iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** X **Month of publication:** October 2019

DOI: http://doi.org/10.22214/ijraset.2019.10073

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue X, Oct 2019- Available at www.ijraset.com

Big Data Analytics: A Concept, Challenges and Research Issue

Nagendra Kumar Tiwari[1], Neelam Sahu[2]
[1, 2]*Dept. of IT, Dr. C.V. Raman University, Bilaspur, India*

Abstract: ***The handling of high dimension data set play very important and challenging task for every organization and institutes. Big data is huge amount of data in structured, unstructured and semi structured format and this huge amount of data generated through various sources like Sensors, Surveillance Systems, Social media, and Networking etc. We know that our daily life is working in machine or devices like reading news paper through mobile , online shopping, and other things. In this paper we have given concept of big data, dimension reduction techniques, security in big data, challenging on big data and big data analytics tools. It is very challenging task to overcome the problem of big data like dimension reduction and its security.***
Keywords: Dimension Reduction, Hadoop, Data Security.

I. INTRODUCTION

Today, every person generate lots of data, sometime its important data sometime not. When many people generating data in this amount than we have to handle it carefully because without handling it, it can be a mess. Not only people but also organisations and companies also generating data not GB or in TB but in PB every day. Some of data is important some of not, to short out this problem we use data mining. But in large amount of data, data mining is very difficult so we use dimension reduction techniques. Dimension reduction techniques reduce the higher dimension of data and make it usable. We know that big data have lots of noises and unwanted data, and without removing it we are unable to use this kind of data. If data is unusable then what is the use of data, its unusable for organisations or for companies.

To solve this problem we use dimension reduction techniques. We can say dimension reduction techniques are a technique to remove unwanted data and remove noises from useful data. In other words we can say that dimension reduction techniques are tool to reduce the complexity of data and make it usable. Dimension reduction is a way to convert vast dimension data in similar dimension data and it contain similar information.

There are various researchers have worked in the field of big data and big data analytics tools and dimension reduction techniques. Fan J., et al., (2018) have used PCA (Principal Component Analysis) and worked on future development in theoretical area of PCA. They have also discussed relationship between PCA and factor analysis. J. Weng, et al., (2017) [4] have discussed an overview of some classic and modern dimension reduction methods, followed by a discussion of how to use the transformed variables in the context of analyzing survey data. M. Pavithra, et al., (2017) [5] have presented a comprehensive classification of different clustering techniques for high dimensional data. D. Lehmann, et al., (2016) [6] have considered the problem of making an appropriate technology selection for a given big data application, and has introduced a corresponding framework, denoted STANDART Selection Framework (SSF). A. Mehmood, et al., (2016) have conducted a comprehensive survey on the privacy issues when dealing with big data and how to deal with it.

M. Habib, et al., (2016) presented a review of methods that are used for big data reduction. It also presents a detailed taxonomic discussion of big data reduction methods including the network theory, big data compression, dimension reduction, redundancy elimination, data mining, and machine learning methods. S. K. Prabhakar, et al., (2016) [2] have concluded the high dimensional EEG data is reduced to a low dimension by techniques such as Independent Component Analysis (ICA) and Principal Component Analysis (PCA). J. Moreno, et al., (2016)[7] have discussed about security and challenges issues in big data and how researchers are dealing with these problems.

S. Mukherjee, et al., (2016) [8] have discussesed big data from its infancy until its current state. It elaborates on the concepts of big data followed by the applications and the challenges faced by it. A. Gholami, et al., (2016) [3] have reviewed several security and privacy issues on big data in the cloud. It described several big data and cloud computing key concepts such as virtualization, and containers.

II. BIG DATA AND DIMENSION REDUCTION TECHNIQUES

Big data is collection of huge amount of data that contains noise, irrelevant information etc in structured, unstructured and semi structured form. Big data is very useful data but we have to extract the useful data. Big data [8] comes in structured, semi structured and unstructured format. Normally big data is defined in 7Vs, 7Vs refer volume, velocity, variety, variability, veracity, visualization and value. Big data comes from various sources like cell phones, satellite, sensors, social media, and weather monitoring system or from internet of things. Due to this reason, we need to dimension reduction techniques. Here are various organizations have used different dimension reduction techniques like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA),Genetic algorithm , particle swarm optimization , raking based feature selection techniques like information gain, gain ratio etc.

