IJMRSETM

e-ISSN: 2395 - 7639

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 10, Issue 6, June 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.580

+91 99405 72462 +9163819 07438 ijmrsetm@gmail.com www.ijmrsetm.com

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |

Data Analytics: Challenges, Open Research Issue and Tools

Miss. Priyanka Chandrashekhar Talekar, Mr. Shripad Shrikant Bhide

Student, Department of M.C.A., P.E.S. Modern College of Engineering. , Pune, India

Assistant Professor, Department of M.C.A., P..E.S. Modern College of Engineering, Pune, India

ABSTRACT: A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of bigdata challenges, open research issues, and various tools associated with it. As a result, this article provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.[1]

KEYWORDS: Big data analytics, Hadoop, Massive data, Structured data, Unstructured Data

I. INTRODUCTION

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data It provides evolutionary breakthroughs in many fields with collection of large datasets In general, it refers to the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing applications.[1]

These are available in structured, semi-structured, and unstructured format in petabytes and beyond. Formally, it is defined from 3Vs to 4Vs.3Vs refers to volume, velocity, and variety. Volume refers to the huge amount of data that are being generated everyday how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi-structured etc. The fourth V refers to veracity that includes availability and accountability. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent Technique. Recent years big data has been accumulated in several domains like health care, public administration, retail, bio-chemistry, and other interdisciplinary scientificresearches. Webasedapplications encounter big data frequently, such as social computing, internet text and documents, and inter-net search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplore, Scopus, Thomson [3]

II. CHALLENGES IN BIG DATA ANALYTICS

Recent year big data has been accumulated interval domain like healthcare, public administration, retail, biochemistry and other interdisciplinary scientific researches

Web-based application encounter big data frequently, such as social computing, internet text and documentation and internet search index, social computing that is included the social network analysis.[2]

Big data that is provided new opportunity in knowledge processing task for the upcoming researchers. that opportunity is always following some challenges

To handle the challenges, we need to know various computational complexity, information security and computational method to analysis big data. for example, many statistical methods that performed well for the small data size do not scale to voluminous data. Similarly, many small data size do not scale to various data.

Various challenges that the health sector face was being researched by much researcher here the challenges of big data analytics are classified into four broad categories namely data storage and analysis; knowledge discovery and computational complexity; scalability and visualization of data and information security. We discuss theses issue below.[1]

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |

A. Data Storage and Analysis

In recent year the size of data has grown exponentially by various means such as mobile devices, aerial sensory technologies, report sensing, radio frequency identification reader etc These data stored on spending much cost where they ignored or deleted finally because there is no enough space to Stord them. Therefore, the first challenges for big data analysis the storage medium and higher input/output speed. In such case the data accessible must be on the top priority for the knowledge discovery and representation.[4]

The prime reason is begun that it must be accessed easy and promptly for further analysis. In past decades, analyst use hard disk drive to stored data bur its slower random input/output performances than sequential input/output performances .to overcome this limitation the concept of solid-state drive and please change memory we introduce. However, the available storage technologies cannot possess the required performances for the processing big data

Another challenge with Bid data is attributed to diversity of the data with the ever-growing datasets, data mining task has significantly increased. The additionally data reduction, data selection feature is essential task especially when dealing with large datasets.

The major challenges in this case is to pay more attention for designing storage system and to elevate efficient data analysis tool that provided guarantee on the output when the data come in different sources.[8]

B. Knowledge Discovery and Computational Complexities

Knowledge discovery and representation is a prime issue in big data.it includes a number of sub field such as authentication, archiving, management, preservation, information retrieval and representation. There are several tools for knowledge's discovery and representation such as fuzzy set. All these techniques are problem dependent. Further some of these are technique may not be suitable for large datasets in the sequential computer. At the sometime some of the techniques has good characteristics of scalability over parallel computer. Since the size of big data keep increasing exponentially, the available tools may not be efficient to process these data for obtaining meaningful information. The most popular approach in case of large datasets management is data warehouse and data that are sources from operational system where data mart is based on data warehouse system where data mart is based on data warehouse and facilities analysis of large datasets required more computational complexities. The major issue is to handle inconsistencies and uncertain present in the datasets. In general, systematic modelling of the computational complexity is used. It may be difficult to establish of comprehensive mathematical system that is broadly applicable to big data. But a domain specific data analytics can be done easily by understanding the particular complexity. A series of such deployment could simulate big data analytics for different area. Much research Ans survey has Beed carry out in the direction using machine learning techniques with the least memory requirement's basic objective in this research is to minimize computational cost processing and complexities [5]

