently, storing data in a DL without any metadata governance will only generate a "data swamp". This paper describes a new architecture implementation for DL systems with optimal management of metadata. This process treats data from heterogeneous data sources and with a combination of Data Warehouse (DW) for better management of structured data. The proposal system discovers, extracts and classifies structural metadata from various data sources using ontologies. A new generation solution that makes available to all users sources of large amounts of information. This new approach offers companies an answer to data management problems and information availability.

Key words : Big Data, Data Lake, Data Warehouse, Metadata, Ontology.

1. INTRODUCTION

Organizations rely on insights derived from data to drive greater profitability, uncover opportunities, detect issues and problems, accelerate product and service innovation, and deliver exceptional customer experiences. Organizations are looking to harness new data processing technologies such as Apache Hadoop to derive previously unattainable insights [1]. The emergence of the DL concept gives organizations the capability of pooling all data, which make it accessible at any time [2, 3].

As organizations increasingly rely on data to power digital transformation, the clamor for faster access to more trusted data is growing. Providing fast access to critical data is fraught with challenges. In addition, as organizations harvest data on premises and in the cloud, fast, flexible, and systematic approaches to data management are essential to ensure repeatable and consistent value.

In the world of Big Data, it is difficult to find, prepare, master and map data due to its variety and large volume. This is the reason why it is essential to put in place systematic approaches to data management so that companies do not risk compromising the quality and speed at which they provide data. Ultimately, due to the very real risk of data security breaches, the credibility of the organization is at stake [9, 10].

Data Lakes are merely means to an end. To achieve the end goal of delivering accurate and consistent business insights repeatedly, the aid of DW and data-driven management will be needed [7]. With structural data algorithms delivered by the ETL(Extract, Transform, Load) data management is now facilitated with fast parsing, discovery, catalog, and preparation of data [12]. Combined with the speed and flexibility of a metadata-based approach to data management, this intelligent DL enables transforming raw Big Data for a variety of consumers without risk [4, 5].

Today, there are several concerns about Data Lakes that have prompted researchers to ask several questions: what is the definition of a DL? How does this help raise the challenges of Big Data? How is it committed to the DW? How can DL and DW be used together? How to launch an integration of a DL in a data management architecture?

This paper aims to answer these questions and discuss the concept of the DL by sharing the steps to establish the architecture of DL and DW complementarity and to create new shapes of the business value with this new technology. It is outlined as follows; second section presented some related work discussing their approaches and proposal solutions. Third section describes how DL and DW work together. Fourth section provides a detailed description of the proposal architecture based on various researches done by authors in each layer. Fifth section suggests best practice to build a robust system able to tackle data issues. Finally, a brief conclusion with some future works.

2. RELATED WORKS

Even though literature is abundant on decision support components, implementation or design nevertheless with regard to DL field, it is still at its beginning. Some suppliers, related to Apache Hadoop technology such as Hortonworks, Cloudera have launched into this term without really explaining what it is, the objectives, the scope and the impact on existing information architectures.

This article aims to highlight this subject and put DL at the heart of information architecture. Clear definition is needed to correctly situate DL place behind information systems and define its links.

Several researches focused on various processes such as extracting, loading or clustering but separately. Few studies discuss global proposal system combining several layers to deal with unstructured data.

Rihan Hai and el. propose Constance, a DL system with metadata management. This system is able to extract and manage metadata for efficient treatment. A global overview is given of the proposal system that is based on three layers. Ingestion layer for importing data from various sources into DL system. Maintenance layer for extracting metadata and explicit data schema based on a component Structural Metadata Discovery (SMD). Another component is the Semantic Metadata Matching (SMM) component, which consists of ontology modeling, attribute annotation.

This paper present the result of several studies were conducted by authors [16], [17]. A proposal ontology model with combination with K-means algorithm based on metadata is setting up for storing data coming from DL into various clusters using MongoDB [18].

To benefit current information systems based on DW of MongoDB structured data, authors propose MQL2SQL (Mongo Query Language To Structured Query Language) a novel algorithm able to transform data from MONGODB to RDBMS [11].

