© 2019 IJSRST | Volume 6 | Issue 4 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology
DOI : https://doi.org/10.32628/IJSRST196422

# Power of Big Data System for Storing and Processing Huge Data

**Dr. S. Natarajan[1], Dr. S. Rajarajesware[2], Suresh Ram R[3]**

[1]Professor, Department of Computer Science, Karpagam Academy of Higher Education (Deemed to be University), Coimbatore, Tamil Nadu, India

[2] HoD, Department of Computer Engineering, Sree NarayanaGuru Polytechnic College, Coimbatore, Tamil Nadu, India

[3]R, B. Tech (EEE) Student, Amrita Vishwa Vidhyapeetam (Deemed to be University), Coimbatore, Tamil Nadu, India

## ABSTRACT

Big data uses storage of huge data with some approaches and techniques to manage and process them. During the past few years the number of persons using internet, email and other internet-based applications has been growing tremendously. Big Data is mainly characterized by 3V's (Volume, Velocity and, Variety). The Big Data Architecture Framework (BDAF) is proposed to address all aspects of the Big Data Ecosystem. BDAF includes components such as Big Data Infrastructure, Big Data Analytics, Data structures & models, Big Data Lifecycle Management and Big Data Security. Nowadays the volume of data used by the people throughout the world is increasing enormously and exponentially. So, the need for storing, processing and protecting large volume of data has been becoming a great challenge in the modern hyper-connected world. On the basis of work from home concept lot of software professionals are doing their jobs with their internet connected systems for development, implementation, testing and maintenance of various softwares. These professionals and experts are sending and receiving lot of data to various locations to their clients, higher authorities and other officials frequently depending upon their requirements. The traditional data management models are not efficient for today's exponentially growing data from variety of industries. This challenging task of storing and managing huge volume of data is achieved in Big Data Systems. In this paper we try to give an overview of Big Data Analytics system for storing and processing huge volume of various types of data. Overwhelming the security threats due to various factors like viruses, worms, etc are also great challenges to protect huge volume of data in a big data system.

**Keywords:** Big Data, volume, variety, velocity, Big Data Analytics System, HDFC, Hadoop

## I. INTRODUCTION

A key success factor for any organization is the availability of relevant data and information at the right time. Big Data Analytics deals with processing huge data using modern methods that uses various algorithms and tools like Hadoop to give required results efficiently. Big Data analytics involves storing, searching, processing and analysing huge data.The input data is not received from a single source. They are received from various sources of different nature known as heterogeneous sources. The input data may

be structured or unstructured or semi structured. These input data as well as the output data may be textual or voice data or moving pictures or video or any other types.

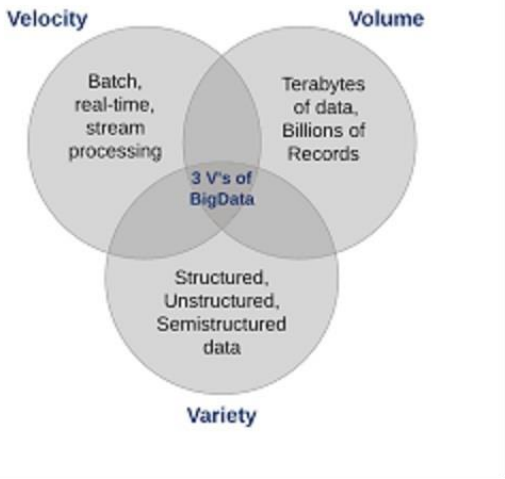



**Figure 1 :** The 3V's of Big Data

Big data refers to huge datasets of various types with high velocity. These datasets have been growing rapidly. Every day, huge volume of new data is being created and stored for further processing to get required information. Big Data Analytics when applied on big data helps in decision making in various areas like business, weather forecasting, scientific data analysis, etc.

In Figure 1 volume refers to the amount of data being created is vast compared to traditional data sources, variety refers to data that comes from different sources and is being created by machines as well as people and velocity refers to data is being generated extremely quickly.

## II. LITERATURE SURVEY

F.Chang, J.Dean, et al. in 2006 suggested that Bigtable is a distributed storage system to manage structured data to scale to a huge size. Lot of projects at Google store data in Bigtables. Bigtable has successfully provided an extensible, high-performance solution for all of the Google products. They describe the simple data model provided by Bigtable.

Azza Abouzeid et al. (2009), concluded that HadoopDB is a hybrid of the parallel DBMS and Hadoop approaches to data analysis, achieving the performance and efficiency of parallel databases, yet still yielding the scalability, fault tolerance, and flexibility of MapReduce-based systems. The ability of HadoopDB to directly incorporate Hadoop and open source DBMS software (without code modification) makes HadoopDB particularly flexible and extensible for performing data analysis at the large scales expected of future workloads. MapReduce-based systems are having superior scalability, fault tolerance, and flexibility to handle unstructured data.

