ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

# A Study on Big Data Mining - Concepts, Platforms, Issues and Challenges, Future Scope

Pebila Shani.S, Pricilla.S

PG Scholar, Dept. of Computer Technology, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India

**ABSTRACT:** Big Data is defined as colossal amount of data or collections of bulky, composite or required data which become tricky or not viable to process using modern methodologies, standard database management or analytical solutions. “Big data” is pervasive, and it has been used to get across all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, concurrent data, and much more. Big Data mining has the potential of discovering knowledge and patterns from the hefty sets of information, streams of data due to its volume, variability, and velocity. Big data brings collectively the large amount of data with assorted data types that formerly never would have been considered. Big data includes structured data, semi structured data and unstructured data. This colossal amount of data is very valuable to get better quality of life and make our world a advanced place by extracting consequential associations, trends and patterns but it is insurmountable. The present paper highlights the concepts of 5 V’s in big data, the emerging platforms and also the issues and challenges with better enhancements for the future.

**KEYWORDS**: big data, concepts, data mining, data types,issues.

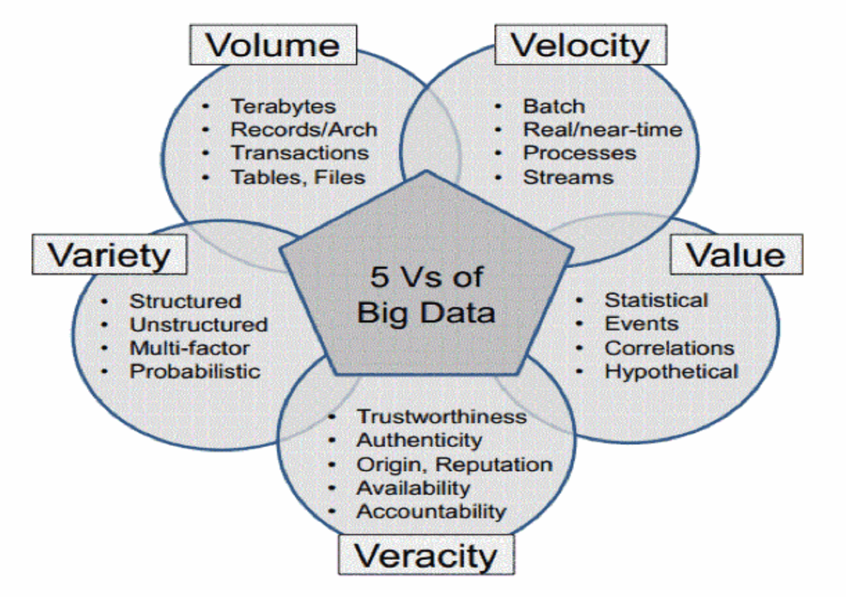
## I. INTRODUCTION

The data fashioned nowadays is probable in the order of zetta bytes, and it is rising around 40% every year. It is notable to understand that what is thought to be big data today won’t give the impression so big in the future. . The size of big data is so outsized, that it is nearly unfeasible to collect, process and store data using traditional database management system and by software techniques. Big data mining is used to extract handy information from large datasets which is habitually unstructured data. There are numerous data sources that are formerly untapped, or at least underutilized. Data becomes big data when individual data stops mattering and only a big collection of it or analyses derived from it are of value. But as said, everything has its flick side as well and big data too has its issues. Security and privacy are immense issues in big data due to its huge volume, high velocity, large variety like wide-ranging cloud infrastructure, assortment in data sources and formats, data acquisition of streaming data, inter cloud relocation and others. Data is generated at exponential rate, though big data has its own reward but the challenge is how to examine and carry out operation on data mining. So, we need a system gear which can handle big data and do mining not only on structured data but also on unstructured data resourcefully.