3. DATA LAKE: NATUREL FIT TO ETL

3.1. Brief presentation

The intelligent DL is an enterprise capability to help organizations embedding new Big Data solutions into existing data landscapes that are often Data Warehouses. This capability will increase organizational performance and competitiveness by setting up a new data management strategy for data insight [11].

It is considered as new approach of analytical insights creation for businesses, from the acceleration of traditional enterprise reporting through new analytics driven by data science. It especially aims to bridge the gap between the rigidity of Data Warehouses/ data marts and the velocity and business needs.

To achieve this, this system must contain the following essential characteristics:

- It covers data processing and data storage at scale for a lower cost compared to previous approaches/solutions.
- It is open to defining data structure at the needed time, and needed context (schema on read)
- It is open to all kinds of data and stores different types of data in the same repository.
- It combines batch layer with streaming layer for both processing large sets of data as well as real time processing.
- It is able to scale over a cluster of machines by distributing both storage and processing capabilities.

In this sense, the proposed architecture is an approach that provides a set of common, core services that are required and useful to create new insights from data .It helps business people and data scientists to take decision from data insights.

3.2. Complementarity overview

DL capability of storing and processing data at a low cost has made it a perfect place for ETL; it is a data preparation process for business use. Data Lakes ingest all kind of data by using robust programming framework and low level coding language. All of these characteristics make DL natural fit for ETL.

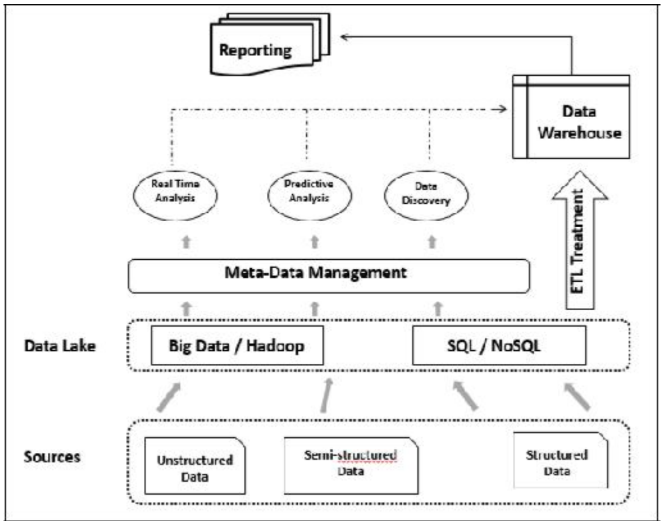

Figure 1: The Architecture of a Hybird System

It is an economical and online data store with powerful data process ability. After processing data in several steps, it will be stored into DW, where it can be analysed and reused for business needs [8, 12].

The emergence of DL in companies that have enterprise data warehouses has led to an inevitable question for CIOs: how can DW and DL work together? The answer is pretty clear, a hybrid data management ecosystem encompasses them, with a powerful data analytics process able to respond user questions with less effort. Over time, DW capabilities have been spread out by consolidated data from a DL in queries using various methods. Over the long term, the data storage location to a certain extent will be unknown from end users. Analysts and end users are limited to ask questions. The hybrid ecosystem composed of DL and DW will define what data to use to answer questions [6].

4. A PROPOSAL ARCHITECTURE

4.1. A detailed architecture of the proposal system

This architecture provides a set of common capabilities that are required and useful to create new insights from data, regardless of its purpose (descriptive, diagnostic, predictive, and prescriptive).