Prashant Kumar, Khushboo Pandey (2013) described the concept of Big Data along with Volume, Velocity and variety of Big Data. The paper also focuses on Big Data processing problems. These technical challenges must be addressed for efficient and fast processing of Big Data. Their paper described Hadoop which is an open source software used for processing of Big Data.

Priya P. Sharma et al, (2014), suggested that in Big Data Era, where data is accumulated from various sources, security is a major concern as there is no fixed source of data. With the Hadoop gaining larger acceptance within the industry, a natural concern over the security has spread

Arushi Gupta, Asmita Sharma et al, (2016),Hadoop is an open source software framework that allows the distributed processing of large data sets across clusters of commodity computers using a single programming mode. Hadoop was created by Doug Cutting and Mike Caffarella in 2005. Framework-Hadoop provides toolset, connections to develop and run

software applications. Hadoop distributes the file in small chunks over thousands of nodes and process the data in parallel way.

## III. OBJECTIVE

In this paper we give an overview of Big Data analytics system for storing and processing huge data. For this purpose, Non-relational databases, like Not Only SQL (NoSQL), were developed to store and manage unstructured and semi structured and other non-relational, data. NoSQL databases are increasingly used in big data and real-time web applications. They may support SQL-like query languages. NoSQL is also known as "Not only SQL" .Open source databases supporting NoSQL are available for BDA. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. Hadoop allows running applications on systems with thousands of nodes with thousands of terabytes of data

## IV. HADOOP

Hadoop is a framework for performing big data analytics that provides reliability and, scalability by providing an implementation for the MapReduce. Hadoop consists of following main components:

i) HDFS for the big data storage and processing.
ii) MapReduce for big data analytics. The HDFS storage function provides a redundant and reliable distributed file system that is optimized for large files, where a single file is split into number of blocks and distributed across other nodes. The data is provided to the other nodes by a replication mechanism to protect the data from node failures. Data Nodes and the Name Nodes are two types of HDFS nodes. Data is stored across the multiple Data Nodes, and the Name Node directs the client to the particular Data Node which contains the requested and required data.

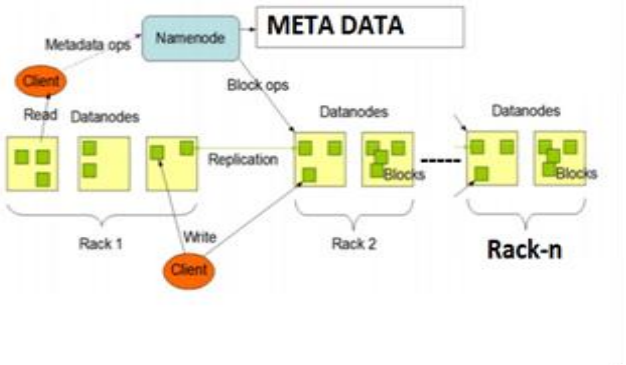



**Figure 2:** Map Reduce architecture.

### A. Big Data Storage and Management

Few years back data processing is done by storing and retrieving data in databases for processing. The main drawback here is that this system cannot be used for processing huge data of different varieties. In Big data analytics, data of different varieties should be stored and retrieved quickly for analysis.

The following are some of the available databases that we can use with NoSQL

i) Neo4J.

It is a Java-based open source NoSQL graph database

ii) Hypertable

It is a high performance, open source, massively scalable database which is modelled after Bigtable,

ii) Accumulo

This is based on Google's BigTable design and it is built on the top of Apache Hadoop.

iii) Hadoop/Hbase

Hadoop uses Apache HBase when we need random, real-time read/write access to the Big Data. Apache HBase is modelled after Google's Bigtable which is an open-source, distributed with versioned non-

relational database Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.

iv) OrientDB

It is an Open Source NoSQL DBMS having both Document and Graph DBMS.It uses Java and can store up to 150,000 records per second on normal hardware. Even for a Document based database, the relationships are managed as in Graph Databases with direct connections among records. We can traverse either the parts or the entire trees and graphs of records within a few milliseconds.

vi) InfiniteGraph

This is a distributed graph database implemented in Java, and is from a class of NOSQL data technologies focused on graph data structures. Graph data typically consist of objects (nodes) and their relationships (edges) connecting two or more nodes. We may use InfiniteGraph to build web and mobile applications and services that need to solve problems on graphs.

vii) Flink

Apache Flink is an open source system very much useful in data analysis. It has the scalability and programming flexibility of distributed MapReduce-like platforms with query optimization capabilities as in parallel databases.

