ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

A SURVEY OF BIG DATA ANALYTICS – IT'S CHALLENGES

F. Amul Mary, Lecturer in Computer Science
JMJ College for Women (Autonomous),
Tenali, Guntur (Dt), Andhra Pradesh
Mail Id: amulmary@gmail.com
Mobile: 9700165855

ABSTRACT:

The arrival of new technologies, devices, and communications, the amount of data producedby mankind is growing rapidly every year. This gives rise to the new era of big data. The term big data comeswith the new challenges to input,process and output the data. This paper focuses on limitation oftraditional approach, the techniques available to manage the data and the components that are useful in handling big data. One ofthe approaches used in processing big data is Hadoop framework, the paper presents the majorcomponents of the framework and working process also introduces some of the challenges on Big Data.

KEYWORDS:*Data Mining, Big Data, Data analytics, Hadoop Framework.*

1. INTRODUCTION:

The digitization process is tremendously fast increasing and due to that the production of data is almost in digital form and the data generated is increasing in size exceeding Exabyte. In accordance to data generation the computer systems are much faster than the old systems, yet analyzing of large scale data is a critical factor. Big data and data science can provide a significant path to value for organizations. These technologies, methodologies, and skills can help organizations gain additional insight about customers and operations; they can help make organizations more efficient, be a new source of revenue, and make organizations more competitive.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

2. TRADITIONAL APPROACH OF DATA MINING:

The process of data mining includes the operations like selection, preprocessing, transformationand evaluation of data [1] in the discovery of knowledge. The first task in data mining is datainput which includes collecting, selecting, preprocessing the data. Preprocessing includescleaning and filtering the data to make it useful for further activities. After the data is in cleanedand reduced from various data mining methods like clustering, classification, association rules and sequential patterns can be applied for data analysis.

Most of the methods cannot be applied to big data because of the following reasons:

- They are designed to work with a single machine with all the data in the memory. Mostof the methods are not for huge and complex data.
- Most of the methods cannot produce the analysis dynamically based on the input.
- Most of the methods work with the same format of input.

After applying the methods, evaluation and interpretation are applied to generate the output. Theyprovide the mechanism to measure the results. The output can be measured for the operators likenumber of errors, accuracy of results, computation speed, computation cost, response time,utilization of memory, etc. The knowledge generation becomes complex and need to be moreversatile for handling the big data.

2.1 BIG DATA: The term big data is used to describe any voluminous amount of data in structured, semi structured and unstructured format that has the potential to be mined to get relevant information.Big data involves the data produced by different devices and applications.It is huge volume of data both structured and unstructured format which is difficult to be processed using traditional database and software techniques.

Different Fields that Generates Big Data: Some of the fields that generate big data are,

- **Social Sites Data:** Social Media such as Facebook, Twitter, etc. carry information,suggestions, invitations, etc. posted by several people across the world. The responses fortheir campaigns, advertising mediums are also known.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

- **Search Engine Data:** Search engines retrieve lots of data from different databases.
- **Medical History Data:** Hospitals can generate medical history of patients for varioushealth issues.
- **Online Shopping Data:** Shopping of various products online can help to know thepreferences and product perception of the customers on different products at differentintervals.
- **Stock Exchange Data:** The stock exchange data holds information about the shares ofvarious companies. These data given an insight on the decisions taken by shareholders forthe trading activities.
- **Vehicle Booking Data:** Booking of vehicles like train, bus, flight, cab, etc. can generatethe data of booking a vehicle based on model, size, distance and availability of a vehicle.
- **Aviation Data:** Audio and Video data recordings, the performance information of theaircraft, etc.

2.2 CHARACTERISTICS OF BIG DATA: The big data has threecharacteristicsknown asVolume,Velocity, and Variety.It is also known as 3Vs [2],which means that the size of data islarge, the data is generated very fast, and the data exists in heterogeneous formats which can beamong structured data, semi structured data and unstructured data captured from differentsources. The common concept of 3Vs is given below:

1. **Volume:**The name 'Big Data' itself is related to a size which is enormous. Size of data plays very crucial role in determining value out of data. Whether the data actually be considered as Big Data or not, is dependent upon **Volume**of data.
2. **Variety:**Variety refers to data collected from heterogeneous sources. Earlier, spreadsheets and databases were the only sources of data that was used by most of the applications. Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio is also being used in the analysis applications. This **Variety**of unstructured data has certain issues for storage, mining and analyzing the data.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

3. **Velocity:**The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, Mobile devices, etc.

Apart from the 3Vs, more components were added to explain big data [5][6] like value, venue, vagueness, veracity, validity, vocabulary and variability.

Observing the above information it is good and necessary to have an insight of big data with theknowledge of tool selection.

3. TOOL SELECTION:

Big data involves various tools, techniques and frameworks to managethe data. There aremany big data platforms available with different characteristics. But, the selection of platform [3] depends on the capability of the platform and various dimensions as listed intable below.

Table1: Dimensions of Big Data

Dimension	Description	Issues in selection o dimension
Big data solution	Can be implemented as software, appliance or cloud based. Can be implemented as hybrid solution also.	➢ Data locality ➢ Privacy and regulation ➢ Human resources ➢ Project requirements
Data Transfer	Can be implemented as processing by transporting the data or processing without transporting the data.	➢ Cost of time, money to transfer data. ➢ Size of data to transfer. ➢ Locality of data especially with rapidly updating data.
Data Structurization	Can be implemented as performing the data acquisition and cleaning by ourselves or make your data available for the market place.	➢ Cost of time, money for data cleaning ➢ Quality of data ➢ Selection of market place.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

Technologies used to handle big data plays an important role in data analysis which leads toaccuracy of decision making resulting in cost reduction, faster services and operational efficiencies. To manage the large volume of data,selecting correct infrastructure is important.

3.1 TECHNOLOGIES USED TO HANDLE THE BIG DATA:

The following technologies can be used in capturing andstoring the big data and analyzing big data.

NoSQL database:

For capturing and storing the data, NoSQL database are commonly used. MongoDB providesfacilities for capturing and storing data. Other databases like Redis, BigTable, Hbase, Hypertable, ZooKeeper etc, are also used. NoSQL providesupport for cloud computing architectures which gives the benefit in reduction of cost, increasein speed of computing and increase in efficiency. It also provides the facility to generate patternsand trends without need for additional infrastructure.

Massively Parallel Processing (MPP) and MapReduce

For big data processing, Massively Parallel Processing (MPP) and MapReduce arecommonly used.MPP includes multiple processors, and each processor works on different partsof the program and has its own operating system and memory to utilize. MPP processorscommunicate using a messaging interface. Data paths are interconnected to allow messagepassing among processors. Usually partitioning of common database and assigning work among processors is quite complex process.MapReduce is a computational approach that involves breaking large volumes of data down intosmaller batches, and processing them separately. MapReduce is a programming model, Googlehas used successfully, processing its big data sets. The computation is done in terms of mapand a reduce function. A cluster of computing nodes which are built on commodity hardware willscan the batches and aggregate their data. Then the multiple nodes' output gets merged togenerate the final result data.MPP has many things in common with MapReduce. But, for a variety of reasons, MPP and MapReduce are used in rather different scenarios as listed below.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

Table 2: MPP and MapReduce Characteristics

MPP	MapReduce
MPP uses expensive, specialized hardware for CPU, storage &network performance.	MapReduce is deployed to clusters of commodity servers that in turn use commodity disks.
MPP products are queried with Structured Query Language (SQL)	MapReduce's native control mechanism is Java Code.
Loading of data is slower.	Loading of data is faster.
Querying is easier.	Forming maps and reducing is complex.
For structured data MPP is good in data refining and transformations.	For semi-structured and unstructured data refining and transformations is faster depending on file type and transformationlogic.

Storage: Amazon Simple Storage Service (Amazon S3) and Hadoop Distributed File System (HDFS) are used for storage purpose. S3 provides developers and IT teams with secure,durable, highly-scalable object storage. It is easy to use as it provides a simple web serviceinterface to store and retrieve any amount of data from anywhere on the web. There is no setupcost.

HDFS is a Java-based file system that provides scalable and reliable data storage, and it was designed to cover large clusters of commodity servers. HDFS has proven good in scaling and forming clusters of servers, which can support billions files and blocks.

Hadoop is one of the frameworksto operate on Big Data. The next section describes the fundamentals of the framework.

4. HADOOP FRAMEWORK:

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models.

In traditional approach, an enterprise will have a computer to store and process big data as shown in figure1.

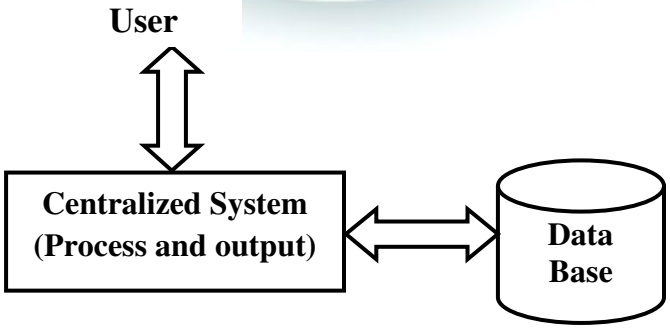

Figure1: Traditional approach to store and process big data

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated softwares can be written to interact with the database, process the required data and present it to the users for analysis purpose.This approach works well, when there is smaller amount of data. But when the data is in huge amount, it becomes time-consuming and tedious task to process the data through traditional database server. [4] discussed about creating Obstacles to Screened networks. In today's technological world, millions of individuals are subject to privacy threats. Companies are hired not only to watch what you visit online, but to infiltrate the information and send advertising based on your browsing history. People set up accounts for facebook, enter bank and credit card information to various websites. Those concerned about Internet privacy often cite a number of privacy risks events that can compromise privacy which may be encountered through Internet use. These methods of compromise can range from the gathering of statistics on users, to more malicious acts such as the spreading of spyware and various forms of bugs (software errors) exploitation.

So, Google adapted a new approach to handle the large volume of data using an algorithm calledMapReduce. This algorithm divides the input into small parts and assigns those parts to manycomputers connected over the network, and collects the results to form the final result as output.In this algorithm, the data is processed in parallelon different CPU nodes as shown in figure2.

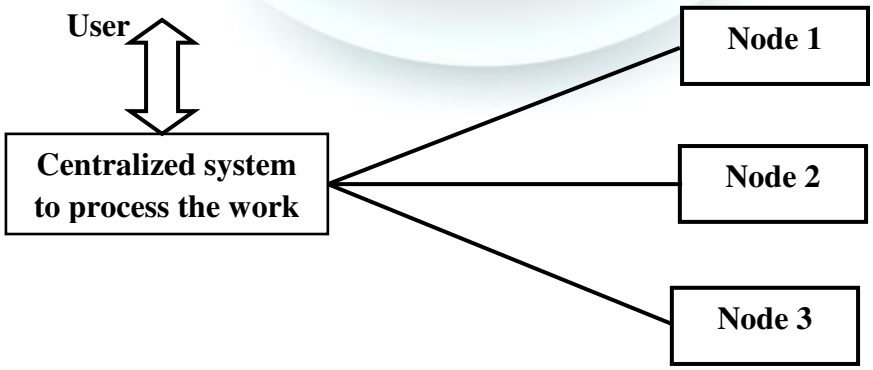

Figure2: working of MapReduce Algorithm to store and process data in parallel

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

Hadoop framework is capable enough to develop applications capable of running on clusters ofcomputers and they could perform complete statistical analysis for huge amounts of data. This is illustrated in the figure3.

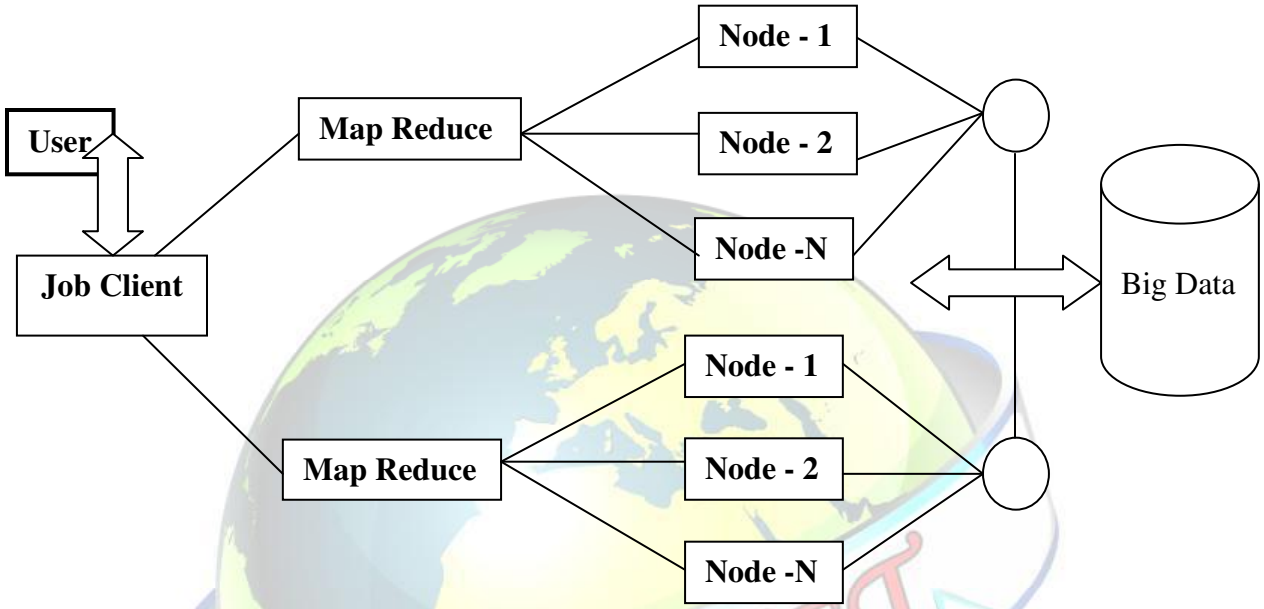

Figure3: HadoopFramework

4.1 WORKING OF HADOOP: Hadoop working can be divided into phases:

Phase 1:A user or application can submit ajob to Hadoop job client with specification of location of input and output files in the distributedfile system, jar files containing map and reduce functions and configuring job by setting variousparameters related to the job.

Phase 2: Hadoop job client submits the job andconfiguration to the MapReduce master called as JobTracker, which distribute thejars/executables to the slaves, it schedules tasks monitor them.

Phase 3: The slaves on variousnodes execute task as per MapReduce implementation and output of the reduce function is storedinto the output files on the file system.

The advantages of Hadoop includes: It is compatible on all platforms, distributes the data and tasks across the nodes automatically, the library has the API to detect and handle failures at the application layer, and the framework continues to operate smoothly with the addition and removal of serversdynamically.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

With the increase in big data, it becomes mandatory to process the data timely and accurately inorder to gain value from it. Hadoop is one of the frameworks which provide a good mechanism to handle the distributed processing of data.

5. CHALLENGES IN BIG DATA ANALYTICS:

1. **The Practical Issues of Storing All the Data:**The data in the large organizations are growing day by day.So, simply storing the data is becoming a real challenge. Companies are looking at options like data lakes, which will allow them to collect and store massive quantities of unstructured data in its native format. The problem is, data storage have to be constructed wisely. Otherwise they become a useless wasteland where data stored will never be retrieved again.
2. **Understanding and Utilizing Big Data:**It is a heavy task in most of the industries and companies to deal with data using traditional processing comparing to big data analytics. These types of analyses need to be performed on an ongoing basis as the data landscape changes at an ever-increasing rate, and as executives develop more and more of an appetite for analytics based on all available information.
3. **New, Complex, and Continuously Emerging Technologies:**Since much of the technology is required to utilize big data, which is new to most organizations, it will be necessary for these organizations to learn about these new technologies at an ever-accelerating pace, and engage with different technology providers and partners than they have used in the past. Firms entering into the world of big dataneed to balance the business needs associated with big data with the associated costs fordata capture, storage, processing, and analysis.
4. **Cloud Based Solutions:**New business software applications have emerged whereby company data is managed and stored in data centers around the globe. While these solutions range from ERP, CRM, Document Management, Data Warehouses and Business Intelligence to many others, the common issue remains the safe keeping and management of confidential company data. It also raises a new dimension related to data security and the overall management of an enterprise's Big Data paradigm.
5. **Privacy, Security, and Regulatory Considerations:**With the volume and complexity of big data, it is challenging for most of the company's to obtain a reliable content of all of their

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

data and secure it adequately, so that confidential and private business and customer data are not accessed by unauthorized parties. The costs of a data privacy breach can be enormous. It will be very important for most forms to tightly integrate their big data, data security/privacy, and regulatory functions.

6. **Archiving and Disposal of Big Data:**Since big data will lose its value to current decision-making over time, and since it is voluminous and varied in content and structure, it is necessary to utilize new tools, technologies, and methods to archive and delete big data, without sacrificing the effectiveness of using your big data for current business needs.
7. **Handling the Various Sources of Data Available:** Handling the volume of storage and the velocity at which data is increasing is one thing. Managing enormous streams of data from various disparate sources, both inside and outside of the organization, is another matter totally. When the enterprise's own data sources like finance, operational, marketing, and other data are combined with external sources such as social media and industry data, it becomes truly diverse as well as exceptionally massive.

6. CONCLUSIONS:

This paper presents the significance of big data, limitations of traditional approach in analyzing big data, the tools used in handling the big data and also explored that Hadoop is one of the most suitable frameworks which provide a good mechanism to handle the distributed processing of data.The major issues in processing big data are computation speed, accuracy, security and privacy. But, there are also challenges in Big Data.

REFERENCES:

[1] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag.1996; 17(3):37–54.

[2] Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001. [Online].

[3] Agneeswaran VS, Tonpay P, Tiwary J (2013) Paradigms for realizing machine learning algorithms. Big Data 1(4):207–214.

[4] Christo Ananth, P.Muppidathi, S.Muthuselvi, P.Mathumitha, M.Mohaideen Fathima, M.Muthulakshmi, “Creating Obstacles to Screened networks”, International Journal of

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 21, August 2017

Advanced Research in Biology, Ecology, Science and Technology (IJARBEST), Volume 1,Issue 4,July 2015, pp:10-14

[5] Van Rijmenam M. Why the 3v's are not sufficient to describe big data, BigData Startups, Tech. Rep. 2013. [Online].Available: http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/.

[6] Borne K. Top 10 big data challenges a serious look at 10 big data v's, Tech. Rep. 2014. [Online].Available: https://www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.