www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 5 Issue -02 February, 2016 Page No. 15835-15840

Determination of cloud adoption factor and Subordinate Virtual Cloud mechanism to overcome the performance issues in bigdata

Krishnamoorthy.V $^{\#1}$ ***Guhan.A*** ^{*2} ***Jai Vignesh.T.R*** ^{*3}

UG Scholar , Department of CSE

Paavai Engineering College, Namakkal , Tamilnadu, India

krishnamoorthy.95.vijayan@gmail.com

UG Scholar , Department of CSE

Paavai Engineering College

UG Scholar , Department of CSE

Paavai Engineering College.

***Abstract*--Any large unstructured data sets with sizes beyond the ability of the software tools to manage and process within a tolerable elapsed time is rightly observed as bigdata. Cloud computing is delivery of on demand computing resources from application to data center over internet. Combining these two strong reliable platforms helps in tackling extraveneous real time problems and obtaining solutions for it. Cloud embedded bigdata supports inexpensive reliable storage and tools for analyzing structured and unstructured, semi streaming, click streaming and various types of data. The existing system tends to be more costlier because of cloud deployment costs and it is not elastic in nature. The subjective nature of cloud delivery to incoming data streams pulls back the efficiency of the system. The paper aims to minimize the cost for cloud adoption by determining the cloud adoption factors from Net present value computation and derives a mathematical expression for 'α'(Cloud adoption factor). It also addresses the issues that affect the performance issues of bigdata by implementing subordinate virtual cloud mechanism to overcome the addressed bottlenecks.**

Keywords--
Bigdata, cloud computing, cloud adoption, map reduce, virtual cloud, Subordinate virtual cloud, datasets.

I. INTRODUCTION

Data sets grow to greater size because of cameras, microphones, radio-frequency identification (RFID) readers, information-sensing mobile devices, aerial sensory technologies remote sensing, software logs, and wireless sensor networks.[1] Relational database management systems and desktop statistics and visualization packages often have difficulty in handling bigdata. This work in may require massively parallel software running on hundreds or even thousands of servers which cannot be provided practically. What is considered "bigdata", varies depending on the capabilities of the users and their tools.[2]The above fact makes bigdata a moving target. Thus, what is considered to be "Big" in one year will become ordinary in later years. Nowadays cloud applications process large amount of data to provide the desired results. Data volumes to be processed by cloud applications are growing much faster than computing power. After a deep analysis on information storage capacity, Hilbert et al submitted a report stating that 90 % of the data in the world today has been created in the last two years alone. The data growth rate forecast for images as said by Hilbert is shown in figure 1.

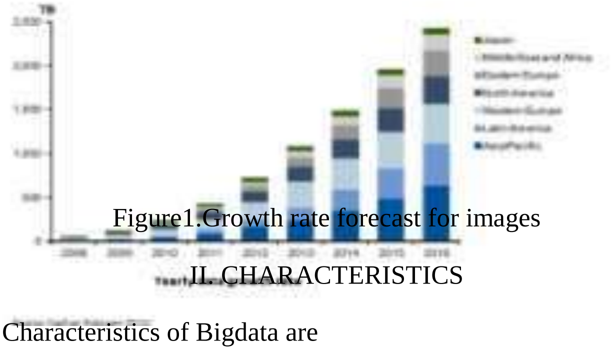


Figure1.Growth rate forecast for images

II. CHARACTERISTICS

Characteristics of Bigdata are
Veracity: Various types of data faced.
Velocity: Rate at which data is changing.
Volume: Amount of data to be managed.
Value: Deals with descriptor of data.
Variability: Ensuring the consistency of the data.
Veracity: Describes the provenance of the data [3].
Visualization: Presentation of data in a pictorial format. [4]

DOI: 10.18535/ijecs/v5i2.29

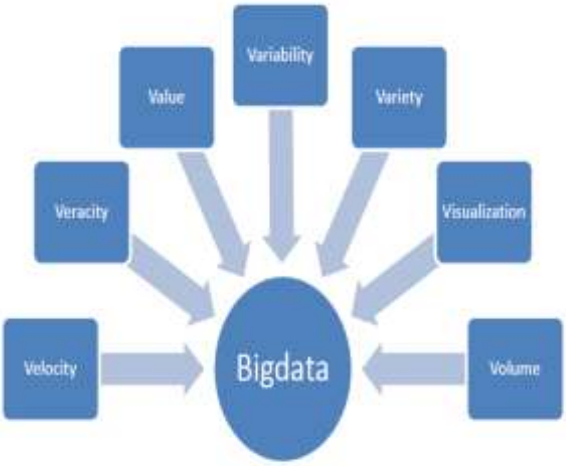


Figure 2.Characteristics of Bigdata

III. RELATED WORKS

In 2001 Research report by META Group Doug Laney after his immense research in data addressed that data growth has three dimensions which paved way for Gartner to develop his “3Vs” model for describing big data.[6]

In 2004, Google published a paper on a process called MapReduce.It is a programming model and an associated implementation for processing and generating large data sets. [6].

In 2012, Gartner updated its definition as “Bigdata is high volume, high velocity, and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”.[7]

IV. RECENT LITERATURE SURVEY

In ASEE 2014 Zone I Conference, University of Bridgeport, Bridgeport, CT, USA Dr.Amir Esmailpour in his paper [8] investigated the key features of big data as formation of clusters and their interconnections along with their connections to the databases. Jenani nelson in his content[9] described about the algorithms for bigdata. They are sketching and streaming, dimensionality reduction, Numerical linear algebra, Map Reduce, Apache hadoop, hadoop framework. Eswara Krishna iyer in his paper Cloud computing and modeling of cash flows for Full versus Fractional Adoption of cloud. Found out a mathematical equation with which Net Present Value for cloud adoption can be computed. This supported the paper to a greater extent. Based on this NPV (Net Present Value), the cloud adoption factor ‘α’ is derived.

V. METHODOLOGY IN EXISTING SYSTEM

As mentioned by Jelani nelson , the existing system first fetch the incoming data stream with the aid of sketching and streaming algorithm and reduces the number of random variable under consideration by applying it in dimensionality reduction.

Numerical linear algebra helps in decomposition and facilitates map reduce. Hadoop is a powerful open source platform that assist in handling large data volumes .For the adoption of cloud, the existing system uses the net present value computation by Eshwara Krishna Iyer that helps to adopt cloud according to the requirement either in a full (or) in a fractional manner.

A. Determination of Net Present Value

The derivation of Net Present Value is as follows.

NPV Modeling: Adoption of cloud

NPV Modeling for Full vs. Fractional Adoption of Cloud

$$NPV = PV\text{-}I \qquad (1)$$

NPV = Net present value

PV = Present value of Future Cash Flows

I = Total Upfront Total Capital Investment (CAPEX)

$$I = I_{total} - \alpha\delta I_{it} \qquad (2)$$

I_{total} = Maximum Upfront Total Capital Investment in the absence of cloud (i.e. α=0)

α = Percentage adoption of cloud by the market/firm (from the technology that can be moved to cloud)

δ = Percentage of what can actually be moved to the cloud today from the total universe of IT assets

I_{it} = Maximum Upfront IT Capital investment in the absence of cloud (i.e. α=0)

$\alpha\delta I_{it}$ = is the fraction of total IT investment that can be deferred because of moving to cloud

$$PV = \frac{(R-O_{nc}-O_{c})}{r} \qquad (3)$$

R = Annualized perpetual cash inflow (excluding IT operation costs)

O_{nc} = Traditional in-house IT Operational Costs which continue even after cloud adoption

O_{c} = IT Operational costs incurred because of cloud adoption

r = Discounting rate computed using Weighted Average Cost of Capital (WACC)

$$O_{nc} \approx b\alpha^{2} - 2b\alpha + a \qquad (4)$$

Where $O_{nc(\alpha=0)} = a$ (maximum value of O_{nc} in absence of cloud adoption)

$O_{nc(\alpha=1)} = a - b$ (minimum value of O_{nc} on complete cloud adoption)

$$O_c = \alpha\,(Y_k + Y_{uk}) \qquad (5)$$

Y_k = Annualized pay-outs to to the vender for

cloud utilization (at α=1)

Y_{uk} = Non cash-yet 'monetizable'– unknown

risk component associated with cloud adoption (at α=1)

Substituting equation (2), (3), (4) and (5) in equation (1)

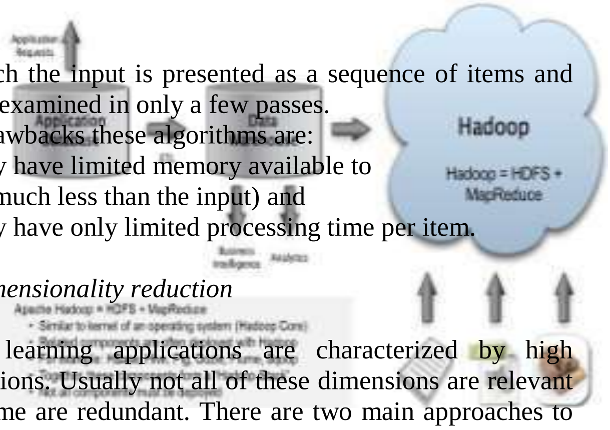
$$NPV = \frac{-b}{r}\alpha^2 + \frac{[r\delta I_{it} - (Y_k + Y_{uk}) + 2b]}{r}\alpha + \frac{(R - a - rI_{total})}{r} \qquad (6)$$

This is the equation for determining the Net Present Value

Equation (6) is the Net Present Value for cloud utilization. It could be noted in that the NPV is inversely proportional to the discounting rate 'r' and directly proportional to'α', the cloud adoption factor.

B. Sketching and Streaming

A sketch is with respect to some function f, and a sketch of data set x is a compressed representation of x from which one can compute f(x). [10]Streaming algorithms are algorithms for processing data streams in which the input is presented as a sequence of items and can be examined in only a few passes.
The drawbacks these algorithms are:
1) They have limited memory available to them (much less than the input) and
2) They have only limited processing time per item.


Application Requests
Application
Data
Hadoop
Hadoop = HDFS + MapReduce
Business Intelligence
Analytics
Apache Hadoop = HDFS + MapReduce
Similar to kernel of an operating system (Hadoop Core)

C. Dimensionality reduction

Many learning applications are characterized by high dimensions. Usually not all of these dimensions are relevant and some are redundant. There are two main approaches to reduce dimensionality feature selection and feature transformation. Dimensionality reduction can be achieved either by feature selection or transformation to a low dimensional space. Feature selection also known as variable selection is the problem of selecting a subset of the original features.[11] Feature extraction transforms the data in the high –dimensional space to a space of fewer dimensions.
The drawbacks of this algorithm is
1. High computational costs.
2. A feature that is not useful by itself can be very useful when combined with others.

D. Numerical linear algebra

Numerical linear algebra deals numerical algorithms for solving problems in Linear Algebra, such as linear algebraic systems and corresponding matrix eigen value problems. It includes computing algorithm of LU decomposition, QR decomposition, eigen values. It helps in signal processing and handling computational science problems. [12].

E. Map reduce

Jeffrey Dean and Sanjay Ghemavat [6] in their paper "MapReduce" defined it as an programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key as shown in figure 3.

Map Reduce runs on a large cluster of commodity machines and is highly scalable a typical Map Reduce computation processes many terabytes of data on thousands of machines. [6]

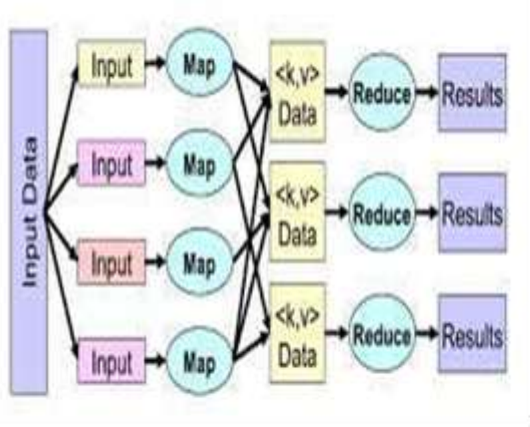


Figure 3 Map Reduce

F. Hadoop common

Hadoop common also known as Hadoop core is the set of common utilities and libraries that support other Hadoop modules. It is an essential part or module of the Apache Hadoop framework, along with the Hadoop Distributed File System (HDFS), Hadoop YARN and MapReduce [13]

G. Hadoop Distributed File Systems

HDFS is a distributed file system that provides high-performance access to data across Hadoop clusters.

DOI: 10.18535/ijecs/v5i2.29

Figure 4 Lay-out of HDFS

Like other Hadoop-related technologies, HDFS has become a key tool for managing pools of big data and supporting bigdata analytics applications and the layout of HDFS is as shown in the figure 4.[14]

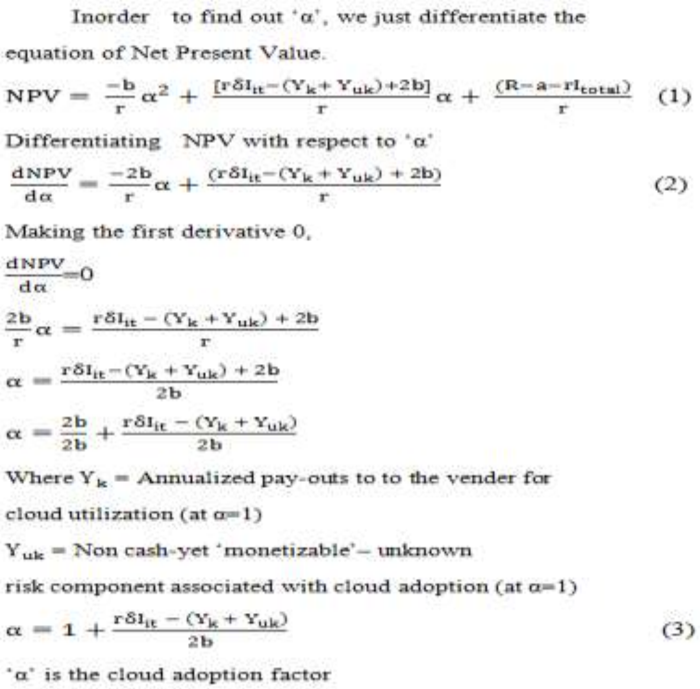
VI. PROPOSED SYSTEM

The existing system bores drawbacks in sketching and streaming algorithm, dimensionality reduction and exists as a subjective system. These addressed bottlenecks could be overcome by adopting following strategies

1. Selecting suitable organisation scope with appropriate cloud delivery deployment model to aid implementation of master.
2. The determination of a mathematical expression for α cloud adoption factor helps in minimising the cost of cloud to a certain extent.
3. Master segregate the different types of data streams and routes into appropriate v-clouds with the help of Map Reduce
4. The aggregated and reduced data stream is fed to implemented master cloud which maintains Forward index table.
5. Deploying SVC mechanism enables dynamic establishment of autonomous, elastic, consistent and scalable system.

A. Determination of cloud adoption factor –'α'

From the net present value equation, derived by Eshawara krishna Iyer, it is very clear that the Net Present Value depends upon the factor called 'α' which is called cloud adoption factor.



Inorder to find out 'α', we just differentiate the equation of Net Present Value.

$$\mathrm{NPV} = \frac{-b}{r}\alpha^2 + \frac{[r\delta I_{it}-(Y_k+Y_{uk})+2b]}{r}\alpha + \frac{(R-a-rI_{total})}{r} \quad (1)$$

Differentiating NPV with respect to 'α'

$$\frac{d\mathrm{NPV}}{d\alpha} = \frac{-2b}{r}\alpha + \frac{(r\delta I_{it}-(Y_k+Y_{uk})+2b)}{r} \quad (2)$$

Making the first derivative 0,

$$\frac{d\mathrm{NPV}}{d\alpha}=0$$

$$\frac{2b}{r}\alpha = \frac{r\delta I_{it}-(Y_k+Y_{uk})+2b}{r}$$

$$\alpha = \frac{r\delta I_{it}-(Y_k+Y_{uk})+2b}{2b}$$

$$\alpha = \frac{2b}{2b} + \frac{r\delta I_{it}-(Y_k+Y_{uk})}{2b}$$

Where Y_k = Annualized pay-outs to to the vender for cloud utilization (at α=1)

Y_{uk} = Non cash-yet 'monetizable'– unknown risk component associated with cloud adoption (at α=1)

$$\alpha = 1 + \frac{r\delta I_{it}-(Y_k+Y_{uk})}{2b} \quad (3)$$

'α' is the cloud adoption factor

With the aid of the equation (3),the cloud adoption factor could be determined by knowing the value such as Y_k, the annualized payouts to the under for cloud utilization and Y_{uk}, non-cash get monetizable.

B. Master cloud implementation

The Master cloud is the heart of the entire system. It is the biggest virtual cloud of the system. It act as the primary support for accepting the incoming data streams. It is ever ready to intake any voluminous amount of data that are uploaded to the cloud. It accepts the incoming data streams and segregates it based on the types of data and store it in corresponding unique virtual cloud allotted for it. The another important work of the master cloud is to maintain the log of data stream. The master cloud after segregation of the data will route it to the corresponding cloud. The master has to keep track of the history of forwarded information. It also maintain a table called forward index table which contains the details such as where the information are forwaded and stored.

DOI: 10.18535/ijecs/v5i2.29

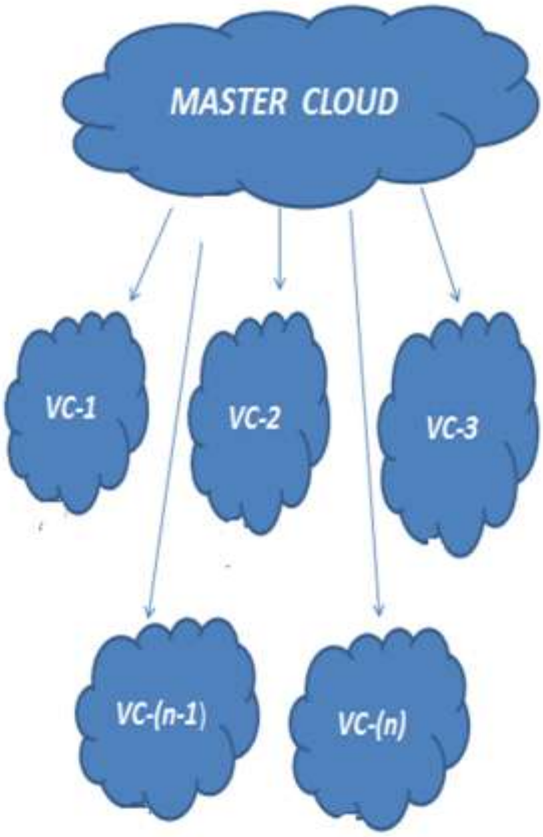


Figure 5..Master Cloud implementation

C. SVC mechanism

Subordinate Virtual Cloud Mechanism is a mechanism that makes the system more reliable. It helps to manage the memory of the Virtual Cloud as and when it is being filled. Sometimes, there may be a chance of meeting insufficient memory space in virtual clouds. This insufficiency in space may delay the rate of acceptance of data streams by master. If this situation persists unabated, there is every chance for the system to get collapsed. So inorder to overcome this anamoly, Subordinate Virtual Cloud mechanism is introduced.

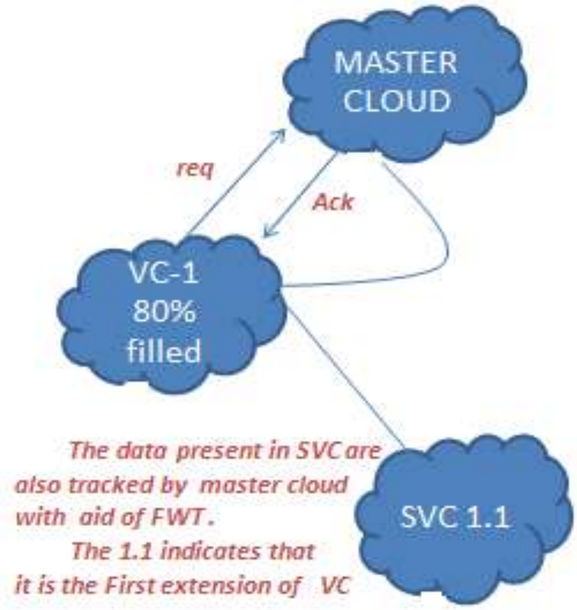

MASTER CLOUD
req
Ack
VC-1 80% filled
SVC 1.1
The data present in SVC are also tracked by master cloud with aid of FWT.
The 1.1 indicates that it is the First extension of VC

Figure 6. Generation of Virtual Cloud

It is an automated request generating and resource provisioning mechanism which solve the above said anamolies. The diagramatic representation of Subordinate Virtual Cloud mechanism is shown in figure This mechanism continually observes the free space of the virtual cloud that is being filled. Once if eighty percent of the total memory of the virtual cloud is filled, a trigger is generated and sent to the master. Now the master allocates another virtual cloud under the virtual cloud which generated the request. The virtual cloud so created in response for the generation of the request by another virtual cloud is called as Subordinate Virtual Cloud (SVC) and this phenomenon is called Subordinate Virtual Cloud mechanism. A system with Virtual Cloud and Subordinate Virtual Cloud is as shown in figure.7

DOI: 10.18535/ijecs/v5i2.29

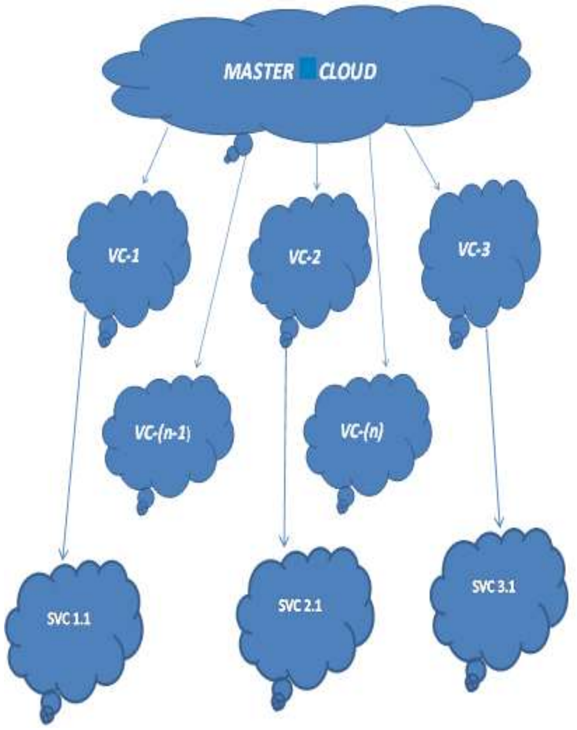


Figure 7. A system with Virtual Cloud and Subordinate Virtual Cloud

This can be illustrated with an example Consider pictures are stored in virtual cloud (vc-1) .Once if virtual cloud 1(vc-1) is 80% filled it sends an automatic request to master and the master allocates another Subordinate Virtual Cloud (SVC) to store the rest of pictures as the Virtual Clouds that are created by the master clouds whenever it is essential.

D. FWT(Forward index Table)

Forward index table is a data structure that contains details about where the information is being directed by the master cloud. It gives the exact location of data that is being searched and facilitating the retrieval. It keeps track of all data entries. Virtual Cloud-1 (VC-1) is filled after inserting pic 3 so pic 4 inserted into Subordinate Virtual Cloud-1.1(SVC-1.1)Keeps track of data present in which virtual and sub ordinate virtual cloud

Figure 8. Forward index Table

Data forwarded	**Location**
pic 1	MC:VC-1
pic 2	MC:VC-1
pic 3	MC:VC-1
pic 4	VC-1:SVC-1.1

VII. CONCLUSION

The present system bores the problem of limited space, limited processing per time and selection features. The proposed system overcomes these drawbacks and improves the processing speed to certain extent. Moreover rapid inflow of data and faster retrieval can be accomplished. The subjectivity of the existing system is converted into an autonomic system. This helps in provisioning and de-provisioning of resources as and when required. In this paper we have given an overview of Bigdata and algorithms used in existing system. The Net Present Value computation for cloud adoption is taken as base and cloud adoption factor ‘α’ is determined which will serve as guideline for minimizing the cost of cloud adoption. Above all the subordinate virtual cloud mechanism will make the system autonomic and performance could be improved to a notable level.

VIII. REFERENCES

[1] *Hilbert, Martin. “Big Data for Development: A Review of Promises and Challenges” In the proceedings of Digital Technology and Social Change on January 3 2015 ¸ university Of California.*

[2] Adam Jacobs. “The Characteristics that Define Bigdata “*In tbe proceedings of The Pathologies of Big data July 6, 2009 Volume 7, issue 6*

[3]IBM Watson. “Characteristics of Bigdata” *In the proceedings of Data discovery tools for every day business uses by IBM.*

[4] M. Ali-ud-din Khan, Muhammad Fahim Uddin, Navarun Gupta. “Seven V’s of Big Data Understanding Big Data to extract Value”.*In Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1) 978-14799-5233-5/14/$31.00 ©2014 IEEE*

[5] Doug Laney. “3d Data management: Controlling Data volume, velocity, variety”. *In the proceedings of Application delievery strategies* by META Group on Feb 6 2001

[6] Jeffrey Dean and Sanjay Ghemawat. “Map Reduce”. *In the proceedings of Simplified data processing on large clusters.*

[7] Shen yin , Okyay kaynak . “ Big Data for Modern Industry Challenges and Trends” *In the Proceedings of the IEEE Vol. 103, No. 2, February 2015*

[8] Dr.Amir Esmailpour, Gautam Siwach.. “Encrypted Search & Cluster Formation in Big Data”.*In the proceedings of ASEE 2014 Zone I Conference, April 3-5, 2014,* University of Bridgeport, Bridgpeort, CT, USA

[9**]** Prof. Jelani Nelson CS 229r Algorithms for Big Data available at http://people.seas.harvard.edu/~minilek/cs229r/

DOI: 10.18535/ijecs/v5i2.29

[10] Jelani nelson. “Sketching and streaming algorithms for processing massive data”. *In the proceedings of the International Conference on Business Management & Information Systems, 2012*

[11] Mahdokht Masaeli, Glenn Fung, Jennifer G. Dy.
“From Transformation-Based Dimensionality Reduction to Feature” *In the Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010

[12] R. Rannacher: Numerische Mathematik. 0 (Einf. in die Numerische Mathematik),
Lecture Notes, Heidelberg University, http://numerik.uni-hd.de/_lehre/notes/

[13] B.Thirumala Rao, N.V.Sridevi, V.Krishna Reddy L.S.S.Reddy. “Performance Issues of Heterogeneous Hadoop Clusters in Cloud Computing” *In the proceedings of Global Journal of Computer Science and Technology, Volume XI Issue VIII May 2011*

[14] "What is the Hadoop Distributed File System (HDFS)?". *ibm.com*. IBM. Retrieved Oct 30 2014