ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Big Data and Its Applications: A Survey.

Magesh G[1*], and P Swarnalatha[2].

[1]School of Information Technology and Engineering, VIT University, Vellore, India
[2] School of Computer Science and Engineering, VIT University, Vellore, India

ABSTRACT

Huge amount of data is generating and increasing in our daily use of various applications in all the fields. The future challenges are to extract the useful and meaningful information from such a huge repository of the data. Applications designed for traditional databases are not so efficient for such a complex and huge databases. In this paper we have addressed few challenges in big data and its applications in various fields. It also describes the popular Hadoop environment and future trends in Big data and its applications. Data mining approaches and scheduling in big data is also reviewed. This paper will facilitate future researchers to develop new approaches and algorithms to solve few challenges in big data.

Keywords: Big Data, Hadoop, Internet, Information Technology (IT)

**Corresponding author*

INTRODUCTION

The inception of the biggest computer network, Internet, has revolutionised the world with seamless/effortless communication and data sharing. “Everything is connected and nothing is lost”, is the beauty of internet. All the marvels of technology today are a direct representation of the complexity of integrating everything together, and like all good things, this comes at a price, The Data Explosion. The quantity of computer data generated is growing exponentially. The retail shops maintain a product and customer database, social media keeps track of each and every move, and each and every appliance we use generates data. As more smart objects go online, enormous amounts of data is being generated, even various advancements of science, requires huge amounts of data to be stored and processed.

Enter Big Data, “a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them”. It is often characterized using the three v’s, Volume, Velocity and Variety. Here, volume poses a greatest challenge and the greatest opportunity as big data could help many organisations understand people better and allocate resources smarter; however traditional relational databases are not scalable to handle data of this magnitude. Velocity also raises a number of issues, with the rate of which data is flowing into many organisations, exceeding the capacity of their IT Systems, in addition to which, users demand data to be streamed to them in real-time and delivering this is quite of a challenge. Today, right of this moment, hundreds of Terabytes of Photographs, Audio, video, 3d models, simulations and location data are being processed, hence the challenge of Variety.

Today the leading big data technology is Hadoop, an open source library for reliable, scalable distributed computing and provides the first viable platform for Big Data Analytics. Hadoop is already used by many big data pioneers, for example, LinkedIn uses it to generate over 100 billion personalised recommendations every week. Hadoop distributes the storage and processing of large datasets across clusters of servers. Whereas traditions large scale computing solutions rely on expensive hardware with high fault tolerance, Hadoop detects and compensates for hardware failures or other system failures at the application level which allows a phenomenal level service continuity to be delivered by clusters of individual computers each of which is prone to failure.

Hadoop consists of two key components, the first is the Distributed File System which permits high bandwidth cluster based storage, and the second is Map reduce which distributes or maps large datasets across multiple servers. It is a data processing framework in which each server creates a data summary of its allocated data and all of the summarised data is aggravated and stored in a “Reduced” state, Map reduce subsequently allows extremely large raw datasets to be rapid distilled before traditional data analysis tools can be applied.

Problems Faced Today:

Most well off organizations in the 21st century is growing 20-40 percent every year, and with growth comes a steep generation of data and with the sheer rate of data increase it’s really hard to store that data. Companies look at large data storage options like data lakes, which is costly and storing it in cheap raid arrays is not very secure as it is pretty slow and is prone to failure at any time. After data loss comes the problem of security, the data stored in huge data lakes, contains sensitive user data and their tracked activity, and if one knows where to look anyone can find out virtually everything about a person, from what he does to where he lives and nowadays where he goes with a very convenient timestamp, so very convenient for the very wrong people.

Since the Hadoop distributes data in different servers, the integrity of the servers comes to question, for example if data for a client is distributed in eight different locations, one of them might be lower in standards of backup, which might lead to a missing data chunk and most of the stored data is unstructured and rebuilding that data could render it useless.It takes a lot of understanding to get data in the right shape so that you can use visualization as part of data analysis. For example, if the data comes from social media content, you need to know who the user is in a general sense – such as a customer using a particular set of products – and understand what it is you’re trying to visualize out of the data. Without some sort of context, visualization tools are likely to be of less value to the user. Even after knowing this, visualisation is usually overlooked just because of the fact that it is expensive.

The biggest challenge is finding a way to cut through the sheer volume of data and the inherent complexity of different databases and unstructured systems, and make use of that data. As believed by the director of Tata Consultancy Services, a best-practice approach to big data means combining the right technologies and tools, building effective workflows and policies, finding talent that can tap into analytics and predictive analytics software, and applying the technology to delve beyond the surface and address real problems.

Big data:

In this paper **"Big Data-Survey"** by PSG Aruna Sri, Anusha M in Indonesian Journal of Electrical Engineering and Informatics, 2016- In today's world we deal with massive sized data, in different fields such as business, healthcare, banking and so on. Gathering and processing them is very hard but necessary, hence we formulated a new technology which does the above processes which high efficiency and precision. This paper focuses on big data, parallel processing and Hadoop architecture and also different tools used for big data and its security issues.[4]

In this paper **"Empirical Big Data Research: A Systematic Literature Mapping"** by L.W.M. Wienhofen , B.M. Mathisen, D. Roman in information systems, 2015-states that the status of empirical research in Big Data which is a relatively new field in research and technology. This paper has a very effective method of mapping the collected research according to the labels Variety, Volume and Velocity which is described as the 3 V's. The mapping result is also presented in this paper. According to this result the number of publication in the field of big data is well below average in comparison to computer science research as a whole. The authors recommend Variety to be the most promising uncharted area. [5]

