iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** V **Month of publication:** May 2017

DOI:

www.ijraset.com

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

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Big Data Analysis: A comprehensive Study

M. Bhaskar
Department of Computer Science Kakatiya University, Warangal, Telangana

Abstract: ***With advance of E-commerce applications, social media and mobile devices in current era the data explosion has started. IT industry and academia have gained much attention for Big Data. By 2020, 50 billion devices are expected to be connected to the internet. With the data transfer and sharing at light speed on optic fibre and wifi networks, there is a voluminous growth in data generations. In this Big Data World, various fields around the globe are facing a big problem with this large scale data which highly supports in decision making. The traditional relational DBMS's were unable to handle this Big Data. The most classical data mining methods are also not suitable for handling this big data. Efficient algorithms are required to process Big Data. As Big Data is still in its infancy stage, in this paper we try to identify the Issues and Challenges faced by MapReduce in handling Big Data with an overview of providing and identifying better opportunities for further research. The Challenges are grouped into (a) Big Data Management and Storage. (b) Big Data Analysis.(c) Security and Privacy. This study encourages future Big Data research.***
Keywords: Big data, Hadoop, MapReduce, BigData Analytics.

I. INTRODUCTION

Looking forward, experts now predict that 40 zetttabytes of data will be in existence by 2020. Three years ago, the entire World Wide Web is estimated to contain approximately 500 exabytes – which is 5 billion gigabytes, but only half of one zettabyte! 40 zettabytes is, therefore, 400 billion gigabytes.[1] With the technological revolution there is a tremendous increase in the amounts data generated through these mobile devices. Particular remote sensors are continuously producing heterrogenous data that are either structured or unstructured, which is treated as Big Data. Big Data is charactirised by three aspects (a) volume of data generated (b) traditional databases do not support big data (c) Data are generated, captured and processed very quickly However , Big Data is its infancy stage, this study classifies various attributes of big data, including its volume, velocity , variety, management, anlaysis, security and rapid growth rate.[2]

This study presents: (i) a comprehensive study of big data characteristics, (ii) a discussion of Big Data Management and Analysis Tools (iii) an enumeration of issues and challenges associated with Big Data.

II. BACKGROUND

Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity, that cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions.

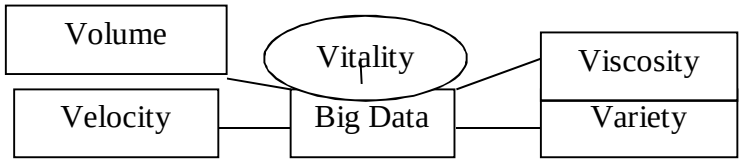


Fig. 1: 5v's of big data

Along with the three V's, there also exists ambiguity, viscosity, and virality.

A. Ambiguity

A lack of metadata creates ambiguity in Big Data.

B. Viscosity

Measures the resistance (slow down) to flow in the volume of data.

C. Virality

Measures and describes how quickly data is shared in a people-to-people (peer) network.

D. Rapid Growth of Data

The data type that increases most rapidly is unstructured data. This data is mostly generated from videos, movies, images, financial transactions, sensors, maps, mails, tweets, facebook data etc. According to Computer World unstructured information may account for more than 70 % to 80% of all data in organizations [3]. Most unstructured data are not modelled and are random which are very difficult to analyze. Table 1 describes the rapid growth of data in various organizations further.

Table.1: Rapid Growth of unstructured data

Source	Production
Youtube[4]	The total number of people who use YouTube - 1,300,000,000. 300 hours of video are uploaded to YouTube every minute! Almost 5 billion videos are watched on Youtube every single day. Youtube gets over 30 million visitors per day
Facebook[5]	• Facebook has more than 500 million active users • Over 300,000 @facebook users helped translate the site to 70 available translations • Facebook alone has 2.5 billion pieces of content, 2.7 billion 'likes' and 300 million photos – all of which adds up to more than 500 terabytes of data.
Twitter[6]	• 100 million users login to twitter daily • Site generates 500 million tweets per day
Google[7]	• Google processes 3.5 billion requests per day • Google stores 10 exabytes of data (10 billion gigabytes!)
Apple[7]	140 billion applications are downloaded per minute
Instagram[7]	Users share 60 million photos per day
Flickr[7]	Users upload 20 million photos per day
Linkedin[7]	2.1 million groups have been created
Amazon[7]	1 billion gigabytes of data across more than 14,00,000 servers

Fig.2: Big Data Analysis Tools

III. BIG DATA MANAGEMENT AND ANALYSIS

Big Data technology aims to minimise hardware and processing costs and to verify the value of Big Data before committing significant company resources. Properly managed Big Data are accessible, reliable, secure, and manageable. Hence, Big Data applications can be applied in various complex scientific disciplines, including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry.