## II. BIG DATA CONCEPTS

With the quantity of data mounting exponentially, improved analysis is requisite to extract information that finest matches user interests. Big data refers to promptly growing datasets with sizes further than the capability of traditional data base tools to store, manage and estimate them. Big data volume is rising 40% per year, and will grow 44x stuck between 2009 and 2020. But while it’s habitually the most visible parameter, volume of data is not the only attribute that matters. Increase of storage capacities, Increase of processing power and accessibility of data are the main cause for the appearance and enlargement of big data. Big data refers to the consumption of large data sets to conduct the collection or reporting of data that serves businesses or other recipients in decision making. In fact, there are five key characteristics that characterize big data.

**A. Volume** -The dimension of data now is larger than terabytes and peta bytes and upsurge of size makes it complicated to store and analyze by means of traditional tools. Machine-generated data is shaped in much larger quantities than non-traditional data which concerns the big quantities of data that is generated incessantly. Primarily, storing such data was knotty because of high storage costs. On the other hand, with diminishing storage costs, this problem has been reserved somewhat at bay as of now. E-Commerce, Smartphone and social networking websites are examples where gigantic amounts of data are being generated. This data can be effortlessly distinguished between unstructured data, structured data and semi-structured data.
**B. Velocity** – Big data should be used to extract large amount of data surrounded by a pre defined period of time. The traditional methods of mining may obtain huge time to excavate such a volume of data. This concept is only realistic when the incoming data rate is slower than the batch processing rate and the delay is much of an obstacle. At current times, the speed at which such colossal amounts of data are being generated is exceptionally prominent and high.
**C. Variety** – Big data comes from a variety of sources which includes both structured and unstructured data. Documents to databases to excel tables to pictures and videos and audios in hundreds of formats, data is now losing structure. Structure can no longer be imposed like before for the analysis of data. Traditional database systems were designed to address smaller volumes of structured and consistent data whereas Big Data is geospatial data, 3D data, audio and video, and unstructured text, including log files and social media. This heterogeneity of unstructured data creates problems for storage, mining and analyzing the data.



**D. Value-** The fiscal value of diverse data varies drastically. Typically there is superior information hidden in the middle of a larger body of non-traditional data; the challenge is identifying what is effective and then renovating and removing that data for analysis.
**E. Veracity-** It means correctness. The raise within the range of values distinctive of a large information set. It is understandable when handling high volume, velocity and selection of information, all the data's aren't going 100% accurate, there will be dirty data. Big information and analytics technologies work with these types of information.

## III. PLATFORMS

Big data platform helps to manage, store and analyze big data to accomplish required business outcomes. Hadoop and HDFS have developed to become the heavy-handed platform for Big Data at large Web companies as well as less traditional corners of traditional enterprises. It is also complex to take out high-value insights which involve deep knowledge and understanding of many fields(machine learning, text analysis, data processing). At the same time, data analysts have grownup drained of the low-level Map Reduce programming. The newer technologies involved in Big Data Analytics involve Hadoop and associated tools such as YARN, Spark, Hive, Map Reduce and Pig as well as NoSQL databases.These languages comprise Pig from Yahoo and Hive from Facebook. Tasks articulated in these languages are compiled down into a sequence of Map Reduce jobs for execution on Hadoop clusters.

### A. Hadoop:

Hadoop is an open-source framework which is of java-based programming that ropes processing and storage of excessively large data sets in dispersed environment. It may run applications on systems with thousands of nodes and terabytes. Other components comprise Hadoop Distributed File System(HDFS) which is capable of storing data across thousands of commodity servers to attain high bandwidth between nodes. It distributes the file between the nodes and allows the system to prolong work in case of a node failure. It also uses Hadoop kernel which provides necessary libraries for the framework. Hadoop is commonly used for distributed batch index building; it is required to optimize the index capability in near real time. The pros of Hadoop is Distributed storage, tremendously climbable optimized for high output, Computational capabilities, tolerant of software and hardware failure, large block sizes.The cons of Hadoop is master processes are on its own points of failure. Hadoop does not deal storage or network level encoding, disorganized for handling small files. Hadoop affords components for storage and analysis for heavy scale processing. Hadoop rapidly emerged as a foundation for big data processing tasks including IoT(Internet of Things). It was encouraged by Google's Map Reduce and Google file system(GFS) , in which an application is packed up in to numerous smaller parts. After years of development with open-source frameworks, in 2012 Hadoop 1.0 became accessible as an Apache project publically by apache software foundation.

**Components of Hadoop:**

**HBase**: It is the open-source, in which the non-relational Hadoop databases are scalable, distributed, big data store.

**Oozie:** A server-based workflow scheduling system deals with hadoop jobs. Oozie use the database to collect the information of progress that is a collection of actions. It manages the Hadoop jobs in a authentic (mannered) method.

**Flume:** A tool to bring together, aggregate and shift massive amount of streaming data into HDFS.

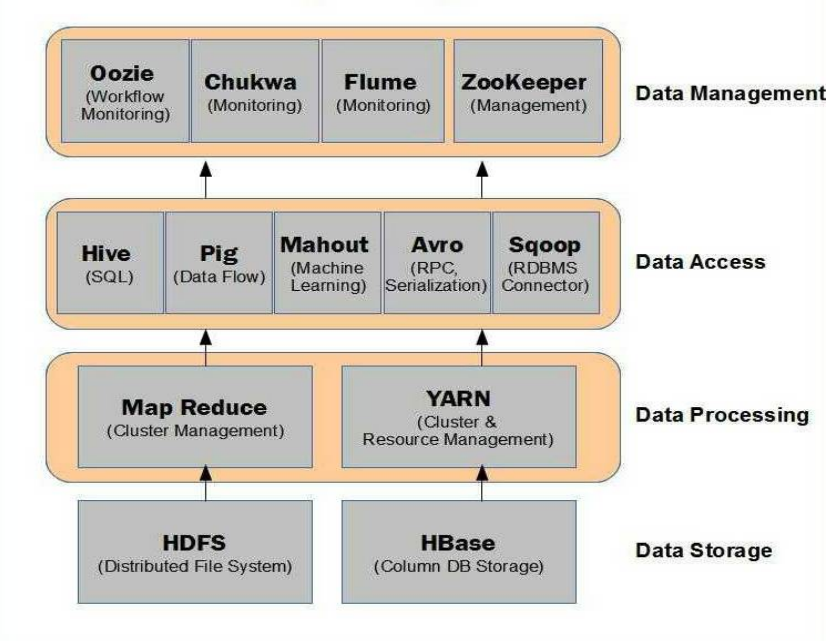
**Pig:** Pig is high-level platform that is engaged to run with Hadoop platform where the Map Reduce framework is created.

**Spark:** A speedy engine for big data processing which is qualified of streaming and supporting SQL, machine learning and graph processing.

**Sqoop:** Sqoop is a command-line interface application. It provides a platform which is employed for renovating information from Hadoop and relative databases or vice versa.

**Avro:** It is a system functionality of service of knowledge exchange and data publishing. It is basically exploited in Apache Hadoop. These services can be used along as well as independently according the information records.

**Chukwa:** Chukwa is a framework that's used for analysis to method and data assortment and analyzes the enormous quantity of logs. It is built on the higher layer of the Map Reduce framework and HDFS.


Oozie (Workflow Monitoring)
Chukwa (Monitoring)
Flume (Monitoring)
ZooKeeper (Management)
Data Management
Hive (SQL)
Pig (Data Flow)
Mahout (Machine Learning)
Avro (RPC, Serialization)
Sqoop (RDBMS Connector)
Data Access
Map Reduce (Cluster Management)
YARN (Cluster & Resource Management)
Data Processing
HDFS (Distributed File System)
HBase (Column DB Storage)
Data Storage

**B. Hive:**

The apache hive is a data warehouse software. It is built on the top of hadoop for query, data summarization, and analysis. It gives an SQL- like interface to query data stored in a variety of databases and file systems that incorporate with hadoop. Amazon maintains a software split of apache hive incorporated in Amazon Elastic Map Reduce on Amazon web services. Applications of apache hive are SQL, IBM DB2, oracle. Architecture is separated into Meta data for data storage, Map-Reduce-oriented execution and an execution half that receives a query from user or applications for execution. The pro of hive is more sheltered and implementations are levelheaded and well tuned. The cons of hive is only for unintentional queries and performance is with a reduction as compared to pig.

