Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 78 (2014) 88 – 95

Procedia Engineering

www.elsevier.com/locate/procedia

Humanitarian Technology: Science, Systems and Global Impact 2014, HumTech2014

Advancing big data for humanitarian needs

Samson Oluwaseun Fadiya[a]*, Serdar Saydam[b], Vanduhe Vany Zira[c]

[a]*MIS Department, Girne American University, TRNC via Mersin 10 Turkey*
[b]*Girne American University, TRNC via Mersin 10 Turkey*
[c]*MIS Department, Cyprus International University, Lefkosa, TRNC via Mersin 10 Turkey*

Abstract

At the present moment, almost every business is witnessing a bursting of data. Big data now available for analytics present complex, daunting challenges due to the vast number of digital data generated daily by different organizations. The public, government, and human rights organizations are adapting to this massively increasing amount of text, images, and video data available to analyze scientifically. For example, the social contribution of conflicts, political violence and disasters are all digitalized through the streaming of data. The vast amount of data has improved the global community's ability to defend and allow for progress of rights of vulnerable people around the globe. And, if big data processing is to improve lives, its existing data gathering methods should assist Humanitarian Affairs, and not replace them. Therefore, finding ways to increase humanitarian services with data, highlighting the importance of big data, are critically important. Numerous organizations have escaped the frustrations of their first-generation data warehouses by replacing older database technologies with significant data which is unstructured in a scalable, error tolerant and efficient way. The purpose of this paper is to propose a big data platform for large-scale data analysis by using the Map Reduce framework for unstructured data stored into integrating distributed-clustered systems such as NoSQL (Not Only SQL) and Hadoop Distributed File System (HDFS).

Selection and peer-review under responsibility of the Organizing Committee of HumTech2014

Keywords: Analytics; Artificial Intelligence; Big data; Business Intelligence; Hadoop; HDFS; MapReduce; NoSQL; Real-time; Scale out NAS;

1. Introduction

In this Information Age, information is unceasingly created all over the world around the clock. Businesses and organizations want to run most of its business processes using big data technology, being created in the class of

* Corresponding author.
E-mail address: samsonfadiya@gau.edu.tr

Selection and peer-review under responsibility of the Organizing Committee of HumTech2014
doi:10.1016/j.proeng.2014.07.043

transactions and interactions. Through emails, videos and images, for example, the huge amount of data is being produced by goods and services, with the Internet becoming a very significant user interface for interactions.

In addition, there is a broad level of data discharge as default in the mechanical structure of database logs, system logs and web server logs. Telecommunications Network providers get a broad degree of data in the form of conversations. Also, social network sites like Facebook have begun acquiring terabytes (TBs) of data every day in the form of comments, blogs, tweets, photos, audio and videos, among others. Internet-based companies in this digital age generate the vast level of on-line streaming data daily. For example, in a medical setting, one can imagine a situation where a patient complained about a particular set of symptoms and the doctor could track the cases accounting for these across all past patients, and understand those comparable symptoms and how this individual patient may respond to different treatment. Consequently, health data about patients, diseases and the data produced by various medical devices will be massive. And, data generated from different machines in the production sectors like transport, war ammunitions, finance and many more are similarly a source of massive data, with each being stored for a different reason for future use. More so, organizations can process this data, analyze it and store it for intelligent decision-making, to gain a highly competitive benefit over their contemporary. For example, social networking site data can support the technical team to understand user navigation and then help determine quality of service and guide essential advances.

More so, the influence question that comes up is how do we process, store and manage such a great size amount of data most of which is Unstructured. Although, there are major categorizations of big data platforms to store, process and manage them on a possible scale, functional, effective and error tolerant and efficient fashion. Unlike other MPP Analytic platforms, which are prone to failure and difficult to manage, Vertica has no leader nodes, and, therefore, no single point of failure. Any node in a Vertica cluster is capable of initiating loads or queries, and will evenly distribute the workload to other nodes when it makes sense to do so. Workload distribution is automatic, no DBA or user intervention required [11]. These Vertica's scales-out MPP Data Warehouses include: 1) Vertica's core analytic surface scale-out to interact significant size of the data. It's designed to unendingly convey the big data size. And its data compression, in-memory potentialities processes, bring back analytic queries in near real-time. (E.g. IBM DB2, Greenplum, AsternCluster, DATAllegro, Kognitio WX2, and Teradata). 2) Appliances: a purpose-built mechanical device, with preconfigured scale-out MPP hardware and software intent towards analytical processing. (E.g. Teradatamachines, Sun's Data Warehousing Appliance, Oracle Optimized Warehouse and Netezza Performance Server). 3) Columnar Storage & Execution: they store data in columns on the contrary of rows; appropriating large compression and more quickly query operation (e.g. ParAccel, Vertica, InfoBright Data Warehouse, Sybase IQ. The majority of them supplies UDFs and SQLs to match the data).

Alternatively, another class offers distributed file systems like Hadoop to store large unstructured data and execute Map Reduce computations on it across a cluster built hardware. The paper is coördinated as follows: In section 2, we show related work and explain previous experience theories such as Big Data Analytic in section 3. In section 4, we bring in our proposed platform for Big Data and then the conclusion in section 5.

2. Related Work

We look over carefully some of the currents, big data platforms for big scale data analysis. There are several types of vendor commodities to consider for big data analytics. Now vendors have contributed analytic platforms established upon Map Reduce, and distributed file system. The Vertica Analytic Platform offers a robust and ever-growing set of Advanced In-Database Analytic functionality. It has a high-speed, relational SQL database management system (DBMS) purpose-built in analytic and business intelligence. It offers a shared-nothing, Massive Parallel Processing (MPP) column-oriented architecture [8].

IBMInfoSphere Biginsights represents a fast, robust, and easy-to-use platform for analytic on Big Data at rest. For example, IBM offers a platform for big data, including IBMInfoSphere Biginsights and IBM InfoSphere Streams. IBMInfoSphere Streams are a powerful analytic computing platform that delivers a platform for analyzing data in real time with micro-latency [1].

And EMC Greenplum is driving the future of data warehousing and analytics with discovery products including the Greenplum Data Computing Appliance, Greenplum Database, Greenplum HD enterprise-ready Apache Hadoop, and Greenplum Chorus. The SAND Analytic Platform is a columnar analytic database platform that achieves linear

data scalability through massively parallel processing (MPP), breaking the constraints of shared-nothing architectures with fully distributed processing and dynamic allocation of resources [5]. Pavlov et al. [10] This described and compared Map Reduce structure and parallel DBMS for large-scale data analysis and defined a benchmark consisting of tasks run on an open source version of MR as well as on two parallel DBMS.

The ParAccel Analytic Database (PADB), the world's fastest, most cost-effective platform for empowering analytic-driven businesses, when combined with the Web FOCUS BI platform. ParAccel enables organizations to tackle the most complex analytic challenges and glean ultra-fast, deep insights from the vast volumes of data. Also, the Netezza, a leading developer of combined server, storage, and database appliances designed to support the analysis of terabytes of data and give companies with a powerful analytic foundation that delivers maximum speed, reliability, and scalability [3].

3. Previous Experience Theory

This section supplies a general examination of big data, big data Analytic, big data solution and big data storage. An elaborated description of them ascertained in [1] [5] [11].

3.1. Big Data

Progressively, establishments today are facing increasingly Big Data stimulating situations; Big Data holds for the data that cannot be processed or analyzed employing traditional processes or carry out such practices. Many organizations have an access to a huge amount of information, because they have no such idea whether it's deserving holding. There are major four dimensions (V's) features of Big Data: volume, variety, and velocity, veracity.

1) Volume: Volume is the first and most notorious feature because many organizations are giving rise to a very large volume of data internally, or assembling other vast amounts of data from the outer side. 2) Variety: In an organization today, there are many ways, whereby data collection has increased. This has caused the rise inside and outside source of data non-structured, such as plain-text documents, electronic sensor, tweets, blogs and social media. 3) Velocity: Data warehouse is traditional types of the result, therefore, in traditional methods, information has been often raw and inevitable to employ and acted according to share time frames to get the best potential value from it and this makes the real-time answers to a common need in advance establishments. There are major two types of big data: the data on balance, e.g. social media, emails, web-logs, and non-structured plain-text documents are all gathering of streamed function. So data gathered in motion, e.g. twitters comments and sensor information (data). 4) Veracity: big data like social media data (e.g. Tweets or Facebook Posts), how much should we put in the data, inaccurate data that is directly different in traditional data warehouses, where it was always the assumption that the data assure, clean, and correct. That is why so practically time on Data Lineage, Master Data Management, ETL/ELT, and Identity Insight/Assertion, etc.

