International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 03, March 2017
www.ijates.com
ijates
ISSN 2348 - 7550

A SURVEY ON BIG DATA ANALYTICS: CHALLENGES, RESEARCH ISSUES AND PLATFORMS

Maddhi Sunitha

Department of Computer Science and Engg,
CVR College of Engineering,Hyderabad,Telangana-501510, India

ABSTRACT

The amount of data produced by mankind is very vast due to the arrival of new technologies, devices, and communication means. A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of this large data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The main objective of this paper is to explore the potential impact of big data challenges, open research issues and provides an in-depth analysis of different platforms available for performing big data analytics.

Keywords:*Big data, Data analytics, Data mining, Hadoop, Datasets*

I. INTRODUCTION

In today's world, due to the advancement of technology huge amount of data is being generated from various sources which led to the growth of big data[1]. The term big data is used as it goes beyond the processing power of traditional database systems .It provides evolutionary advances in many fields with collection of large datasets. It generally refers to the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing applications. These are available in structured, semi-structured, and unstructured format in petabytes and beyond.

Big data has three characteristics data volume, velocity, and variety which is referred as 3V. Large amount of data that is being generated every day is referred as Volume whereas velocity is the rate of growth and how fast the data is gathered for analysis. Variety provides information about the types of data such as structured, unstructured, semistructuredetc. Apart from these 3V there is another characteristics called veracity. It refers to availability and accountability[2]. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques.

Generally, Data warehouses have been used to manage the large dataset. In this case extractingthe precise knowledge from the available big data is a critical issue. The present approaches in data mining are not suitable to handle the large datasetssuccessfully[1]. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such asdata mining and statistical analysis.

These challenges generally arise when we wish to perform knowledge discovery and representation for its practical applications.

II. CHALLENGES OF BIG DATA

Recent years big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches.Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges[3].

To handle the challenges we need to know various computational complexities, information security, and computational method, to analyze big data. For example, many statisticalmethods that perform well for small data size do not scale to voluminous data.

Here the challenges of big data analytics are classified into four broad categories namely data storage and analysis,knowledge discovery and computational complexities, scalability and visualization of data and information security. We discuss these issues briefly in the following subsections.

2.1 Data Storage and Analysis

In recent years the size of data has grown exponentiallyby various means. These data are stored on spending much cost whereas they ignored or deleted finally because there is no enough space to store them. Therefore, the first challenge for big data analysis is storage mediums and higher input/output speed. In such cases, the data accessibility must be on the top priority for the knowledge discovery and representation. In past decades, analyst use hard disk drives to store data but, it slower random input/output performance than sequential input/output[6]. To overcome this limitation, the concept of solid state drive (SSD) and phrase change memory (PCM) was introduced. However the available storage technologies cannot possess the required performance for processing big data.

Another challenge with Big Data analysis is attributed to diversity of data with the ever growing of datasets, datamining tasks has significantly increased[3]. Additionally data reduction, data selection, feature selection is an essential task especially when dealing with large datasets. This presents an unprecedented challenge for researchers. It is because, existing algorithms may not always respond in an adequate time when dealing with these high dimensional data. Automation of this process and developing new machine learning algorithms to ensure consistency is a major challenge in recent years. Recent technologies such as hadoop and mapReduce make it possible to collect large amount of semi structured and unstructured data in a reasonable amount of time. The key engineering challenge is how to effectively analyze these data for obtaining better knowledge. A standard process to this end is to transform the semi structured or unstructured data into structured dataand then apply data mining algorithms to extract knowledge.

The major challenge in this case is to pay more attention for designing storage sytems and to elevate efficient data analysis tool that provide guarantees on the output when the data comes from different sources. Furthermore, design of machine learning algorithms to analyze data is essential for improving efficiency and scalability.

2.2 Knowledge Discovery and Computational Complexities

Knowledge discovery and representation is a prime issuein big data. It includes a number of sub fields such as authentication, archiving, management, preservation, information retrieval, and representation. There are several tools for knowledge discovery and representation. Additionally many hybridized techniques are also developed to process real lifeproblems[4]. All these techniques are problem dependent. Further some of these techniques may not be suitable for large datasets in a sequential computer. At the same time some of the techniques has good characteristics of scalability over parallel computer. Since the size of big data keeps increasing exponentially, the available tools may not be efficient to process these data for obtaining meaningful information. The most popular approach in case of large dataset management is data warehouses and data marts. Datawarehouse is mainly responsible to store data that are sourced from operational systems whereas data mart is based on a data warehouse and facilitates analysis.

Analysis of large dataset requires more computational complexities. The major issue is to handle inconsistenciesand uncertainty present in the datasets. In general, systematicmodeling of the computational complexity is used. It may be difficult to establish a comprehensive mathematical system that is broadly applicable to Big Data. But a domain specific data analytics can be done easily by understanding the particular complexities. A series of such development could simulate big data analytics for different areas. Much research and survey has been carried out in this direction using machine learning techniques with the least memory requirements. The basicobjective in these research is to minimize computational cost processing and complexities.

However, current big data analysis tools have poor performance in handling computational complexities, uncertainty, techniques and technologies that can deal computational complexity,

uncertainty,and inconsistencies in a effective manner.

2.3 Scalability and Visualization of Data

The most important challenge for big data analysis techniquesis its scalability and security. In the last decades researchers have paid attentions to accelerate data analysis and its speed up processors followed by Moore's Law. For the former, it is necessary to develop sampling, on-line, and multiresolution analysis techniques. Incremental techniques have good scalability property in the aspect of big data analysis[5]. As the data size is scaling much faster than CPU speeds, there is a natural dramatic shift in processor technology being embedded with increasing number of cores. This shift in processors leads to the development of parallel computing. Real time applications like navigation, social networks, finance, internet search, timeliness etc. requires parallel computing.

The objective of visualizing data is to present them more adequately using some techniques of graph theory. Graphical visualization provides the link between data with proper interpretation.

However, online marketplace like flipkart, amazon, e-bay have millions of users and billions of goods to be sold everymonth. This generates a lot of data. To this end, some company uses a tool Tableau for big data visualization. It has capability to transform large and complex data into intuitive pictures. This help employees of a company to visualize search relevance, monitor latest customer feeback, and their sentiment analysis.

However, current big data visualization tools mostly have poor performances in functionalities, scalability, and response in time.

2.4 Information Security

In big data analysis massive amount of data are correlated,analyzed[7], and mined for meaningful patterns. All organizations have different policies to safe guard their sensitive information.Preserving sensitive information is a major issue in big data analysis. There is a huge security risk associated with big data. Therefore, information security is becoming a big data analytics problem. Security of big data can be enhanced by using the techniques of authentication, authorization, and encryption.

Various security measures that big data applications face are scale of network, variety of different devices, real timesecurity monitoring, and lack of intrusion system.The security challenge caused by big data has attracted the attention of information security. Therefore, attention has to be given to develop a multi level security policy model andprevention system. Although much research has been carried out to secure big data but it requires lot of improvement. The major challenge is to develop a multi-level security, privacy preserved data model for big data.

II. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction fromdata[8,9]. Applications of big data and data science include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, datawarehousing, and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing, bio inspired computing, and quantum computing. However it is not limited to these issues. More research issues related to health care big data can be found.

3.1 IoT for Big Data Analytics

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things (IoT). Thus, appliances are becoming the user of the internet, just like humans with the web browsers. Internet of Things is attracting the attention of recent researchers for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information, network and communication technology. The new regulation of future will be eventually, everything will be connected and intelligently controlled. The concept ofIoT is becoming more pertinent to the realistic world due to the development of mobile devices, embedded and ubiquitous communication technologies,cloud computing, and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety.

