ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 1, 2025

Big Data Analytics: Challenges, Issues and Tools

*A Bhasha [1], A Bheem raj[2], K Yadagiri[3], V Chandraprakash

[1,2,3,4] Assistant Professor, St. Martin's Engineering College, Secunderabad, Telangana – 500100

*Corresponding Author

Email: abhashait@smec.ac.in

ABSTARCT

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 big data 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.

Keywords—Big data analytics; Hadoop; Massive data; Structured data; Unstructured Data

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. 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 whereas velocity is the rate of growth and 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 techniques [1]. Some of these extraction methods for obtaining helpful information was discussed by Gandomi and Haider[2]. The following Figure 1 refers to the definition of big data. However exact definition for big data is not defined and there is a believe that it is problem specific. This will help us in obtaining enhanced decision making, insight discovery and optimization while being innovative and cost-effective.

It is expected that the growth of big data is estimated to reach 25 billion by 2015 [3]. From the perspective of the information and communication technology, big data is a robust impetus to the next generation of information technology industries [4], which are broadly built on the third platform, mainly referring to big data, cloud computing, internet of things, and social business. Generally, Data warehouses have been used to manage the large dataset. In this case extracting the precise knowledge from the available big data is a foremostissue.

A. Data Storage and Analysis

In recent years the size of data has grown exponentially by various means such as mobile devices, aerial sensory

technologies, remote sensing, radio frequency identification readers etc. These data are stored on spending much cost whereas they ignored or deleted finally because there is no enough space to store them. Therefore, the first challenge for big data analysis is storage mediums and higher input/output speed. In such cases, the data accessibility must be on thetop priority for the knowledge discovery and representation. The prime reason is being that, it must be accessed easily and promptly for further analysis. In past decades, analyst use hard disk drives to store data but, it slower random input/output performance than sequential input/output. To overcome this limitation, the concept of solid state drive (SSD) and phrase change memory (PCM) was introduced. However the available storage technologies cannot possess the required performance for processing big data.

Another challenge with Big Data analysis is attributed to diversity of data. with the ever growing of datasets, data mining tasks has significantly increased. Additionally data reduction, data selection, feature selection is an essential task especially when dealing with large datasets. This presents an

Unprecedented challenge for researchers. It is because, existing algorithms may not always respond in an adequate time when dealing with these high dimensional data. Automation of this process and developing new machine learning algorithms to ensure consistency is a major challenge in recent years. In addition to all these Clustering of large datasets that help in analyzing the big data is of prime concern. Recent technologies such as hadoop and map Reduce make it possible to collect large amount of semi structured and unstructured data in a reasonable amount of time. The key engineeringchallenge is how to effectively analyze these data for obtainingbetter knowledge. A standard process to this end is to transformthe semi structured or unstructured data into structured data, and then apply data mining algorithms to extract knowledge. Aframework to analyze data was discussed by Das and Kumar. Similarly detail explanation of data analysis for public tweets was also discussed by Das et al in their paper.

The major challenge in this case is to pay more attention for designing storage sytems and to elevate efficient data analysis tool that provide guarantees on the output when the data comes from different sources. Furthermore, design of machinelearning algorithms to analyze data is essential for improving efficiency and scalability.

B. Knowledge Discovery and Computational Complexities

Knowledge discovery and representation is a prime issue in big data. It includes a number of sub fields such as authentication, archiving, management, preservation, information retrieval, and representation. There are several tools for knowledge discovery and representation such as fuzzy set , rough set , soft set, near set, formal concept analysis , principal component analysis etc to name a few. Additionally many hybridized techniques are also developed to process real life problems. All these techniques are problem dependent. Further some of these techniques may not be suitable for large datasets in a sequential computer. At the same time some of the techniques has good characteristics of scalability over parallel computer. Since the size of big data keeps 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 dataset management is data warehouses and data marts. Data warehouse is mainly responsible to store data that are sourced from operational systems whereas data mart is based on a data warehouse and facilitates analysis.

C. Scalability and Visualization of Data

The most important challenge for big data analysis tech- niques is its scalability and security. In the last decades researchers have paid attentions to accelerate data analysis and its speed up processors followed by Moore's Law. For the former, it is necessary to develop sampling, on-line, and mul- tiresolution analysis techniques. Incremental techniques have good scalability property in the aspect of big data analysis. As the data size is scaling much faster than CPU speeds, there is a natural dramatic shift in processor technology being embedded with increasing number of cores. This shift in processors leads to the development of parallel computing. Real time applications like navigation, social networks, finance, internet search, timeliness etc. requires parallel computing.

The objective of visualizing data is to present them more adequately using some techniques of graph theory. Graphical visualization provides the link between data with proper inter- pretation. However, online marketplace like flipkart, amazon, e-bay have millions of users and billions of goods to sold each month. This generates a lot of data. To this end, some company uses a tool Tableau for big data visualization. It has capability to transform large and complex data into intuitive pictures. This help employees of a company to visualize search relevance, monitor latest customer feedback, and their sentiment analysis. However, current big data visualization tools mostly have poor performances in functionalities, scalability, and response in time.

D. Information Security

In big data analysis massive amount of data are correlated, analyzed, and mined for meaningful patterns. All organizations have different policies to safe guard their sensitive information. Preserving sensitive information is a major issue in big data analysis. There is a huge security risk associated with big data. Therefore, information security is becoming a big data analytics problem. Security of big data can be enhanced by using the techniques of authentication, authorization, and en- cryption. Various security measures that big data applications 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 has attracted the attention of information security. Therefore, attention has to be given to develop a multi level security policy model and prevention system.

OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are becoming the research focal point in industries and academia.Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include infor- mation science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. Main focus of this section is to discuss open research issues in big data analytics. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing, bio inspired computing, and quantum computing. However it is not limited to these issues. More research issues related to health care big data can be found in Husing Kuo et al. paper.

A. IoT for Big Data Analytics

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things (IoT). Thus, appliances are becoming the user of the internet, just like humans with the web browsers. Internet of Things is attracting the attention of recent researchers for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information,

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 1, 2025

network and communication technology. The new regulation of future will be eventually, everything will be connected and intelligently controlled. The concept of IoT is becoming more pertinentto the realistic world due to the development of mobile de- vices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety.

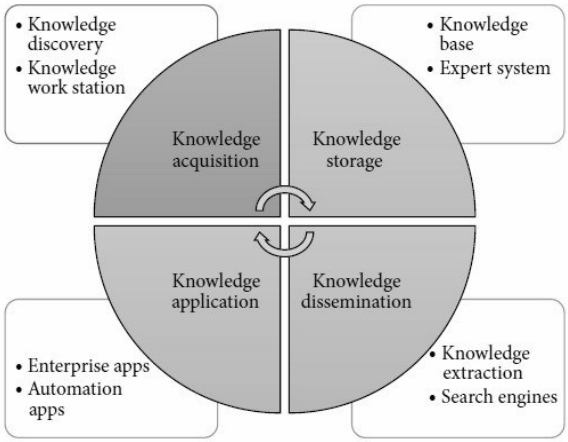

Fig.1 IoT Knowledge Exploration System

B. Cloud Computing for Big Data Analytics

The development of virtualization technologies have made supercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibilityof specification details such as number of processors, diskspace, memory, and operating system. The use of these virtual computers is known as cloud computing which has been oneof the most robust big data technique. Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data. Cloud computing harmonize massive data by on- demand access to configurable computing resources through virtualization techniques. The benefits of utilizing the Cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction. Open challenges and research issues of big dataand cloud computing are discussed in detail by many re- searchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, and resource management. So Cloud computing helpsin developing a business model for all varieties of applicationswith infrastructure and tools.

Big data application using cloud computing should support data analytic and development. The cloud environment should provide tools that allow data scientists and business analysts to interactively and collaboratively explore knowledge acquisition data for further processing and extracting fruitful results. This can help to solve large applications that may arise in various domains. In addition to this, cloud computing should also enable scaling of tools from virtual technologies intonew technologies like spark, R, and other types of big data processing techniques.

C. Bio-inspired Computing for Big Data Analytics

Bio-inspired computing is a technique inspired ny natureto address complex real world problems. Biological systems are self organized without a central control. A bio-inspiredcost 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 computingis 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 are generated from variety of resources across the web since the digitization. Analyzing these dataand categorizing into text, image and video etc will requirelot 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 analyze large and furnish cost effective results.

D. Quantum Computing for Big Data Analysis

A quantum computer has memory that is exponentiallylarger than its physical size and can manipulate an exponentialset of inputs simultaneously. This exponential improve- ment in computer systems might be possible. If a real quantum computer is available now, it could have solved problemsthat are exceptionally difficult on recent computers, of course today's big data problems. The main technical difficulty in building 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 complexvalues to be stored in a classic computer system.

TOOLS FOR BIG DATA PROCESSING

Large numbers of tools are available to process big data. In this section, we discuss some current techniques for analyzing big data with emphasis on three important emerging tools namely MapReduce, Apache Spark, and Storm. Most of the available tools concentrate on batch processing, stream pro- cessing,and interactive analysis. Most batch processing tools are based on the Apache Hadoop infrastructure such as Mahoutand Dryad. Stream data applications are mostly used for real time analytic. Some examples of large scale streaming platformare Strom and Splunk. The interactive analysis process allow users to directly interact in real time for their own analysis.

For example Dremel and Apache Drill are the big data plat- forms that support interactive analysis. These tools help us in developing the big data projects. A fabulous list of big data tools and techniques is also discussed by much researchers [6],. The typical work flow of big data project discussed by Huang et al is highlighted in this section and is depicted.

A. Apache Hadoop and MapReduce

The most established software platform for big data anal- ysis is Apache Hadoop and Mapreduce. It consists of hadoop kernel, mapreduce, hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two

steps such as Map step and Reduce Step. Hadoop works ontwo kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the subproblems in reduce step. Moreover, Hadoop and Map Reduce works asa powerful software framework for solving big data problems.

B. Apache Mahout

Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent dataanalysis applications. Core algorithms of mahout includingclustering, classification, pattern mining, regression, dimen- sionalty 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 avibrant, responsive, diverse community to facilitate discussionson the project and potential use cases. The basic objective of Apache mahout is to provide a tool for elleviating big challenges. The different companies those who have imple- mented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and face book.

C. Apache Spark

Apache spark is an open source big data processing frame- work built for speed processing, and sophisticated analytics.It is easy to use and was originally developed in 2009 in UC Berkeleys AMPLab. It was open sourced in 2010 as an Apacheproject. 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 dataprocessing. Spark runs on top of existing hadoop distributed file system (HDFS) 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 appli- cation on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing inthe form of tasks. Each application will have a set of processescalled 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. Figure 5depicts the architecture diagram of Apache Spark. The various features of Apache Spark are listed below:

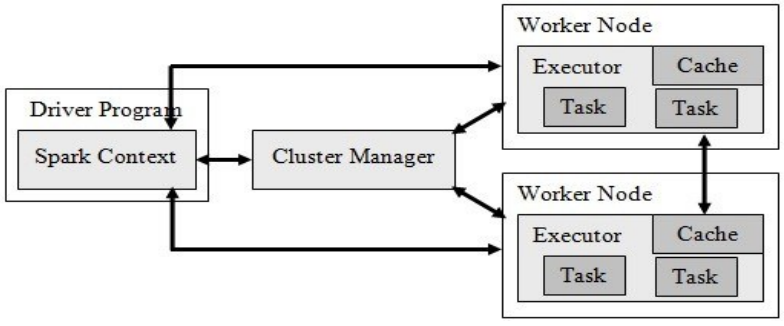

Fig. 2: Architecture of Apache Spark

D. Dryad

It is another popular programming model for implementingparallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and an user use the resources of a computer clusterto run their program in a distributed way. Indeed, a dryaduser 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.

E. Storm

Storm is a distributed and fault tolerant real time com- putation 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 jobsand 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 jobtracker and tasktracker 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.

F. Jaspersoft

The Jaspersoft package is an open source software that produce reports from database columns. It is a scalable big data analytical platform and has a capability of fast data visu- alization on popular storage platforms, including MangoDB, Cassandra, Redis etc. One important property of Jaspersoft is that it can quickly explore big data without extraction, transformation, and loading (ETL). In addition to this, it also have an ability to build powerful hypertext markup language (HTML) reports and dashboards interactively and directly from big data store without ETL requirement. These generated reports can be shared with anyone inside or outside user's organization.

G. Splunk

In recent years a lot of data are generated through machine from business industries. Splunk is a real-time and intelligent platform developed for exploiting machine generated big data.It combines the up-to-the-moment cloud technologies and big data. In turn it helps user to search, monitor, and analyze their machine generated data through web interface. The results are exhibited in an intuitive way such as graphs, reports, and alerts.Splunk is different from other stream processing tools. Its peculiarities include indexing structured, unstructured machine generated data, real-time searching, reporting analytical results,and dashboards. The most important objective of Splunk isto provide metrices for many application, diagnose problems for system and information technology infrastructures, andintelligent support for business operations.

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 analyzed to get useful information. This necessitates

the development of techniques which can be used to facilitate big data analysis. The development of powerful computers isa 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 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 storingit efficiently for future use is another issue. Further, pro- gramming for big data analysis is an important challenging

issue. Expressing data access requirements of applications and designing programming language abstractions to exploit parallelism are an immediate need.

In recent years data are generated at a dramatic pace. Analyzing 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 analyze these big data. From this survey, it is understood that every big data platform hasits 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 belive that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

REFERENCES

[1] M. K.Kakhani, S. Kakhani and S. R.Biradar, *Research issues in big data analytics*, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), 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) (2015), pp.137-144.

[3] C. Lynch, *Big data: How do your data grow?*, Nature, 455 (2008), 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) (2015), pp.59-64.

[5] R. Kitchin, *Big Data, new epistemologies and paradigm shifts*, Big Data Society, 1(1) (2014), 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- mation Sciences, 275 (2014), pp.314-347.

[7] K. Kambatla, G. Kollias, V. Kumar and A. Gram, *Trends in big data analytics*, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.

[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, *On the use of mapreduce for imbalanced big data using random forest*, Information Sciences, 285 (2014), pp.112-137.