iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34590

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: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

Challenges of Big Data for Development

Mr. Zulfikar Ali Ansari[1], Harshit Sinha[2], Nandini Sharma[3]
[1]*Assistant Professor, Department of CSE, BBDNITM, Lucknow, India*
[2]*B.tech Student, Department of ECE, BBDNITM, Lucknow, India*
[3]*B.tech Student, Department of CSE, BBDNITM, Lucknow, India*

Abstract: Big data is the data which is 'big' or 'large' in amount. Where data can be structured, semi-structured, unstructured that may vary in format 8/video. The quality and quantity of big data face challenges in its extraction, processing and analytics. Here, the characteristics of big data ie. 7v's are defined. Day to day increment in data statistics through detailed speculation is proved helpful to observe the trends. With increase in datasets there is surplus usage of big data in various software applications and the latest emerging technologies of the century to handle the growing data. Since the solution cannot be same for all we can have a proper understanding of tools and technologies , their usage. This report highlight the challenges of big data by enlisting and describing them along with toolkits to deal with it. So as to handle any amount of data. Different technologies are proved helpful where HADOOP, H-BASE, SPARK is most trending. Hadoop removes the complexity of high performance computing and can be installed on conventional machines. Nowadays, Emerging technologies and various research findings are a source of many researchers and professionals to work and dig more into this outlet.

Keywords: big data; Hadoop; Hbase; spark

I. INTRODUCTION

Vast, complex, large and diverse unstructured dataset can be termed as Big Data!! Big data is basically data which is big or extremely large. Data can be extracted from various sources and is available in various forms. Internet, internet of things (IOT), artificial intelligence (AI), sensing devices, remote devices, cameras, medical devices and surgical parameter equipment's, electronic media etc. are adversely responsible for data production. The rise in data and its enormous production is very common in this 21^{st} century. The so called "digital era" is carefree about the creation and usage of data available to us. We can categorize this data as structured , semi- structured , unstructured data.

Big data is also known as 'Dirty Data'. The reason behind this is its inconsistency, complexity, inaccuracy, ambiguity and unstructured format of dataset. Social media like facebook, Instagram, emails, youtube etc. is one of a powerful asset of data for its uploading, sorting, searching, synchronizing, creating, collecting different field outlooks and major production of datasets not just in bytes but from bits to quintillion bytes . They can be audio/video/pictures/3d-modellings/documents or any other media.

The 7 V's of big data which plays an important role in handling and processing are :

1) *Volume:* In big data term "volume" is used to describe "large" amount of data generated each day by various platforms , the volume of data generated is so high that gigabyte is not enough to store data , now units like zettabyte , Exabyte and yottabytes are used to measure data.
2) *Velocity:* Velocity refers to the speed at which data is processed , Since the amount of data generated is "big" or large in volume the processing of the data should be fast and hassle free.
3) *Variety:* In terms of big data "variety" refers to various types of data sources , since data generated could be of any form it can be structured , semi-structured or unstructured , it can be made available in various formats like video , image , audio , document etc , so it becomes essential to have a set of tools which defines their type
4) *Variability*: Variability is different from variety . Variability refers to that data which keeps on changing constantly , it is all about understanding and interpreting the correct meaning of raw data.
5) *Veracity:* The concept of Veracity lies in the correctness of the data ,veracity makes sure that the data received is accurate and ensures unuseful data is kept away.
6) *Visualization*: The term Visualization is used for the presentation of the gathered data for the purpose of decision making , since data can be presented in many ways which include word file , excel file , presentation file , graphs , charts and many other ways , it is important to choose the right technique
7) *Value*: It refers to the end result of using all these techniques , every user needs to get a value for all the efforts in using these techniques for gathering and processing the data , so at the end the value obtained by using all the resources is a matter of concern.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