**C. Map reduce:**

it is a framework for processing parallelizable tribulations of extremely large datasets using considerable number of nodes which referred to as clusters or grid. It can take its advantage of processing the locality of data which reduces the communication overhead. Using Map Reduce framework the effectiveness and the time to repossess the data is sort of convenient. It works by isolating the input information into even-sized data blocks for alike load distribution originally. The information block is then allocated to one slave node and is processed by a map task and result is generated. The master node gets interrupted by the slave node when it is idle. The slave node then assigns a new task by the scheduler. The scheduler takes data section and resources into attention once it disseminates data blocks.Scheduling, re-executing failed tasks and Monitoring are taken care of by the framework.

- **"map" step:** master node make certain that the process applied is only one copy of unnecessary data. Each worker node has a local data which is applied to map() function. And the output is written in temporary storage.
- **"shuffle" step:** the worker nodes reorganize the data's based on the output keys such that all the data that belongs to one key is positioned on the same worker node.
- **"Reduce" step:** worker node now process each group of output data per key, in parallel.

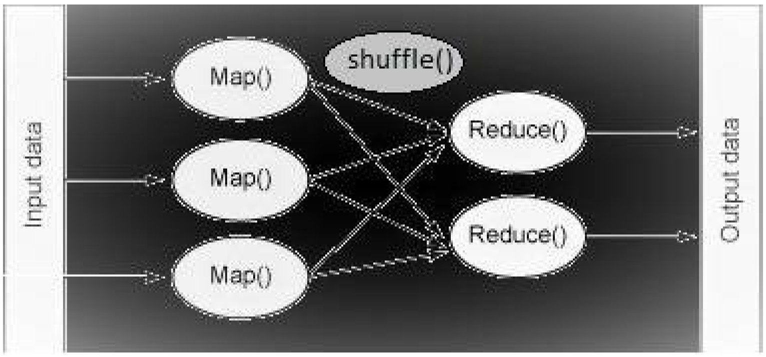
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017



## IV. ISSUES AND CHALLENGES

**A. Characteristics in big data concepts:** With the enhance in volume of data the value of data records have a tendency to diminish in proportion to age, type, richness and quality. The initiation of social networking sites have led to production of data of the order of terabytes on a daily basis which are difficult to be conducted using existing traditional databases.

**B. Human Resource and Man Power Issues:** Digital data arrives from many medium as contented to humans, like documents, drawings, sounds, video recordings, models and user interface designs, with or without metadata relating what the data is and its derivation and how it was collected. Immaturity with these new data types and sources and insufficient data management infrastructure are a big crisis.

**C. Technical Issues:**

- **Fault tolerance:** Fault-tolerant computing is monotonous and requires extremely composite algorithms**.** Whenever failure occurs the harm done must happen within satisfactory threshold rather than the complete work requiring to be redone.
- **Data Heterogeneity:** 80% of data are unstructured in today’s world. Working with unstructured data is problematic and high-priced too. Converting these to structured data is impracticable as well.
- **Data Quality:** There is no point in storing very hefty data sets that are unrelated as better outcome and conclusions can’t be drawn from them.
- **Scalability:** It is capable of aggregating numerous diverse workloads with different performance goals into vastly large clusters. This needs high level of sharing of resources that is fairly high-priced and brings along with it different challenges like executing numerous jobs so that the goal of every workload is met successfully.

**D. Processing issues:** The parallel processing and new analytics algorithms make available rapid information. Often it may be unknown to deal with a very large and varied volume of data and whether all of it needs to be analyzed.

