International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 11 • Issue 1 pp. 796-801 Jan 2019-June 2019 www.csjournals.com

Review Paper on Big Data Analytics

[1]Nisha Choudhary, [2]Dr. Sonal Sharma
[1]MCA Student, [2]Head-Department
Department of Computer Applications, Uttaranchal Institute of Management, Uttaranchal University, Dehradun

Abstract:- Big data analytics is refer to the analyzing of huge amount of the structured or unstructured data, using expert tools and software. On the daily basis people generate nearly 3 to 4 hex byte data. This complete data is called big data which can be in different forms or format. The data cannot be analyzed together which is difficult to process normally. Then we need some expert tools and software to generate some meaningful information. It is a great and unique research topic in the field of research. Although we still need some unique methodology or technique to analyze Big data. In this paper we have discussed about big data, its basic concept, tools, methodology, and research details.
Keywords:- Big data, Application, tools and technologies.

INTRODUCTION

Data is important for every single scenario in today world. We use and store data i.e. generated from various sources. Data is not technology, it is the combination of old and new technology that helps in generating valid ideology and results for profit gain. Big data is facility to manage a huge volume of unequal data, at the right speed and within the right time frame to allow real time analysis and reaction[1]. According to Apache Hadoop Big data is a dataset which could not be captured, managed, and processed by general computers within an suitable scope[2].

Data arrives from different sources used to collect information, posts from social platforms, pictures, videos and other form of data on internet, daily records, and telephonic conversation etc. All this data is Big Data. It is the combination of data sets that is large and complex that is hard to handle becomes of its large content to process using convenient database management tools[3].Therefore we need specialized tool and software to analyze the big data. Below the table shows different categories of data.

TABLE 1

DATA SIZE	DATA TYPE	TOOLS	ANALYTICAL METHODS	EXAMPLE
MEGA BYTE	SMALL DATA	Personal computer, excel	Statics data	Small companies database
GIGA BYTE	LARGE DATA	RDBMS	Data Mining	Database for huge companies
TERA BYTE	BIG DATA	Data warehouse	Map Reduce	Social platforms
PETA BYTE	BIG DATA	NoSQL , Hadoop	Distribute file system	Mobile, Multimedia

International Journal of Electronics Engineering (ISSN: 0973-7383)
Volume 11 • Issue 1 pp. 796-801 Jan 2019-June 2019 www.csjournals.com

CHARACTERISTICS OF BIG DATA

The big data can be broken down into 5 characteristics or the 5 V's.

- **Volume**

This include Large volume of data contains more size of data which cannot be managed or hold through earlier or traditional methods. In all the sectors there are almost 100 terabytes.

- **Velocity**

When the data formation is very quick and cannot be control which is nearly unstoppable and the speed of data incoming and outgoing is difficult to handle.

- **Variety**

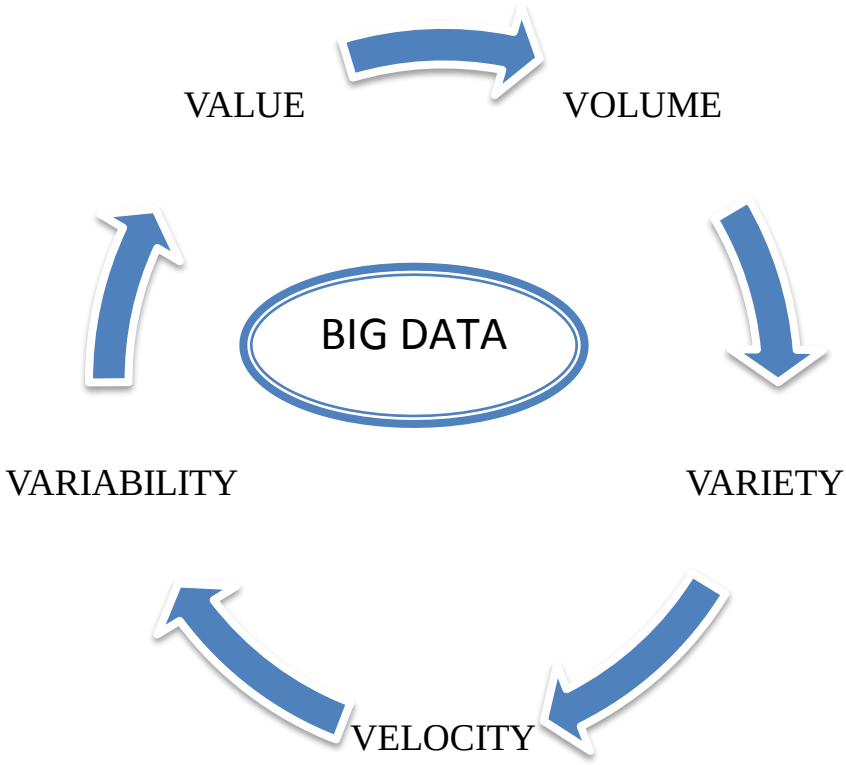
Data which is collected from different sources further include text, images, video and other structured or unstructured data.

- **Value**

It is the mixture of all the unstructured, structured and semi-structured data from which we take out the information needed at time in short insights view of data.

- **Veracity**

The better and Trustworthy quality of data.



TOOLS OF **BIG DATA**

1) **Apache Hadoop**

The most specialized platform for big data analysis is apache hadoop and Mapreduce . It is the combination of hadoop distributed file system (HDFS) , hadoop kernel, apache hive and mapreduce etc . it is the open source platform which has very low hardware requirements . It is the power full platform which helps in solving storage of data, processing and big data problems. Mapreduce is a programming set to process the large number of dataset is based on the divide and conquer methods.

MapReduce
- It is powerful model for parallelism.
- Based on the rigid procedural data.

Pig
- Data-flow language.
- It is used by the programmers and researchers.

Hive
- Used by generating reports.
- Declarative SQLish language.

2) **Apache Spark**

Apache spark is an best alternative open source frame work which is used for processing big data. It is the advance version which almost covers the draw back of hadoop and is extremely well in performance. It was develop 2009 in UC Berkeleys AMPLab. It is 100 times faster then MapReduce which helps in operate real time data. The main feature of apache spark is in-memory cluster computing that is increase the processing speed of application.

Spark SQL
- It is a component on top of spark core and introduces a new data instruction.
- It support for structured and semi-structured data.

Spark streaming
- It ingests data in mini-batches and performs RDD(Resilient Distributed Datasets)
- Perform streaming analytics.

MLlib(Machine Learning Library)
- MLlib is a distributed machine learning framework.
- Spark MLlib is a faster then hadoop.

3) **Apache Storm**

Apache storm directly on methodology of DAG (Direct Acyclic Graph) topology. The generated output files are in JSON format . Nimbus controls distributive code from the over all strong cluster and task perform by the worker nodes by monitoring the whole system.
- It is a open source and the part of apache project.
- Apache storm is real time processing system.
- It helps to process big data.

4) **R programming platform**

R is widely used with Python, Julia, R (JuPyteR satck) for analyzing and visualizing data. The most popular tool for analyzing data tool set is JupyteR Notebook. This platform is highly portable, supported by SQL Server, Windows, Apache Hadoop, Linux Servers and spark.

Data Manipulation
- It helps in accessing data sets and provides shapes and format which can be easily analyze.
- It is amongst one of the few tools which involves indexing techniques.

Data Analysis
- Data can be analyze with the help of R.
- R helps in implementation of various techniques, testing and machine learning.

Data Visualization
- R act as a packets for data visualization.
- R also sports graphic implementation and animation.

5) **Apache SAMOA**

It is part of Apache family of tools used for big data processing. It is specialized tool for successful data mining. It is best for Python platform and Machine Learning include the following:

- CLUSTRING
- CLASSIFICATION
- NORMALIZATION
- REGRESSION

6) **Weka**

Weka is a collection of many machine learning algorithm for data mining task. These algorithm can directly apply to data set for some one who is new to weka, it provides easiest function to work upon with its GUI. Weka is written in java programming language. It supports several data mining task.

7) **NodeXL**

NodeXL supports data visualization and relationship between software. It is the best way for exact calculations. NodeXL is statistical tools to analysis data that contain advance network mythology, social media network mythology and automation.

8) **Dryad**

Dryad provides a huge number of data including generation of job activities, managing of the machines for the required processes, handling transition failure in the cluster, collection of performance metrics, monitoring the job, and updating the job graph in response to these policy decisions without knowing the semantics of the vertices [5]

APPLICATION OF BIG DATA

1. **BANKING AND SECURITIES**

The Security Exchange Commission (SEC) is use big data to observe financial market. They are presently using NLP and network analytics to detect illegal activities in the markets. The industry also is dependent on big data to manage risk which include fraud activities , anti-money laundering etc.

2. **COMMUNICATION , MEDIA AND ENTERTAINMENT**

It is used to create meaning full content for different demand, performance raise by the targeted audiences. Some of the online platforms and software are highly demanded by the worldwide market handle by the million users who uses analyzed data to produced recommended information to the individual users.

3. **HEALTHCARE PROVIDERS**

It means big data plays a key role in healthcare departments as healthcare providers are indulge in large amount of data which is difficult to analyze and can only be analyze by the big data tools. Big data tools are best to meet the requirements and behavior to make the task easier by analyzing the pattern occurrence to achieve better results and will also improve medical facilities all over the world[4].

4. **EDUCATION**

In education big data plays important role like tracking recodes of worker and student as well as considering teacher effectiveness to insure good outcome for organizations related to education field. Not only in education field but also in government sector education technologies using big data to analyze detected data patterns.

5. **GOVERNMENT**

Big data has a high command in applications which includes researchers like health, fraud detection and informant protection. Big data is also wildly use to analyze large amount of social claims. The FDA uses

big data to detect the pattern of food which can further help in detection of the root cause of infection or diseases. It is not only affected for FDA but also for the government agencies and protect the country.

6. **MANUFACTURING AND NATURAL RESOURCES**

Big Data is the huge chance for production, sales and marketing and nowadays internet too because of the priority of the customer need, behavior and satisfaction. Therefore it is the main key to analyze the customer behavior and proven ways to elevate customer experiences.

7. **INSURANCE**

Big data has one of the key feature to analyze the customer behavior with the help of social media therefore it helps in providing transparent and good quality to the customer. The big data also allow for improved customer earnings from insurance companies.

The predictive analytics from big data is very much helpful to offer faster service since large data which can be analyzed particularly in the virtual step including scam detection [4].

8. **TRANSPORTATION**

The applications of big data in following fields include:

- Governments fully use big data for congestion manage, traffic and system planning, intellectual traffic light system.
- Private sector implements big data in transport like cost and income management, technological enhancement, logistics and for viable advantage.
- Individual entity use big data includes path planning and car functions to accumulate fuel for travel planning in tourism etc.

Mode	**Project Type**	**Year**	**Value**	**Technology**
Road	Congestion and traffic management	2010	66 million euro	IBM
Road	Traffic patterns and congestion	2006-2011	218 million euro	IBM
Road	Traffic management and congestion control	2013 ongoing	37 million euro	IBM

CONCLUSION

Big data analytics is a good research topic among many different fields and sector. We applied different methods to analyze big data. We have considered in our paper. But there are different tools, applications and open source software available on the paper. Analyzing data in today scenario is much challenging task. This research paper includes various tools and methods to analyze this data with the help of big data analytics. There is a scope for feature research to find out the best in this particular situation which can provide more storage and security.

REFERENCE

1) Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding,Data Mining with Big Data **IEEE Transactions on Knowledge and Data Engineering, January 2014, Volume 26, Issue 1,pp 97-107**
2) Min Chen, Shiwen Mao and Yunhao Liu (2014). BigData: A Survey, ©Springer Science+Business MediaNew York 2014, published online: 22 january.

3) http://en.wikipedia.org/wiki/Big-data
4) Vikas Upadhyay, Insha Shaikh, Big Data Analytics, Mumbai University, Mumbai, M.S India.
5) H. Li, G. Fox and J. Qiu, Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime, Second International Conference on Cloud and Green Computing, 2012, pp.675-683.