A. Management Tools

With the evolution of computing technology immense volumes can be managed without requiring supercomputers and high cost. Many tools and techniques are available for data management, including Google Big Table, Simple DB, Not Only SQL (No SQL), Data Stream Management System (DSMS) [2] . However, companies must develop special tools and technologies that can store, access, and analyze large amounts of data in near-real time because Big Data differs from the traditional data and cannot be stored in a single machine. The following section describes Hadoop and MapReduce in further detail, as well as various projects /

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

frameworks that are related to and suitable for the management and analysis of Big Data.

B. Big Data Analysis Tools

As much of the big data generated today is in unstructured format, big data analysis is a major challenge. Fig 2 illustrates the Big Data Analysis tools that are used for efficient data handling generated from various sources at varying speeds.

<table>
<tr><td colspan="6">DATA ANALYSIS
Business Intelligence and Data Analytic Tools
(Query, Reporting, Data Mining, Predictive Analysis)</td></tr>
<tr><td colspan="2">Oozie
(workflow)</td><td>Chukwa
(Monitoring)</td><td>Flume
(Monitoring)</td><td>Zookeeper
(Management)</td><td>.....</td></tr>
<tr><td colspan="6">DATA ACCESS</td></tr>
<tr><td>Hive
(SQL)</td><td>Pig
(Data Flow)</td><td>Avro</td><td>Mahout
(Machine Learning)</td><td>Sqoop
(Data Connector)</td><td>.....</td></tr>
<tr><td colspan="6">DATA PROCESSING</td></tr>
<tr><td colspan="6">MapReduce Framework</td></tr>
<tr><td colspan="6">DATA STORAGE</td></tr>
<tr><td colspan="2">HDFS</td><td>Hbase</td><td></td><td>Cloud Store</td><td>.....</td></tr>
</table>

1) *HDFS:* This paradigm is applied when the amount of data is too much for a single machine. HDFS is more complex than other file systems given the complexities and uncertainties of networks.
2) *HBase:* HBase is a management system that is open-source versioned and distributed based on the Big Table of Google. For example, read and write operations involve all rows but only a small subset of all columns. HBase is an accessible through application programming interfaces (APLs).
3) *ZooKeeper:* ZooKeeper maintains, configures, and names large amounts of data. It also provides distributed synchronization and groups services. This instance enables distributed process to manage and contribute to one another through a name space of data registers.
4) *Hive:* Hive is a sub platform in the Hadoop ecosystem and produces its own query language (HiveQL) . This language is complied by Map Reduce and enables user- defined functions (UDFs) . The Hive platform is primarily based on three related data structures: tables partitions, buckets .
5) *Pig:* The pig framework generates a high level scripting language and operates a run-time platform that enables users to execute MapReduce on Hadoop .
6) *Mahout:* Mahout is a library for machine – learning and data mining . It is divided into four main groups: collective filtering, categorization , clustering , and mining of parallel frequent patterns .
7) *Oozie :* In the Hadoop system, Oozie coordinates , executes, and manages job flow . Oozie combines actions and arranges Hadoop tasks using a directed acyclic graph (DAG). This commonly used for various tasks.
8) *Chukwa :* Currently , Chukwa is a framework for data collection and analyses that is related to Map Reduce and HDFS . As an independent module, Chukwa is included in the distribution of Apache Hadoop.
9) *Flume:* Flume is specially used to aggregate and transfer large amounts of data (i.e, log data) in and out of Hadoop. It utilizes two channels, namely, sources and sinks. Sources include Avro, files, and system logs, whereas sinks refer to HDFS and HBase . Flume transforms each new batch of Big Data before it is shuttled into the sink.
10) *Hadoop:* Hadoop[8] is written in Java and is a top-level Apache project that started in 2006. Map Reduce programming environment could be applied in a distributed system. Presently, it is used on large amounts of data With Hadoop, enterprises can harness data that was previously difficult to manage and analyze. In particular, Hadoop can process extremely large volumes of data with varying structures. Hadoop is composed of HBase, HCatalog , Pig, Hive, Oozie , Zookeeper , and Kafka; however , the most common components and well-known paradigms are Hadoop Distributed File System (HDFS) and MapReduce for Big Data.
11) *MapReduce:* MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