**E. Security and privacy issues:** A major reason for security and privacy involve in big data is because big data is now extensively accessible. Data are shared on worldwide by scientists, business officials, government agencies have been urbanized till date to handle these colossal volumes of data are not resourceful enough to provide enough security and privacy to data. There is a lack of fundamental understanding about how to offer security to these massive volumes of data and adequate training is not provided on the subject of how to provide security and privacy to these big scale data. Big data lacks adequate policies that make certain agreement with current approaches to security and privacy.Reassessing and updating current approaches to prevent data leakage has to be done on a incessant basis.

## V. FUTURE SCOPE

Big data enormously has the capability to modify the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives. Companies like Google, General Electric, Yahoo!, Cornerstone, Kaggle, Microsoft, Facebook, Amazon that are investing a lot in Big Data research and projects. IDC predicted the value of Big Data market to be ―about $ 6.8 billion in 2012 mounting almost 40% every year to $17 billion by 2015. By 2017, Wikibon‗s Jeff Kelly expects the Big Data market will top $50 billion and the fusion will happen across information management, analysis, and search technology. By 2018, investing in this Business Intelligence(BI) enabler of end-user self-service will become a requirement for all enterprises. 70% of huge organizations by now pay for external data and 100% will do so by 2019. Acceptance of technology to continuously analyze streams of events will speed up in 2015 as it is applied to Internet of Things (IoT) analytics, which is predictable to produce at a five-year compound annual growth rate (CAGR) of 30%. By 2018 partly of all consumers will act together with services based on cognitive computing on a habitual basis. Taking an average of all the figures recommended by leading big data market analyst and research firms, it can be accomplished that approximately 15 percent of all IT organizations will move to cloud-based service platforms, and between 2015 and 2021, this service market is predictable to raise about 35 percent.

## VI. CONCLUSION

From a historical perspective, big data can be out looked as the latest generation in the evolution of decision support data management.By the 1990s, there was a necessity to hold up a wide variety of BI and analytic applications (e.g., reporting, executive information systems) with data. Having detached databases (i.e., independent data marts) for each application was expensive, resulted in data inconsistencies across applications, and failed to hold enterprise-wide applications. The outcome was the emergence of enterprise data warehouses (the second generation), which represented a data-centric approach to data management. The next generation (the third) was real-time data warehousing. Technology had recovered by 2000 so that it was likely to capture data in real time and supply it into the data warehouse. The capability to capture, store, and analyze high-volume, high-velocity, and high-variety data is permitting decisions to be maintained in new ways. It is also creating new data management disputes. As organizations make larger use of big data, it is likely that there will be increased concerns and legislation about specific privacy issues and there are business users, analysts, and data scientists who can work with and use big data.

## REFERENCES

1. Samiddha Mukherjee1, Ravi Shaw, "Big Data – Concepts, Applications, Challenges and Future Scope" Vol. 5, Issue 2, February 2016
2. Poonam G. Sawant, Dr. B.L.Desai, "Big Data Mining: Challenges and Opportunities to Forecast Future Scenario", Vol. 3, Issue 6, June 2015
3. Tejaswini U. Mane, Asha M. Pawar, "Big Data Mining Platforms: A Survey", Vol. 4, Issue 6, June 2016
4. Getaneh Berie Tarekegn, Yirga Yayeh Munaye, "Big Data: Security Issues, Challenges and Future Scope", International Journal of Computer Engineering & Technology (IJCET) Volume 7, Issue 4, July–Aug 2016
5. J.Vasuki, S.Priyadarshini, "A Study on Basics of Data mining, Machine Learning and Big data", Vol. 5, Issue 1, January 2017
6. Tejaswini U. Mane, Asha M. Pawar, "Big Data Mining Platforms: A Survey Tejaswini", Vol. 4, Issue 6, June 2016
7. Abhinav Kathuria, "Issues and Challenges in the Era of Big Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 6, June 2016
8. Subaira.A.S, Gayathri.R, Sindhujaa, "Security Issues and Challenges in Big Data Analysis", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 2, February 2016.