C. Scalability and Visualization of Data

The most important challenges of big data analysis techniques is its scalability and security, In the last decade have paid attentions to accelerates data analysis and its speed up processors followed by Moore's law. For the former it is necessary to develop sampling on line and analysis techniques. Incremental techniques have good scalability property in the accept of big data analysis. As the data size is scaling much faster the CPU Speed. There is natural dramatics shift in processor technology being embedded with increasing number of parallel computing. Real time application like navigation, social network, finance, internet search, timeliness etc required parallel computing.

The objective of virtualizing is to present them more adequately using some techniques of graph theory. Graphical visualization provides the link between data with proper interpenetration However online marketplace like flip cart, amazon. Have millions of users and billions of good to sold each month. This generates a lot od data To this end, some company uses a tool tableau for big data visualization. It has capability to transfer large and complex dta into intuitive picture. This help employees of company to visualize search relevance's monitor latest current big data visualization tools mostly have poor performances in functionality, scalability and response in time.[6]

We can observer the big data produced many challenges for the development of the hardware and software lead to parallel computing, cloud computing, distributed computing, visualizations process, scalability. To overcome this issue, we need to correlate more mathematical model to computer science.

D. Information Security

In big data analysis massive amount of data are corelated analysed a mined for meaningful pattern. All organization have different policies to safe guarded their sensitive information. Preserving sensitive information is major issue in big data analysis. There is huge security risk associated with big data therefore information security is becoming of big data analysis problem. Security of big data can be enhanced by using the techniques of authentication, authorization and

encryption. Various security measured the big face are scale of network, variety of different devices, real time security monitoring and lack of intrusion system. The security challenge caused by big data attracted the attention of information security. Therefore, attention has to be given to develop a multi-level security policy model and prevention system.

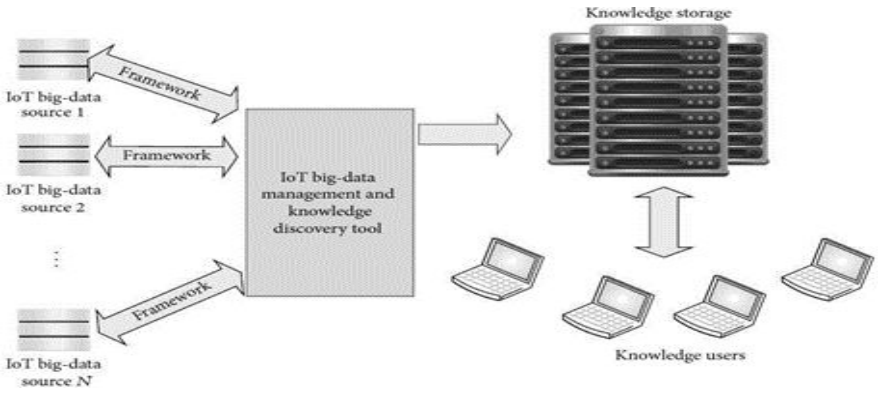
Although much research has been carried out to secure big data but it required lot of improvement. The major challenges is to devolved a multi-level security, privacy preserved data model of big data.[5]

III. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are becoming the research focal in industries and academia. Data science aims at researching big data and knowledge extraction from data Application of big data and data science include information science, uncertainly modelling, uncertain data analysis, machine learning, statical learning, pattern recognition of technologies and analysis will result in predicting the further drift of events. Main focus of the section is to discuss open research issue in big data analysis. The research issues pertaining to big data analysis are classified into three broad categories namely IOT, cloud computing, bio inspired computing and quantum computing.
However, it is not limited to this issue. More research issues related to health care big data can be found.

IOT for Big Data Analytics

Internet has revolution global interrelation the arts of business, culture revolution and unbelievable number of personal characteristics. Current, machine is getting in on the acts to control innumerable autonomous gadgets via internet and create internet of things Thus appliances are becoming the user of internet just like human with the web browser. Internet of things is attracting the attention of recent research for the most promising opportunity and challenged.it has an imperceptive economic and social impact foe the further construction of information, network and communication technology. The new regulation is further will be eventually everything will be connected and intelligent controlled. The concept of IOT is becoming more pertinent to the realistic world due to the development of mobile devices embedded and ubiquitous communication technologies cloud computing and data analysis. Moreover, IoT presents challenges in combination of volume, velocity and verity.
In a broad sense just like the internet IOT is the devices to exist in myriad of places and facility applicational range from trivial to the current place. Knowledges acquired from IoT data is the biggest challenges that big data professionals are fighter fore it is essential to develop infrastructure to analysis the data An IOT devices gent=rated continuous stream of data the researcher can delved tools to extract meaningful information from these data using machine learning tourniquet. Under standing these streams of data foeneration from IOT devices and analysis them to gest meaningful information is challenging issue and it lead to big data analysis. Machine learning algorithms and computing intelligent technique is only solution to handle big data from IOT prospectives.[6]

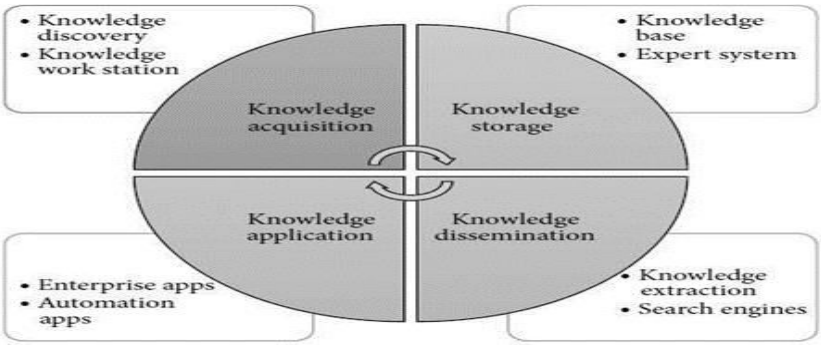


IoT Big Data Knowledge Discovery

Knowledge exploration system have organised from the theories of human processing such as frame, rules, tagging and semantics network. In general, it consists of four segment such as knowledge acquisition,
Knowledgebase, knowledge dissemination and knowledge application [6]

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |



B. Cloud Computing for Big data Analytics

The development of virtualization technology has made supercomputing more accessible and afforded. cloud computing infrastructure that hidden in virtualization software make the system to behave like a true computer, but with flexible of specification details such number of processors, disk space, memory, and operation system. The use of these virtual computer is known as cloud computing which has bees on of the most robust bid data technology. Bid data and cloud computing technology are the development and important developing scalable and data. cloud computing harmonizes massive data by demand access

The cloud computing harmonizes massive data on demand access to configurable computing resource through virtualization technique. The benefit of utilization the cloud computing included offering resource when there are in demand and pay only for the resource which is need to developed the product. Simultaneously improve availability and cost reduction. Open challenges and research issue of big data and cloud computing discussed in details by many researchers which highlight the challenges data management, data verity and velocity data storage, data processing and resource management so cloud computing help in developing business model for all verity application with infrastructure and tools

Big data application using cloud computing should support data analytic and development. The cloud environment should

C. Bio-inspired Computing for Big Data Analytics

Bio-inspired computing is a technique inspired Ny nature to address complex real-world problems. Biological systems are self-organized without a central control. A bio-inspired cost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and

processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance.

These systems are more suitable for big data applications. Huge amount of data is generated from variety of resources across the web since the digitization. Analysing these data and categorizing into text, image and video etc will require lot of intelligent analytics from data scientists and big data professionals. Proliferations of technologies are emerging like big data, IoT, cloud computing, bio inspired computing etc whereas equilibrium of data can be done only by selecting right platform to analyse large and furnish cost effective results.

D. Quantum Computing for Big Data Analysis

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously [33]. This exponential improve-

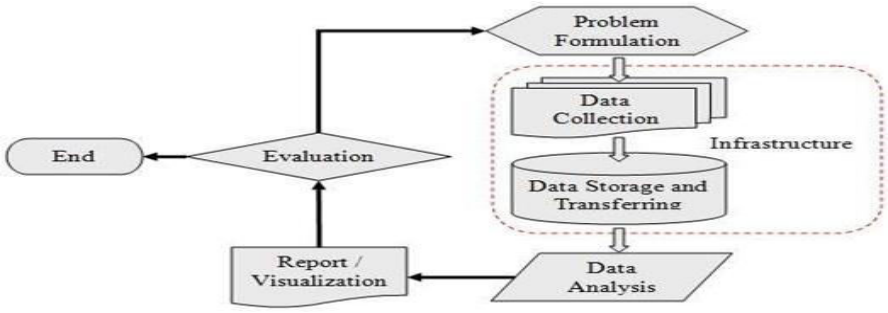
mint in computer systems might be possible. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computers, of course today's big data problems. The main technical difficulty inbuilding quantum computer could soon be possible. Quantum

computing provides a way to merge the quantum mechanics to process the information. In traditional computer, information is presented by long strings of bits which encode either a zero or a one. On the other hand, a quantum computer uses quantum bits or qubits. The difference between qubit and bit is that, a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomena of superposition and entanglement. It is because qubits behave quantumly. For example, 100 qubits in

quantum systems require 2100 complex values to be stored in a classic computer system. It means that many big data problems can be solved much faster by larger scale quantum computers compared with classical computers. Hence it is a challenge for this generation to build a quantum computer and facilitate quantum computing to solve big data problems.[6]

IV. TOOLS FOR BIG DATA PROCESSING

Large number of tools are available to process big data.in thief section we discuss some current technique for analyses big data with emphasis on three important emerging tools namely mapreduce,apche spark and storm Most of available tools concentrate on batch processing, stream processing, and interactive analysis. Most batch processing tools are based on Apache Hadoop infrastructure such as mahout and dryad Steam data application are most used for the scale streaming platform are analysis Strome. The interactive analysis process allow user to directed interacting real time for own analysis



A. Apache Hadoop and MapReduce

The most established platform for big data analysis is Apache Hadoop and MapReduce. It consists of Hadoop kernel MapReduce, Hadoop distributed file system. Map reduce is the programming model for processing large datasets is base on division and conquer method. The device and conquer method are implemented in two step such as map reduce step Hadoop work n two kind of node such as master node and worker node in map step. There fore the master node is the combination in reduce step, moreover Hadoop and MapReduce work as powerful software frame soar solving big data problem[6]

B. Apache Mahout

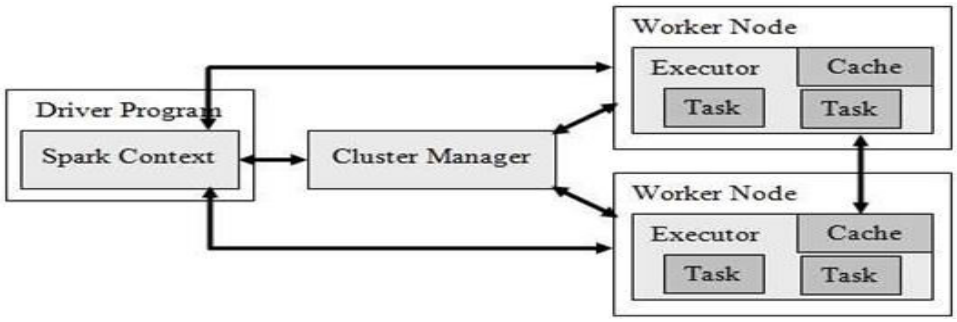
Apache mahout aims to provide scalable and commercia machine learning techniques for large scale and intelligent data analysis applications. Core algorithms of mahout including clustering, classification, pattern mining, regression, dimensionality reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework. The goal of mahout is to build a vibrant, responsive, diverse community to facilitate discussions
on the project and potential use cases. Thebaic objective of Apache mahout is to provide a tool for alleviating big challenges. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and Facebook[8]

C. Apache Spark

Apache spark is an open-source big data processing framework built for speed processing, and sophisticated analytics. It is easy to use and was originally developed in 2009 in UC Berkeley's AMP Lab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, Scala, or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of existing Hadoop distributed file system (HDES) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the
resources and the worker nodes to do the data processing in the form of tasks. Each application will have a set of processes called executors that are responsible for executing the tasks. The major advantage is that it provides support for deploying spark applications in an existing Hadoop clusters.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |



Architecture of Apache Spark

D. Dryad

It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and a user use the resources of a computer cluster to run their program in a distributed way. Indeed, a dryad user use thousands of machines, each of them with multiple processors or cores. The major advantage is that users do not need to know anything about concurrent programming. A dryad application runs a computational directed graph that is composed of computational vertices and communication channels. Therefore, dryad provides a large number of functionalities including generating of job graph, scheduling of the machines for the available processes, transition failure handling in the cluster, collection of performance metrics, visualizing the job, invoking user defined policies and dynamically updating the job graph in response to these policy decisions without knowing the semantics of the vertices [8]

E. Storm

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with Hadoop which is for batch processing. Additionally, it is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances. The storm cluster is apparently similar to Hadoop cluster. On storm cluster users run different topologies for different storm tasks whereas Hadoop platform implements map reduce jobs for corresponding applications. There are number of differences between map reduce jobs and topologies. The basic difference is that map reduce job eventually finishes whereas a topology processes messages all the time, or until user terminate it. A storm cluster consists of two kinds of nodes such as master node and worker node. The master node and worker node implement two kinds of roles such as nimbus and supervisor respectively. The two roles have similar functions in accordance with job tracker and task tracker of map reduce framework. Nimbus is in charge of distributing code across the storm cluster, scheduling and assigning tasks to worker nodes, and monitoring the whole system. The supervisor complies tasks as assigned to them by nimbus. In addition, it start and terminate the process as necessary based on the instructions of nimbus. The whole computational technology is partitioned and distributed to a number of worker processes and each worker process implements a part of the topology.[8]

E Apache Drill

Apache drill is another distributed system for interactive analysis of big data. It has more flexibility to support many types of query languages, data formats, and data sources. It is also specially designed to exploit nested data. Also it has an objective to scale up on 10,000 servers or more and reaches the capability to process petabytes of data and trillions of records in seconds. Drill use HDFS for storage and map reduce to perform batch analysis.[8]

V. SUGGESTIONS FOR FUTURE WORK

The amount of data collected from various applications all over the world across a wide variety of fields today is expected to double every two years. It has no utility unless these are analysed to get useful information. This necessitates the development of techniques which can be used to facilitate big data analysis. The development of powerful computers is a boon to implement these techniques leading to automated systems. The transformation of data into knowledge is by no means an easy task for high performance large-scale data processing, including

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 10, Issue 6, June 2023 |

exploiting parallelism of current and upcoming computer architectures for data mining. Moreover, these data may involve uncertainty in many different forms. Many different models like fuzzy sets, rough sets, soft sets,
neural networks, their generalizations and hybrid models obtained by combining two or more of these models have been found to be fruitful in representing data. These models are also very much fruitful for analysis. More often than not, big data are reduced to include only the important characteristics necessary from a particular study point of view or depending upon the application area. So, reduction techniques have been developed. Often the data collected have missing values. These values need to be generated or the tuples having these missing values are eliminated from the data set before analysis. More importantly, these new challenges may comprise, sometimes even deteriorate, the performance, efficiency and scalability of the dedicated data intensive computing systems. The later approach sometimes leads to loss of information and hence not preferred. This brings up many research issues in the industry and research community in forms of capturing and accessing data effectively. In addition, fast processing while achieving high performance and high throughput, and storing it efficiently for future use is another issue. Further, programming for big data analysis is an important challenging

VI. CONCLUSION

In recent years data are generated at a dramatic pace. Analysing these data is challenging for a general man. To this end in this paper, we survey the various research issues, challenges, and tools used to analyse these big data. From this survey, it is understood that every big data platform has its individual focus. Some of them are designed for batch processing whereas some are good at real-time analytic. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis,
machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing. We believe that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.[8]

REFERENCES

[1] M. Kahani, S. Kakhani and S. R.Biradar, Research issues in bigdata analytics, International Journal of Application or Innovation inEngineering & Management, 2(8) , pp.228-232.
[2] A. Gandomi and M. Haider, Beyond the hype: big data concepts, meth-ods, and analytics, International Journal of Information Management, 35(2), pp.137-144. [31 C. Lynch, Big data: How do your data grow?,Pp.28-29.
[4] X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research, 2(2) , pp.59-64.
[5] R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1) , pp.1-12.
[6] C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Infor-motion Sciences, 275, pp.314-347.
[7] K. Kambale, G. Kollias, V. Kumar and A. Gram, Trends in big dataanalytics, Journal of Parallel and Distributed Computing, 74(7)

Impact Factor: 7.580

INTERNATIONAL JOURNAL
OF MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com

www.ijmrsetm.com