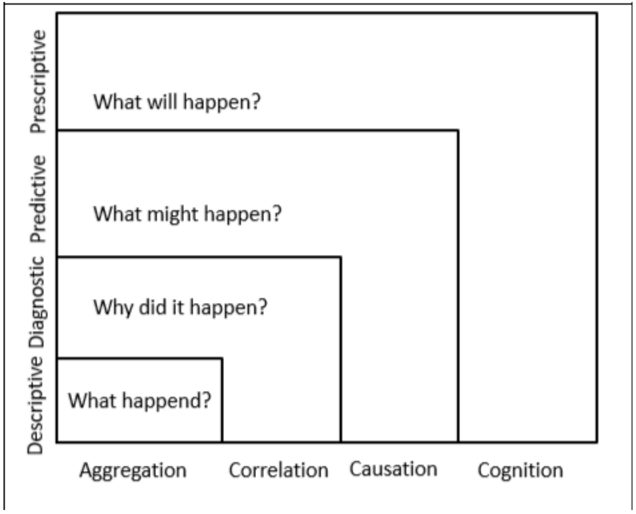

Figure 2: Data analytics types

The proposed architecturehas been implemented to respond to system failures and human error tolerance dueto the growing risk of software and hardware failures and thus to meet specific end user needs. Figure 3 depicts the proposal architecture of new Data Lake system, as well as its key components. This architecture is composed of four layers: acquisition layer, exploration layer, semantic layer and insight layer. Incoming data is present in both acquisition layer and exploration layer. In acquisition layer that implements the interface between data sources and the proposal system, the basic concept is that the data is immutable in the dataset, data is never updated in this layer, and the new data is provided at the end of file. Then, there are dataviews that meet business requirements. In exploration layer, data flows coming from previous layer are processed and data schemas are grouped. Stream and batch processing frameworks such as Stream Spark or Storm are used for data preprocessing and schema parsing. The nature of this architecture is to process a large amount of data that is routed to semantic layer for categorization and clustering based on K-means algorithm and ontologies. Finally, comes the role of insight layer, that represents the major human machine interaction, to display user request by compromising statistical and data visualization tools.

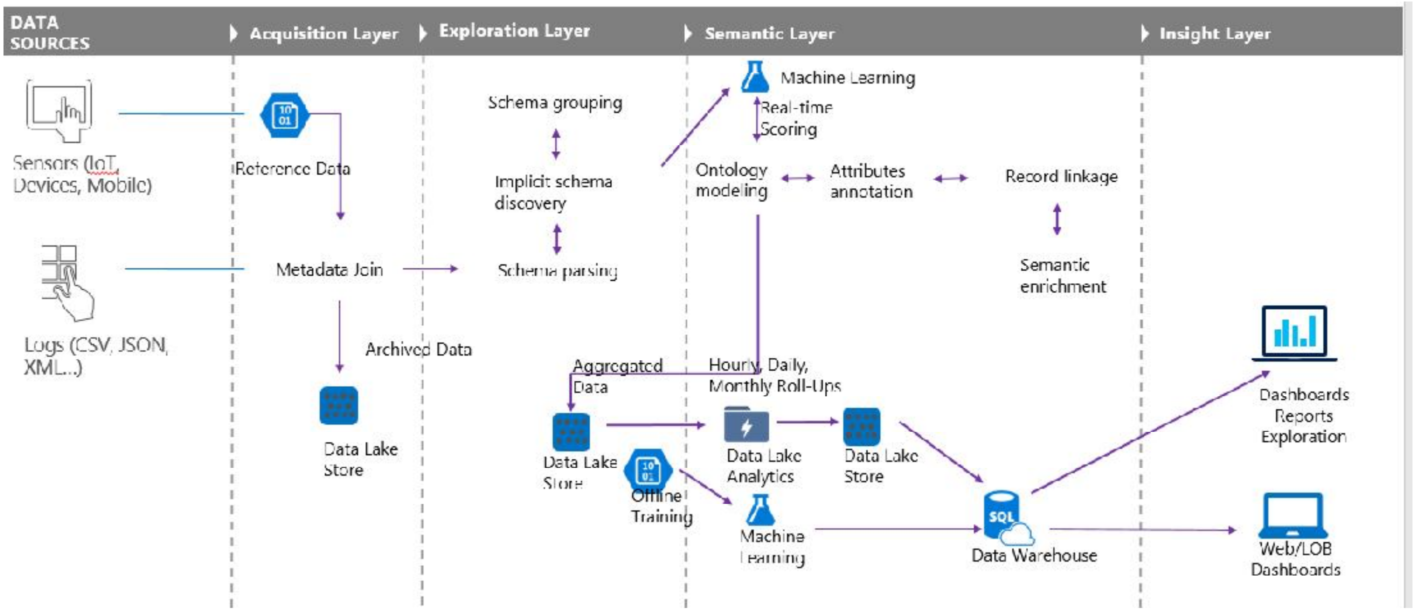

Figure 3: Intelligent Data Lake Process Overview

A. Data acquisition layer

Exploiting data from current systems represents a competitive challenge for businesses. In particular, data acquisition, the first phase of this exploitation, is particularly critical. Several intelligent systems and algorithms development have been set up which make it possible to select useful data and thus, to save considerable time for the analysis phases, or to carry out a data processing on line.

Data acquisition layer plays a key role in Data Lake; it provides precious data in its raw format to improve productivity. Once collected, the data is processed and analyzed and thus serves as the basis for customer requests.

B. Data Exploration layer

Data exploration is the process of automated sorting through huge data sets to identify trends and patterns, and build relationships. It consists on exploring and analyzing big amount of data in order to discover relevant rules and patterns. It is considered a discipline in data science field studies and is distinguished from predictive analytics, which describes historical data.

Organizations today are collecting ever-increasing volumes of information from a variety of sources, including websites, business apps, social media, mobile devices and, increasingly, Internet of Things (IoT). Data exploration layer allows discovering and grouping data schemas based on the metadata, then redirecting them to the next layer (Semantic layer) for advanced treatment in machine learning.

C. Semantic layer

The semantic layer acts as an interface between database and the users. This interface presents to end users a panel with business-oriented elements, which allows them to generate requests to access data without knowing the language of requests. Information access security is based on controlling access to data and controlling rights in applications.

Data is restructured, enriched, aggregated, reformatted, nomenclatured to be presented to the user in a semantic form (meaningful business views) which allows decision-makers to interact with the data without having to know their physical storage structure.

In This system, this layer consists of analyzing data to respond to users' requests through predefined ontologies and returning a response in structured data format interpreted by DW.

D. Insight layer

Data insight value as a vector for transmitting information and as a tool for analyzing and exploring data no longer needs to be demonstrated. It is obviously not the only possible approach to study data, statistics also play a major role in this area and the two approaches are generally combined. Insight, however, holds a major place in the human cognitive system and represents a real tool of interpretation and understanding for humans. The problems raised by Big Data, previously introduced, have passed on insight systems and techniques. Indeed, these are only rarely adapted for representation or exploration of datasets so large that cannot be envisaged for storing them on a simple machine, or even treating them with traditional approaches. The transition between the two methods, for large amount of data or not, must be done within insight Pipeline. In this process, the first step is to analyze data, resulting in creation of a data abstraction. Once data has been abstracted, it is filtered to keep only pertinent data. Then comes the creation of geometric data, that is to say the transformation of data into a structure that can be used to generate representation. Finally, the image is created which will be rendered for user observation and analysis.

5. DEPLOYING THE INTELLIGENT DATA LAKE

Many companies have intended to use DL. Almost of them use Hadoop that is considered as the most adequate platform dealing with DL needs, but technologies do not always respond to DL challenges. DL concept tries to work out with two difficulties, one old and one recent. The old one is information silos. Data sources can be joined in DL, instead of having many independent data collections. The consolidation is explained by increased use and information sharing, while reducing server and license costs. The new constraint of conceptualized DL approaches is Big Data initiatives. The concept is that Big Data projects involve a large amount of varied information. The variation of information makes it unknown while receiving and compelling it in structured format as a DW or RDBMS (Relational Database Management System) constrains deep analysis. DL permits to deal with these issues which benefits IT in the short run. With this concept, data is simply ingested into DL in its raw format to help IT users to save time and effort that was spent in understanding how information is used. However, getting sense from information and extracting value from it, is business end user tasks. In addition to that, DL itself is not able to set up an inherent mechanism for adding meaning or reconciling semantics. Technology could be applied on DL to deal with this issue, however, without any data governance policy, DL will finish being a collection of isolated disconnected data pools. With all of that, Data Lakes have many weaknesses. The first one is data quality, DL is unable to determine lineage of findings. Another issue is access control and security. Data can be dumped into the Lake with no contents supervision. After all, performance aspects should not be neglected. Only simple tools and data interfaces cannot perform as well face of a general-purpose store as they can face of optimized and special design infrastructure [15].