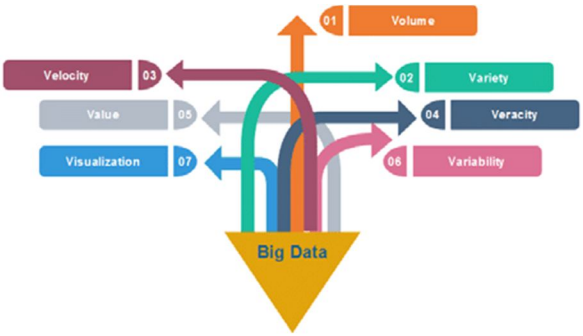


Fig. 1 7v's of big data

II. DETAILED SPECTULATION

IDC (international data co-corporation) is a wholly owned subsidiary which is globally providing market trends, stats, ser- vices, events, datasets related to information technology. IDC is connected to 110 countries across the globe (Asia/Pacific, Australia and New Zealand, Benelux, Canada Central & Eastern Europe & Israel, Central Asia, China, Colombia, France, Ger- many, India, Italy, Japan, Latin America, Middle East, Africa & Turkey, Nordic, Portugal, Russia & CIS, South Korea, Spain Taiwan, United Kingdom, ASEAN). IDC globally satisfies the purpose of providing/collecting the facts as raw or furnished , trends as latest or historic , media as advanced/local/global, data as structured , unstructured or semi-structured , marketing as stocks and business trades that relates or co-relates each other influencing people or professionals who desires for buying the content. Likewise, taking one such example of one big data producing platform is IRCTC (Indian Railway Catering and Tourism Corporation) which generates approximately 100 terabytes of data each year which is gradually increasing at a variable rate of 10% - 25 % per year. Similarly if we talk about other examples of Big data producing industries , first we have the world's biggest search engine that is google , there are more than 3.5 billion Searches daily done by users on google ,holding biggest market share of 87.35% among Search engine market google resolves 1.2 trillion search queries yearly, WhatsApp which is the largest messaging app worldwide has a platform where the users exchange up to 65 billion messages daily , WhatsApp has a separate application for business purpose which has a user base of 5 million apart from that there are over 1 billion WhatsApp.

III. AIM AND OBJECTIVE

The aim of this study is to discover the challenges of big data in development with suggested technologies to partially or fully rectify the issues dealing with big data.

DATA ISSUES	*ANALYSIS ISSUES*
• Data production/source • Data storing • Data transferring (uploading/downloading) • Data querying • Data identification • Data professionals scarcity	• Searching and sorting • Integration • Interpretation • Visualization • Validation • Classification/ Categorization • Monitoring (growth/shrink) • Tool-selection

1) *Data Production/Source:* Creation and source of data varies in its format, size, time, order. Where different formats can be audio (wav, mp3, flac, aac, etc.)/video (avi, mov, mp4, etc.)/image (png, jpeg, gif, etc.)/text(pdf, word, etc.). At instant, its difficult to manage all the data formats simultaneously.
2) *Data Storing:* Storing big data is a challenging task worldwide. Capacity of storage/memory depends upon the platform used such as databases, disks, libraries, external/internal storage, cloud storage, storage software's etc.
3) *Data Transferring:* Heavy servers/networks that allows uploading and downloading may crash sometimes , leading to data loss.
4) *Data Querying:* Software's and tools that allows managing , storing of data in database becomes slow, unresponsive and may create unexpected errors dealing with large dataset.

5) *Data Identification:* It helps to categorize the data which is a complex process for unsorted data.
6) *Data Professionals Scarcity:* To run and execute the data technologies/software's for its well functioning and practising, data scientist, data engineers, data analyst, data architect and professionals that must have adequate field knowledge about traditional as well as trending technologies are difficult to get in numbers to work with big data.
7) *Searching and Sorting:* While data searching through various algorithm and different methods, complexity of the same should be considered and sorting through various sort techniques is a matter of concern and expertise.
8) *Integration:* summing up the required/retrieved data in order to perform sorting or categorization requires methodolo- gies and tools to interact with such large data items.
9) *Interpretation:* Meaningful, complete, definite data understanding is must for proper analysis.
10) *Visualization:* Prediction on patterns and trends, about planning with future ideology depends totally upon designed models and methods followed.
11) *Validation:* Correctness and conciseness forms the checklist of accurate or precise datasets.
12) *Classification/Categorization:* Dividing data into classes and subclasses for their identification and interpretation plays a significant role in categorizing only if done properly.
13) *Monitoring:* Gradual increment or decrement of datasets affect data trends if observation is mis-understood which ultimately leads to wrong consequence.
14) *Tool-selection:* Unavailable, non-affordable, inefficiently accessible technologies arise issues working with big data.

IV. REQUIREMENT IN SOLVING BIG DATA ISSUES

1) *Apache Spark:* Apache spark is a fast flexible and developer friendly platform for large-scale SQL , batch processing, Stream processing and machine learning , it is an open source platform and can be used free of cost by anyone, apart from that it can be altered to fulfill individual needs. spark provides comprehensive unified framework to manage big data processing requirements . Sparks allows us to quickly write applications in Java , Scala or python . Apache spark uses master / slave Architecture i.e. one central coordinator and many distributed workers in this context the central coordinator is called The Driver. The Drivers runs communicates with a potentially large number of distributed workers which are called Executors. A spark application is a combination of driver and its own executors, the spark application is launched on set of machines. standalone cluster manager is the default build in cluster manager of spark.
2) *Apache Hadoop:* Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving big data , it is an ecosystem, composed of framework , open source software libraries and methodologies for data analysis. Hadoop and it's ecosystem have become attractive to companies of all sizes. Hadoop removes the complexity of high performance computing and can be installed on conventional machines. Hadoop is a set of open source programs written in Java which can be used to perform operations on a large amount of data. Hadoop is a scalable, distributed and fault tolerant ecosystem. The main components of Hadoop are - Hadoop YARN = manages and schedules the resources of the system, dividing the workload on a cluster of machines. Hadoop Distributed File System (HDFS) = is a clustered file storage system which is designed to be fault-tolerant, offer high throughput and high bandwidth. It is additionally able to store any type of data in any possible format.

Hadoop MapReduce = is used for loading the data from a database, formatting it and performing a quantitative analysis on it.

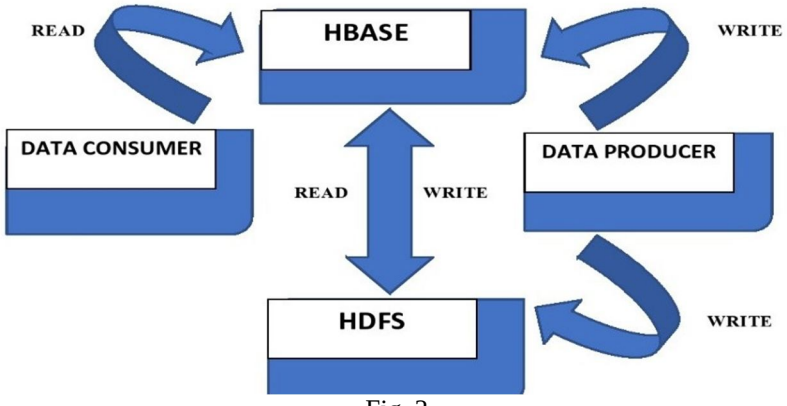


Fig. 2

3) *Apache H-BASE:* HBase is an open-source, column-oriented distributed database system in a Hadoop environment Apache HBase is needed for real-time Big Data applications. HBase can store massive amounts of data from terabytes to petabytes. It is based on Google's Big Table. It has set of tables which keep data in key value format. Hbase is well suited for sparse data sets which are very common in big data use cases It is meant to host large tables with billions of rows with potentially millions of columns and run across a cluster of commodity hardware. But beyond its Hadoop roots, HBase is powerful database in its own right that blends real-time query capabilities with the speed of a key/value store and offline or batch processing via MapReduce. In short, HBase allows you to query for individual records as well as derive aggregate analytic reports across a massive amount of data.HBase is optimized for sequential write operations, and it is highly efficient for batch inserts, updates, and deletes. HBase works seamlessly with Hadoop, sharing its file system and serving as a direct input and output to Hadoop jobs.

V. APPLICATION OF BIG DATA

A. Operational Technologies (ex- social media, online shopping)
B. Analytical Technologies (ex- space technology, weather forecasting)

The applications of big data is never ending and long-lasting. Depending upon the requirement and development, its uses are in number. Without data there will be no useful information and without information there will be no production/generation of assets. Emerging trends in data stats help in monitoring growth and changes. Business , marketing , finance, goods production and manufacturing looks for pattern through which they analyze the rise and fall in their budgets ,trades, profit or loss. E-tickets are generated and made available by keeping the track record of all the passengers and vacancies respectively. Medical diagnosis and various test reports , health surveys are obtained by big data. Commodities are searched and explored worldwide when presented in number of choices. Bunch of network and servers are associated to use this enlarged entangled data for telecom services. Also, government functioning for the wellfare of humankind is fully dependenciable on large datasets of persons detail. So, the uses or applications is now trending and are in great demand. This will be more betterly furnished in upcoming future scope .