III. DATA ANALYTICS TOOLS

Analysis of data is very essential and challenging task for each and every organization. In this research work we have explored Apache Hadoop big data analytics tools for analyzing of big data. Hadoop is (Beakta R., 2015) [1] open source application that can be use for process the Big data. Hadoop is very popoular for every organizations, researchers and industries.. Hadoop can process large data sets in distributed computer systems. Apache Hadoop contains Hadoop kernel, Map reduce system, HDFS and other parts. Hadoop is very essential component for data analysis .There are following Hadoop component available for high dimensional data set.

1) *HDFS (Hadoop Distributed File System):* HDFS is the storage layer of Hadoop.
2) *Map Reduce:* MapReduce is the data processing layer of Hadoop. It processes huge amount of data in parallel by dividing the job (submitted job) into a set of independent tasks.
3) *HBase:* HBase is a column-oriented database that runs on top of HDFS. It is a NoSQL database which provides random real-time read/write access to data in the Hadoop File System.
4) *Pig:* Pig enables writing complex data processing operators in Hadoop using Pig Latin programming.
5) *Hive:* Apache Hive is a data warehousing software on Hadoop that facilitates reading, writing, and managing large datasets residing in distributed storage using SQL.
6) *Mahout:* A library of scalable machine-learning algorithms, implemented on top of Apache Hadoop and using the MapReduce paradigm. Once big data is stored on the Hadoop Distributed File System (HDFS), Mahout provides the data science tools to automatically find meaningful patterns in those big data sets.
7) *Flume:* Flume is a reliable system for collecting large amounts of log data from many different sources in real-time.
8) *Oozie:* Oozie is a workflow scheduler system that is used to schedule Apache Hadoop jobs. It combines multiple jobs sequentially into one logical unit of work.
9) *Sqoop:* Sqoop is a data collection tool design to transport huge volumes of data between Hadoop and RDBMS.
10) *Zookeeper:* ZooKeeper is a high-performance coordination service for distributed applications. It provides a centralized service for maintaining configuration information, providing distributed synchronization, and providing group services.

IV. BIG DATA SECURITY

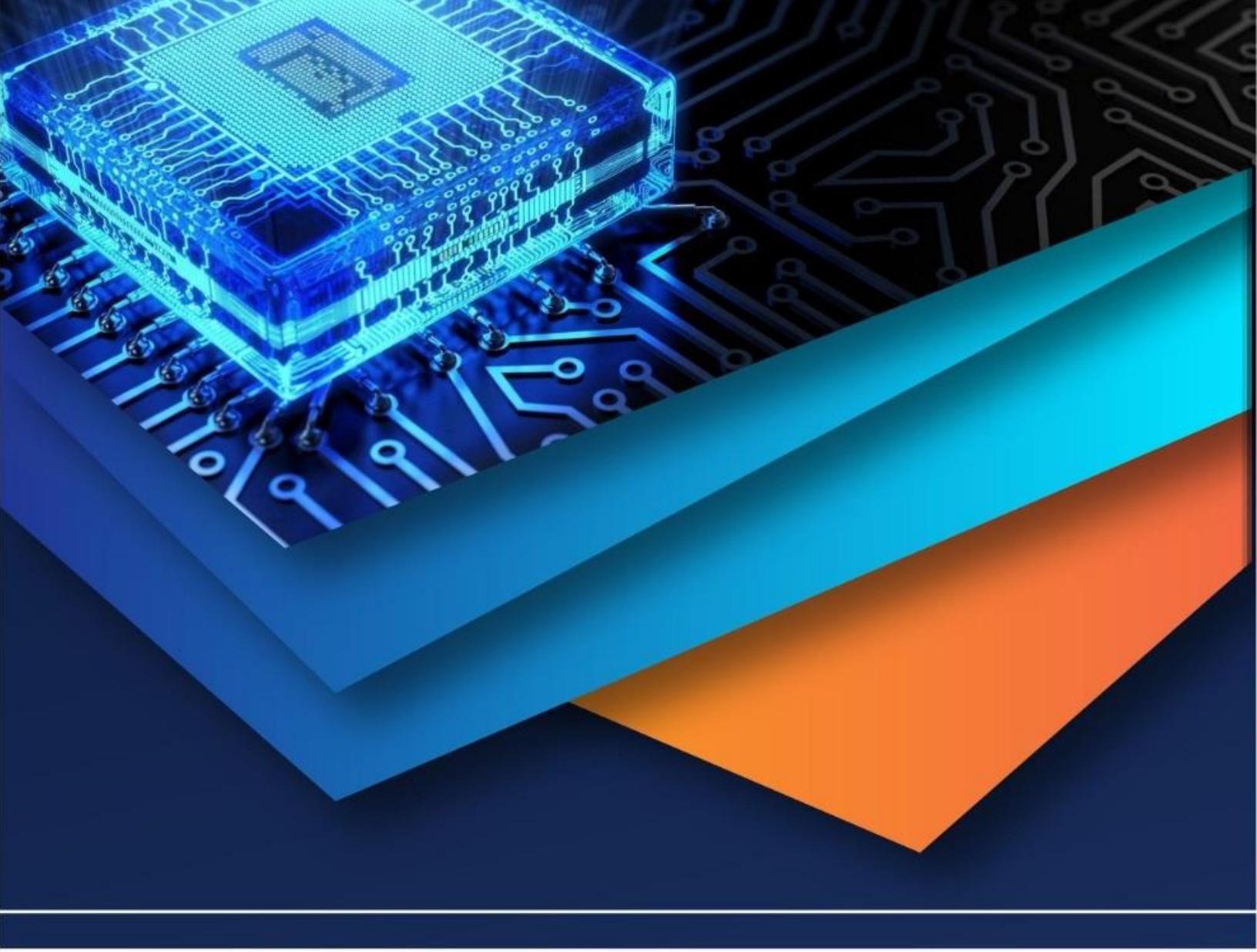
Security is very important role because most of the transaction is doing through online like online shopping, buying anything , hiring cars and many things. This task can generates huge amount of data in every organization. The security is challenging task for big data. Big Data security[7] is categorizes into infrastructure security, data privacy, data management, and integrity and reactive security Infrastructure security is first issues of big data security, it means how data is secure where is generates huge amount of data in various organizations. Data privacy is second issue of big data security, where how the data are isolated from other person Data management is third issue of big data security; where private data is managed. The last security issue is Integrity and Reactivate Security which comes from various sources to check its integrity and secure it.

