ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MESURING THE EFFICENY OF USING HADOOP TO ANALYZE BIG DATA- A CASE STUDY ON TWITTER DATA SET

YOUSEF K. SANJALAWE[1], MOHAMMED ANBAR[2]

[1] National Advanced IPv6 Centre of Excellence, Universiti Sains Malaysia, Penang, Malaysia.
[2] National Advanced IPv6 Centre of Excellence, Universiti Sains Malaysia, Penang, Malaysia.
Email: josseph_hfs@hotmail.com[1], anbar@usm.my[2]

ABSTRACT

In last decades, the continuous enhancements of computational power have produced a massive data flow. Big data has been becoming more understandable as well as becoming more available. For instance, the famous online social networks, such as Facebook, and twitter, serve about 560 billion page views everyone's month. They also store new number of images near 3 billion each one month. This research emphasizes the solution and importance of big data problems using cloud computing (CC). Knowledge implanted in big data can be generated by personal computers PCs, mobile devices, and sensors, but it requires spending millions or billions of dollars in order to solve knowledge and information extraction problems to make suitable and critical decisions. In this research, we use Hadoop tool to analyze the big online data set with 8 GB size. The results show the importance of cloud approach to analyze the huge data with less efforts, cost, and time.

Keywords: *Big Data; Big Data Analytics*; *Cloud Service*; *Hadoop*; *Analyze Data Of Twitter.*

1. INTRODUCTION:

The Information Technology industry usually produces new technologies. Big data is considered one of them. With the huge developments of the cloud computing storage, big data has clearly attracted more attention. Because of the emergence of the Internet, the technology of big data will improve the innovation of the enterprises, lead the rapid revolution within the business field and build unlimited commercial opportunities.

Recently, we have been drowning in the ocean of data as a result of the development of the Internet, the Internet of Things (loT), Mobile Internet, and the Social Networks. An image that uploaded to Instagram has a size about 1MB; a video that uploaded to YouTube is about dozens of Mega sizes. Browsing websites, chatting online, shopping online, playing online games will be also stored as a data in any corner in the world. Hence, what is the volume of data in our daily life? According to IBM, 2.5 quintillion bytes of data have been created every day. 90% percent of that data was generated in the recent two or three years. That means, just in one day, we need 168 million DVDs to fill the data that appears on the Internet; also in one day, we sent about 294 billion emails, which equal to the numbers of printed newspaper in US for recent two years [1].

The volume of data, by 2012, has increased from the level of Terabyte to level of Petabyte. Producing supercomputers, and reducing price of computer hardware make it possible and easy, to some extent, to deal with huge and complex data. The data can be categorized into four main types: structured (like: trading data), semi-structured (like: blogs), unstructured (like: audio, and video), and multi-structured data.

Nowadays, a large quantity of data generated by multiple sources has been appeared. Dealing and working with this data has clearly a risen the so called "the problem of big data," that can be faced only and only with new computing platforms and paradigms. Different vendors compete in this sector, but on this day the standard platform for dealing with big-data is the Apache Hadoop that is an open-source framework.

Inspired by cluster platform of Google's, independent developers build Hadoop and following the published structure by Google's team, a complete group of components for the elaboration of big data has been developed. The Hadoop Distributed File System is one of these components, as it is considered as one of the core components.

In this project work, we will provide a comprehensive overview of big data and analyze Hadoop's behavior. With this overview, some problems and limitations in literature will be explored, then we will suggest some action points that can be applied to enhance the behavior of Hadoop. Finally, we will analyze a big data set taken from taken from KAIST institution [6] by using Hadoop, then interpret the results.

The main contribution of these research is to propose efficient model to be used in order to analyze big data. Hence, the main goals of this project is to explore big data technology and to provide a full review about the Hadoop as a platform for analyzing this type of data. Therefore, the project will:

- Explore the big data technology;
- Analyze Hadoop behavior;
- Discuss some of limitations in the big data analytics in the literature;
- Enhance the behavior of Hadoop by providing some action points; and
- Analyze the big data set taken from KAIST institution by using Hadoop and interpret the results.

This research will include the definitions of big data, big data attributes, and related works of using big data, our proposed model, analysis and results, and the conclusion and future works will be presented at the end of this article.

2. RELATED WORKS

2.1 Definitions of Big Data:

Big data fundamentally means not only a huge amount of data, but there are also other attributes that differentiate that concept from the other concepts of very large data and massive data. In fact, various definitions to the term of big data are existed in the previous literature, and three kinds of definitions play a significant role in viewing big data as the followings [1, 2, 3, and 4]:

1. *Attributive Definition*: IDC is a prospector in studying big data and its influences and impacts. It mainly defines big data as: ''big data describe a new evolution of architectures and technologies, economically designed and modeled to extract an important value from the very large amount of data, by using high-velocity discovery, capture, and analysis.'' This definition includes the four main salient characteristics of big data, (variety, velocity, amount, and value).
2. *Comparative Definition:* Mckinsey's report, in 2011, defined the term of big data as ''huge data sets which have size beyond the capability of Database software tools to analyze, manage, store, and capture.'', this definition does not define the term of big data with respect to any particular metric.
3. *Architectural Definition:* The NIST institution suggests that, ''Big data is where the data representation, acquisition velocity, or data volume limits the ability to perform efficient analysis using traditional relational techniques or requires horizontal scaling for effective processing.''

Big data can be particularly classified into big data frameworks and big data science. Big data frameworks are considered as software libraries with the related algorithms and procedures that provide distributed analysis and processing of big data problems, whereas big data science is defined as the study of techniques that provides the evaluation ,conditioning, and acquisition of big data. Table 1 summarizes the differences between the traditional data and big data.

Table 1: The differences between traditional and big data [3].

	Big data	Traditional Database
Generated rate	More rapid	Per hour, or day
Volume	TB or PB	GB
Data integration	Difficult	Easy
Data source	distributed	centralized
Structure	Semi structured or unstructured	structured
Access	Real time and batch	interactive
Data store	HDFS	RDBMS

2.2 Big Data Attributes

The three Vs (attributes) – volume attribute, velocity attribute and variety attribute - are widely used to describe various aspects of big data. Figure 1 summarizes these attributes. These three attributes make it simple to realize the nature of the data and the required platforms to analyze it.

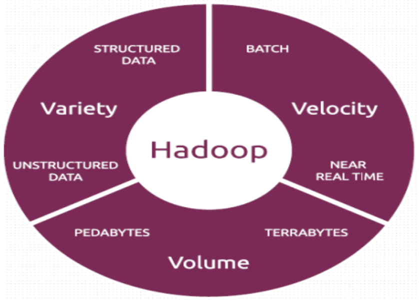

Figure 1: Big data attributes [5].

1. Volume

The most challenging aspect which related to Big Data is volume since it brings a need for scalable power and storage, as well as a distributed technique to querying. Large enterprises have already its own huge amount of data archived and captured over the years. That data can be in the form of record keeping, system logs, etc. The amount of this captured data gets easily to the point where classical DBMSs cannot handle it. Data warehouse may not always have the ability to analyze or process this type of data because of the lack of parallel processing framework. A lot of meaningful knowledge can be derived from log files, text data, and locations. For instance, consumer preferences, security investigations, patterns of email communications, and trends in transactions. Time-stamped and spatial data absorb quickly storage space. Different technologies of big data offer a convenient solution to earn value from this huge, difficult to use, or unused data [5].

2.3.2.2 Velocity

Data is rapidly flowing into organizations. Mobile and Web technologies have largely enabled gathering a data flow back to its origin (the providers). Online shopping has clearly revolutionized the interactions between provider and consumer. Also, online retailers can easily now keep logging of customer's interactions and can easily maintain the history and can quickly utilize this data in recommending products, thus will keep the organization on a suitable leading edge. Organizations of online marketing are currently deriving many advantages with the ability to instantaneously earn insights. With the current invention of the smartphones era, more location-based data captured, and it can be used to produce several advantages from this huge amount of data [5].

2.3.2.3 Variety

The data that gathered with digital and social media is seldom structured data. Unstructured audio and video data, text documents, financial transactions, images, interactions on social media are different examples of unstructured data. Traditional DBs support 'large objects' (LOB's), but it could face several limitations and challenges if not distributed. This data is complicated and complex to fit in conventional elegant relational DBMSs structures. On the other hand, big data tends to keep all the information since most of this is read many times and write once type of data. Big Data usually believes that there are internal insights hidden partially or completely in every bit of data [5].

2.3 Previous Works

MapReduce and Hadoop can be used to handle big data. The clusters with multiple nodes could be set up using these tools (i.e. MapReduce and Hadoop). Different files with different sizes can be saved in this cluster. A shared analysis of the Hadoop is performed. Size of Terabytes is generated and stored and analyzed as a file in the Hadoop cluster. Teragen, Teravalidate, and Terasort are used for generating, validating, and sorting the Tera file.

Cloud computing is also provided Big data as a service (Data-as service). Analysis becomes more important with this big in the environment of cloud computing. Traditional DBMS, Hadoop, appliance, and SSD are used to manage of data in memory. Hadoop offers the optimal service within the context of scalability, cost, and unstructured data [6].

In the past two decades, the continuous improvements of computational strength have produced massive flow of data. This massive flow of data is called as "big data" which cannot be analyzed by using existing techniques and tools. For instance, any website for social network like Facebook or twitter, monthly, operate about 570 billion accesses for its web page, accumulate three billion photos each month, and deal with twenty five billion pieces of matter [6].

Flickr, You Tube, Facebook, Google's search, and LinkedIn work based on a wrap-up of AI activities; as these web pages parse huge quantities of data and generating decisions instantly. Cloud computing is considered as one of the most powerful architecture for computing big data that effectively executes complex and large scale computing by integrating the available resources and presenting a one system view [7].

Big data is the captured of information from our life day by day. Data from Mobile devices, the Internet, Finance, Streaming, Sensors, and Science are the top six data drivers. With the fast rises in computing storage capacities and power, different organizations have improved their skills to deal with these massively diverse, noisy, large data sets generated from different sources [7]. Many researchers and organization define big data as the data that they are not able to work with using convenient approach, theory, and technology. Thus bring challenges for managing and analyzing data, and also for the whole IT industry [8].

3. RESEARCH METHOD

3.1 Proposed framework

In general, the framework of Hadoop and MapReduce can be used to deal with a variety of problems. Representing the input and output as a pair of <key-value> is the only one issue in the formulation to the problem. As data sets are contentiously increased in size, scalability of processing and computational time is still the main strengths of Hadoop. Performance can easily be enhanced with respect to time by adding additional nodes to the computer cluster.

Recently, telecommunication providers have seen a huge increase in the volume of data related to user activity generated in their own networks. This generated data, if utilized properly, can be used as an opportunity to earn competitive advantage. The proper utilization includes convert huge volume of data into information as well as analyze this information. Specifically, page ranking or subscribers ranking with respect to some specific predefined attribute, for instance, influence within the network, is very important applications.

Large number of analysis methods and algorithms of social network can be used to gain valuable information about the social network. One method is centrality method that can be used on the network to find the most central nodes. A common way of analyzing social networks information is by looking into separate for these centrality measures. By performing this only one single aspect of influence is taken into consideration, and some existing information is ignored within the network. A more accurate and precise analysis could be done by including as much available information as possible such as hidden patterns.

On the other hand, it is a big challenge for the operator to choose algorithms in order to be used as well as to evaluate the measures from all available algorithms to pinpoint the influential users or to pinpoint the interesting segments of users or the network as a whole. To deal with these challenges, algorithms of machine learning ML are implemented. The main idea is to produce a social network graph SNG from the given data set's information. After that, by using the produced SNG, different metrics is calculated and ML algorithm is implemented to create a model for handling each page. For example, this can be a classification model that aims to classify each page according to specific attribute. To build this model, for example, we need a training set that is constituted of the network's subset while the predefined attribute is exactly known. After that, this model can be easily used to get information about the pages of the network.

A parallel implementation is more beneficial to perform this process in very large networks to ensure scalability. Thus, we use Hadoop framework to implement the prototype, the proposed prototype is described below in Figure 2. Precisely, every box included in Figure 2 can be efficiently performed by using a MapReduce job, or at least in a few cases one or more iteration of several jobs.

The proposed prototype can be mainly divided into three parts:

- Pre-Processing;
- Naïve Algorithm; and
- Influence Weight Estimation.

The pre-processing includes filtering each record and keeping only the relevant records. This typically involves removing records that are having irregularly formatted information or having been missing elements. After that, filtered records will be an input to the procedure of generating. SNG, both a directed and weighted SNG and an undirected and unweighted SNG.

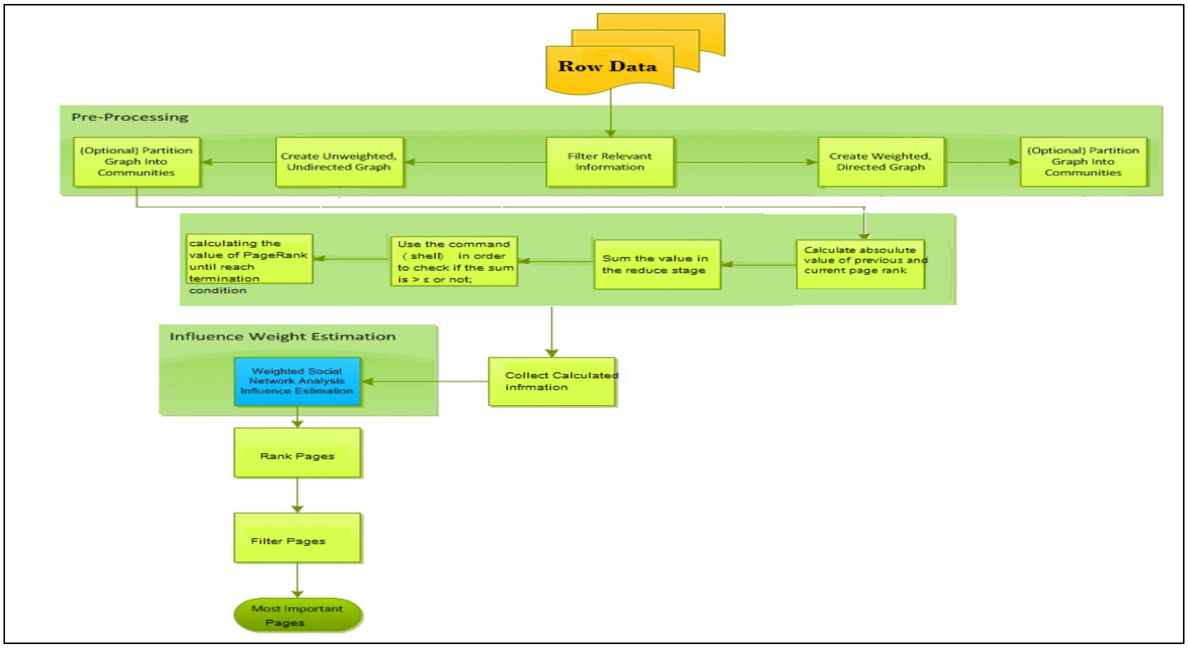
Row Data
Pre-Processing
(Optional) Partition Graph Into Communities
Create Unweighted, Undirected Graph
Filter Relevant Information
Create Weighted, Directed Graph
(Optional) Partition Graph Into Communities
calculating the value of PageRank until reach termination condition
Use the command (shell) in order to check if the sum is > ε or not;
Sum the value in the reduce stage
Calculate absoulute value of previous and current page rank
Influence Weight Estimation
Weighted Social Network Analysis Influence Estimation
Collect Calculated infrmation
Rank Pages
Filter Pages
Most Important Pages

Figure 2: The Proposed Prototype.

Influence Weight Estimation step performs data interpretation as much as possible, in order to mitigate this responsibility from the user. This detailed interpretation can be performed by common techniques of data mining, sometimes referred to as ML.

3.2 Data Set to Be Analyzed

Twitter provides an API which is Application Programming Interface that offers easy to gather and crawl data. We are so fortunate to have Twitter that gives the huge number of user profiles that available for processing. It was gathered and crawled by analyst group of social media from KAIST institution [11].

- This data set is available for free and public download at the following website http://an.kaist.ac.kr/traces/WWW2010.html
- FOLLOWERS as well as USERS are represented by ID as numeric integer. So we can directly access any users' profile (such as: XYZ) by using its ID like: http://api.twitter.com/1/users/show.xml?user_id=XYZ.
- In order to collect user profiles, a researcher that called Kwak started with Perez Hilton profile that has more 1 million followers. Twitter rate limit's id 20,000 requests per hour per whitelisted IP. Twenty machines were used with different IPs with different collection rate at 10K requests per one hour. The data set was collected 6th of July, 2009 to 31st July, 2009 with total size of 8 Giga bytes.

3.3 Hadoop

The Apache's definition of Hadoop is: "The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly available service on top of a cluster of computers, each of which may be prone to failures" [8].

Hadoop was basically inspired by researches published by Google, focusing on its mechanism to manage an avalanche of data, after that, it became the standard for processing, analyzing, and storing hundreds of Terabytes, and Petabytes of data.

Hadoop has basically drawn the revelation from GFS (Google's File System). In 2006, Hadoop was initially spun from Nutch to be considered a Lucene's sub-project and was directly renamed after that to Hadoop. Yahoo has been a contributor to the evolution of Hadoop. By 2008, the index of yahoo search engine was being generated by ten thousands core cluster of Hadoop [9].

Hadoop has invented a new manner of processing and storing data. It does not rely on high efficiency and expensive hardware. In addition, it leverages on several benefits from using parallel and distributed processing of very huge volume of data across low-cost servers. This infrastructure processes as well as stores the data, and can scale to easily changing needs. Hadoop provides limitless capability of scale up and no data (theoretically) is too big to process with distributed architecture [10].

Hadoop is mainly designed to run effectively on hardware and can scale down or up without interruption. Hadoop consists of three functions:

- Processing;
- Resource management; and
- Storage.

Now, Hadoop is used by big organization such as LinkedIn, Twitter, Facebook, Yahoo, and eBay. Traditional analytics systems and data storage were not initially built keeping in mind the main needs of dealing with big data.

4. ANALYSIS AND RESULTS

Different measure used in these research including: Ego betweeness centrality, Ego betweeness cluster, cluster relative speed-up, eigenvector centrality, and computational time.

The degree centrality is a local property and the ego degree centrality of ego is the same as the degree of the actor in the whole network there is no issue. At the other extreme closeness is about the connections from an actor to all other actors in the network and so is simply not applicable to ego networks" [12]. Eigenvector centrality is "similarly problematic, the power of eigenvector centrality is that it takes account of the connectivity of the alters, but these are unknown outside of the ego network and so this undermines the fundamental principles behind the measure" [12]. Betweenness examines |the extent to which an actor is between all other actors within the network. If an actor is between two other actors then it follows that there is not a connection between the alters on the path connecting them (otherwise this would form a shorter path)" [13]. On the other hand, the computational time is "the efficiency and velocity in both receiving and processing the data".

Results about analyzing of the Dataset using the platform of Hadoop are discussed below.

4.1 Running on One Single Cluster (Machine)

Hadoop's tests have been initially performed on one single cluster (machine) such ego centrality betweenness test. The results of Betweenness Test are presented below in figure 3.

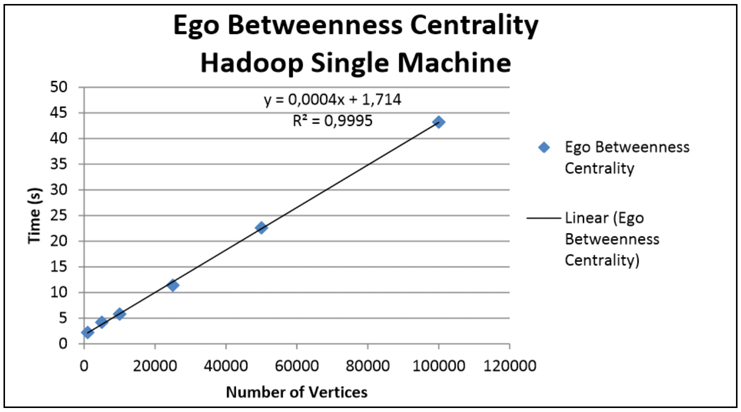

Figure 3: Betweenness Test.

We can notice the computational time and the relation between Time and Number of Vertices

4.2 Applying Hadoop Application in Platform of Multi-Clusters Mode

4.2.1 Scalability characteristic based on number of used machines

To determine the reliability characteristic that could be result by adding further machines to a computer cluster in order to minimize computational time, more than one test with a different number of machines must be performed.

for finding ego betweenness in Hadoop which actually ran on one single machine.

Moreover, the computational time of a function of the count of used processors in the cluster is shown in figure 4 in sixteen for ego centrality for degree centrality. A 41.7 million user profiles with 106 million tweets generated using the Erdos-Renyi random graph model. For the ego centrality test, an even decrease is very clear.

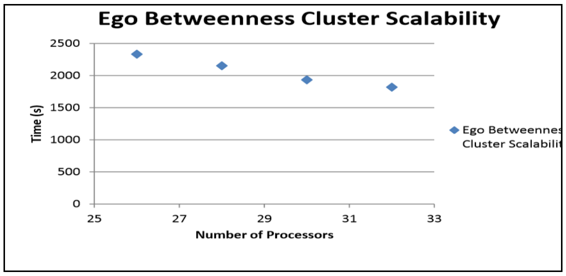

Figure 4: Scalability of cluster.

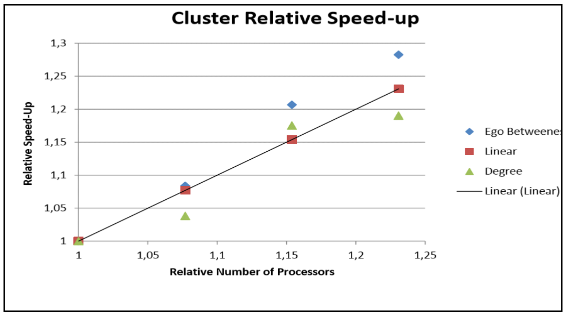

Figure 5: Relative Speed Up.

4.2.2 Full cluster tests

In addition, for further test of the scalability characteristics, different experiments have run in a cluster. In these tests, all user profiles have been used as input by using Erdos-Renyi algorithm or uniform distribution. Different algorithms have been tested (ego betweenness algorithm, degree centrality algorithm, betweenness centrality algorithm, and eigenvector centrality algorithm). The results are summarized below in figure 6.

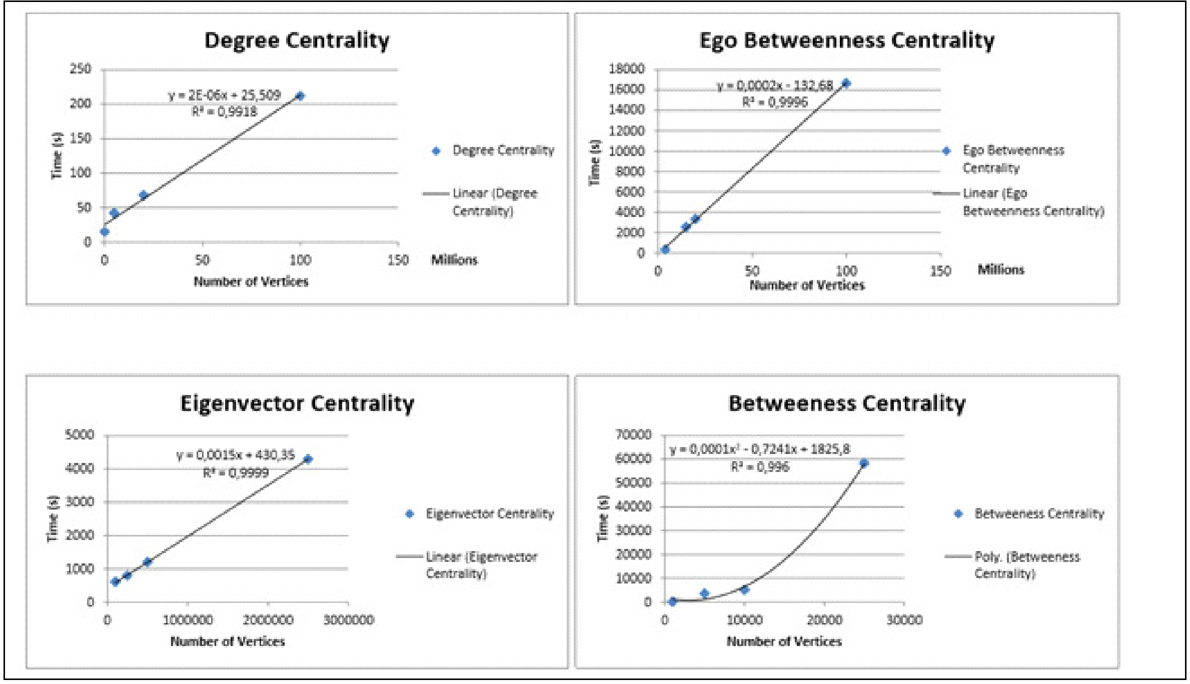

Figure 6: Test Results.

The results imply that the fastest algorithm is the degree centrality is. In order to calculate this measure for a 41.7 million user profile's vertices, we just need less than four minutes. Figure 4 indicates that ego centrality algorithm is also able to process a 41.7 million user profile's vertex size. However, the computation's time are larger than, by 2 orders of magnitude, for the degree centrality, as it needs 5 hours for 41.7 million user profiles vertices. On the other hand, Eigenvector centrality indicates even more tests and time have not been applied on user profiles more than 250 thousand vertices. A scalability issue is clearly linear for all selected metric algorithms, depending on number of vertices, is apparent in the dataset. This linearity feature is not, at all, shared by the selected centrality algorithm.

4.3 Applying Proposed Prototype

4.3.1 Computational time

This subsection presents the obtained results when performing the prototype on our data set. The data consists of 41.7 million user profiles, about 1.47 billion social relations, 4,262 trending topics, and 106 million tweets. Table 2 presents most of the tasks applied and the required computational time to each task.

Table 2: Computational Time, and # of Iteration for each process in Prototype

Process	Time (S)	# of Iteration
Building unweighted, and undirected graph	1101	1
Compute Ego centrality, and Unweighted degree for unweighted graph	4300	1
Compute Eigenvector centrality for unweighted graph	123432	110
Compute Ego centrality, and Unweighted degree for weighted graph	2001	1
Compute Eigenvector centrality for weighted graph	42	1
Compute Eigenvector centrality for unweighted graph	134233	46
Total	265109	160

Most of these tasks can be executed in a time of one hour or less. However, the iterative jobs are the two exceptions, computing of weighted and unweighted eigenvector centrality that need more than twenty-four hours for performing these tasks.

4.3.2 Testing different algorithms of Machine Learning (ML)

This sub-sections is mainly devoted to the performance of the ML mechanisms used to build models for ranking user's profile of the Twitter. Three categories of techniques have been used before ranking pages, namely decision trees, regression techniques, and neural networks. The result for each technique is shown below in figure 7. The training set contains of 2/3 of the user's profile in the selected data set. These training dataset were randomly chosen, and the remaining portions of data were been used as a validation dataset. The optimal technique in terms of F-value from the decision tree methods was the REP Tree, with a %49.6 F-value. However, if we want to classify any user profile's instances, the optimal techniques are ADTree and J48 algorithms with accuracy about 73.4931 %.

According to F-value, fewer accurate results have been obtained from the regression technique. Logistic has adjusted with more than a %44 F-value. In addition, Simple-Logistic algorithm gives the highest accuracy for classifying any instances of user's profile with % 73.8193 accuracy.

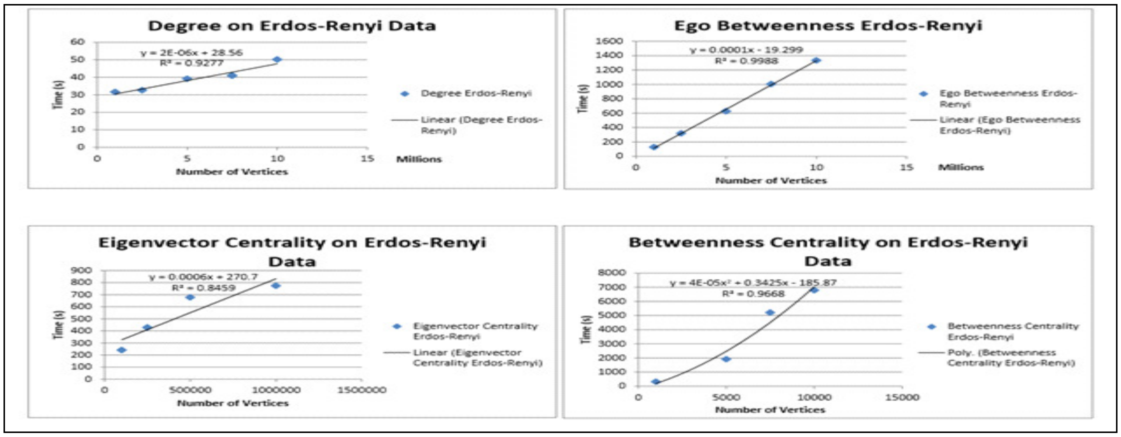

Figure 7: the required time for each ML technique.

4.3.2 Applying naïve method

In TERM of efficiency, we firstly focused on the performance of naïve method on task workload and time cost. In respect to evaluation of time cost, each method requires less time to completely finish the process, and producing is absolutely considered as a best method.

Moreover, the least workload will absolutely produce a faster result. Figure 4 presents the plotted value from the log of residual for every iteration. Residual is computed as the total difference between current values of PageRank from previous one. For better visual representation, the log value is used instead of the real value. This representation helps to check how the naïve method reaches convergence as well as the rate of convergence.

From figure 8, we can see that the naïve method has faster convergence rate in first 8 iterations and after this iterations, it get slower in gradual way. Furthermore, naïve method needs longer time to converge.

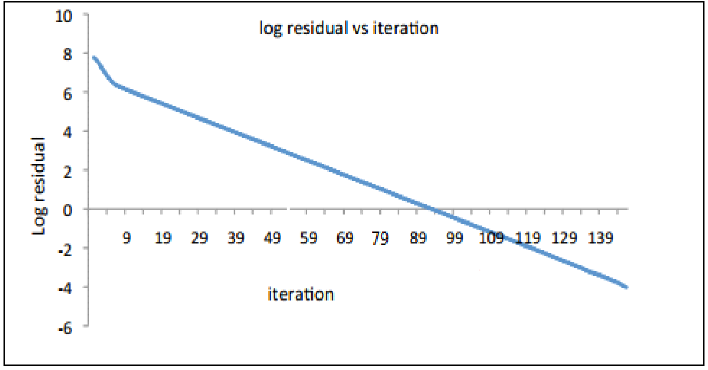

Figure 8: residual per iteration.

By looking to Table 3, we can take a look on how many iteration could be stored in order to calculate the vector of PageRank. This method needs 147 iterations to calculate the value of PageRank within 162,846 seconds.

The overhead cost measures how much is cost to get fewer number of iteration and less time to finish the overall task. Overhead for naïve convert from edge to adjacency list initialize the value of PageRank. It consumes about 45 hours to complete calculating the vector of PageRank.

Table 3: Statistics of Naïve method.

# of iterations	Saved iterations	Time to complete (Seconds)	Overhead Cost (Seconds)
147	0	162,846	1,142

The descriptive statistics analysis on time for each iteration is presented below in table 4. On the other hand, the size of sample is strictly limited to number of iterations for naive method.

Table 4: descriptive statistics analysis.

Mean	Std Deviation	Std Error	Confidence Interval		Minimum	Maximum
			Lower Bound	Upper Bound		
778.67	78.060	6.438	765.95	791.40	615	947

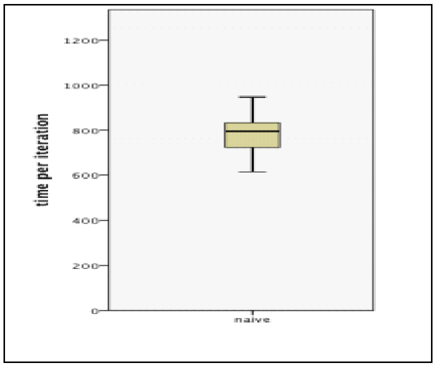

Figure 9: Time per iteration.

Figure 9 illustrates box plot to compare dataset side by side. Box plot graph is useful to get early indication about the data skewness and symmetry.

Table 5 shows the required input into recalculation step. Naïve method requires 41,589,858 inputs for every recalculation step, because recalculation step is considers as an iterative implementation. In order to get high-efficiency score and faster results, we have to minimize the workload.

Table 5: Recalculation.

Number of inputs into recalculation stage
41,589,858

To check the distribution to the data, we use Normal Q-Q plot, because some statistic tests might be misleading. Figure 10 shows the plotted Normal Q-Q plot for naïve method.

We could say that the data had normal distribution if it has a straight line of distribution. In our case, the data not lie perfectly on the main line, but some of the collected data are on the main line, so there are some outliers appear out of the line. Those outliers can effect upon the results of statistical test. They may come from the machine that used for processing, from other applications, or and from other Hadoop tasks that are running at the same time or in parallel with our experiment. This situation is considered normal, in reality, because the cluster of Hadoop could be used for another job too not exclusively for the calculation of PageRank. As result, the time per iteration for naïve method come from normal distribution.

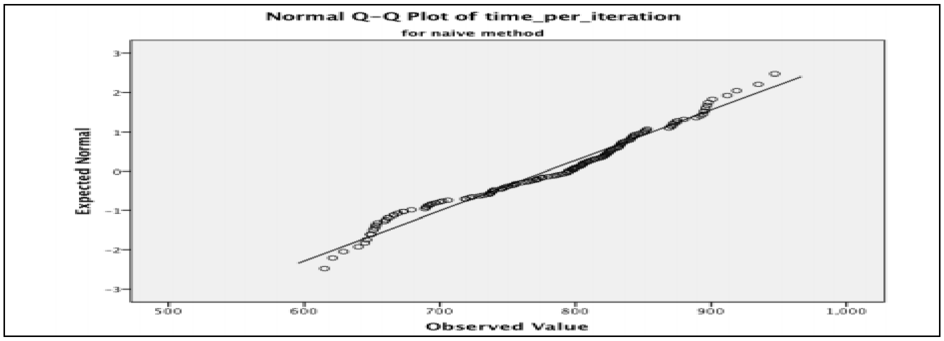

Figure 10: Q-Q plot

After calculating PageRank value, the pages will be sorted according to this value. Table 6 presents the top ten pages according to its PageRank value.

Table 6: The top 10 Ranked pages

Rank	Page name
1	Barack Obama
2	Number 10 Gov
3	Whole food
4	Britney spears
5	Omnomnome
6	Zappos
7	The onion
8	BJ Mendelson
9	Threadless
10	Radio blogger

As result, Big Data analytics is a new research area and there is lot of scope of research in this area. Big Data is a very challenging research area. Data is too big to process using conventional tool of data processing. Academia and industry has to work together to design and develop new tools and technologies which effectively handle the processing of Big Data. Big Data is an emerging trend and there is immediate need of new machine learning and data mining techniques to analyze massive amount of data in near future.

5. CONCLUSION

An urgent need for using advanced and accurate data acquisition, analysis, and management mechanism's management has become important with the new era of big data. In this research, we have proposed the concept of big data, the value chain of big data, which totally covers the entire life cycle of big data. The value chain of big data contains four phases: data generation, acquisition, storage, and analysis.

In addition, from the system perspective, we have presented a literature for management systems of big data. Those systems can be distributed file systems or semi-structured and non-structural data storage. The distributed file systems provides high performance using and access to data , it is considered as tool for managing big data pools and supporting analytics applications of big data. On the other hand, semi-structured storage is a repository product to store data that doesn't reside inside a relational DB, but that does have organizational characteristics that allow easier analysis.

Furthermore, in this research we analyze large-scale data about twitter created by KAIST institution by using Hadoop Framework. The main objective of this study is to use Hadoop to analyze this data. Naïve method was used to compute PageRank value for these data sat. The findings of this research focus on the role of cloud software for analyzing big data. Therefore, IT organizations, whether large or small scales are encouraged to use big data infrastructure, as it can use cloud-based techniques to analyze and solve their big data problems in less cost, less time and efforts.

Hadoop showed significant promises of a capability to handle and analyze very huge data sets and to reduce computational time in a linear manner. Following that, more than one machine was used, forming a cluster for the Hadoop's platform. Test results indicate that for sufficiently huge tasks, Hadoop basically manages to use more of the capability of every cluster and gain a speed increase as well as an increase of computational power.

6. FUTURE WORK

One is using multi instances (clusters) Hadoop for load balancing in social network's servers.

These enhancements can easily be implemented by using the enhanced algorithm of path resolver to running paths that have a specification of which cluster to use. The interface to the file system can efficiently implement the logic for linking to specified instance. Another advantage is to support other types of file systems. This is simply achieved by implementing the interface of file system for that system. Also, we can develop a novel algorithm for Sentiment analysis to determine the sentiment value of social media datasets related to platform usages such as movie reviews personal blogs.

REFERENCES:

[1] Longbing Cao, *"Big Data Analytics - Innovation and Practices",* IEEE, 2015.

[2] Huansheng Ning, and David G. Belanger, "*Guest Editorial Special Issue on Big Data Analytics and Management in Internet of Things",* IEEE internet of Things Journal, Vol. 2, no. 4, August 2015.

[3] Omar El-Gayar, and Prem Timsina, "*Opportunities for Business Intelligence and Big Data Analytics in Evidence Based Medicine",* 47th Hawaii International Conference on System Science, 2014.

[4] Dominic Breuker, "*Towards Model-Driven Engineering for Big Data Analytics – An Exploratory Analysis of Domain-Specific*

[5] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li," *Big Data Processing in Cloud computing Environments*", 2012 International Symposium on Pervasive Systems, Algorithms and Networks.

[6] http://an.kaist.ac.kr/traces/WWW2010.html

[7] "*Big Data: Science in the Petabyte Era*", Nature 455 (7209): 1, 2008.

[8] http://hadoop.apache.org/.

[9] Blazhievskv S., *"Introduction to Hadoop, MapReduce, and HDFS for Big Data Applications"*, SNIA education.

[10] Fernández A, et a;.(2014), "*Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks",* WIREs Data Mining Knowledge Discovery.

[11] http://an.kaist.ac.kr/traces/WWW2010.htm

[12] Everetta M, Borgattib S, (2004), "*Ego network betweenness*", science direct.

[13] Freeman, L.C., 1982. Centered graphs and the construction of ego networks. Mathematical Social Sciences 3, 291–304.