viii) MongoDB

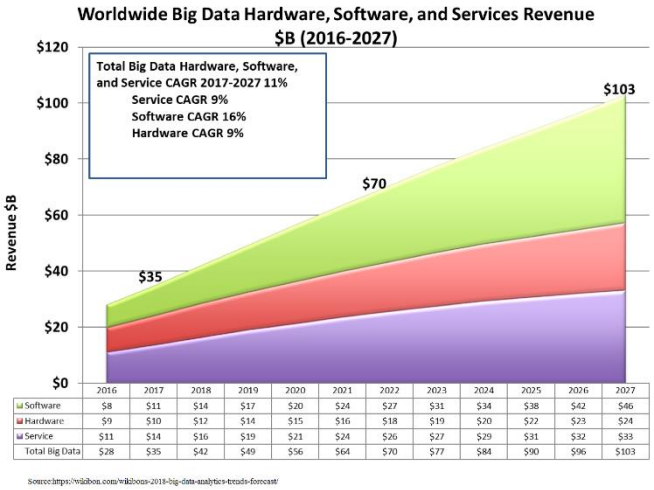
MongoDB is an open-source database used by companies of all sizes for various kinds of applications. MongoDB is useful for its high scalability, performance (read as well as write) reliability, operational flexibility and availability.

ix)DensoDB

DensoDB is a new NoSQL document database highly useful for .Net environment in c# language. It's not so complex but simple and reliable. It is fast as it provides direct access to the DataBase memory and one can do manipulations quickly. It provides the power of a distributed scalable fast database, with or without a server environment.

Map/Reduce is a term commonly thrown about these days, in essence, it is just a way to take a big task and divide it into discrete tasks that can be done in parallel. A common use case for Map/Reduce is in document database.Mapping and reducing is done to increase the efficiency of storage and retrieval process.

The following Big data market forecast chart gives importance of big data in current scenario.




SOURCE : *https://wikibon.com/wikibons-2018-big-data-analytics-trends-forecast/*

**Figure 3 :** Big data market forecast chart

## V. CONCLUSION

The volume of data usage in today's life is increasing enormously. Storing and processing huge data are very challenging tasks. In this paper we gave an overview of storing and retrieving data in a Big data analytics system. We can do this efficiently with the usage of HDFS along with MapReduce. In future some other databases can be tested for different environments to increase the efficiency of the system.

## VI. REFERENCES


[1] Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments www.securosis.com.

[2] S. Natarajan and S. Rajarajesware, "Computer Virus: A Major Network Security Threat," International Journal of Innovative Research & Development, vol. 3, no. 7, pp. 229-302, 2014.

[3] Arushi Gupta, Asmita Sharma, AsthaSahu, Anjali Mukati and AshleshaPanse, (2016), 'Study Of Pros And Cons In The Education System Using Big Data', International Journal Of Engineering Sciences & Research Technology.

[4] Miss Gurpreet Kaur Jangla and Mrs. Deepa.A.Amne, 'Development of an Intrusion Detection System based on Big Data for Detecting Unknown Attacks', ISSN (Online) 2278-1021ISSN (Print) 2319 5940 International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, December 2015.

[5] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, (2014), 'A Review Paper on Big Data and Hadoop', International Journal of Scientific and Research Publications, Volume 4, Issue 10, ISSN 2250-3153.

[6] Priya P. Sharma et al, (2014), 'Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution', (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2126-2131

[7] Nishu Arora and Rajesh Kumar Bawa, (2014), 'A Review on Cloud to Handle and Process Big Data', International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 3, Issue 5.

[8] Big Data Analytics for Security Intelligence September 2013, CLOUD SECURITY ALLIANCE.

[9] . Seungwoo Jeon, Bonghee Hong, Joonho Kwon, Yoon-sik Kwak and Seok-il Song, (2013)

[10] 'Redundant Data Removal Technique for Efficient Big Data Search Processing', International Journal of Software Engineering and Its Applications Vol. 7, No. 4.

[11] Prashant Kumar B and Khushboo Pandeya, (2013), 'Big Data and Distributed Data Mining: An Example of Future Networks', Volume 1, Issue 2 (2013) 36-39 ISSN 2347 - 3258 International Journal of Advance Research and Innovation.

[12] Azza Abouzeid, Kamil BajdaPawlikowski, Daniel Abadi, AviSilberschatz and Alexander Rasin, (2009),' HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads'.

[13] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach

[14] Mike Burrows, Tushar Chandra, Andrew Fikes and Robert E. Gruber, (2006), 'Bigtable: A Distributed Storage System for Structured Data'.

[15] https://www.ey.com/Publication/vwLUAssets/EY__Big_data:_changing_the_way_businesses_operate/%24FILE/EY-Insights-on-GRC-Big-data.pdf

[16] https://wikibon.com/wikibons-2018-big-data-analytics-trends-forecast/ YuriDemchenko, Canh Ngo, Peter Membrey., Architecture Framework and Components for the Big Data Ecosystem Draft Version 0.2


### Cite this article as :

Dr. S. Natarajan, Dr. S. Rajarajesware, Suresh Ram R, "Power of Big Data System for Storing and Processing Huge Data", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 6 Issue 4, pp. 138-142, July-August 2019. Available at doi : https://doi.org/10.32628/IJSRST196422 Journal URL : http://ijsrst.com/IJSRST196422