3.2. Big Data Analytic

Big data analytics are a boosting analytic practical method to a very big data sets, because both structured and unstructured of big data can meet data from product data sources, which include network and mobile devices, and the web technologies. Progress analytic is a gathering of practical techniques, which includes data mining, complex SQL, data visualization, artificial intelligence, predictive analytics. For example, database techniques that hold analytic, such as, Map Reduce, columnar data stores and, in-memory database.

3.3. Big Data Storage

3.3.1. Big Data Storage Towards Big Data Analytic

Shared storage in Big Data analytic marketers and the storage community in general, have a shared storage in Big Data analytic, although this is a case made. For example, we can see the incorporation of the NetAppSANstorage in

the ParAccel's Analytic Database (PADB). Big Data professionals should see the shared storage surroundings as one in which they can notice potentially valuable data services, such as data protection and usage, availability, change management, cost savings, the reduced time to deployment for new applications and automated processes and Life-cycle management [9].

3.3.2. Big Data Storage Conditions

Lots of establishments are finding it difficult to discuss the increasing data intensities. In other word, big data plainly cause the problem more. And in order to solve this difficulty, big data, establishments need to cut the measure of data being stored and manipulated to the benefit of the new storage technologies that better the operation and storage usage. Apparently, in a big data view, there are four important directions:

1) Reducing data storage requirements Using data compression and new physical storage structures such as columnar storage.
2) Potentially, big data (using an index that combines several quantitative metrics), will underpin new waves of productivity growth and consumer surplus.
3) Improving input/output (I/O) performance using solid-state drives (SSDs).
4) Increasing storage use by using tiered storage [2]

3.4. Big Data Outcomes

The new emergence of a big data solution shows the expert means to carry out operations with greater degree volumes of data in a limited period, with the power to interact with many types of data from different sources. A good example of big data solutions for humanitarian is the timely intervention of a life-threatening condition that it can offer, when take a full advantage of a massive data available. And big data can make positive risk decisions based on its ability to offer a real-time trace of data.

Furthermore, big data has the intensify ability to show threats and criminals different stations or a stream of data, audio, and video feeds, It can also predict weather patterns to design optimal wind turbine use, or multi-channel customer analysis and optimize capital expenditure on asset placement. The big data solution can offer these abilities for sectors like educations, hospitals, and governments for humanitarian purposes. They are:

1) Deep Analytic — a fully parallel, extensive and extensible toolbox full of advanced and unique statistical and data mining capabilities
2) High Agility — the ability to create temporary analytic environments in an end-user driven, yet secure and scalable environment to deliver new and unique insights to the working business
3) Massive Scalability — the ability to scale analytics and sandbox to before unknown scales while leverage previously untapped data potential.
4) Low Latency — the ability to act based on these advanced analytic in working, production environments [7].

4. Proposed System Architecture

4.1. Big Data Storage Architecture

At the present moment, big data includes multiple entity types, traditional database models, and processing techniques. Big data needs have breached the traditional data storage processes, and produced the need for novel architectures for storage and integration with the analytic systems that do the analysis. Older storage architectures could not scale to the size required or hold the diverse data types. Data that could be stored in an array was in the 100s of terabyte range, but the file systems they provided could not scale beyond 16 terabytes [8].

4.1.1. Traditional Data Architecture Potentialities

Current advances in architectures use the key approach of scale up versus scale out. To interpret the high-level architecture faces of Big Data, we first review well-formed information architecture for structured data. In the example, you see two data origins that use incorporation (ELT/ETL/Change Data Capture) efficiencies to transfer data into a DBMS data warehouse and then offer a broad variety of analytical capacities to bring out the data.

Roughly, these analytic capabilities include semantic interpretations of textual data — EPM/BI applications, dashboards, summarizing, reporting, and a statistical query — and visual perception tools for high-density data.

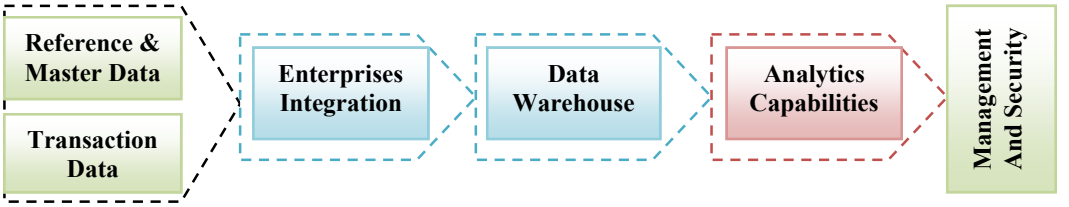

Fig. 1. Traditional Potentiality (Structured).

4.1.2. Big Data Architecture Potentialities

The working potentialities for big data architectures are to cope with the volume, velocity, variety, and veracity necessities. There are differing technology schemes for real-time and batch processing demands. For example, real-time, key-value data stores, such as heftiness allow for high performance index-based retrieval, for the bunch processing, a method known as "Map Reduce." This is a Filter data according to a specific data breakthrough scheme and after the filtered data; it analyzed instantly filled into other unstructured or semi-structured databases, sent to mobile devices and dependent on structured data. Figure 2 describes the large-scale shared storage system architecture for big data.

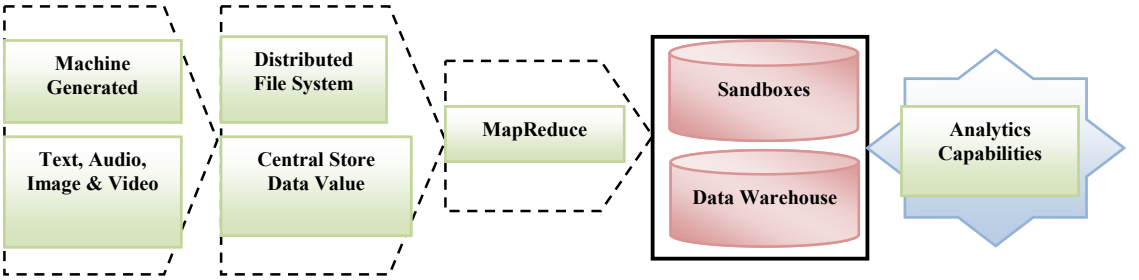

Fig. 2. Big Data Potentiality (Unstructured).

Of benefit to new unstructured data domains, there are two central differences for big data. First, in reference to the size of the data sets, we do not position the raw data to a data warehouse. Nevertheless, after Map Reduce processing, the "reduction result" integrates into the data warehouse environment so that we can gain traditional BI reporting, semantic, correlation and statistical capabilities. Ultimately, it is ideal to have analytic capabilities that blend a traditional BI platform with big data visualization and query capabilities. The second difference relates to the ease of analysis in the Hadoop environment and sandbox environments.

4.2. Proposed Big Data Platform

For big data analytics, generally, there are three major advances: 1) direct analytic over massively parallel processing data warehouses, 2) indirect analytic over Hadoop and 3) direct analytic over Hadoop. The proposed approach performs analytic over the Hadoop Map Reduce framework and distributed-clustered systems, such as NoSQL and Hadoop Distributed File System (HDFS). Altogether, the queried file for analytic are performed as Map Reduce problems across big unstructured data placed into NoSQL and Hadoop Distributed File System (HDFS). This approach can cause, low-cost big data, result, highly scalable, and fault tolerant achieved. Figure 2 describes the proposed significant data approach.