To deploy a successful Data Lake, three prerequisite keys should be discussed:

- The right platform
- The right data
- The right interfaces

5.1. The right platform

The most popular platforms for DL are represented by several technologies: Hadoop in Big Data, Cloud like Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform. There are several advantages in common between these technologies that are considered important and relevant:

Volume: these platforms have been designed to support an expanding volume of stored data and to establish a significant increase in performance without degradation.

Cost: There are several storage methods that have capacity to store a lot of data and that are less expensive. Such as, WORM (Write Once, Read Many)disks and hard disks. However, there is only one technology that is possible to both store and process huge amounts of data that are inexpensive. The cost of these storage methods is generally between one tenth and one hundredth of that of a commercial relational database.

Variety: These platforms are based on file systems that makes them able to store all kinds of files: Hadoop HDFS (Hadoop Distributed File System), MapR FS (MapReduce FileSystem), AWS's (Amazon Web Services) Simple Storage Service (S3), etc.

In contrast to a relational database that requires predefined structure of Data (schema on write), a file system does not really give importance to what is written. Obviously, in order to process data effectively, knowledge of the schema when using data is essential. This approach represents one of the most important advantages of Big Data technology allowing frictionless ingestion and it is known as the schema on read. In fact, unlike a relational database where the data cannot be loaded until it is converted to the schema expected by the database, this approach allows the data to be loaded without any treatment [19].

Future Proofing: Because our world continues to evolve and so do our needs, we must guarantee that the data we currently have will be accessible for possible use in the future. In fact, data that is stored in a relational database is only accessible via this database. However, the different Big Data platforms like Hadoop are very modular in the sense that the same file can be used by several engines and processing programs. In other words, Hive provides an SQL interface to Hadoop files that allow you to perform queries, also Pig scripts to Spark and MapReduce custom tasks and all types of tools and systems are able to use the same file.

As a result, as Big Data Technology develops quite quickly, this ensures that future projects will always be accessible in a data Lake.

5.2. The right data

Nowadays, most of companies discard their collected data. Indeed, they only keep a small percentage of this data in a DW for a few years later, but most of the detailed operational data, machine-generated data and old historical data are either aggregated or discarded. Therefore, performing analysis is almost impossible. For example, if an analyst identifies the value of certain discarded data in a traditional way, he can spend months, even years, to gather enough history and make meaningful analysis.

DL technology guarantees that we can store as much data as possible for future use. In addition, data does not need to be converted or processed prematurely, as its uses need is not defined [20].

In order to get the right data, another challenge presents itself: Data silos. Data accumulation can be done by several departments at the same time because of the difficulty and the high cost they face. In any business, if a group needs data from another group, it must explain what data needs, and then the group owning the data must implement ETL jobs that extract and aggregate the required data. This process is expensive, difficult, and wastes a lot of time and effort, so teams must clearly understand data requests and then take the time it takes to deliver the need. This hard work is often used as a pretext to not share data [14]. By opting for a DL as a solution, this effort can be avoided, thanks to the easy ingestion that allows data to be centralized and erased in a raw way and without any processing. Data governance under a DL also provides a transparent process for enterprise users on how to obtain data, so that ownership no longer becomes an obstacle.

5.3. The right interface

After setting up the right platform and loading the right data, the hard part now is to choose the right Interface. This part represents the heaviest aspect for a business in a DL, where most fail. In Order to obtain widespread adoption and gain multiple benefits by helping business users make data-driven decisions, solutions offered by businesses must be self-service. In other words, users should not need the help of IT (Information Technology) experts and should be able to find, understand and use the data themselves.

6. CONCULSION AND FUTURE WORKS

Today, it is no longer a question of knowing if a DL is necessary or not, but of identifying which solution to use and how to implement it. Companies should determine the DL capability they want to work with based on their current data process systems. Driving the business result and gain values is the main goal for Data Lakes. Moreover, the hybrid data management ecosystem made up by DL and DW will be the right decision for companies dealing with big data challenges [13]. Our future work consists on determining a new algorithm of big data categorization in the aim of getting data insights by end users and put into practice the proposal architecture.