computers (nodes), collectively referred to as a cluster (if all nodes are on the same local network and use similar hardware) or a grid (if the nodes are shared across geographically and administratively distributed systems, and use more heterogeneous hardware). Processing can occur on data stored either in a files system (unstructured) or in a database (structured). MapReduce can take advantage of locality of data, processing it on or near the storage assets in order to reduce the distance over which it must be transmitted. [9]

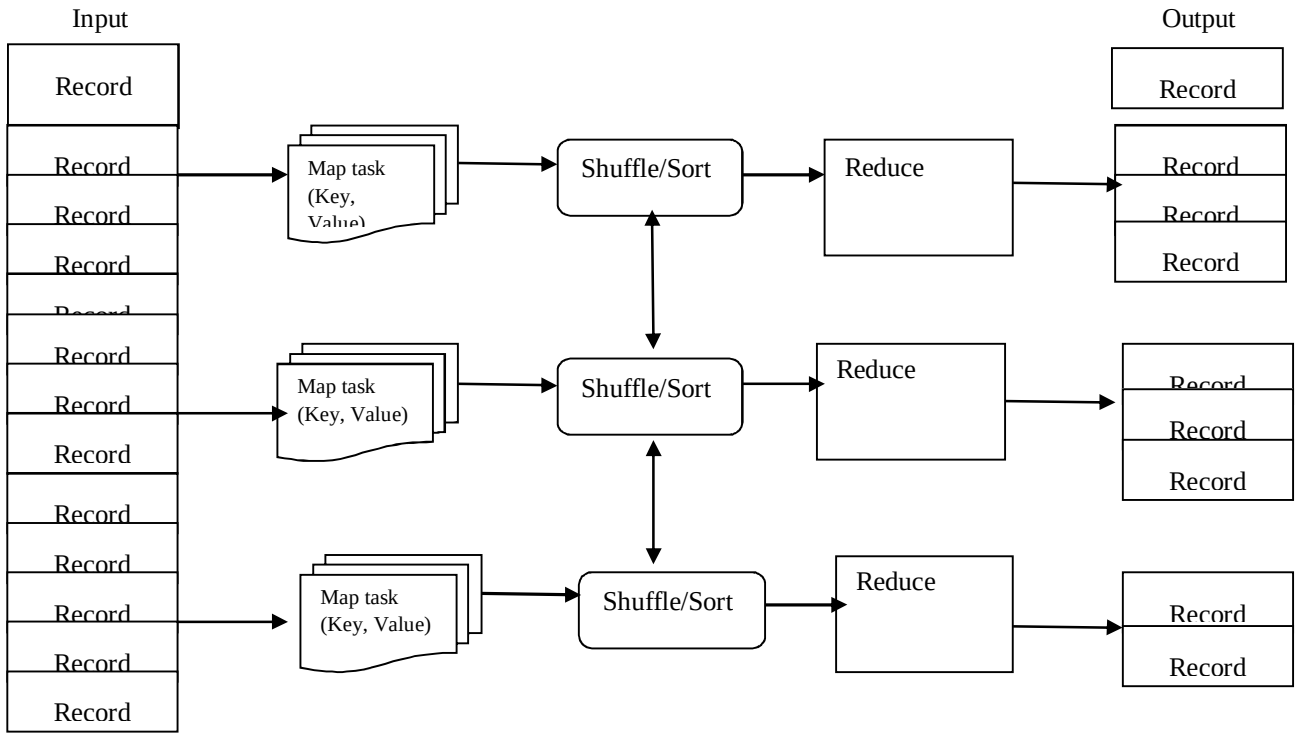


Fig. 3 : Map Reduce Framework and Working

IV. OPPORTUNITIES, OPEN ISSUES AND CHALLENGES

Big Data is the basic competitive strategy of current enterprise. New competitors must be able to attract employees who posses critical skills in handling Big Data. By harnessing Big Data, businesses gain many advantages, including increased operational efficiency, informed strategic direction, improved customer service, new products, and new customers and markets.