In this paper **"Geospatial Big Data: Challenges and Opportunities"** by Jae-GilLee, MinseoKang, in Elsevier 2015- has detected starts off by stating what is geospatial big data and why is significant. Then it discusses about the challenges and opportunities brought by geospatial big data including revenue increase, urban planning etc. Then, it introduce emerging platform for sharing the collected geospatial big data and for tracking human mobility via mobile devices. This paper presents the current research activities toward the analytics of geospatial big data, especially on interactive analytics of real-time or dynamic data. [12]

In this paper **"Research of Big Data Based on the Views of Technology and Application",** by Zan Mo, Yanfei Li, American Journal of Industrial and Business Management, 2015- shows that how big data affects different industries, general life, work, study and economic development. It helps in solving big problems by analysing huge chunk of information's. This paper also describes how to improve the already available features of big data, specially analysis algorithms and storage solutions. It also describes the modern trends in big data and how it is used in different domains. And finally it shows various challenges faced by it. [13]

In this paper **"Data-intensive applications, challenges, techniques and technologies: A survey on Big Data"** by C.L. Philip Chen, Chun-Yang Zhang, in Elsevier 2014- According to the authors, big data has already achieved huge attention in various domains but it still has a lot of potential. Overabundance of information is a huge issue as well as extracting needed datasets. A new scientific paradigm is born as data intensive scientific discovery (DISD), also known as Big Data problems. Though big data is essential for improving productivity of businesses but big data problems are still a bottleneck. This paper discusses about some methods to deal with them like granular computing, cloud computing, bio-inspired computing, and quantum computing. [16]

In this paper **"Application of Big Data in Education Data Mining And Learning Analytics – A Literature Review"** by Katrina Sin and Loganathan Muthu, ICTACT JOURNAL ON SOFT COMPUTING, in july 2015- The usage of learning management systems in education has been increasing in the last few years. Students have started using mobile phones, primarily smart phones that have become a part of their daily life, to access online content. Student's online activities generate enormous amount of unused data that are wasted as traditional learning analytics are not capable of processing them. This has resulted in the penetration of Big Data technologies and tools into education, to process the large amount of data involved. This study looks into the recent applications of Big Data technologies in education and presents a review of literature available on Educational Data Mining and Learning Analytics. [19]

In this paper **"A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES"** by S. Justin Samuel, Koundinya RVP, etc., ARPN Journal of Engineering and Applied Sciences, in May 2015- There has been an ever-increasing interest in big data due to its rapid growth and since it covers diverse areas of applications. Hence, there seems to be a need for an analytical review of recent developments in the big data technology. This paper aims to provide a comprehensive review of the big data state of the art, conceptual explorations, major benefits, and research challenging aspects. In addition to that, several future directions for big data research are highlighted. [20]

In this paper **"Survey of Research on Information Security in Big Data"** by Zhang Hongjun, Hao Wenning, etc., workshop on social networks, 2014- With the advancement of information technology, big data application prompts the development of storage, network and computer field. It also brings new security problems. This security challenge caused by big data has attracted the attention of information security and industrial community domain. This paper summarizes the characteristics of big data information security, and focuses on conclusion of security problems under the big data field and the inspirations to the development of information security technology. Finally, this paper outlooks the future and trend of big data information security. [23]

In this paper **"Big Data Security Issues and Challenges"** by Raghav Toshniwal, Kanishka Ghosh Dastidar, Asoke Nath, in IJIRAE, Feb 2015- deals with the issues like over-saturation of data. Traditional database systems are not able to capture, store and analyse this large amount of data. Big data analytics provide new ways for businesses and government to analyse unstructured data. This paper defines Big Data and discusses the parameters along which Big Data is defined. This includes the three V's. The authors also look at processes involved in data processing and review the security aspects of Big Data and propose a new system for Security of Big Data and finally present the future scope of Big Data. [26]

In this paper **"Big data analytics in healthcare: promise and potential"** by Wullianallur Raghupathi and Viju Raghupathi, in Health Information Science and Systems 2014- describes that promise and potential of big data analytics in healthcare and the benefits of implementing it. It also outlines an architectural framework and methodologies. It also briefly discusses about the challenges and concludes with a possible solution. Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome. [28]

In this paper "The Impact of Big Data on the Healthcare Information Systems" by Kuo Lane Chen, Huei Lee, in Transactions of the International Conference on Health Information Technology Advancement, 2013- Big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware; nor can they be easily managed with traditional or common data management tools and methods. This article explores the possible impact of big data on healthcare information systems. Possible research issues include: 1). what applications in healthcare information systems are impacted most? 2). what algorithm/programs will be used for big data? 3). what privacy, security, and ethical issues are there for big data? In the biology area, big data becomes the newest technology for genomics. For the big data scientist, there is, amongst this vast amount and array of data, opportunity. By discovering associations and understanding patterns and trends within the data, big data analytics has the potential to improve care, save lives and lower costs. Thus, big data analytics applications in healthcare take advantage of the explosion in data to extract insights for making better informed decisions. [32]

In this paper **"BIG Data and Methodology"** by Shilpa, Manjit Kaur, in IJARCSSE, oct 2013- The rapid evolution and adoption of big data by industry has leapfrogged the discourse to popular outlets, forcing the academic press to catch up. Academic journals in numerous disciplines, which will benefit from a relevant discussion of big data, have yet to cover the topic. This paper presents a consolidated description of big data by integrating definitions from practitioners and academics. The paper's primary focus is on the analytic methods used for big data. Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. Big data analytics is the process of examining large amounts of data. Big Data is characterized by the dimensions volume, variety, and velocity, while there are some well-established methods for big data processing such as Hadoop which uses the map-reduce paradigm. [34]

In this paper **"Big-Data Security"** by Kalyani Shirudkar, Dilip Motwani, IJARCSSE, Mar 2015- Big Data is an immensely popular talking point, but from a security perspective, there are two distinct issues: securing the organisation and its customers' information in a Big Data context; and using Big Data techniques to analyse, and even predict, security incidents. Many businesses already use Big Data for marketing and research, yet may not have the fundamentals right – particularly from a security perspective. As with all new technologies, security seems to be an afterthought at best. Big Data breaches will be big too, with the potential for even more serious reputational damage and legal repercussions than at present. A growing number of companies are using the technology to store and analyse petabytes of data including web logs, click stream data and social media content to gain better insights about their customers and their business. [35]

In this paper **"Web technologies for environmental Big Data",** by Claudia Vitolo , Yehia Elkhatib, etc., in Elsevier, 2015- Web services are essential in the orchestration of internet-based workflows. In essence, a web service is an application that enables access to its functions using established internet standards. As such they provide seamless cross-platform interoperability between different loosely coupled systems. Currently, two main architectural styles are most commonly used: SOAP and REST. Recent evolutions in computing science and web technology provide the environmental community with continuously expanding resources for data collection and analysis that pose unprecedented challenges to the design of analysis methods, workflows, and interaction with data sets. In the light of the recent UK Research Council funded Environmental Virtual Observatory pilot project. [36]

In big data security, various researchers have proposed several cryptography algorithms under various environment constraints such as Cloud-centric and IoT (Internet of Things). Some of the recent variants et. al [54-64] gives the asymmetric based secured communication among the client and server ends. As the number of clients gets increases, the complexity to generate the keys also gets increases. So before the cryptography design one has to consider both the time-memory trade-offs.

In this paper **"Bootstrapping Smart Cities through a Self-Sustainable Model Based on Big Data Flows"** by Ignasi Vilajosana, Jordi Llosa, etc.,in IEEE, jun 2013- To propose a viable approach to scale business within that ecosystem. We also describe the available ICT technologies and finally exemplify all findings by means of a sustainable smart city application. Over the course of the article, we draw two major observations, which are seen to facilitate sustainable smart city development. First, independent smart city departments (or the equivalent) need to emerge, much like today's well accepted IT departments, which clearly decouple the political element of the improved city servicing from the underlying technologies. Second, a coherent three-phase smart city rollout is vital, where in phase 1 utility and revenues are generated; in phase 2 only-utility services is also supported; and in phase 3, in addition, a fun/leisure dimension is permitted. [39]

In this paper **"Optimizing Big Data Processing Performance in the Public Cloud: Opportunities and Approaches"** by Dan Wang, Jiangchuan Liu, in National Natural Science Foundation of China- A huge amount of data is generated these days and it is now important to make this generation of data more efficient for processing. This requires huge demands on the computing and networking infrastructures. State-of-the-art tools, most notably MapReduce, are generally performed on dedicated server clusters to explore data parallelism. For grass root users or non-computing professionals, the cost for deploying and maintaining a large-scale dedicated server clusters can be prohibitively high, not to mention the technical skills involved. On the other hand, public clouds allow general users to rent virtual machines and run their applications in a pay-as-you-go manner with ultra-high scalability and yet minimized upfront costs. This new computing paradigm has gained tremendous success in recent years, becoming a highly attractive alternative to dedicated server clusters. [40]

In this paper **"Big Data Analytics: A Literature Review Paper"** by Nada Elgendy and Ahmed Elragal, in Springer International Publishing Switzerland, 2014- Data analytics is used to examine raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories. Big data analytics is the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Big data analytics can help organizations to better understand the inforation contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with big data basically want the knowledge that comes from analyzing the data. [41]

In this paper **"Adoption of Big Data Technology for the Development of Developing Countries"** by Remya Panicker, in National Conference on New Horizons in IT - NCNHIT 2013- Big Data can be extreamly influential in the development of a nation. Some of the major infrastructures in a nation are dependent on or can be improved using information from big data technologies. We can explore the various possibilities that Big Data can deliver with prospect to improve decision-making in critical development areas such as health care, employment, economic productivity, crime and security, natural disaster and resource management. The various assumptions based on the fact available are made to create a model for the nation building. We can find the outcome of implementing Big Data in various area. A conceptual model is created that can be implemented or build in future. It shows the opportunities that are useful in policy making and decision making with use of Big Data. This can be immensly influential in the development of a Developing Country. [42]

In this paper **"The Use of Big Data in Education"** by Athanasios S. Drigas and Panagiotis Leliopoulos, in IJCSI, Sep 2014- Data and analytics increasingly are being used by forward-thinking teachers and school administrators to gain an overview of how the provision of services is going in their districts. In Wisconsin's Menomonee Falls School District, data has been put to use for everything from improving classroom cleanliness to planning school bus routes, after department leaders were encouraged to attend classes themselves on how to gain insights from data and analytics. Products and services are coming onto the market to automate many processes based on big data analytics. Eduvant, for example, is a tool that allows teachers and school administrators to get everything from an overview of the school's performance against its targets around academic achievement and discipline, to warnings when an individual pupil's learning is not progressing as expected. [44]

In this paper **"Perspectives on Big Data and Big Data Analytics"** by Elena Geanina, Florina Camelia, Anca, Manole, in Database Systems Journal, 2012- In order to keep up with the information based society we are a part of now, we need to manage and analyze huge quantities of data and fast. This data, also known as Big data, has repositories that include information deposited by various sources across the enterprise. This variety of data makes analytics management a challenge. Each data source will likely have its own access restrictions and security policies, making it difficult to balance appropriate security for all data sources with the need to aggregate and extract meaning from the data. The large quantity of data is better used as a whole because of the possible correlations on a larger amount, correlations that can never be found if the data is analyzed on separate sets or on a smaller set. A larger amount of data gives a better output but also working with it can become a challenge due to processing limitations. [48]

Big data analytics:

In this paper ***"Beyond the hype: Big data concepts, methods, and analytics" by*** Amir Gandomi, Murtaza Haider published in the year 2015 - The evolution of big data is quite a new topic. Though the various outlets are trying to catch up with this technology to get a broader perspective, the fast and rapid evolution is acting as a bottleneck. This paper presents us with a brief grasp of the defining characteristics by focusing primarily on the analytic method used in big data. The conclusion of this paper is to improve the analytical algorithms related to the unstructured data as it constitutes 95% of big data. [1]

In this paper **"Big data analytics: a survey"** by Chun-Wei Tsai, Chin-Feng Lai, etc., in Tsai *et al. Journal of Big Data, 2015*- Big data is a huge success in terms of data management, but the traditional data analytics are not so efficient at handling such huge chunks of data. This paper deals with the idea of making a high performance platform to efficiently analyse big data. It also describes the uses of appropriate mining algorithms to find useful things from big data. The overview of this paper is the idea on big data and analytics and also the next step of big data analytics. [9]

In this paper **"Big Data Analytics for Healthcare"** by Jimeng Sun, Chandan K. Reddy in ACM, 2013- Healthcare industry is huge, and thus the relevant information's or datasets are overwhelmingly large. The enormity and complexity of these datasets present great challenges in analyses and subsequent applications to a practical clinical environment. The authors present us a tutorial on how to deal with these big medical data.

This tutorial is based on various surveys performed to obtain information on key problems and trends in healthcare. This tutorial also includes several case studies. [24]

In this paper **"A Review Paper on Big Data Analytics"** by Ankita S. Tiwarkhede, Prof. Vinit Kakde, in International Journal of Science and Research (IJSR), Apr 2015- The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet clickstream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. The goal of Big Data analysis is to extract useful values, suggest conclusions and/or support decision making. In this topic, we provide an extensive survey of big data analytics research, while highlighting the specific concern in big data world. According to Application evolution, we discuss six types of big data application such as structured data analytics, Text analytics, Web analytics, Multimedia analytics, and Mobile analytics. [31]

In this paper **"Trends in big data analytics"** by Karthik Kambatla , Giorgos Kollias, etc., in Elsevier, 2014- Continued growth of systems that support non-relational or unstructured forms of data as well as massive volumes of data. These systems will evolve and mature to operate well inside of enterprise IT systems and standards. This will enable both business users and data scientists to fully realize the value of big data. Data repositories for such systems currently exceed exabytes and are rapidly increasing in size. Beyond their sheer magnitude, these datasets and associated applications' considerations pose significant challenges for method and software development. Datasets are often distributed and their size and privacy considerations warrant distributed techniques. Data often resides on platforms with widely varying computational and network capabilities. Considerations of fault-tolerance, security, and access control are critical in many applications (Dean and Ghemawat, 2004; Apache hadoop). Analysis tasks often have hard deadlines, and data quality is a major concern in yet other applications. [33]

In this paper **"Starfish: A Selftuning System for Big Data Analytics"** by Herodotos Herodotou, Harold Lim, etc., in 5th Biennial Conference on Innovative Data Systems Research, Jan 2011- Starfish is a self-tuning system for analytics on big data. An important design decision is to build Starfish on the Hadoop stack. Hadoop, as observed, has useful primitives to help meet the new requirements of big data analytics. In addition, Hadoop's adoption by academic, government, and industrial organizations is growing at a fast pace. Web search engines and social networks capture and analyze every user action on their sites to improve site design, spam and fraud detection, and advertising opportunities. Powerful telescopes in astronomy, genome sequencers in biology, and particle accelerators in physics are putting massive amounts of data into the hands of scientists. Key scientific breakthroughs are expected to come from computational analysis of such data. [37]

Big data in cloud:

In this paper **"The rise of "big data" on cloud computing: Review and open research issues"** by Ibrahim Abaker Targio Hashem a,n , IbrarYaqoob in Elsevier 2015- One of the main issues faced by modern technologies is scarcity of storage. The use of local storage systems for performing massive-scale computing is very inefficient. The solution to this issue is cloud computing which eliminates the need of unnecessary hardware and software. The rise of big data in cloud computing is reviewed in this study. The primary focus is on the relationship and systems like Hadoop. The challenges are also studied to help in targeted research efforts. [2]

In this paper **"Big Data computing and clouds: Trends and future directions"** by Marcos D. Assunção a, Rodrigo N. Calheiros, etc., in Elsevier 2015-states that approaches and environments for carrying out analytics on Clouds for Big Data applications. It revolves around four important areas of analytics and Big Data, namely data management and supporting architectures; model development and scoring; visualization and user interaction; and (iv) business models. Through a detailed survey, we can identify possible gaps in technology and provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions. [11]

In this paper **“Security Issues Associated With Big Data in Cloud Computing”** by Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, International Journal of Network Security & Its Applications (IJNSA), May 2014- discusses about security issues for cloud computing, big data, Map Reduce and Hadoop environment. The main focus is on the security issues of cloud computing that are associated with big data. It also highlights on different issues of cloud computing and Hadoop ecosystem and how to solve them. It ends with the role of cloud computing in various fields and the advantages of using it. [17]

In this paper **“Schedule optimization for big data processing on cloud”** by Ibrahim Abaker Targio Hashem, Nor Badrul Anuar and Abdullah Gani, in Big Data Analysis and Data Mining November 30-December 01, 2015- The major enabler for underlying many big data platforms is certainly the MapReduce computational paradigm. MapReduce is recognized as a popular programming model for the distributed and scalable processing of big data and is increasingly being used in different applications mostly because of its important features that include scalability, flexibility, ease of programming, and fault-tolerance. Scheduling tasks in MapReduce across multiple nodes have shown to be multiobjective optimization problem. The problem is even more complex by using virtualized clusters in a cloud computing to execute a large number of tasks. The complexity lies in achieving multiple objectives that may be of conflicting nature. For instance, scheduled tasks may require to make several tradeoffs between the job performance, data locality, fairness, resource utilization, network congestion and reliability. [38]

In this paper **“Cloud-Based Big Data Aanalytics – A Survey Of Current Research And Future Directions”** by Samiya Khan1, Kashish Ara Shakil and Mansaf Alam- Data analytis plays an important role in the current data based systems. A lot of information can be extracted by finding out how the data for a particular application is being handled in the cloud. The advent of the digital age has led to a rise in different types of data with every passing day. In fact, it is expected that half of the total data will be on the cloud soon. This data is complex and needs to be stored, processed and analyzed for information that can be used by organizations. Cloud computing provides an apt platform for big data analytics in view of the storage and computing requirements of the latter. This makes cloud-based analytics a viable research field. However, several issues need to be addressed and risks need to be mitigated before practical applications of this synergistic model can be popularly used. This paper explores the existing research, challenges, open issues and future research direction for this field of study. [46]

In this paper **“Security Issues with Big Data in Cloud Computing”** by Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, in IJNSA, May 2014- As the amount of data being collected continues to grow, more and more companies are building big data repositories to store, aggregate and extract meaning from their data. Big data provides an enormous competitive advantage for corporations, helping businesses tailor their products to consumer needs, identify and minimize corporate inefficiencies, and share data with user groups across the enterprise. With a growth rate of 58 percent recently, these technologies and their benefits are here to stay. The main focus is on security issues in cloud computing that are associated with big data. Big data applications are a great benefit to organizations, business, companies and many large scale and small scale industries. We also discuss various possible solutions for the issues in cloud computing security and Hadoop. These tools take steps to ensure the security in Big Data and related fields. [47]

Big data and Hadoop:

In this paper **“A Review Paper on Big Data and Hadoop ”** by Harshawardhan S. Bhosale , Prof. Devendra P. Gadekar in International Journal of Scientific and Research Publications, 2014- describes that the characteristics of big data and defines it. It also describes how to manage and control big data in orderly fashion, like the use of parallelism. It’s focus is the open source software Hadoop which is the core platform for structuring, and solving problems of big data. Hadoop is also used to distribute processing of large data sets across clusters of commodity servers to scale up from a single server to thousands of machines. [3]

In this paper **“Big Data and Hadoop: A Review Paper”** by Rahul Beakta, RIEECE, 2015- defines that why it is popular and why we should master new techniques to analyse and process it. Mastery of data analysis is required to get the information from unstructured data. This paper also describes the advantage and scope of big data. An overview on opportunities to healthcare, technology etc. is given. It also gives an introduction to Hadoop and data mining. [8]

In this paper **"A survey of open source tools for machine learning with big data in the Hadoop ecosystem",** by Sara Landset, Taghi M. Khoshgoftaar, etc., in *Journal of Big Data, 2015*- It deals with the idea of selecting the best machine learning tool for big data. The available tools have many advantages and drawbacks, this paper is intended to aid the researchers to build tool for distributed and real-time processing. It also highlights Hadoop ecosystem. It also discusses the advantages and disadvantages of three different processing paradigms along with a comparison of engines that implement them. It's main goal is to make the process of selecting the best machine learning tool smoother by providing information as much as possible.[10]

In this paper **"Big Data Analytics with Hadoop to analyse Targeted Attacks on Enterprise Data",** by Bhawna Gupta Dr. Kiran Jyoti, International Journal of Computer Science and Information Technologies, in 2014- describes big data analytics as the process of analysing and mining unstructured or too fast changing information's. The big data security analytics system has many flaws since it rely on untrustworthy data. Attackers have become more adapt at highly targeted, complex attacks that overtake static threat detection measures. This paper discusses about the techniques of how big data is analysed and how to make the process more secure and why it is so important. [18]

In this paper **"MRPR: A Map Reduce solution for prototype reduction in big data classification"** by Isaac Triguero , DanielPeralta, in October 2014- emphasizes on how difficult it is to analyse and extract knowledge from large set of datasets. The authors propose a novel distributed partitioning methodology for prototype education techniques in nearest neighbour classification. Their main purposes are to speed up the classification process and reduce the storage requirements and sensitivity to noise of the nearest neighbour rule. Still this model has some underlying issues which are solved by a Map Reduce-based framework. [21]

In this paper **"A Comprehensive View of Hadoop Map Reduce Scheduling Algorithms"** by Seyed Reza Pakize, in International Journal of Computer Networks and Communications Security, Sep 2014- It starts off by explaining what Hadoop framework along with Map Reduce model is. It also explains the role of Map Reduce model in big data and how it is used to paralyze the job execution across multiple nodes for execution. Then it discusses about the three important scheduling issues in Map Reduce such as locality, synchronization and fairness. It then discusses about the common objectives of scheduling algorithms. Finally, highlighting the implementation Idea, advantages and disadvantage of these algorithms. [22]

In this paper **"Efficient Big Data Processing in Hadoop Map Reduce"** by Jens Dittrich, JorgeArnulfo Quian´eRuiz, in The 38th International Conference on Very Large Data Bases, August 27th 31st 2012, Istanbul, Turkey- states that all about handling big data volumes efficiently, and how it can be done by using advanced tool like Hadoop Map Reduce. It also describes the technique to boost performance, job optimization and physical data organization. This paper highlights on many topics such as detailing Hadoop Map Reduce and other system like Parallel DBMS, and the similarities and differences between them. Finally, it touches on unresolved research problems and open issues. [25]

In this paper "Big Data Analytics using Hadoop" " by Bijesh Dhyani, Anurag Barthwal, in International Journal of Computer Applications, Dec 2014- describes big data and its usefulness to an organization from the performance perspective. Also, the important parameters and attributes that make this emerging concept attractive to organizations have been highlighted. The paper also evaluates the differences in the approach and treatment of big data by small, medium and large organizations. The second part of the paper deals with the technology aspects of big data for its implementation. The paper deals with the overall architecture of Hadoop along with the details of its various components. [27]

In this paper **"Big Data challenges and Hadoop as one of the solution of big data with its Modules"** by Tapan P. Gondaliya, Dr. Hiren D. Joshi, International Journal of Scientific & Engineering Research, June-2014- This entire mainly depending upon zettabytes of information's. But problem is it is huge and over-saturated. It is very hard to organize or manage this kind of big data. Apache Hadoop is one of the most popular frameworks which is very efficient in handling big data. This paper describes the main purpose of Apache Hadoop and how can it organizes or manages the different kinds of data and what are the main techniques used for doing so. [29]

In this paper **"A Review Paper on Big Data and Hadoop"** by Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, in International Journal of Scientific and Research Publications, Oct 2014- Big data is large volumes of unstructured data and the goal is to analyze this. For others, it encompasses all of their companies' data and how they may better monetize or analyze the data they have, whether structured or unstructured. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur on multiple dimensions. [45]

Data Mining in Big Data:

In this paper **"Data Mining with Big Data"** by Xindong Wu, Gong-Qing Wu, and Wei Ding in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2014- states that raises issues like its rapid expansion in all science and engineering domains, like physical, biological and biomedical science. It presents a HACE theorem that characterizes the features of big data revolution, and proposes a big data processing model, from data mining perspective. It analyses the data-driven model and its involvement in big data. It also analyses the challenges faced by this model. [6]

In this paper **"Mining Big Data to Predicting Future"** by Amit K. Tyagi, R. Priya,A. Rajeswari, Int. Journal of Engineering Research and Applications, 2015- The paper starts off with the importance of big data and the tools used to manage it and extracting datasets. According to the study the current methodologies or data mining software tools are not sufficient. But if done correctly these huge datasets can be used to predict the future. This paper discusses about this topic in a broad overview like; its current status; controversy; and challenges to forecast the future. This paper defines at some of these problems, using illustrations with applications from various areas. Finally this paper discusses secure management and privacy of big data as one of essential issues. [14]

In this paper **"Data, DIKW, Big data and Data science"** Data, DIKW, Big data and Data science by Gu Jifa, Zhang Lingling, 2nd International Conference on Information Technology and Quantitative Management, ITQM, 2014- has stated the relationship between data and DIKW, that the data only evolves to knowledge, which may have some value, but if without the wisdom we still could let the knowledge be really useful to people. Now the big data occupies much attention in some extent for his volume, velocity, and variety or the 3 V's. But in practical use the value plays more important role. Finally to judge the value for data not necessary for big, in some cases the small data also may lead to big value. So we appreciate the data science, which may consider more inherent value from data. [15]

Scheduling and Big data:

In this paper **"A Survey on Job Scheduling Algorithms in Big Data Processing"** by Jyoti V Gautam, Harshadkumar B Prajapati, etc.- The Apache Hadoop framework is considered as the gold standard when it comes to distributed data processing especially because of it being open source. This paper highlights fundamental issues in job scheduling, present's classification of Hadoop schedulers, and discusses presented survey of existing scheduling algorithm. It describes the Hadoop-MapReduce model in details. It also discusses how to customize Map Reduce feature for improving performance. This paper is very efficient for beginners and researchers to understand scheduling in big data processing. [7]

In this paper **"A Survey on Big Data Management and Job Scheduling"** by Sreedhar C, N. Kasiviswanath, P. Chenna Reddy, in International Journal of Computer Applications, Nov 2015- Big data management focuses in handling the huge quantities of data acquired through the various sources and its storage and transportation. It has been noticed though that in the recent years there is an increased need for sophisticated method to collect, process, analyse and visualize huge volumes of data generated by our digital and computing world. This has helped the advent of Big Data. Big data management is the organization, administration and governance of large volumes of both structured and unstructured data. The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications. Corporations, government agencies and other organizations employ big data management strategies to help them contend with fast-growing pools of data, typically involving many terabytes or even petabytes of information saved in a variety of file formats. To be effective in big data management proper job scheduling channels have to be used and utilized. [30]

In this paper **"A Survey on Job Scheduling in Big Data"** by M. Senthilkumar, P. Ilango, in CYBERNETICS AND INFORMATION TECHNOLOGIES, 2016- Scheduling has become a major part of Big Data application in the last few years. When distributed data processing comes to mind, Hadoop is a big player in the area. It is extreamly popular and the most used framework for the use. Hadoop is also open source software that allows the user to effectively utilize the hardware. Various scheduling algorithms of the MapReduce model using Hadoop vary with design and behavior, and are used for handling many issues like data locality, awareness with resource, energy and time. This paper gives the outline of job scheduling, classification of the scheduler, and comparison of different existing algorithms with advantages, drawbacks, limitations. In this paper, we discussed various tools and frameworks used for monitoring and the ways to improve the performance in MapReduce. The pager is a step towards understanding the applications of Big Data Scheduling systems and its mechanisms. [44][51][52][53]

CONCLUSION

Big data is the emerging research field; this paper is an attempt to present an overview of big data, its applications in various fields. This paper also brings out the challenges and the research opportunities in big data.

REFERENCES

[1] "Beyond the hype: Big data concepts, methods, and analytics" by Amir Gandomi, Murtaza Haider in Elsevier 2015.

[2] "The rise of "big data" on cloud computing: Review and open research issues" by Ibrahim Abaker Targio Hashem a,n , IbrarYaqoob in Elsevier 2015.

[3] "A Review Paper on Big Data and Hadoop" by Harshawardhan S. Bhosale , Prof. Devendra P. Gadekar in International Journal of Scientific and Research Publications, 2014.

[4] "Big Data-Survey" by PSG Aruna Sri*, Anusha M in Indonesian Journal of Electrical Engineering and Informatics, 2016.

[5] "Empirical Big Data Research: A Systematic Literature Mapping" by L.W.M. Wienhofen , B.M. Mathisen, D. Roman in information systems, 2015.

[6] "Data Mining with Big Data" by Xindong Wu, Gong-Qing Wu, and Wei Ding in IEEE transactions on knowledge and data engineering, 2014.

[7] "A Survey on Job Scheduling Algorithms in Big Data Processing" by Jyoti V Gautam, Harshadkumar B Prajapati, etc.

[8] "Big Data And Hadoop: A Review Paper" by Rahul Beakta in RIEECE, 2015.

[9] Big data analytics: a survey" by Chun-Wei Tsai, Chin-Feng Lai in Tsai et al. Journal of Big Data, 2015.

[10] "A survey of open source tools for machine learning with big data in the Hadoop ecosystem" by Sara Landset, Taghi M. Khoshgoftaar, etc., in Journal of Big Data, 2015.

[11] "Big Data computing and clouds: Trends and future directions" by Marcos D. Assunção a, Rodrigo N. Calheiros, etc., in Elsevier 2015.

[12] "Geospatial Big Data: Challenges and Opportunities" by Jae-GilLee, MinseoKang, in Elsevier 2015

[13] "Big Data computing and clouds: Trends and future directions" by Marcos D. Assunção , Rodrigo N. Calheiros, etc., in Elsevier 2015.

[14] "Research of Big Data Based on the Views of Technology and Application" by Zan Mo, Yanfei Li, etc., American Journal of Industrial and Business Management, 2015.

[15] "Mining Big Data to Predicting Future" by Amit K. Tyagi, R. Priya,A. Rajeswari, Int. Journal of Engineering Research and Applications, 2015

[16] Data, DIKW, Big data and Data science by Gu Jifa, Zhang Lingling, 2nd International Conference on Information Technology and Quantitative Management, ITQM, 2014

[17] Data-intensive applications, challenges, techniques and technologies: A survey on Big Data" by C.L. Philip Chen, Chun-Yang Zhang, in Elsevier 2014.

[18] “Security Issues Associated With Big Data In Cloud Computing” by Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, International Journal of Network Security & Its Applications (IJNSA), in May 2014

[19] “Big Data Analytics with Hadoop to analyse Targeted Attacks on Enterprise Data” by Bhawna Gupta Dr. Kiran Jyoti, International Journal of Computer Science and Information Technologies, in 2014

[20] “Application of Big Data In Education Data Mining And Learning Analytics – A Literature Review” by Katrina Sin and Loganathan Muthu, ICTACT Journal on Soft Computing, in july 2015

[21] “A Survey On Big Data And Its Research Challenges” by S. Justin Samuel, Koundinya RVP, etc., ARPN Journal of Engineering and Applied Sciences, in May 2015

[22] “MRPR: A MapReduce solution for prototype reduction in big data classification” by Isaac Triguero , DanielPeralta, in Elsevier, October 2014.

[23] “A Comprehensive View of Hadoop MapReduce Scheduling Algorithms” by Seyed Reza Pakize, in International Journal of Computer Networks and Communications Security, Sep 2014.

[24] “Survey of Research on Information Security in Big Data” by Zhang Hongjun, Hao Wenning, etc., workshop on social networks, 2014.

[25] “Big Data Analytics for Healthcare” by Jimeng Sun, Chandan K. Reddy in ACM, 2013.

[26] “Efficient Big Data Processing in Hadoop MapReduce” by Jens Dittrich, JorgeArnulfo Quian´eRuiz, in The 38th International Conference on Very Large Data Bases, August 27th 31st 2012, Istanbul, Turkey.

[27] “Big Data Security Issues and Challenges” by Raghav Toshniwal, Kanishka Ghosh Dastidar, Asoke Nath, in International Journal of Innovative Research in Advanced Engineering (IJIRAE) , Feb 2015.

[28] “Big Data Analytics using Hadoop” by Bijesh Dhyani, Anurag Barthwal, in International Journal of Computer Applications, Dec 2014.

[29] “Big data analytics in healthcare: promise and potential” by Wullianallur Raghupathi and Viju Raghupathi, in Health Information Science and Systems 2014.

[30] “Big Data challenges and Hadoop as one of the solution of big data with its Modules” by Tapan P. Gondaliya, Dr. Hiren D. Joshi, International Journal of Scientific & Engineering Research, June-2014.

[31] “A Survey on Big Data Management and Job Scheduling” by Sreedhar C, N. Kasiviswanath, P. Chenna Reddy, in International Journal of Computer Applications, Nov 2015.

[32] “A Review Paper on Big Data Analytics” by Ankita S. Tiwarkhede, Prof. Vinit Kakde, in International Journal of Science and Research (IJSR), Apr 2015.

[33] “The Impact of Big Data on the Healthcare Information Systems” by Kuo Lane Chen, Huei Lee, in Transactions of the International Conference on Health Information Technology Advancement, 2013.

[34] “Trends in big data analytics” by Karthik Kambatla , Giorgos Kollias, etc., in Elsevier, 2014.

[35] “BIG Data and Methodology-A review” by Shilpa, Manjit Kaur, in International Journal of Advanced Research in Computer Science and Software Engineering, oct 2013.

[36] “Big-Data Security” by Kalyani Shirudkar, Dilip Motwani, in International Journal of Advanced Research in Computer Science and Software Engineering, Mar 2015.

[37] “Web technologies for environmental Big Data” by Claudia Vitolo , Yehia Elkhatib, etc., in Elsevier, 2015.

[38] “Starfish: A Selftuning System for Big Data Analytics” by Herodotos Herodotou, Harold Lim, etc., in 5th Biennial Conference on Innovative Data Systems Research, Jan 2011.

[39] “Schedule optimization for big data processing on cloud” by Ibrahim Abaker Targio Hashem, Nor Badrul Anuar and Abdullah Gani, in Big Data Analysis and Data Mining November 30-December 01, 2015.

[40] “Bootstrapping Smart Cities through a Self-Sustainable Model Based on Big Data Flows” by Ignasi Vilajosana, Jordi Llosa, etc.,in IEEE, jun 2013.

[41] “Optimizing Big Data Processing Performance in the Public Cloud: Opportunities and Approaches” by Dan Wang, Jiangchuan Liu, in National Natural Science Foundation of China.

[42] “Big Data Analytics: A Literature Review Paper” by Nada Elgendy and Ahmed Elragal, in Springer International Publishing Switzerland, 2014

[43] Adoption of Big Data Technology for the Development of Developing Countries” by Remya Panicker, in National Conference on New Horizons in IT - NCNHIT 2013.

[44] Senthilkumar, M., Ilango, P. A survey on job scheduling in big data (2016) Cybernetics and Information Technologies, 16 (3), pp. 35-51.

[45] “A Survey on Job Scheduling in Big Data” by M. Senthilkumar, P. Ilango, in Cybernetics and Information Technologies, 2016.

[46] “The Use of Big Data in Education” by Athanasios S. Drigas and Panagiotis Leliopoulos, in International Journal of Computer Science, Sep 2014.

[47] “A Review Paper on Big Data and Hadoop” by Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, in International Journal of Scientific and Research Publications, Oct 2014.

[48] “Cloud-Based Big Data Analytics – A Survey of Current Research And Future Directions” by Samiya Khan1, Kashish Ara Shakil and Mansaf Alam.

[49] “Security Issues Associated With Big Data In Cloud Computing” by Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri, in International Journal of Network Security & Its Applications (IJNSA), May 2014.

[50] “Perspectives on Big Data and Big Data Analytics” by Elena Geanina, Florina Camelia, Anca, Manole, in Database Systems Journal, 2012.

[51] M. Senthikumar and P. Ilango, 2016. Big Data Optimization for Social Networking Tweet. International Journal of Soft Computing, 11: 305-311. DOI: 10.3923/ijscomp.2016.305.311

[52] Senthilkumar, M., Ilango, P. Analysis of DNA data using hadoop distributed file system (2016) Research Journal of Pharmaceutical, Biological and Chemical Sciences, 7 (3), pp. 796-803.

[53] M.Senthilkumar, P.Ilango. "Weather Data Analysis Using Hadoop" International Journal of Pharmacy and Technology 8.4 (2016): 21827-21834.

[54] Chandrasegar Thirumalai, “Physicians Drug encoding system using an Efficient and Secured Linear Public Key Cryptosystem (ESLPKC),” International journal of pharmacy and technology, Vol. 8 Issue 3, Sep. 2016, pp. 16296-16303

[55] Chandrasegar Thirumalai, Senthilkumar M, “Secured E-Mail System using Base 128 Encoding Scheme,” International journal of pharmacy and technology, Vol. 8 Issue 4, Dec. 2016, pp. 21797-21806.

[56] Chandrasegar Thirumalai, Senthilkumar M, Silambarasan R, Carlos Becker Westphall, “Analyzing the strength of Pell’s RSA,” IJPT, Vol. 8 Issue 4, Dec. 2016, pp. 21869-21874.

[57] Chandrasegar Thirumalai, “Review on the memory efficient RSA variants,” International Journal of Pharmacy and Technology, Vol. 8 Issue 4, Dec. 2016, pp. 4907-4916.

[58] Chandrasegar Thirumalai, Senthilkumar M, Vaishnavi B, “Physicians Medicament using Linear Public Key Crypto System,” in International conference on Electrical, Electronics, and Optimization Techniques ICEEOT, IEEE & 978-1-4673-9939-5, March 2016.

[59] T Chandra Segar, R Vijayaragavan, “Pell's RSA key generation and its security analysis,” in Computing, Communications and Networking Technologies (ICCCNT) 2013, pp. 1-5

[60] Chandrasegar Thirumalai, Senthilkumar M, “Spanning Tree approach for Error Detection and Correction,” IJPT, Vol. 8, Issue No. 4, Dec-2016, pp. 5009-5020.

[61] Chandrasegar Thirumalai, Senthilkumar M, “An Assessment Framework of Intuitionistic Fuzzy Network for C2B Decision Making”, International Conference on Electronics and Communication Systems (ICECS), IEEE & 978-1-4673-7832-1, Feb. 2016

[62] Vaishnavi B, Karthikeyan J, Kiran Yarrakula, Chandrasegar Thirumalai, “An Assessment Framework for Precipitation Decision Making Using AHP”, International Conference on Electronics and Communication Systems (ICECS), IEEE & 978-1-4673-7832-1, Feb. 2016

[63] Vinothini S, Chandra Segar Thirumalai, Vijayaragavan R, Senthil Kumar M, “A Cubic based Set Associative Cache encoded mapping,” International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 02 May -2015 pp. 360-364

[64] Nallakaruppan, M.K., Senthil Kumar, M., Chandrasegar, T., Suraj, K.A., Magesh, G., “Accident avoidance in railway tracks using Adhoc wireless networks,” 2014, International Journal of Applied Engineering Research, 9 (21), pp. 9551-9556.