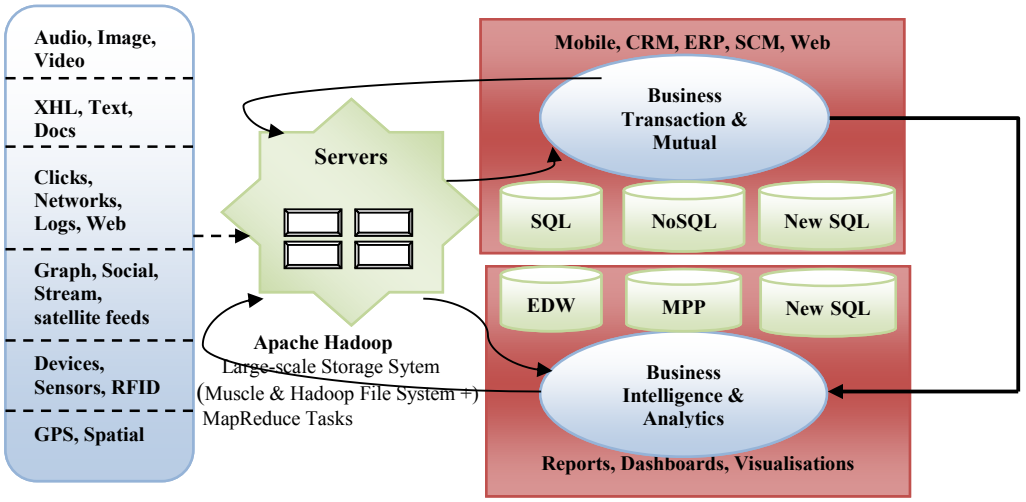

Fig. 3. Future Big Data Architecture.

In Figure 3, Apache Hadoop acts as the Big Data servers. It is big at storing, combining, and translating multi-structured data into greater useful and valuable arrangements. For instance, Apache Hive is a Hadoop-associated element that corresponds inside the Business Intelligence & Analytic class since it is usually used for querying and analyzing data within Hadoop in an SQL-like fashion. Apache Hadoop can also be included with the implied EDW, MPP, and NewSQL components such as HP Vertica, Teradata, EMC Greenplum, Aster Data, IBM Netezza, SAP Hana and many others.

Furthermore, Apache HBase is a Hadoop-related NoSQL Key/Value store that is usually employed for building extremely reactive future-generation applications. Apache Hadoop can also be included with other SQL,NoSQL, and NewSQL technologies such as MySQL, IBM DB2, PostgreSQL, Oracle, MongoDB, Terracotta,GemFire, SQLFire, VoltDB, Microsoft SQL Server and many others. In conclusion, data trends and integration applied science help in assuring data flows smoothly between the systems in the above plots.

4.2.1. Hadoop and MapReduce Framework

The emergence volume of big data types outmatched the capabilities of formal technologies such as relational databases that are the unstructured data, which has posed a unique challenge. Now, organizations are looking into future generation technologies for data analytics. One of the most anticipating technologies is the Apache Hadoop software and the Map Reduce framework for contending with this “big data” problem. A Map Reduce framework typically divides the input dataset into independent tasks that map tasks in a completely parallel way. The framework sorts the outputs of the maps, which are then input to the cut tasks. Typically, both the input and the output of the jobs are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re executing the failed tasks [6].

4.2.2. NoSQL and Hadoop Distributed File System

NoSQL and HDFS file system is capable of being a scaled open source clustered file system that extended a global namespace, distributed front end, i.e., hundreds of petabytes with no problem. It is also a software-only, highly available, scalable, centrally managed storage pool of unstructured data. It is also scaling-out file storage software for NAS, object, big data. There are lots benefits of NoSQL and HDFS over any other file systems. These benefits are: 1) An Apache open source distributed file system. 2) Anticipated running on high-performance value hardware. 3) Experienced for highly scalable storage and automatic data reproduction across three nodes for Fault is tolerance. 4) Automatic data reproduction across three nodes carries off the need for backup. 5) It writes once, read

many times. 6) Dynamic and flexible schema design.7) Able to process data without a row and column structure, allows many records reads in a single API call. 8) Highly scalable multi-node, many data centers, fault tolerant, ACID operations. 9) Simple programming model, random index reads and writes. 10) Not Only SQL but simple pattern queries and custom-developed solutions to get access to data such as Java APIs.

NoSQL and HDFS file system allows software-based data security and practicality to the Hadoop cluster and removes the single point of failure issue. Existing Map Reduce based applications can use NoSQL and HDFSseamlessly. This novel practicality opens up data within Hadoop deployments to any file-based or object-based application [3]. Figure 4 describes NoSQL and HDFS file system compatibility for Apache Hadoop.