With Big Data, users not only face numerous attractive opportunities but also encounter challenges[10]. Such difficulties lie in data capture, storage, sharing, analysis, and visualisation. Currently a limited number of tools are available to completely address the issues in Big Data analysis. Challenges in Big Data analysis include data inconsistency and incompleteness, scalability, timeless, and security[11].

Prior to data analysis, data must be well constructed. However, considering the variety of datasets in Big Data, the efficient representation, access, and analysis of unstructured or semi structured data are still challenging. Hence, current real-world databases are highly susceptible to inconsistent, incomplete, and noisy data. Therefore, numerous data pre-processing techniques, including data cleaning, integration, transformation, and reduction, should be applied to remove noise and correct inconsistencies.[12] Thus, future research must address the remaining issues related to confidentiality. These issues include encrypting large amounts of data, reducing the computation power of encryption algorithms, and applying different encryption algorithms to heterogeneous data.

Privacy is major concern in outsourced data. Policies that cover all user privacy concerns should be developed. Furthermore, rule violators should be identified and user data should not be misused or leaked. The major challenges in integrity are that previously developed hashing schemes are no longer applicable to such large amounts of data. Integrity checking is also difficult because of the lack of support given remote data access and the lack of information regarding internal storage.

Big Data has developed such that it cannot be harnessed individually. Big Data is characterized by large systems, profits, and

challenges. Thus, additional research is needed to address these issues and improve the efficient display, analysis, and storage of Big Data. To enhance such research, capital investments, human resources, and innovative ideas are the basic requirements.[12]