REFERENCES

1. H. Alrehamy and C. Walker. **Personal data lake with data gravity pull**. In Proc. IEEE BDCloud, 2015.
2. J.Kachaoui and A.Belangour.**Challenges and Benefits of Deploying Big Data Storage Solution**. *In Proc. New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Societ, Article No.: 22, 2019, pp 1–5.* https://doi.org/10.1145/3314074.3314097
3. M. Chessell, F. Scheepers, N. Nguyen, R. van Kessel, and R. van der Starre.**Governing and managing big data for analytics and decision makers**.www.redbooks.ibm.com/redpapers/pdfs/redp5120.pdf, 2014.
4. J. Dixon. **Data lakes** revisited. https://jamesdixon.wordpress.com/2014/09/25/data-lakes-revisited/, 2014.
5. Foshay, N. et al. **Does data warehouse end-user metadata add value?** Commun. ACM, 2007. https://doi.org/10.1145/1297797.1297800
6. Kalousis, A. et al. **Using meta-mining to support DM workflow planning and optimization**. JAIR, 2014.
7. M. Jarke, M. A. Jeusfeld, C. Quix, and P. Vassiliadis. **Architecture and Quality in Data Warehouses: An Extended Repository Approach**. Information Systems, 24(3):229–253, 1999.

8. D. Kensche, C. Quix, X. Li, Y. Li, and M. Jarke. **Generic schema mappings for composition and query answering. Data Knowl**. Eng., 2009. https://doi.org/10.1016/j.datak.2009.02.006
9. C. Quix, R. Hai, and I. Vatov. Gemms: **A generic and extensible metadata management system for data lakes**. In CAISE FORUM, 2016.
10. I. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino. **Data wrangling: The challenging yourney from the wild to the lake**. In Proc. CIDR, 2015.
11. J.Kachaoui and A.Belangour. **MQL2SQL:A Proposal Data Transformation Algorithm from MongoDB to RDBMS**, International Journal of Advanced Trends in Computer Science and Engineering, in progress.
12. E. Boci and S. Thistlethwaite. **A novel big data architecture in support of ads-b data analytic**. In Proc. ICNS, 2015. https://doi.org/10.1109/ICNSURV.2015.7121281
13. J.Kachaoui and A.Belangour, **A Multi-criteria Group Decision Making Method for Big Data Storage Selection**, *In Proc. International Conference on Networked Systems, 2019, pp 381-386.* https://doi.org/10.1007/978-3-030-31277-0_25
14. A.Gorelik. **The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science**, 2019.
15. H.Fang**. Managing Data Lakes in Big Data Era: What's a data lake and why has it became popular in data management ecosystem**, 2015. https://doi.org/10.1109/CYBER.2015.7288049
16. J.Kachaoui and A.Belangour. **Enhanced Data Lake Clustering Design based on K-means Algorithm**. International Journal of Advanced Computer Science and Applications, in progress.
17. J.Kachaoui and A.Belangour. **An Adaptive Control Approach for Performance of Big Data Storage Systems**. *In Proc. International Conference on Advanced Intelligent Systems for Sustainable Development, 2019, pp 89-97.*
18. J.Kachaoui, J.Larioui and A.Belangour.**Towards an Ontology Proposal Model in Data Lake for Real-time COVID-19 Prevention Cases**. International Journal of Emerging Technologies in Learning (iJET), unpublished.
19. M.Govindarajan.**Ensemble of Classifiers in Text Categorization**, International Journal of Emerging Trends in Engineering Research. Volume 8, No. 1, 2020, pp. 41-45. https://doi.org/10.30534/ijeter/2020/09812020
20. A. D. M. Africa, G. Ching, K. Go, R. Evidente and J. Uy, **A Comprehensive Study on Application Development Software Systems**, International Journal of Emerging Trends in Engineering Research, Volume 7, No.8, 2019, pp.99-103. https://doi.org/10.30534/ijeter/2019/03782019