In a broader sense, just like the internet, Internet of Things enables the devices to exist in a myriad of places and facilitates applications ranging from trivial to the crucial. Conversely, it isstill mystifying to understand IoT well, including definitions, content and differences from other similar concepts. Several diversified technologies such as computational intelligence, and big-data can be incorporated together to improve the data management and knowledge discovery of large scale automation applications. Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing[10]. Therefore, it is essential to develop infrastructure to

analyze the IoT data. AnIoT device generates continuous streams of data and the researchers can develop tools to extract meaningful information from these data using machine learning techniques. Understandingthese streams of data generated from IoT devices and analysing them to get meaningful information is a challenging issue and it leads to big data analytics. Machine learningalgorithms and computational intelligence techniques is theonly solution to handle big data from IoT prospective.

In knowledge acquisition phase, knowledge is discovered by using various traditional and computational intelligence techniques. The discovered knowledge is stored in knowledgebases and expert systems are generally designed based on the discovered knowledge[8]. Knowledge dissemination is important for obtaining meaningful information from the knowledgebase. Knowledge extraction is a process that searches documents,knowledge within documents as well as knowledge bases. The final phase is to apply discovered knowledge in various applications. It is the ultimate goal of knowledge discovery. The knowledge exploration system is necessarily iterative with the judgement of knowledge application. There are many issues, discussions, and researches in this area of knowledge exploration.

3.2. Cloud Computing for Big Data Analytics

The development of virtualization technologies have made supercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been one of the most robust big data technique[11]. Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data. Cloud computing harmonize massive data by on demand access to configurable computing resources through virtualization techniques. The benefits of utilizing the Cloud computing include offering resources when there is a demandand pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction. Open challenges and research issues of big dataand cloud computing are discussed in detail by many researchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, andresource management. So Cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools.

Big data application using cloud computing should support data analytic and development. The cloud environment should provide tools that allow data scientists and business analysts tointeractively and collaboratively explore knowledge acquisition data for further processing and extracting fruitful results.This can help to solve large applications that may arise in various domains. In addition to this, cloud computing should also enable scaling of tools from virtual technologies intonew technologies like spark, R, and other types of big dataprocessing techniques.

3.3. Bio-inspired Computing for Big Data Analytics

Bio-inspired computing is a technique inspired by nature to address complex real world problems. Biological systems are self-organized without a central control. A bio-inspiredcost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These

techniques are developed bybiological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance[7].

These systems are more suitable for big data applications. Huge amount of data are generated from variety of resources across the web since the digitization. Analyzing these data and categorizing into text, image and video etc will require lot of intelligent analytics from data scientists and big data professionals. Proliferations of technologies are emerging like big data, IoT, cloud computing, bio inspired computing etcwhereas equilibrium of data can be done only by selecting right platform to analyze large and furnish cost effective results.

Bio-inspired computing techniques serve as a key role in intelligent data analysis and its application to big data. These algorithms help in performing data mining for large datasetsdue to its optimization application. The most advantage is its simplicity and their rapid convergence to optimal solutionwhile solving service provision problems.

From the discussions, we can observe that the bio-inspired computing models provide smarter interactions, inevitable data losses, and help is handling ambiguities.Hence, it is believed that in future bio-inspired computing may help in handling big data to a large extent.

3.4. Quantum Computing for Big Data Analysis

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously. This exponential improvementin computer systems might be possible. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computers, of course today's big data problems. The main technical difficulty in building quantum computer could soon be possible. Quantumcomputing provides a way to merge the quantum mechanics toprocess the information. In traditional computer, information is presented by long strings of bits which encode either a zero or a one[13]. On the other hand a quantum computer uses quantum bits or qubits. The difference between qubit and bit is that, a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomena of superposition and entanglement. It is because qubits behave quantumly. Forexample, 100 qubits in quantum systems require 2100 complex values to be stored in a classic computer system. It means that many big data problems can be solved much faster by largerscale quantum computers compared with classical computers.

Hence it is a challenge for this generation to built a quantum computer and facilitate quantum computing to solve big data problems.

IV. PLATFORMS

4.1. Horizontal scaling platforms

Some of the prominent horizontal scale out platforms include peer-to-peer networks andApache Hadoop. Recently, researchers have also been working on developing the nextgeneration of horizontal scale out tools such as Spark to overcome the limitations ofother platforms. We will now discuss each of these platforms in more detail in this section.

4.1.1. **Peer-to-peer networks**

Peer-to-Peer networks involve millions of machines connected in a network. It isa decentralized and distributed network architecture where the nodes in the networks(known as peers) serve as well as consume resources. It is one of the oldest distributedcomputing platforms in existence[15]. Typically, Message Passing Interface (MPI) is thecommunication scheme used in such a setup to communicate and exchange the databetween peers. Each node can store the data instances and the scale out is practicallyunlimited (can be millions of nodes).

The major bottleneck in such a setup arises in the communication between differentnodes. Broadcasting messages in a peer-to-peer network is cheaper but the aggregationof data/results is much more expensive. In addition, the messages are sent over the networkin the form of a spanning tree with an arbitrary node as the root where thebroadcasting is initiated.

MPI, which is the standard software communication paradigm used in this network,has been in use for several years and is well-established and thoroughly debugged. Oneof the main features of MPI includes the state preserving process i.e., processes can liveas long as the system runs and there is no need to read the same data again and againas in the case of other frameworks such as MapReduce (explained in section "Apachehadoop"). All the parameters can be preserved locally. Hence, unlike MapReduce, MPIis well suited for iterative processing. Another feature of MPI is the hierarchicalmaster/slave paradigm. When MPI is deployed in the master–slave model, the slavemachine can become the master for other processes. This can be extremely useful fordynamic resource allocation where the slaves have large amounts of data to process.

MPI is available for many programming languages. It includes methods to send andreceive messages and data. Some other methods available with MPI are 'Broadcast',which is used to broadcast the data or messages over all the nodes and 'Barrier', whichis another method that can put a barrier and allows all the processes to synchronizeand reach up to a certain point before proceeding further.

Although MPI appears to be perfect for developing algorithms[14] for big data analytics,it has some major drawbacks. One of the primary drawbacks is the fault intolerancesince MPI has no mechanism to handle faults. When used on top of peer-to-peer networks,which is a completely unreliable hardware, a single node failure can cause theentire system to shut down. Users have to implement some kind of fault tolerancemechanism within the program to avoid such unfortunate situations. With other frameworkssuch as Hadoop (that are robust to fault tolerance) becoming widely popular,MPI is not being widely used anymore.

4.1.2. **Apache Hadoop**

Apache Hadoop is an open source framework[12] for storing and processing large datasetsusing clusters of commodity hardware. Hadoop is designed to scale up to hundredsand even thousands of nodes and is also highly fault tolerant. The Hadoop platform contains the followingtwo important components:

i) Distributed File System (HDFS) is a distributed file system that is used to storedata across cluster of commodity machines while providing high availability andfault tolerance.

ii) Hadoop YARN is a resource management layer and schedules the jobs acrossthe cluster.

4.1.3. **Spark: next generation data analysis paradigm**

The major feature of Spark that makes it unique is its ability to perform in-memorycomputations. It allows the data to be cached in memory, thus eliminating theHadoop's disk overhead limitation for iterative tasks. Spark is a general engine forlarge-scale data processing that supports Java, Scala and Python and for certain tasks itis

tested to be up to 100× faster than Hadoop MapReduce when the data can fit in the memory, and up to 10× faster when data resides on the disk. It can run on HadoopYarn manager and can read data from HDFS. This makes it extremely versatile to run
on different systems.

4.2. **Vertical scaling platforms**

The most popular vertical scale up paradigms are High Performance Computing Clusters(HPC)[10], Multicore processors, Graphics Processing Unit (GPU) and Field ProgrammableGate Arrays (FPGA). We describe each of these platforms and their capabilities in thefollowing sections.

4.2.1. **High performance computing (HPC) clusters**

HPC clusters , also called as blades or supercomputers, are machines with thousandsof cores. They can have a different variety of disk organization, cache, communicationmechanism etc. depending upon the user requirement. These systems use well builtpowerful hardware which is optimized for speed and throughput. Because of thetop quality high-end hardware, fault tolerance in such systems is not problematic since hardware failures are extremely rare. The initial cost of deploying such a system can be very high because of the use of the high-end hardware. They are not as scalable as Hadoop or Spark clusters but they are still capable of processing terabytes of data. The cost of scaling up such a system is much higher compared to Hadoop or Spark clusters. The communication scheme used for such platforms is typically MPI. We already discussed about MPI in the peer-to-peer systems (see section “Peer-to-peer networks”).

Since fault tolerance is not an important issue in this case, MPIs’ lack of fault tolerancemechanism does not come as a significant drawback here.

4.2.2. **Multicore CPU**

Multicore refers to one machine having dozens of processing cores. They usually have shared memory but only one disk. Over the past few years, CPUs have gained internal parallelism. More recently, the number of cores per chip and the number of operations that a core can perform has increased significantly[15]. Newer breeds of motherboards allow multiple CPUs within a single machine thereby increasing the parallelism.Until the last few years, CPUs were mainly responsible for accelerating the algorithmsfor big data analytics.

The parallelism inCPUs is mainly achieved through multithreading . All the cores share the samememory. The task has to be broken down into threads. Each thread is executed in parallelon different CPU cores. Most of the programming languages provide libraries tocreate threads and use CPU parallelism. The most popular choice of such programminglanguages is Java. Since multicore CPUs have been around for several years, a largenumber of software applications and programming environments are well developedfor this platform. The developments in CPUs are not at the same pace compared toGPUs. The number of cores per CPU is still in double digits with the processing powerclose to 10Gflops while a single GPU has more than 2500 processing cores with 1000Tflops of processing power. This massive parallelism in GPU makes it a more appealing option for parallel computing applications.

The drawback of CPUs is their limited number of processing cores and their primarydependence on the system memory for data access. System memory is limited to a fewhundred gigabytes and this limits the size of the data that a CPU can process efficiently.

Once the data size exceeds the system memory, disk access becomes a huge bottleneck.Even if the data fits into the system memory, CPU can process data at a much fasterrate than the memory access speed which makes memory access a bottleneck. GPUavoids this by making use of DDR5 memory compared to a slower DDR3 memory usedin a system. Also, GPU has high speed cache for each multiprocessor which speeds upthe data access.

4.2.3. **Graphics processing unit (GPU)**

Graphics Processing Unit (GPUs) is a specialized hardware designed to accelerate thecreation of images in a frame buffer intended for display output. Until the past fewyears, GPUs were primarily used for graphical operations such as video and image editing,accelerating graphics-related processing etc. However, due to their massively parallelarchitecture, recent developments in GPU hardware and related programmingframeworks have given rise to GPGPU (general-purpose computing on graphics processing units). GPU has large number of processing cores as compared to a multicore CPU. In addition to the processing cores, GPU hasits own high throughput DDR5 memory which is many times faster than a typical DDR3memory. GPU performance has increased significantly in the past few years compared tothat of CPU. Recently, Nvidia has launched Tesla series of GPUs which are specifically designedfor high performance computing. Nvidia has released the CUDA framework whichmade GPU programming accessible to all programmers without delving into the hardwaredetails. These developments suggest that GPGPU is indeed gaining more popularity.

It usually has two levels of parallelism. At the first level, there are several multiprocessors (MPs) and within each multiprocessor there are severalstreaming processors (SPs). To use this setup, GPU program is broken down into threadswhich execute on SPs and these threads are grouped together to form thread blockswhich run on a multiprocessor. Each thread within a block can communicate with eachother and synchronize with other threads in the same block. Each of these threads has accessto small but extremely fast shared cache memory and larger global main memory.

Threads in one block cannot communicate with the threads in the other block as theymay be scheduled at different times. This architecture implies that for any job to be runon GPU, it has to be broken into blocks of computation that can run independently withoutcommunicating with each other [32]. These blocks will have to be further brokendown into smaller tasks that execute on an individual thread that may communicate withother threads in the same block.

GPUs have been used in the development of faster machine learning algorithms.Some libraries such as GPUMiner implement few machine learning algorithms onGPU using the CUDA framework. Experiments have shown many folds speedup usingthe GPU compared to a multicore CPU.

GPU has its own drawbacks. The primary drawback is the limited memory that itcontains. With a maximum of 12GB memory per GPU (as of current generation), it isnot suitable to handle terabyte scale data. Once the data size is more than the size ofthe GPU memory[14], the performance decreases significantly as the disk access becomesthe primary bottleneck. Another drawback is the limited amount of software andalgorithms that are available for GPUs. Because of the way in which the task breakdownis required for GPUs, not many existing analytical algorithms are easily portableto GPUs.

4.2.4. **Field programmable gate arrays (FPGA)**

FPGAs are highly specialized hardware units which are custom-built for specific applications. FPGAs can be highly optimized for speed and can be orders of magnitudefaster compared to other platforms for certain applications. They are programmedusing Hardware descriptive language (HDL). Due to customized hardware, the developmentcost is typically much higher compared to other platforms. On the softwareside, coding has to be done in HDL with a low-level knowledge of the hardware whichincreases the algorithm development cost[6]. User has to carefully investigate the suitabilityof a particular application for FPGA as they are effective only for a certain set ofapplications.

FPGAs are used in a variety of real-world applications .One example whereFPGA was successfully deployed is in the network security applications. In onesuch application, FPGA is used as a hardware firewall and is much faster than the softwarefirewalls in scanning large amounts of network data. In the recent years, thespeed of multicore processors is reaching closer to that of FPGAs.

V. CONCLUSION

This paper surveys about big data and its challenges. It also focuses on open issues in big data analytics and the platforms in big data analytics. This paper deals with different issues its advantages and disadvantages. The various platforms and its pros and cons have been discussed in detail.

REFERENCES

[1]. Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.

[2]. Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech.Rep. 2001. [Online]. Available:http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[3]. Beyer MA, Laney D. The Importance of 'BigData': A Definition, Gartner, Stam-ford, CT; 2012.

[4]. Gartner IT Glossary (n.d.). Available from http://www. gartner.com/it-glossary/big-data/

[5]. Krishnan K. Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence. Elsevier Science; 2013

[6]. Hu H, Wen Y, Tat-Seng C, Li X. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access, Practical Innovations: Open Solutions. 2014 Jul; 1–36.

[7]. Gantz J, Reinsel D. Extracting value from chaos. Proc. IDC iView. 2011; 1–12.

[8] .Cooper M, Mell P. Tackling Big Data. 2012. Available from http://csrc.nist.gov/groups/SMA/forum/documents/june-2012presentations/f%csm_june2012_cooper_mell.pdf

[9]. Wang C, Rayan IA, Schwan K. Faster, larger, easier: reining real-time big data processing in cloud. Proceedings of the Posters and Demo Track, Middleware '12, ACM; 2012; New York, NY, USA. pp. 4:1–4:2

[10]. Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. Commun. ACM. 2008; 51(1):107113.

[11]. Emani CK, Cullot N, Nicolle C. Understandable Big Data: A survey. Computer science review. 2015; 17:70–81.

[12]. Zaharia M, Das T, Li H, Hunter T, Shenker S, Stoica I. Discretized streams: Fault-tolerant streaming computation atscale. Proc 24th ACM Symp Operating Syst Principles. 2013; 423–38.

[13]. Neumeyer L, Robbins B, Nair A, Kesari A. S4: Distributed stream computing platform. Proc IEEE IntConf Data Mining Workshops; 2010. pp. 170–7.

[14]. Zhang H, Chen G, Ooi BC, Kian-Lee T, Zhang M. In-Memory Big Data Management and Processing: A Survey IEEE transactions on knowledge and data engineering. 2015 Jul; 27(7).

[15]. Yu Y, Wang X. World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fan's tweets. Computers in Human Behavior. 2015; 48:392–400