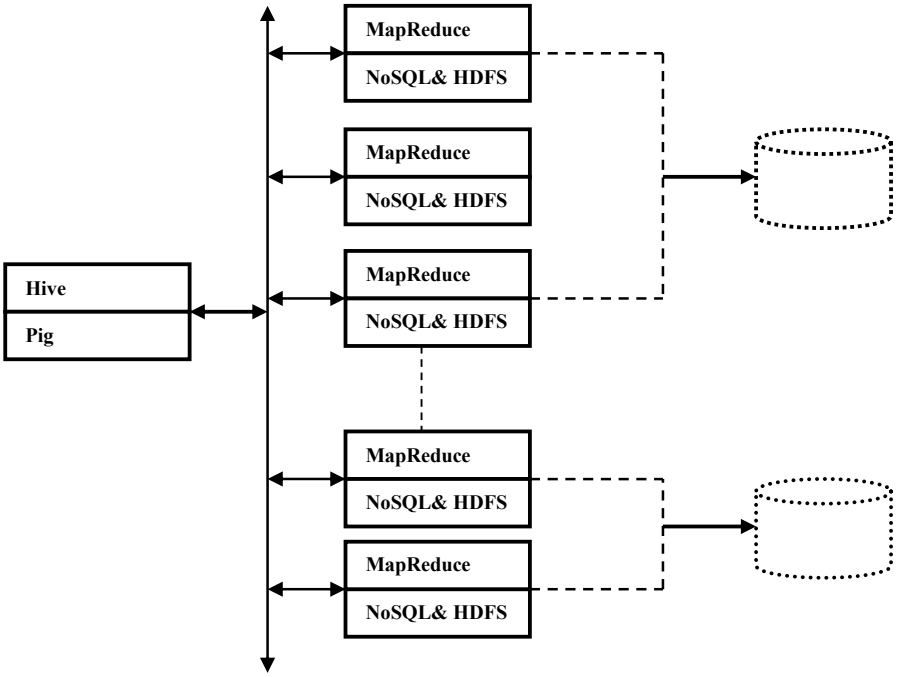

Fig. 4. NoSQL and HDFS file system compatibility for Apache Hadoop.

5. Conclusion

Big Data is an emerging problem for large companies and organizations, as massive volumes of data are being generated, examined, stored and analyzed. The demanding problems of big data can be categorized as issues pertaining to data variety, velocity (speed), volume, and veracity.

To handle these very demanding challenges, many vendors have grown and modernized a big data platform. However, in this paper, we have suggested a big data platform for large-scale data analysis by using the Hadoop/Map Reduce Framework and NoSQL and HDFS file system over scale out NAS. Simply, Hadoop/Map Reduce is batch-like, and not instantaneously desirable for real-time analysis, and not desirable to ad hoc queries. Hadoop figures out the volume and variety issues and so we still need to solve the speed outcome.

References

[1] C. Eaton, T. Deutsch, D. Deroos, G. Lapisand P. Zikopoulos, “Understanding Big Data: Analytics for Enterprise Class Headband Streaming Data”, McGraw-Hill, 2011

[2] Colin White, (July 2011). BI Research. 1st Ed. England: IBM Corporation.

[3] IBM Software, (August 2011). IBM Netezza Analytics. 1st Ed. NY 10589 U.S.A: IBM Corporation.
[4] M. Janson, (November 2011). Big Data and the New StorageArchitecture.1st Ed. New York: OnX Enterprise Solutions.
[5] Philip Russom, (2011). Big Data Analytics. 4th Ed. Renton, WA: TDWI.
[6] Apache Hadoop, Hadoop, HDFS, Avro, Cassandra, Chukwa, HBase, Hive, Mahout, Pig, Zookeeper (2013). Hadoop release. [ONLINE] Available at: http://hadoop.apache.org/. [Last Accessed 29 January 14].
[7] Alex Popescu (e.g. 2011). Achieve the Impossible in Real-Time. [ONLINE] Available at: http://nosql.mypopescu.com/post/6312810458/big-data-achieve-the-impossible-in-real-time. [Last Accessed 2 February 14].
[8] Hewlett-Packard (2011). Sustainable Business Advantage is using Vertica Analytics. [ONLINE] Available at: http://www8.hp.com/ch/de/pdf/IM_A_Advanced_Services_Vertica_4AA3-8467ENW_tcm_179_1247469.pdf. [Last Accessed 10 February 14].
[9] John Webster (2011). Understanding Big Data analytics. [ONLINE] Available at: http://searchstorage.techtarget.com/feature/Understanding-Big-Data-analytics. [Last Accessed 3 February 14].
[10] Kyar Nyo Aye (2013). Big Data Analytics on Large Scale Shared Storage System. [ONLINE] Available at: https://www.academia.edu/3502343/Big_Data_Analytics_on_Large_Scale_Shared_Storage_System. [Last Accessed 8 February 14].