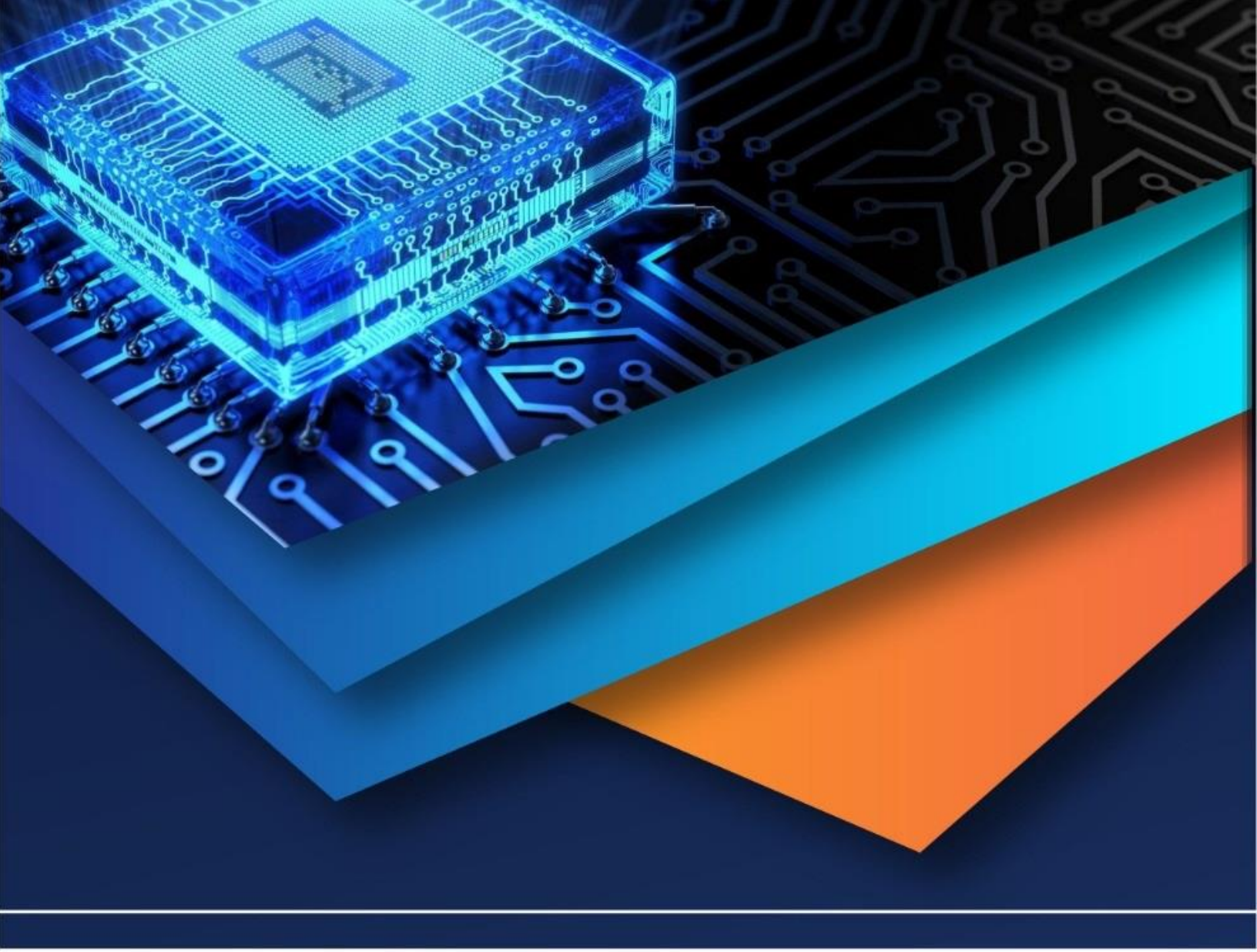
V. CONCLUSION

This paper presents the fundamental concepts of Big Data. These concepts include the increase in data, the progressive demand for HDDs, and the role of Big Data in the current environment of enterprise and technology. Organizations often face teething troubles with respect to creating, managing, and manipulating the rapid influx of information in large datasets. Given the increase in data volume, data sources have increased in terms of size and variety. Data are also generated in different formats, which adversely affect data analysis, management, and storage. This variation in data is accompanied by complexity and the development of additional means of data acquisition.

The extraction of valuable data from large influx of information is a critical issue in Big Data. Qualifying and validating all of the items in Big Data are impractical; hence, new approaches must be developed. From a security perspective, the major concerns of Big Data are privacy, integrity, availability and confidentiality with respect to outsourced data. Large amounts of data are stored in cloud platforms. Big Data involves large systems, profits and challenges. Therefore, additional research is necessary to improve the efficiency of integrity evaluation on line, as well as the display, analysis, and storage of Big Data.

REFERENCES

[1] https://cloudtweaks.com/2015/03/surprising-facts-and-stats-about-the-big-data-\industry
[2] M. Chen, S. Mao, and Y. Liu, “Big data: a survey,” MobileNetworks and Applications,vol. 19, no. 2, pp. 171–209, 2014.
[3] S. Sagiroglu andD. Sinanc, “Big data: a review,” in Proceedings of the International Conference on Collaboration Technologies and Systems (CTS ’13), pp. 42–47, IEEE, San Diego, Calif, USA,May 2013.
[4] YouTube, “YouTube statistics,” 2014, http://www.youtube.com/yt/press/statistics.html.
[5] Facebook, Facebook Statistics, 2014, http://www.statisticbrain.com/facebook-statistics/.
[6] Twitter, “Twitter statistics,” 2014, http://www.statisticbrain.com/twitter-statistics/.
[7] Marcia, “Data on Big Data,” 2012, http://marciaconner.com/blog/data-on-big-data/.
[8] A. Hadoop, “Hadoop,” 2009, http://hadoop.apache.org/.
[9] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” in Proceedings of OSDI’04: Sixth Symposium on Operating System Design and Implementation, December 2004.
[10] J. Ahrens, B. Hendrickson, G. Long, S. Miller, R. Ross, and D.Williams, “Data-intensive science in the US DOE: case studies and future challenges,” Computing in Science and Engineering,vol. 13, no. 6, pp. 14–23, 2011
[11] N. S. Kim, T. Austin, D. Blaauw et al., “Leakage current:Moore’s law meets static power,” Computer, vol. 36, no. 12, pp. 68–75,2003.
[12] D. E. O’Leary, “Artificial intelligence and big data,” IEEE Intelligent Systems, vol. 28, no. 2, 2013.
[13] K. Michael and K. W. Miller, “Big data: new opportunities and new challenges” Editorial: IEEE Computer, vol. 46, no. 6, pp. 22–24, 2013
[14] D. Usha, A.P.S. AslinJenil, “ A Survey of Big Data Processing in Perspective of Hadoop and Mapreduce**”,** International Journal of Current Engineering and Technology, Vol.4, No.2,April 2014.
[15] Nada Elgendy, Ahmed Elragal, “Big Data Analytics: A Literature Review Paper”,P.Perner(Ed.):ICDM 2014, pp.214-227, 2014©Springer International Publishing Switzerland 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY