

SURVEY ON BIG DATA: ISSUES, CHALLENGES, TOOLS

M.Ashok Kumar
Ph.D Research Scholar
Dept. of Computer Science
Periyar University
Salem-11
williamashok@gmail.com

Dr. I.Laurence Aroquiaraj
Assistant Professor
Dept. of Computer Science
Periyar University
Salem-11
laurence.raj@gmail.com

P.Surya
Ph.D Research Scholar
Dept. of Computer Science
Periyar University
Salem-11
suriyaa14@gmail.com

ABSTRACT-- In this paper, we introduce about the big data and big data analytics as a basis. Big data is a popular term, It describes about the exponential growth and availability of data, both structured and unstructured. Big data analytics is the process of collecting , organizing large data sets containing a variety of data types.

Keywords--Big data; Map reduce;Hadoop

I. INTRODUCTION

In the world there are 7.28 billion people are there in Dec 2014.Data are increasing from various sources like social medias, youtube, users, applications and so on. So that the data are simultaneously increasing [WOR,2014]. From this statistic there are 5 billion people having mobile phones, and 2 billion people using internet [CHE, 2013]. When coming into big data it throws a great challenge to manage the information.Because it is inducing a great amount of data to examine.And also it will become more complex when it is coming into streaming analysis, which is questioning the information in the query. Whenever the volume of data is increased, the speed will be lessened[25]. This is the main disadvantage of big data. The big data having 3 V's mainly.They are Variety, Velocity, Volume.And it has 2 extra V's, they are value and veracity. There is no standard definitions for big data. Big data is a buzz word. Here some definitions are there, given by some organization. The Big Data Commission at theTechAmerica Foundation offers the following definition:

“ Big Data is a term that describes large volumes of high- velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information ” (TechAmerica Foundation, 2012) [JEA, 2013]

In Jul y 2000 by Francis Diebold of University of Pennsylvania in his work of Econometrics and statistics (2000):

“Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. In this new and exciting world, sample sizes are no longer fruitfully measured in “Number of observations,” but rather in, say, mega bytes.Even data accruing at the rate of several gigabytes per day are not uncommon.”

Researchers at McKinsey propose an intentionally subjective definition:

“ Big data refer to datasets whose size is beyond the ability of the typical database software tools to capture, store, manage, and analyze ” (McKinsey Global Institute, May 2011).

Meanwhile, Jerry Smith, Data Scientist Insights contributor, developed a mathematically sound

Definition of big data:

“‘Big Data’ represents the historical debris (observed data) resulting from the interaction of at between 70 and 77 independent variable/subjects, from Which non- random samples of unknown populations, shifting in composition With a targeted time frame, can be taken” (Smith, 2012). Fig1 represents what is big data as simple.

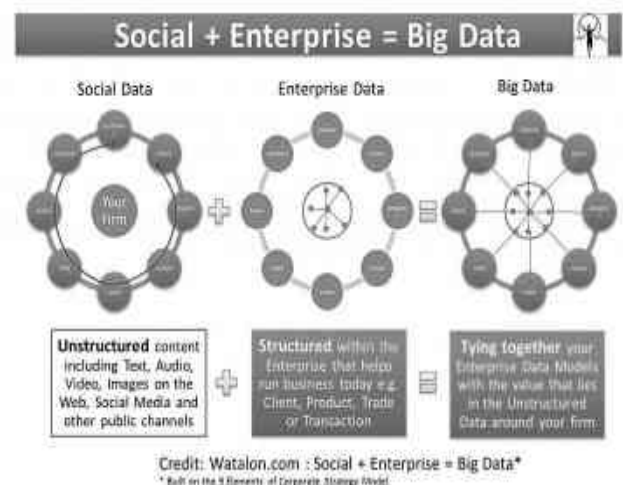


Fig1-Big data

Evolution of big data:

1) 1919- The U.S department of commerce partners and with IBM to perform an agriculture census. They want to keep 5000 employers record per month. By 15 month using more than 200 million IBM-supplied punched cards and processing equipment.

2) The Los Angeles police department uses IBM tabulating equipment to process criminals and arrest records.

3) 1934- The IBM flagship tabulating produced 405 electric machine. The 405 cloud process 150 80characters punched cards a minute, and print 80 alpha-numeric characters a minute.

4) 1937- "The biggest accounting part of all time" the U.S government partners with IBM implement the U.S social security act of 1935, which requires to maintain the 26 million American s.

5) 1948- The IBM 604 electronic punching calculating punch is introduced, and convinces the fast data calculation.

6) 1952- IBM introduces the revolutionary magnetic tape drive vacuum column to viable data storage medium.

7) 1965- more than 1800000 IBM punch records were stored and in that 144000000 pieces of information about the test children.

8) 1956- Introduces RAMAC(Random Access Method of accounting and control). It is the 1st magnetic hard disk, which stores 2000 bits per data square.

9) 1962- Two IBM 7090 mainframe form the backbone of the SABRE reservation system, this is the 1st airline reservation system which stores all the passengers data.

10)1969- IBM technology guides the Apollo mission to the system. Its processing the data and controlling the Saturn rocket until the Apollo safely headed on the moon.

11) 1970- Relational DataBase.

12) 1988- Increase the cpu capacity as 35%.

13) 2002- Oxford university join with IBM and the UK government to build sophisticated computing grid for discovering the breast cancer.

14)2005- In a collaboration with the national geographic society IBM lanches the landmark geographic project to find the moving of the place by ancient people.

15) 2007- IBM lanches MANY EYE, the data visualization tool that user allow to upload the data and then produce graphic notaion.

16) 2009- IBM and the marine institute Ireland developed the SmartWay project, which collected the vsrious streams of environment data.

17) 2011-watson is a computing system which understand the natural language and can analze data.

18) 2013- IBM helps to close skills gap by patterning with more than 1000 global universities on big data and analytics circulum.

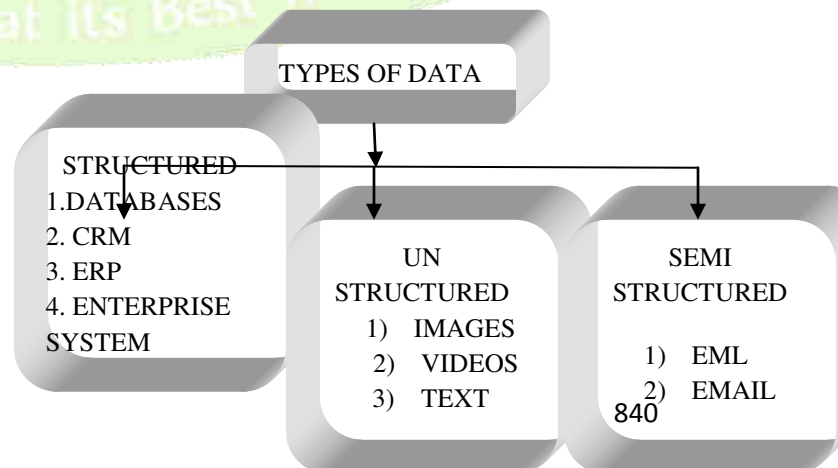
A. Why big data?

Big data is "hot" due to the potential value that it brings to us. With competition for resources becoming more intense than ever, organizations – both public and private sectors – have been searching for ways to differentiate themselves from their competitors by diving into the wealth of information to improve their competitiveness, efficiency, profitability and more.

II. CHARACTERISTICS

A. Variety:

Data being produced is not of a single category as it not only includes the traditional data, but also the semi structured data from various resourceslike web Pages, Web Log Files, social media sites, electronic mail, text files, sensor device data both from activepassive devices. All this information is totally different consisting ofraw,structured,semistructured and even unstructured data whichis unmanageableto behandled by the existing traditional analytic systems[SER, 2013].



store and this step won't be depressed enough. And then our traditional systems are not capable enough of making out the analytics on the data which is constantly in doubt [SER, 2013].

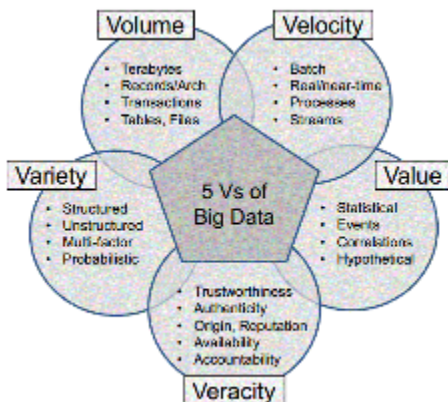
Fig 2. Types of data

Variety	Volume	Velocity	Value	Veracity
Types of data	Amount of data	Speed of data	Data value	Trustworthiness of data

B. Volume:

The Big news on big data itself defines the volume. At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes every day and this amount of data is definitely difficult to be done by using the existing traditional systems [AMR, 2014].

Fig 2.1 Characteristics of big data



C. Velocity

Velocity in Big data is a concept which deals with the swiftness of the data arriving from various sources. This characteristic is not being set to the velocity of incoming data, but also the speed at which the data flows. For example the data from the sensor devices would be constantly moving to the database

D. value

This is also important V in big data, because there are so much amount of money is spending in storing the data. At the end if we cant to extract the value from the data there is no use of the money. So that this is slao a main V in the big data.

E. Veracity

The quality of the databeing captured can vary greatly. Accuracy of the data being measured by the veracity of the source. And also it refers the trustworthiness of the data.

BIG DATA CONTROVERCIES

It seems about the current status of big data. IBM Watson where they reported that their Jeopardy winning performance was based upon only 3000 observations, but obviously very high dimension data. This was possible because they dramatically reduced error with their ensemble modeling. processing trillions of credit card transactions per year and detecting/preventing fraud in real time. Another success story is Google, which uses big data (web content and user-generated signals) to transform itself from a zero-revenue to \$59 billion-revenue company in 16 years. Amazon is another example that they generating real-time recommendations for 164 million active customers, resulting in over 300 purchases per second at its peak.

ISSUES AND CHALLENGES

1. Privacy, security and trust
2. Data management and sharing
3. Technology and analytical system

The video game industry is using big data for tracking during gameplay and after, predicting performance, and analyzing over 500GB of structured data and 4 TB of operational logs each day [26].

APPLICATION AREA

1. FINANCIAL INDUSTRY

All the financial industry will spend the money to want to use of data, to make better profit. Hadoop is being used in the industry for sentiment analysis, predictive analysis, and the financial trades.

2. AUTOMOTIVE INDUSTRY

In the Ford modern industry, they are developing 25GB of data per hour. Because they want to know the behavior of the driving and want to reduce the accidents.

3. SUPPLY CHAIN, LOGISTICS AND INDUSTRIAL ENGINEERING

companies are using telematics and big data to streamline trucking fleets and how they can improve fuel usage and routes. Union Pacific Railroad use thermometers, microphones, and ultrasound to capture data about their engines and send it for analysis to identify equipment at risk for failure.

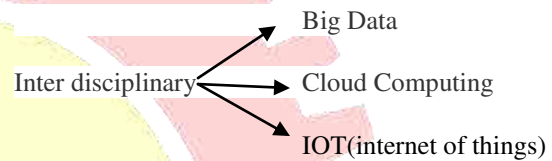
4. RETAIL

Walmart is using big data from 10 different sites to provide for shopper and transaction data into an analytical system. Amazon uses 1 million Hadoop clusters to support their affiliate network, risk management, machine learning, website updates, and more.

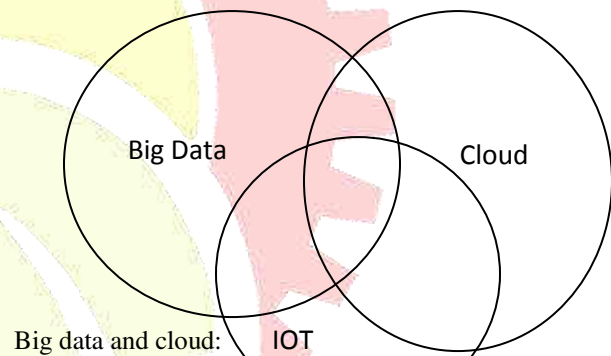
5. ENTERTAINMENT

INTERDISCIPLINARY

Inter disciplinary



```
graph LR; A[Inter disciplinary] --> B[Big Data]; A --> C[Cloud Computing]; A --> D[IOT (internet of things)];
```



[CHA, 2012] In this paper the author told about the survey of big data in the flow of cloud computing. Nowadays all the data are storing in the cloud.

III. RELATED WORK:

In paper [PAR, 2014] the author discussed about big data analytics, characteristics, issues, challenges in big data are discussed. In a paper [PUN, 2013] the authors suggest various methods for catering to the problems at hand through Map Reduce framework over Hadoop Distributed File System (HDFS). Map Reduce is a Minimization technique which makes usage of file indexing with mapping, screening, mixing and finally melting off. MapReduce techniques have been studied at in this paper which is implemented for Big Data analysis using HDFS. In paper [AMR, 2014] provides an introduction to Hadoop HDFS and MapReduce for storing large number of files and retrieve information from these files. The author shows his experimental work done in Hadoop by applying a number of files as input to the

system and then analyzing the public display of the Hadoop system. We have studied the MapReduce. We have analyzed the behavior of the map method and the reduce method with increasing number of files and the amount of bytes written and learn from these tasks. On paper [SER,] the author discussed about the privacy challenges in real time analytics. On paper [ANI,] the author discussed about Cluster File Systems employing a shared storage model also includes other benefits like stability & robustness, a rich set of features and compatibility with traditional analytics applications. On paper [SAC,] covers Big Data adoption trends, entry & exit criteria for the marketer and merchandise selection, best uses, customer success story, benefits of Big Data analytics. The insights derived from the user generated online contents and collaboration with customers is critical for success in the age of social media. In this paper [MAD,] first, the different problem definitions related to data stream clustering in general; second, the specific difficulties encountered in this expanse of research; third, the varying assumptions, heuristics, and intuitions forming the base of different attacks; and how several prominent solutions tackle different problems. In this paper [BOG,] written for (social science) researchers seeking to analyze the wealth of social media now available. It presents a comprehensive inspection of software tools for social networking media, wikis, real simple syndication feeds, network logs, newsgroups, chat and news feeds. On paper [DEW,] this paper discussed about the data stream availability, dependability, performance analysis. In this paper [UMU,] In this report, we propose an approach based on self-adjusting computation that can dramatically better the efficiency of such deliberations. In this paper [YIX,] the author introduced about the algorithm, calling as a D - stream which is very efficient the technique makes high-speed data stream Clustering feasible without degrading the clustering quality. In this report [KAR,] architecture includes a memory hierarchy optimized for Big-Data streams and implements modern one-grained power management techniques over all the dissimilar types of cores allowing then minimum energy expenditure for each example of executing classifier. In this paper [KEV,] a declarative query language for network traffic processing that Bridges the gap between powerful intrusion detection Systems and a simple, platform-independent SQL syntax. In a paper [JAM, 2013] the author described about 3 fundamental issues: they are (i) storage issue

(ii) management issue (iii) processing issue. In paper [STE, 2013] the author explained about the types of data. They are structured, unstructured and semi structured. Structured--Data that resides in fixed fields (for example, data in relational databases or in spreadsheets), Unstructured – Data that do not reside in fixed fields (for example, free-form text from articles, email messages, untagged audio and video data, etc.), Semi-structured – Data that does not reside in fixed fields but uses tags or other markers to capture elements of the data (for example, XML, HTML-tagged text). In paper [SAM,] the author explained about the sources of big data. How the data are created and all. In paper [JEA,] the author discussed about the types of data. They are structured, unstructured and semi structured data

Structured	The data which are having in the same format
Unstructured	The data which are having in the different format
Semistructured	Combination of both

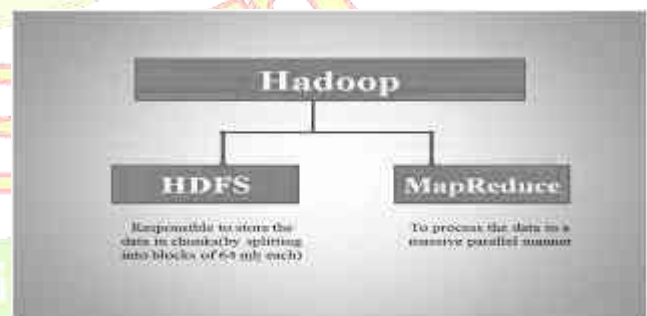
Sources of big data and its range per minute

S.no	Sources	Production
1.	YouTube [21]	Upload 48 hours videos per minute
2.	Facebook [22]	Share 684478 pieces of content per minute
		Likes (34722)
		100 terabytes of data uploading
		Share 3600 new

3.	Instagram [23]	photos uploading
4.	Tumblr[23]	Sees 27778 posts
5.	Web sites	571 new web sites are created
6.	Twitter [24]	(i) This site has over 645 million users (ii) The site generates 175 million tweets per day

of information. Though the memory capabilities of the rides have increased massively but the pace of reading data from them hasn't shown that considerable improvement. The reading process takes a great amount of time and the process of composition is also denser. This fourth dimension can be reduced by understanding from multiple disks at once. Just using one hundredth of a disk may seem wasteful. Only if in that respect are one hundred datasets, each of which is one terabyte and providing shared access to them is also a solution. There come many problems as well with using many pieces of hardware as it increases the prospects of bankruptcy. This can be avoided by Replication i.e. creating redundant copies of the same information on different devices so that in case of failure the copy of the data is usable. The primary problem is combining the data being read from different devices. Many methods are available in distributed computing to handle this problem, but even so it is rather challenging. All the problems discussed are well managed by Hadoop. The problem of failure is handled by the Hadoop Distributed File System and problem of combining data is handled by Map reduce programming Paradigm. Map Reduce basically reduces the problem of disk records and writes by providing a programming model dealing in computation with keys and values.

Fig 4.2 major components of Hadoop



Hadoop thus provides: a reliable shared storage and analysis organization. The depot is provided by HDFS and analysis by MapReduce.

IV. TOOLS AND TECHNIQUES:

A. Hadoop:

Hadoop is an open source project hosted by Apache. These are areas where HDFS is not a good fit: Low-latency, Software Foundation. It consists of many small sub projects which belong to the category of infrastructure for distributed computing. Hadoop mainly consists of:

File System (The Hadoop File System)

Programming Paradigm (Map Reduce)

The other subprojects provide complementary services or they are built along the kernel to add higher-level generalizations. There exist many problems in dealing with warehousing of large sums

B. Hadoop Components in detail:

Hadoop Distributed File System: Hadoop comes with a distributed File System called HDFS, which stands for Hadoop Distributed File System. HDFS is a File System designed for storing very large files with

streaming data access patterns, running on clusters on commodity hardware. The HDFS block size is a lot bigger than that of normal file system, i.e. 64 MB by default. The grounds of this large size of blocks is to scale down the number of disk seeks.

Fig 4.1 Hadoop architecture



Hadoop components

Components	Functions
HDFS	Storage
Map reduce	Distributing
HBASE	Read/write
Pig	Scripting
Hive	SQL
Ooze	Workflow
Zookeeper	Coordination
Kafka	Messaging
Mahout	Machine learning

An HDFS cluster has two cases of clients, i.e. name node (the captain) and number of data nodes (workers). The name node manages the file system namespace, maintains the file system tree and the metadata for all the files and directories in the tree. The datanode stores and retrieve blocks as per the instructions of clients or the namenode. The data retrieved is reported back to the namenode with lists of blocks that they are storing. Without the name, it is not possible to access the file. Thus, it becomes really significant to make namenode resilient to failure. Information access, Lots of small files, multiple writers and arbitrary file modifications.

C. MapReduce:

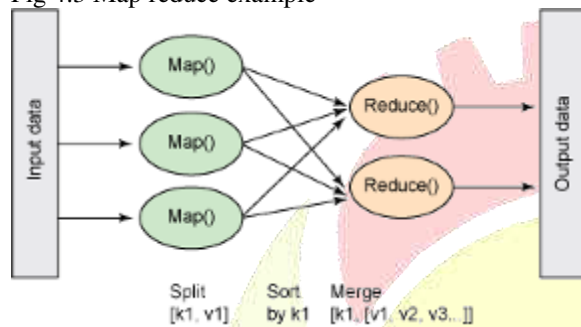
MapReduce is the programming paradigm allowing massive scalability. The MapReduce basically performs two different projects, i.e. Map Task and Reduce Task.

Steps	Tasks
Input	(i) Data are loaded into HDFS in blocks and distributed to data nodes (ii) Blocks are replicated in case of failures (iii) The name node tracks the blocks and Data nodes
Job	Job details submission
Job initialization	(i) The Job Tracker interacts with the Task Tracker on each data node (ii) All tasks are scheduled
Mapping	(i) The Mapper processes the data blocks (ii) Key value pairs are listed
Sorting	The Mapper sort the list of key value pairs
Shuffling	(i) The mapped output is transferred to the reducers (ii) Values are rearranged in a sorted format
Reduction	Merge the list of key pairs
Result	Client read the result to the HDFS

A map-reduce computation executes as follows: Map projects are dedicated input from distributed file system. The map tasks produce a succession of key-value pairs from the input and this is practiced according to the code written for map function. These values generated are collected by master controller and are sorted by key and divided among reduce tasks. The sorting basically assures that the same key values end with the same reduce tasks. The Reduce tasks combine all the values associated with a key working with one key at a time. Once more the

combination process depends on the code written for reduce job.

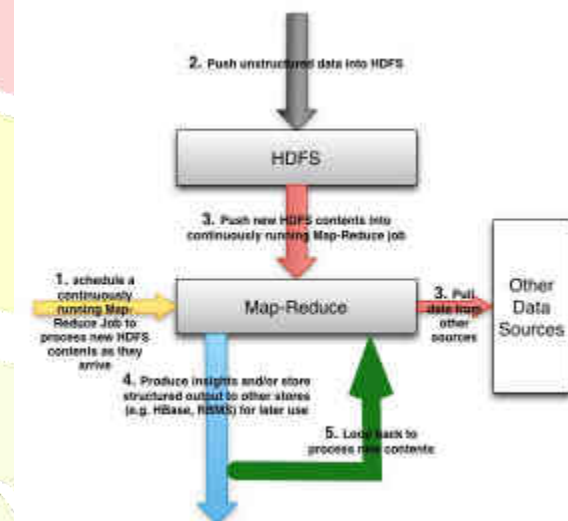
Fig 4.3 Map reduce example



The Master controller process and whatever act of worker processes at different compute nodes are forked by the user. Worker handles map tasks (MAP WORKER) and reduce tasks (REDUCE WORKER) but not both.

The Master controller creates some number of map and reduce tasks which is usually determined by the user program. The projects are attributed to the worker nodes by the superior controller. Racecourse of the status of each Map and Reduce task (idle, executing at a particular Worker or completed) is maintained by the Master Process. At the culmination of the work assigned the worker process reports to the captain and master reassigns it with some chore. The loser of a compute node is observed by the captain as it periodically pings the worker nodes. All the Map tasks assigned to that client are restarted even if it had completed and this is ascribable to the fact that the consequences of that calculation would be usable on that node only for the reduce tasks. The position of each of these Map tasks is set to idle by the Master. These get scheduled by Master on a Worker only when one becomes available. The Master must also inform each Reduce task that the localization of its input from that Map task has shifted.

Fig 4.4 HDFS and Mapreduce architecture



CONCLUSION

This paper described about big data, big data analytics, real time streaming analysis and those tools, issues challenges and tools, technique. These will be helping to prepare the business organization successfully. And we have discovered about the streaming concept. Whenever the volume is increased the velocity will be diminished. And also this includes the Hadoop and map reduce techniques.

REFERENCES

- [1] [WOR, 2014] Worldometers, "Real time world statistics," 2014, <http://www.worldometers.info/world-population/>
- [2] [CHE, 2013] D. Che, M. Saffron, and Z. Peng, "From Big Data to Big Data Mining: challenges, issues, and opportunities," in Database Systems for Advanced Applications, pp. 1–15, Springer, Berlin,

- Germany, 2013.
- [3] [JEA, 2013] Jean Yan, U.S. General Services Administration "Big Data, Bigger Opportunities", April 9, 2013.
- [4] [SER, 2013] Serif SAGIROGLU and Duygu SINANC, Dept of computer science, Gazi University, Ankara, Turkey @2013.
- [5] [PAR, 2014] Parth Chandarana, V.E.S.I.T, Chembur M. Vijayalakshmi, Department of Information Technology, "Big Data Analytics Frameworks", 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA).
- [6] [PUN, 2013] Puneet Singh Duggal, Department of Computer Science & Engineering Birla Institute of Technology, Sanchita Paul, Department of Computer Science & Engineering Birla Institute of Technology, "Big Data Analysis: Challenges and Solutions", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.
- [7] [AMR, 2014] Amrit Pal, Pinki Agrawal, Kunal Jain, Sanjay Agrawal, "A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data Using Hadoop", 2014 International conference on communication system and mesh technologies.
- [8] [SER,] Serif SAGIROGLU and Duygu SINANC, Gazi University, Ankara, Turkey Department of Computer Engineering, Faculty of Engineering, "Big Data: A Review".
- [9] [ANI,] Anirban Mukherjee, Joydip Datta, Raghavendra Jorapur, Ravi Singhvi, Saurav Haloi, Wasim Akram, "Shared Disk Big Data Analytics with Apache Hadoop".
- [10] [SAC,] Sachchidanand Singh, Business Analytics Division, IBM India Software Lab (ISL), Nirmala Singh
- [11] [MAD,] Data Warehouse Division, "Big Data Analytics".
- [12] [BOG,] Madjid Khalilian, Norwati Mustapha "Data Stream Clustering: Challenges and events".
- [13] [DEW,] Bogdan Batrinca, Philip C. Treleaven "Social media analytics: a survey of techniques, tools and platforms".
- [14] [UMU,] Dewey Sun, Guangyan Zhang, Weimin Zheng, and Keqin Li, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, "Key Technologies for Big Data Stream Computing".
- [15] [YIX,] mutA. Acar Carnegie Mellon University, Yan Chen Max Planck Institute for Software Systems, "Streaming Big Data with Self-Adjusting Computation".
- [16] [KAR,] Yixin Chen, Department of Computer Science and Engineering Washington University in St. Louis St. Louis, USA, "Density-Based Clustering for Real-Time Stream Data".
- [17] [KEV,] Karim Kanoun, Martino Ruggiero, David Atienza and Mihaela van der Schaar, University of California "Low Power and Scalable Many-Core Architecture for Big-Data Stream Computing".
- [18] [JAM, 2013] Kevin Borders, Jonathan Springer, Matthew Burnside, "Chimera: A Declarative Language for Streaming Network Traffic Analysis".
- [19] [STE, 2013] jamyang, "big data bigger opportunities" U.S. General Services Administration April 9, 2013.
- [19] [STE, 2013] Stephen Kaisleri_SW Corporation, Frank Armour American University, J. Alberto Espinosa, American University, William Money, George Washington University, "Big Data: Issues and Challenges Moving Forward",

- 2013 46th Hawaii International Conference on System Sciences.
- [20] [SAM,] SameeraSiddiquiCse, Rkgit, Ghaziabad, India Deepa Gupta Amity Institute of Information Technology, Noida, India “Big Data Process Analytics: A Survey”.
- [21] [JEA, 2013] Jean Yan, U.S. General Services Administration “Big Data, Bigger Opportunities”, April 9, 2013.
- [22] YouTube, “YouTube statistics,” 2014, <http://www.youtube.com/yt/press/statistics.html/>.
- [23] Facebook, Facebook Statistics, 2014, <http://www.statisticbrain.com/facebook-statistics/>.
- [24] Marcia, “Data on Big Data,” 2012, <http://marciaconner.com/blog/data-on-big-data/>.
- [25] Twitter, “Twitter statistics,” 2014, <http://www.statisticbrain.com/twitter-statistics/>.
- [26] www.youtube.com/user/prashundas89, “big data basics overview”.
- [27] <http://blog.pivotal.io/pivotal/news-2/20-examples-of-getting-results-with-big-data>
- [28] ChangqingJi, WenmingQiu, UchekukwuAwada, Keqiu Li, 2012 International Symposium on Pervasive Systems, Algorithms and Networks, “Big Data Processing in Cloud Computing Environments”