V. CONCLUSION AND FUTURE WORK

In this research work we have explored about big data, Dimension reduction techniques, Security , Apache Hadoop, components of Hadoop and challenges of big data analytics and big data mining tools. We have reviewed the aspects of big data and how dimension reduction method or techniques work to reduce the dimension of big data. It is very important to reduce the dimension of big data to remove noises and irrelevant information from it. Security is also a main concern in big data and without security it can harm single person or organization. We can say that big data is growing rapidly and we have to improve the purity of database. Also Apache Hadoop is useful tool for analyzing the large amount of data.

REFERENCES

[1] R. Beakta, "Big Data And Hadoop: A Review Paper", RIEECE India, vol.2(2), pp 13-15, 2015
[2] Prabhakar S.K. and Rajaguru H., (2016), Performance Analysis Of ICA, PCA AS Dimensionality Reduction Techniques And Approximate Entropy, SRC As Post Classifier For The Classification Of Epilepsy Risk Levels Form EEG Signals, International Journal of Advanced Engineering Technology India, 7(1):486-489.
[3] Gholami A. and Laure E., (2016), Big data Security and Privacy Issues in the CLOUD, International Journal of Network Security & Its Applications (IJNSA) Sweden, 8(1):59-79.
[4] Weng J. and Young D.S., (2017), Some dimension reduction strategies for the analysis of survey data, Journal of Big data, 4(43):1-27.
[5] Pavithra M. and Parvathi R.M.S., (2017), A Survey on Clustering High Dimensional Data Techniques, International Journal of Applied Engineering Research India, 12(11):2893-2899.
[6] Lehmann D., Fekete D., Vossen G., (2016), Technology Selection for Big data and Analytical Applications, European Research Center for Information Systems, 27(1):1-37.
[7] J. Moreno, M.A. Serrano, E. Fernández Medina, "Main Issues in Big Data Security", Alarcos Research Group, University of Castilla-La Mancha, 13005 Ciudad Real, Spain, vol.8(44), pp1-16, 2016
[8] S. Mukherjee & R. Shaw, "Big Data – Concepts, Applications, Challenges and Future Scope", International Journal of Advanced Research in Computer and Communication Engineering India, vol.5(2), pp66-74, 2016



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY