IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Big Data Analytics and its Tools

Deepak Kumar Verma[1*], Ashakti[2]

[1,2]Dept. of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Vidya Vihar, Raebareli Road, Lucknow, India -*226025*

[*]*Corresponding Author: deepak.discrete@gmail.com, Tel.: +91-9452834077*

DOI: https://doi.org/10.26438/ijcse/v7i4.876880 | Available online at: www.ijcseonline.org

Accepted: 19/Apr/2019, Published: 30/Apr/2019

***Abstract*—** In current scenario the Big Data has a rapid growth and has become a most popular term in the world of internet. The size of generated data is so huge and complex that traditional data processing application tools and platforms are inadequate to deal with it. This article is a result of a systematic analysis that discusses Big Data concepts and applications in various domains. The goal is to explore and understand the current research, opportunities, and challenges relating to the utilization of Big Data and analytics**.** The contribution of this paper is to provide an analysis of the available literature on big data analytics. Accordingly, some of the various big data tools, methods, and technologies which can be applied are discussed, and also shows the strength and limitations of Hadoop and HPCC systems based on some specific criteria**.**

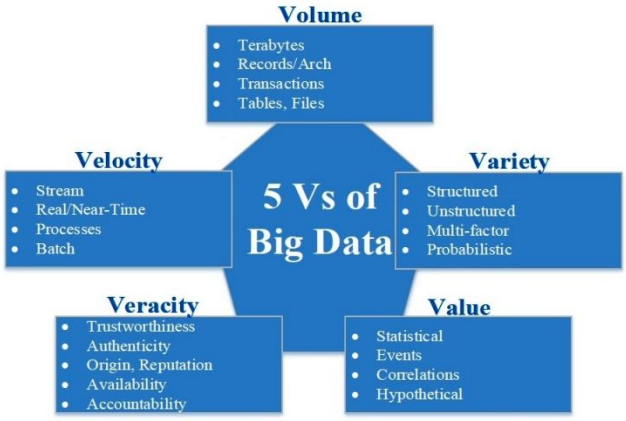
***Keywords*—**Bigdata, Bigdata Analytics, IoT.

I. INTRODUCTION

The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. In big data technology having four Phases with two computing techniques 5V's separately (Volume, Velocity, Variety, Veracity and Value). The Figure shows the overview of the big data phases and their characteristics.

Volume – The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, 'Volume' is one characteristic which needs to be considered while dealing with Big Data.

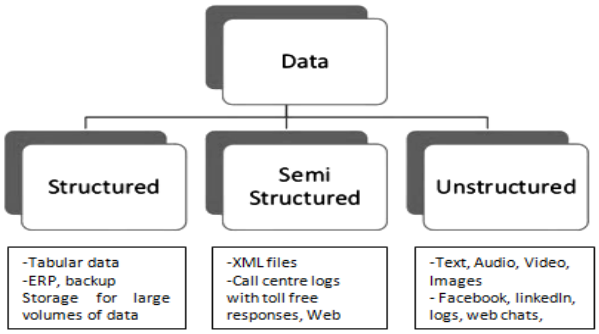
Variety **–** The next aspect of Big Data is its variety. Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.



Value- Value discourses to this approach are capable of turning data into towards value. It is important that businesses make a situation for any challenge to leverage and collect big data. It is easy to drop into the embark and buzz trap on big data enterprises without a clear kind of the business value it will take [7].

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. The flow of data is massive and continuous.

- **Types of Big Data:** BigData could be found in three forms: Structured, Unstructured and Semi-structured.



Structured: Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

Unstructured: Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format. Examples of un-structured data is the output returned by 'Google Search'

Semi-structured: Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

- **Strengths of Big Data**

• **Big Data is Timely** – 60% of each workday, knowledge workers spend attempting to find and manage data.

• **Big Data is Accessible** – Half of senior executives report that accessing the right data is difficult.

• **Big Data is Holistic** – Information is currently kept in silos within the organization. Marketing data, for example, might be found in web analytics, mobile analytics, social analytics, CRMs, A/B Testing tools, email marketing systems, and more… each with focus on its silo.

• **Big Data is Trustworthy** – 29% of companies measure the monetary cost of poor data quality. Things as simple as monitoring multiple systems for customer contact information updates can save millions of dollars.

• **Big Data is Relevant** – 43% of companies are dissatisfied with their tools ability to filter out irrelevant data. Something as simple as filtering customers from your web analytics can provide a ton of insight into your acquisition efforts.

• **Big Data is Secure** – The average data security breach costs $214 per customer. The secure infrastructures being built by big data hosting and technology partners can save the average company 1.6% of annual revenues.

• **Big Data is Authoritive** – 80% of organizations struggle with multiple versions of the truth depending on the source of their data. By combining multiple, vetted sources, more companies can produce highly accurate intelligence sources.

• **Big Data is Actionable** – Outdated or bad data results in 46% of companies making bad decisions that can cost billions.

II. RELATED WORK

Shilpa, Manjit Kaur (2013), in this paper author discussed that Big Data is characterized by the dimensions volume, variety, and velocity, while there are some well-established methods for big data processing such as Hadoop which uses the map-reduce paradigm. Using MapReduce programming paradigm the big data is processed and Big data analytics is the process of examining large amounts of data [1].

Judith Hurwitz, Alan Nugent, Dr. Fern Halper, and Marcia Kaufman (2013), Bigdata is new to many people, so it requires some investigation and understanding of both the technical and business requirements. Many different people need knowledge about big data. Some of you want to delve into the technical details, while others want to understand the economic implications of making use of big data technologies. Other executives need to know enough to be able to understand how big data can affect business decisions. Implementing a big data environment requires both an architectural and a business approach — and lots of planning.

Nada Elgendy, Ahmed Elragal (2014), This paper aims to analyze some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains. In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high

in variety and velocity, which makes them difficult to handle using traditional tools and techniques.

Mrs. Mereena Thomas (2015), In this paper author discussed about Big data refers to datasets high in variety and velocity, so that very difficult to handle using traditional tools and techniques. The process of research into massive data to reveal secret correlations named as big data analytics. Big Data is a data whose complexity requires new techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. We need a different platform named Hadoop as the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes[3].

Kevin Taylor-Sakyi (2016), However Big Data Analytics has a few concerns including Management of Data lifecycle, Privacy & Security, and Data Representation. This paper reviews the fundamental concept of Big Data, the Data Storage domain, the MapReduce programming paradigm used in processing these large datasets, and focuses on two case studies showing the effectiveness of Big Data Analytics and presents how it could be of greater good in the future if handled appropriately [4].

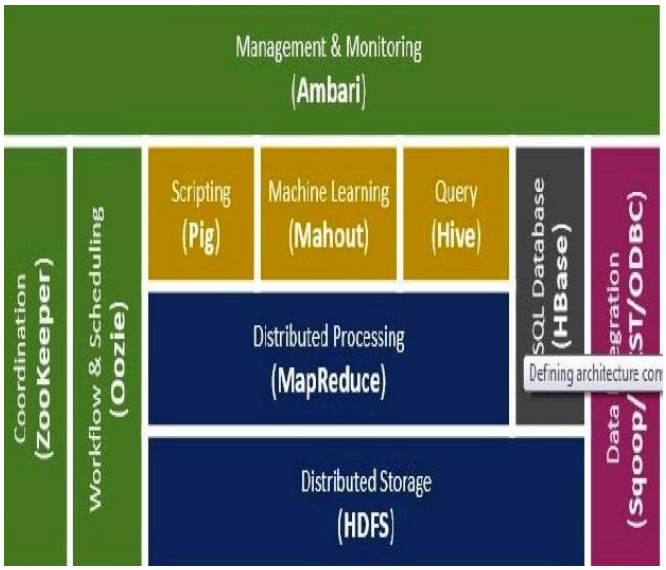
AK Bharti, Neha Verma, Deepak Kumar Verma (2019), have analyzed various big data analytics tools in context with scalability and realized that when number of available options increases, the task of selecting big data analysis tools becomes difficult as the available tools have their own advantages and disadvantages. Because the world is developing rapidly, the ordinary tools of data mining have become ineffective, which led to the tendency to use data processing tools that can be scalable and distributed. In this paper authors have reviewed the Hadoop structure and a look at the projects accompanying it. Spark was also highlighted as one of the most important tools in-memory computing and its various components were reviewed. In the end, a brief overview of H2O was presented as one of the most important tools in the process of analyzing the big data [5].

III. TECHNIQUES AND METHODOLOGY

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data. There are many solutions to handle the Big Data, but the Hadoop and HPCC are two most widely used technologies. We have studied both of these techniques to identify the strength and limitations of these techniques.

3.1 Hadoop

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's Mapreduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, Mapreduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. The base of Hadoop paradigm consists of the following four modules which are shown in Figure [5].



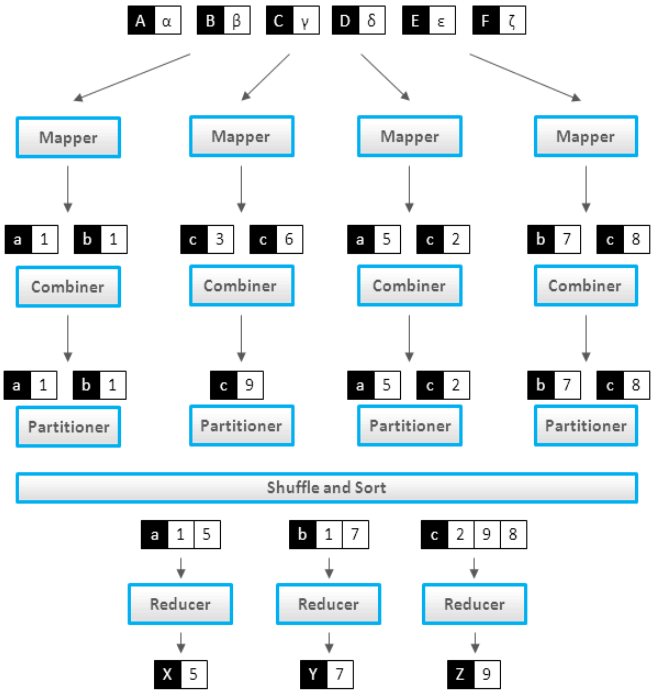
- ***HDFS*:** HDFS is a block-structured distributed file system that holds the large amount of Big Data. In the HDFS the data is stored in blocks that are known as chunks. HDFS is client-server architecture comprises of NameNode and many DataNodes. The name node stores the metadata for the NameNode. NameNodes keeps track of the state of the DataNodes. NameNode is also responsible for the file system operations etc. [5].When Name Node fails the Hadoop doesn't support automatic recovery, but the configuration of secondary nod is possible. HDFS is based on the principle of "Moving Computation is cheaper than Moving Data".

- ***MapReduce:*** MapReduce is a programming framework for distributed computing which is created by the Google in which divide and conquer method is used to break the large complex data into small units and process them .MapReduce have two stages which are:

  Map ():- The master node takes the input, divide into smaller subparts and distribute into worker nodes. A worker node further do this again that leads to the multi-level tree structure. The worker node process the m=smaller problem and passes the answer back to the master Node.

  Reduce ():- The, Master node collects the answers from

all the sub problems and combines them together to form the output.



- **HBase:** it is open source, Non-relational, distributed database system written in Java. It runs on the top of HDFS. It can serve as the input and output for the MapReduce.
- ***Pig:*** Pig is high-level platform where the MapReduce programs are created which is used with Hadoop. It is a high level data processing system where the data sets are analyzed that occurs in high level language.
- ***Hive:*** it is Data warehousing application that provides the SQL interface and relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query and analysis.
- ***Sqoop:*** Sqoop is a command-line interface platform that is used for transferring data between relational databases and Hadoop.
- ***Avro:*** it is a data serialization system and data exchange service. It is basically used in Apache Hadoop. These services can be used together as well as independently.
- ***Oozie:*** Oozie is a java based web-application that runs in a java servlet. Oozie uses the database to store definition of Workflow that is a collection of actions. It manages the Hadoop jobs.
- ***Chukwa:*** Chukwa is a data collection and analysis framework which is used to process and analyze the large amount logs. It is built on the top of the HDFS and MapReduce framework.
- ***Flume:*** it is high level architecture which focused on streaming of data from multiple sources.
- ***Zookeeper:*** it is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information etc.

3.2 HPCC

HPCC (High Performance Computing Cluster) is a open source computing platform and provide the services for management of big data workflow. HPCC' data model is defined by the user. HPCC system is designed to manage the most complex and data-intensive analytical problems. HPCC system is a single platform, a single architecture and a single programming language used for the data processing. HPCC system is based on Enterprise control language that is declarative, on-procedural programming language HPCC system was built to analyze the large volume data for the purpose of solving complex problem. The table shows the strength and limitations of Hadoop and HPCC systems based on some specific criteria:

Criteria	**HPCC Systems**	**Hadoop Systems**
Origin	Thor was invented in 1999 by LexisNexis specifically to solve large graph problems. The business model is based upon consuming large volumes of structured and unstructured data and converting it into a massive social graph of people and businesses.	Hadoop was invented at Yahoo to index web data. The project was started in 2006. In 2008 Yahoo released the source as open source.
High Level Scripting Languages	ECL (Enterprise Control Language) is a language built specifically to tackle the complexities around MPP data problems. KEL (Knowledge Engineering Language) is dedicated to tackle graph problems around big data.	Pig, Hive. Both the languages convert high level scripts to MapReduce jobs
Real-time Query	Yes. Thor data can be indexed and deployed to ROXIE for high performance Real-time query. The ROXIE and Thor components form the HPCC Systems platform and are completely open source and available for free. The ROXIE architecture was specifically built for random access performance and low latency and highly concurrent queries.	No. Third party vendors have built integrations on top of HDFS to provide this feature. Examples: Pivotal HD & Teradata. However, HDFS being block oriented was built with sequential access in mind
Monitoring	Tightly integrated with world class monitoring tools - Ganglia & Nagios. Available as Open Source and packaged as part of the platform	Mostly vendor dependent
	The Thor architecture was	The Hadoop

Parallelism Architecture	based on the data flow paradigm that supports three types of parallelism: 1. Data Parallelism = Where data is divided in parts and distributed across multiple nodes. Compute occurs on each node in parallel. 2. Pipeline Parallelism = Two consecutive operations can work on the same dataset at the same time. As soon as parts of the data is processed by the first it becomes available to the next operation. 3. System Parallelism = If the system detects that two or more operations are independent, the system will try to execute all of them in parallel.	architecture is based on the Map Reduce paradigm that was originally used by Google to index large volumes of content on the web. The (only) parallelism is derived by mapping (Map phase) the data into multiple parts, processing the maps and then consolidating (Reduce phase) the data into an output format.
Shared Nothing Architecture	Yes	Yes
License	Apache 2.0	Apache 2.0
Native Binaries	Yes. Coded in C++ and compiled to native binaries	No. Based on a JVM.
Enterprise Licensing Required	No. There is exactly one package available - The HPCC Systems Platform. It is enterprise ready and is the platform running the LexisNexis 1.4 billion dollar business and several Reed Elsevier projects.	Multiple packages and vendor dependent. Most vendors offer a separate enterprise license.

IV. CONCLUSION AND FUTURE SCOPE

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software's has produced a unique moment in the history of data analysis. Because the world is developing rapidly, the ordinary tools of data mining have become ineffective, which led to the tendency to use data processing tools that can be scalable and distributed. In this paper we have analyzed various big data analytics tools and realized that when number of available options increases, the task of selecting big data analysis tools becomes difficult as the available tools have their own strength and limitations. We have reviewed the Hadoop structure and HPCC Systems (High Performance Computing Cluster) to find out the strengths and limitations of both data analytics tools based on few popular criterias. As we can see that table shows the comparison between both and we can say that the selection of big data analysis tool depends upon the type of data, amount of data and the operations/queries to be performed on the data. Much cannot be said until data to be processed is not analysed.

REFERENCES

[1] Shilpa, Manjit Kaur," BIG Data and Methodology-A review", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, PP.991-995, October 2013.

[2] Smorodin, G, "Big Data-driven world needs Big Data-driven ideology", Big Data as the Big Game Changer, PP.991-995,2015.

[3] Mrs. Mereena Thomas," A Review paper on BIG Data", International Research Journal of Engineering and Technology (IRJET), Volume: 02,PP.1030-1034, Dec-2015.

[4] Kevin Taylor-Sakyi," Understanding Big Data", https://www.researchgate.net/publication/291229189,PP.01-08, 13 June 2016.

[5] AK Bharti, Neha Verma, Deepak Kumar Verma, "A Review on Big Data Analytics Tools in Context with Scalability", International Journal of Computer Sciences and Engineering, Vol.7, Issue.2, pp.273-277, 2019. DOI: https://doi.org/10.26438/ijcse/v7i2.273277

[6] Aditya B. Patel, Manashvi Birla, Ushma Nair,"Addressing Big Data Problem Using Hadoop and Map Reduce" PP.01-21,6-8Dec,2012.

[7] AK Bharti, Rashmi Negi, Deepak Kumar Verma, "A Review on Performance Analysis and Improvement of Internet of Things Application", International Journal of Computer Sciences and Engineering, Vol.7, Issue.2, pp.367-371, 2019. DOI: https://doi.org/10.26438/ijcse/v7i2.367371

[8] Xiaomeng Su,"Introduction to Big Data",Institutt for informatikk og e-loering ved NTNU,PP.01-11.

[9] Sun, D., G. Zhang, S. Yang, Zheng W., S. U.Khan and K. Li, "Re-stream: Realtime and Energy-efficient Resource Scheduling in Big Data Stream Computing Environments", Information Sciences, No. 319, pp. 92-112, 2015.

[10] Liu, X., N. Iftikhar and X. Xie, "Survey of Real-Time Processing Systems for Big Data", 18th Int. Database Engineering and Applications Symposium, New York, pp. 356–361, USA, 2014

[11] Bakshi, K,"Considerations for big data: Architecture and approach" 2012.

Authors Profile

Dr. Deepak Kumar Verma, MCA, UGC-NET, Ph.D.(CS), pursed Bachelor of Science and Master of Computer Application from University of Lucknow, India in year 2011. Dr Verma completed his doctrate in Computer Science in the year 2016 and currently working with Department of Computer Science, Babasaheb Bhimrao Ambedkar University(A Central University), Lucknow, India. His research interests are Artificial intellegence, data security and Software Engineering. He has published number of research papers in reputed national/international journals and conferences.

Ms. Ashakti Agnihotri pursed Bachelor of Computer Application (BCA) from MJP Rohilkhand University, Bareilly. She is currently pursuing Master of Computer Application (MCA) from Department of Computer Science, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, India. Her Research Interests are BigData, Internet of Things and cloud computing.