Fig.3 Applications of Big Data

VI. CONCLUSIONS

The study of big data and making it useful in day to day life has solved many problems in data and information sector. Big data can be huge and is made available in structured, semi-structured and unstructured format, the generation of data is increasing everyday and is estimated to The study of big data and making it useful in day to day life has solved many problems in data and information sector. Big data can be huge and is made available in structured, semi-structured and unstructured format, the generation of data is increasing everyday and is estimated to be around 2.5 quintillion bytes per day , which is a massive number . In this research paper we have highlighted the challenges we are facing in handling and processing the big data and what are the solution or technologies available to tackle this huge amount of data. Further we have also discussed the applications of big data and how this data is generated through various sources. Our aim is to categorise various formats of data and which technology is useful in handling big data issues of particular sectors . Since the solution cannot be same for all we can have a proper understanding of tools and technologies , their usage. So as to handle any amount of data.

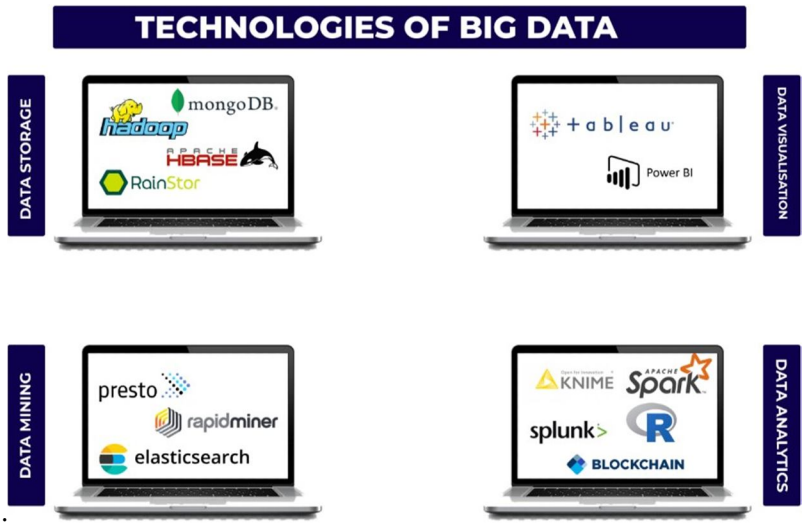
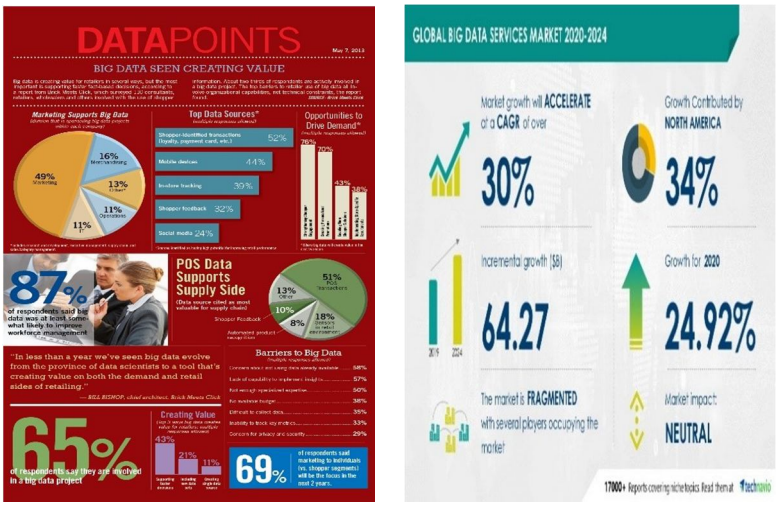


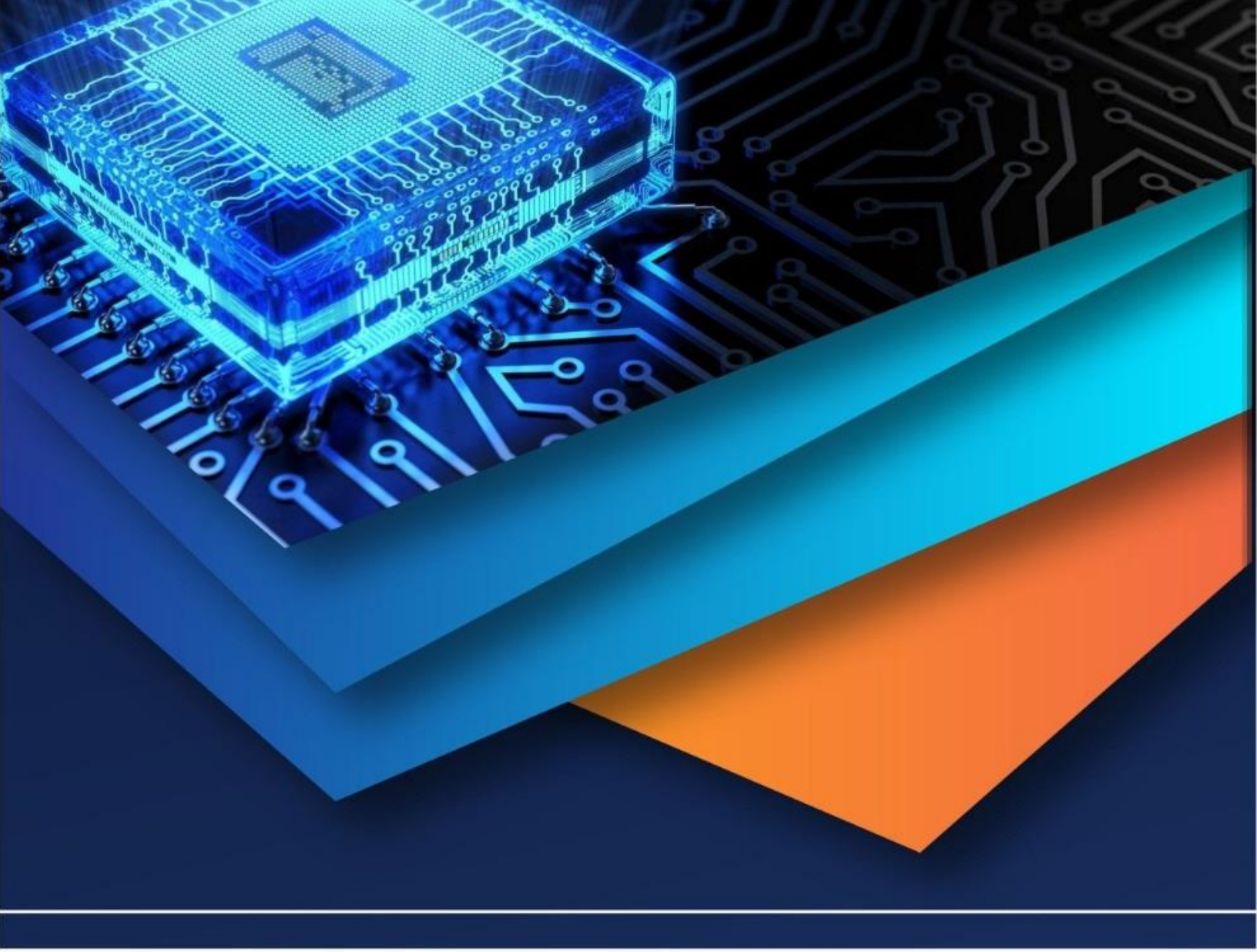
Fig. 4

VII. RESEARCH FINDINGS



REFERENCES

[1] en.wikipedia.org/wiki/Big-data, “big data introduction”
[2] Eileen mucnulty/dataconomy, “understanding big data : the seven V’s”, may 22, 2014
[3] Comsource media metrix, desktop unique visitors, worldwide ,January 2017
[4] Christo petrer, “big data statistics”, march18, 2021
[5] Ravi kiran, “big data technology”, nov 25, 2020
[6] G Bello-Orgaz, JJ Jung, D Camacho - Information Fusion, 2016 – Elsevier
[7] Esri/Arcnews,vol.40,no.2, 2018
[8] Singapore edition technewsletter, “technews”, may7, 2013
[9] Technavio, “global market survey”, 2020
[10] Frost and Sullivan, IBM, cisco, “sub trends”, m82c-mt,
[11] Seb murray/MBA careers, “MBA turns to india for forming big data jobs”, 23rd July, 2014
[12] Jin, BW Wah, X Cheng, Y Wang - Big Data Research, 2015 – Elsevier
[13] T Nasser, RS Tariq - J Comput Eng Inf Technol 4: 3. doi: http://dx …, 2015 - researchgate.net
[14] AA Tole - Database systems journal, 2013 - dbjournal.ro
[15] N Khan, I Yaqoob, IAT Hashem, Z Inayat… - The scientific world …, 2014 - hindawi.com
[16] S Sagiroglu, D Sinanc - 2013 international conference on …, 2013 - ieeexplore.ieee.org



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY