DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.5102

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

BIG DATA ANALYTICS: AN EPITOME

Vishal Batvia
Computer Science and Technology
Babaria Institute of Technology
Vadodara, India

Drashti Patel
Computer Science and Technology
Babaria Institute of Technology
Vadodara, India

Parth Goel
Devang Patel Institute of Advance Technology and Research
Charotar University of Science and Technology
Anand, India

Dr. Mamta Padole
Department of Computer Science and Engineering
The Maharaja Sayajirao University of Baroda
Vadodara, India

***Abstract*—** Big Data is one of the most popular buzzword nowadays. There is a huge demand for Data analyst and other Big Data professionals. As the enormous amount of data is making its space inside the world there is a new evolution of data. Data nowadays comes from small components as small as sensors to as big as Multinational companies do. In addition analyzing the data helps in business expansion and making profits.In this paper, there is brief Introduction to big data analyses. It starts from the most basic thing that is % V's of Big data then we will move on to the frameworks and platforms of Big Data Analytics after that we study the Issues, opportunities, Application and Challenges of Big Data and Analytics

***Keywords*—** Big-Data, Data Analysis, Volume, Velocity, Hadoop, Mapreduce, Spark

.

1. INTRODUCTION

Big Data has become an interesting and most important topic not only in the field of Computers but also in the Industries and Society. The main reason for this is the ability it has to produce, protect and process huge amount of data. As, the use of Technology is increasing daily most of the data coming in the market are digital or internet born. It is very difficult to get useful information from large-scale data then to just store it. Even though we have lots of improvement in technologies and hardware but analyzing large-scale data is still a headache.

We have been searching for some technique for data analysis and one of them is Map Reduce. There are many Map Reduce frameworks available like Hadoop, Spark, and Flink etc. The performance of these frameworks depends on factors such as HDFS block size, input data size, network connection etc [1]. A detailed information about big data analytics, its frameworks, Tools, Issues and Challenges are given in the document

2. BIG DATA ANALYTICS

Big Data is very important concept in today's world as it works differently from traditional databases. Big Data works together with some of the key technologies like Hadoop, HBASE, NoSQL, HDFS, and Cassandra etc. to extract important information from raw data, which was considered not possible before [2]. It is not only including traditional relational database but also includes all the properties and structures of unstructured data. Unstructured data means the data not in the relational format or Xml format. It includes text, pdf, images, audio video and other all types.

Big Data Analytics deals with data that are too large, fast, unstructured and not in the range of traditional databases. In every field nowadays, we can observe huge amount of data coming from Business and Research Institution to all Government offices. [2] It is very important for organizations to take out meaningful information from these data about Market Trends, Customer needs, and Competition in the market for their progress. However, it is very difficult to take out meaningful information from that data quickly. Moreover, that is why Big Data Analytics has become extremely important for Business organization for their progress and to increase their market share. Tools, which help in Big Data Analytics, have improved greatly in recent years. Thus, the use of these technologies is not extremely expensive but also many software are open source too. Hadoop – the most commonly used technology combines many open source software with the commodity hardware. It uses multiple disks to distribute the source data and performs analysis on it with the help of analytics tools such as JasperSoft BI Suite, Pentaho Business Analytics etc. [2]. The skill required to do this task is very new in most of the IT departments and it needs hard work to combine all the internal and external data.

TABLE I.

Understanding BigData	
Traditional Data	**Current Data**
Documents	Photos
Finances	Audio and Video
Stock Records	3D Models
Others	Simulations
	Location Data
	Web Data
	Sensor Data
	Social Networks Data

Fig. 1. Comparison of types of Data

Currently the data that is coming to us is not just huge data but it also consists of different data types and file formats as well as streaming data. Big Data analytics deals with all these issues but one common misconception is that more the data better the analytic information taken out from it, but more data can also mean more noise in data or ambiguous data. For example, we can say that if a person is using several bank accounts or many persons are using one single bank account. Thus, there are several new problems such as security, privacy, storage, fault tolerance and authentic data

3. V'S OF BIGDATA

The amount of data is growing everywhere nowadays. Major data is coming from sensors (IOT Devices, Sensors and Social Media), internet, traffic analysis, current trends, Business etc. Big Data mostly deals with volume but it is said that there are several V's of Big Data on which it depends. Although a Major part is Volume, there have been lots of talks and confusion about V's of Big Data. Volume is inseparable from bigdata and if we remove Volume from bigdata then it will not be big enough. It will be small set of data, which can be easily used, in traditional system. In general, the V's accepted everywhere are Volume, Velocity, Veracity, Value and Variety [3]. All the V's of Big Data are equally important and removing any one will create a huge difference.

A. Volume

Volume is huge set of Data collected from various sources, which increases exponentially. It is very difficult to store and manage zettabytes or more data. The data that is collected manually and automatically in Databases and Data warehouses needs to be maintained. The service providers can ignore not a single bit of data, as those data are highly important to the clients. They have to deal with problems like security, fault tolerance, Confidentiality, disaster Management, System crash etc.[3] Volume is the term used for very huge amount of data, which is not in the reach of traditional system to process also it, is very difficult to store and manage such data.

B. Velocity

As we discussed earlier that Volume is very important part of Big Data but when the data is increasing at an exponential rate it is very important to note at what rate that data is created. In many cases, Data creates another data and in such cases, the velocity of data creation is too high. We can consider two types of velocity that is the rate at which data is being created and the rate at which data is transferred [4].

C. Veracity

Veracity is another most important factor of Big Data. It deals with the authentication and truthfulness of data. It does not matter if we have a small data or a huge data but if the data is inaccurate then it is of no use. We can analyze huge data but it is not possible to crosscheck the authentication of such a huge data. In addition, we are supposed to take some decision based on the result we got after data analysis but if the data is not authentic then the results are ambiguous. Thus, it is very important to see if the data is trustworthy or not.

D. Variety

Variety includes the different forms of data that are to be managed namely Structured, Semi-Structured and Unstructured. The structured data includes the traditional table format data while Semi-Structured data includes Log data and XML data and unstructured data comprises of all the text, images, audio, video etc. As the data is increasing at a very high rate with a very high speed, it also comes with a variety of data.

Fig. 2. 5 V's of Bigdata [5]

E. Value

Although Data analytics mostly deals on extracting value from huge data but Value is also one of the important V's. As it is known that with the help of big data analytics we extract useful information out of the given huge amount of data, which adds value to the data, Thus value is an integral part of BigData analytics. [3]

Thus, we can say that all the V's of Big Data are equally important and it is not possible to analyze data ignoring even one of it.

4. FRAMEWORKS AND PLATFORMS OF BIG DATA ANALYTICS

A. Frameworks

i. Apache Hadoop

Hadoop is one of the most popular framework for Big Data analytics. It is a collection of Hadoop and MapReduce ecosystem tools like Pig, Flume, Hive, HDFS etc. A number of frameworks are available nowadays but still Hadoop is used, as it is very simple to use. Other frameworks such as Spark can also use many Hadoop tools such as YARN-The resource management layer.

ii. MapReduce

MapReduce is the programming model most commonly used with Hadoop [7]. It works with mappers and

reducers. Mappers deals with collection of data and analyzing it and producing Intermediate data, which is then passed to reducers [9]. Reducers deals with aggregation of the results and give proper output.

iii. Apache Pig and Hive

These are the two wrappers, which provide easy way instead of dealing with MapReduce. Apache Pig is an SQL like environment, which helps in analyzing of Data [10]. It is lying on top of Hadoop, which allows higher-level languages to use Hadoop's MapReduce library. While Hive is a technology which turns Hadoop into a data warehouse with a help to use SQL Queries [7].

iv. Spark

Spark and Hadoop are often considered same but in reality, it is not so [9]. Spark can be used in Hadoop ecosystem in place of MapReduce and the tools for both can be used to perform certain actions. Spark performs in-memory processing and allows pipelined construction for data flow unlike Hadoop and MapReduce.

v. Apache Flink

It is a Streaming data flow engine, which helps us to perform distributed operation on stream of data. Flink consist of many API such as stream API, Static API and SQL like query API [10]. It also has its own Machine Learning and Graph Libraries. It works with Stream flow in real time also, which is not possible with Hadoop or Spark.

vi. Apache Strom

It is a Distributed Computational System, which designs its problem as DAG (directed acyclic graph). It can be used with any programming languages with all its application. Some of its application are Real-Time analysis, Distributed Machine Learning etc. It can run on top of YARN and thus can work with Hadoop ecosystem [9]. It is a stream Processing Engine unlike Spark.

vii. Apache Flume

It is a distributed and reliable system for collecting large amounts of log data to a centralized data space. There are three main components of Flume: sinks, sources and channels. Sink is mostly a distributed file system like HDFS. Sources does the task of listening and consuming the data. Channels are the mechanism by which Flume transfers data from sources to sinks.

viii. Apache Samza

It is another distributed stream processing framework. It is built on YARN for cluster Resource management to work with Hadoop.

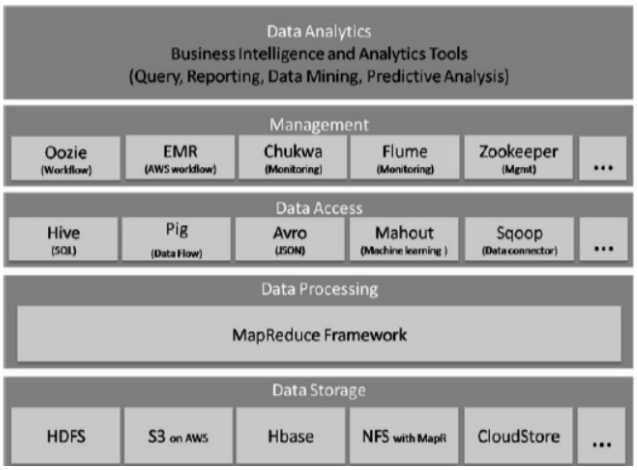

Fig. 3. BigData Tools [8]

B. Platforms

i. High performance computing (HPC) clusters

These are the cluster having thousands of cores to process data. It consists of different architecture, Cache, etc. which depends on the requirement of users. It is capable of processing huge data but it is not scalable. It can also be compared with Supercomputer [7].

ii. Multicore CPU

Multicores CPU consists of Various Cores, which helps the user to deal with huge data. It can process a huge amount of data. With the advancement in technology, the size of processors has decreased and now processors can be kept in one Single system. CPU had been a major part in Data analytics since last few years.

iii. Graphics processing unit (GPU)

GPU is an architecture used to create the frames of images and videos with a much higher speeds. GPU has been used in doing images related operations in recent times. GPU is having much more cores than Multicore CPU and it is having its own DDR5 memory, which is much faster than our DDR3 memory. Thus, Graphic programming is much more popular nowadays. The CUDA programming has given the opportunity of GPU programming to all the programmers to without having much knowledge about the hardware. As GPU has much, more cores that Multicore CPU it will have a huge impact in Big Data Analytics to analyze data in minimum time.

5. APPLICATION OF BIGDATA

Data Analytics is used in most of the fields nowadays. It is very rare that we do not see application of data analytics anywhere around, from schools and colleges to Multi-National Companies all are using it in some or the other way. Some of the applications are given below-

1. Big data in healthcare

The use of Big Data in Healthcare has been greatly helpful. We can collect the data of all the patients such as patient's name, diseases suffered, medicine given etc. With analysis of data, we can get remedies and detection of diseases very easily.

2. Market and business

Big Data has been a real boon to sales and market. With a huge amount of data available about customer behavior and needs, it has become very easy for companies to fulfil customers' demands and increase their sales.

3. Sports

This area is hugely benefited by the analysis of data. Sports is the heart of every nation. A huge number of athletes and players join the field. It is very important to keep data about the game and about each player for the sake of preparation, competition and rehabilitation. By analysis of data, we can know who needs rest, training, support, guidance and who are suffering from injury. It can be very useful in selecting young talents by keeping an eye on their data always.

4. Banking and Security

A huge amount of data is available in this sector and many crimes can be stopped by using Big Data Analytics. We can detect fraud and some illegal transactions by the analysis of general behavior of data. In addition, theft of cards or any other Security keys can also be detected by doing processing on data.

5. Communication, Media and Entertainment

One of the biggest amount of data comes from communication industry. The analysis of these data is quite necessary to efficiently use the network and for providing customers good services. Nowadays huge amount of data comes in form of video and audio and it is very important to analyze everything from it to get to a decision in real time.

Big data analytics is nowadays important and applicable to most of the sectors like Education, Transportation, Insurance, Energy sources etc.

6. CHALLENGES IN BIGDATA

All the factors of Big data i.e. Volume, Velocity, Veracity, Variety are all in itself a challenge. Volume is huge amount of data which was previously in gigabytes are now in tera and peta bytes. The Velocity with which these data comes from different sources are different and so to manage data at a variable speed is very difficult. In addition, there is a lot of Noise in the data and there comes Veracity in the frame, which affects the decision taken from the output. In addition, these data are of different forms and this shows Variety of data coming at a variable speed in a huge amount containing Noise, which cannot be analyzed by traditional system.

i. Data Representation

Most of the datasets have different level of heterogeneity of data and different meanings of data types. It is very important to have proper data Representation to have effective data analysis results. The way we represent our data affects the output of our results. [4]

ii. Data Compression and Redundancy Reduction

It is very important to see that the data we have does not have redundancy as it increases overall cost of both storage and analysis. Also a variety [3] of data comes from different sources at different speeds so it is very important to recognize the type of data and their compression level for instance video data occupy more space so we can use Compression techniques on it.

iii. Data privacy. Data security.

This is one of the major threat to both individuals and companies. Companies to make marketing strategies use the data posted by the users online. However, the privacy of personal information is very important as it can create many problems for individuals. In addition, the data companies are having should be secure as any malfunction to that data can create noise in it, which affects the results.

iv. Analytical process and Result Delivery

As we are supposed to work with a huge amount of structured and unstructured data simultaneously in real time, it is very important to work with proper analysis process as the structured data can be processed by RDBMS but the real problems come with the introduction of unstructured data. Also after successful processing of the data, it is very important to represent the result in a form, which can be easily understood by the client, as they might not be from technical background. [4]

v. Distributed Storage and Search

The data comes from geographically diverse location and are stored in a distributed manner and thus to store and maintain the data properly is very important.[4] In addition, as the data are stored in a distributed manner that might be at different locations Searching is also the property to be kept in mind.

vi. Infrastructure faults

Storing and analyzing such a large data requires proper hardware compatibility. As the data increases the need of more and more hardware is there and these hardware systems will be affected by the time and usage. Over a period of time system will be crashed of malfunctioned which makes loss of data. Companies cannot afford to lose their past data and thus they need backup hardware infrastructure, which just stores the data for backup.

7. REFERENCES

[1] Jasmine Zakir, Tom Seymour, Kristi Berg "BIG DATA ANALYTICS" Issues in Information Systems Volume 16, Issue II, pp. 81-90, 2015

[2] Tsai, Chun-Wei, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V. Vasilakos. "Big data analytics: a survey", Journal Of Big Data, 2015.

[3] Ripon Patgiri, Arif Ahmed. "Big Data: The V's of the Game Changer Paradigm", 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016.

[4] Neetu Chaudhari, Satyajee Srivastava. "Big data security issues and challenges", 2016 International Conference on Computing, Communication and Automation (ICCCA), 2016 .

[5] Sabia,Sheetal Kalra "Applications of big Data: Current Status and Future Scope" International Journal on Advanced Computer Theory and Engineering (IJACTE).

[6] Adiba Abidin, Divya Lal, Naveen Garg, Vikas Deep "Comparative Analysis on Techniques for Big Data Testing," 2016 InCITe.

[7] Dilpreet Singh, Chandan Reddy "A survey on platforms for big data analytics" Journal of Big Data 2014 1:8.

[8] Jaseena K.U, Julie M. David "ISSUES, CHALLENGES, AND SOLUTIONS: BIG DATA MININGNatarajan Meghanathan et al. (Eds) : NeTCoM, CSIT, GRAPH-HOC, SPTM – 2014.

[9] Justin Ellingwood. Hadoop, Storm, Samza, Spark, and Flink: Big Data FrameworksCompared[online].Available https://www.digitalocean.com/community/tutorials/hadoop-storm-samza-spark-and-flink-big-data-frameworks-compared.

[10] Matthew Mayo. Top Big Data Processing Frameworks [Online]. Available http://www.kdnuggets.com/2016/03/top-big-data-processing-frameworks.html.

[11] Peter Wayner. 7 top tools for taming big data [Online]. Available http://www.infoworld.com/article/2616959/big-data/7-top-tools-for-taming-big-data.html.