

## PROBLEMS IN DATA ANALYTICS AND ITS SOLUTIONS

<sup>1</sup>Sayantana Talapatra, <sup>2</sup>Nazeef Ahmed, <sup>1</sup>Soham Chakraborty, <sup>1</sup>Soham Roy, <sup>3</sup>Ayan Basu,  
<sup>2</sup>Arindit Guha Sinha

<sup>1</sup>*Department of Computer Science & Engineering  
Institute of Engineering & Management, Kolkata-700091*

*Email: sayantan.officialpurp@gmail.com*

<sup>2</sup>*Department of Mechanical Engineering  
Institute of Engineering & Management, Kolkata-700091*

*Email: nazeefahmed14@gmail.com*

<sup>3</sup>*Department of Information Technology  
Institute of Engineering & Management, Kolkata-700091*

*Email: basu\_ayan@yahoo.in*

### Abstract

In the information era, a huge amount of terabytes is generated from our day to day lives. This mainly comes from our modern information system and advanced digital technologies such as internet of this and cloud computing. Since the data are huge the analysis of these data requires a large amount of efforts at multiple levels and for multiple purposes. Therefore to cope up with this big data analysis is a current area of research. Due to rapid growth of such data solutions need to be studied and found. The ones taking the decisions also called as decision makers need to have a detailed insights about the data they are handling otherwise how can they come up with correct and best decisions, which can only be provided using data analytics. As a result this article provides a platform or rather a solution or we can also call it a brief analysis of big data challenges open research issues and also some of the different methods and tools which can be used for data analytics. It also opens a new horizon for the researchers to come up with the solution.

**Keywords:** *Data analytics, Big Data analytics, Quantum computing, Tools for big data analytics, IOT, Cloud Computing, Apache Hadoop.*

### INTRODUCTION

Data helps us improve processes. It helps us understand and improve business processes so that we can reduce wasted money and time. Every company feels the effects of waste. It uses up resources that could be better spent on other things, squanders people's time and ultimately impacts your bottom line. Data analytics help in analyzing the value chain of business and gain insights. The use of analytics can enhance the industry knowledge of the analysts. Data analytics experts provide the organizations a chance to learn about the opportunities for the business.

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping business operate more effectively and efficiently.

Every second huge amounts of data is created from modern technologies like Internet of things and cloud computing. Due to the growth of big data we can analyze and break into many fields utilizing this collection of this enormous datasets. The prime motive of big data analysis is to process data of huge volume, velocity, variety and veracity using various traditional and computational intelligent techniques. Volume is the large amount of data that is generated everyday, velocity is the rate of growth and how fast the data are collected for analysis. Variety refers to the types of data such as structured, unstructured, semi-structure, etc.

### **CHALLENGES IN BIG DATA ANALYTICS**

In recent times big data has been accumulated in several sectors like health care, public administration, retail bio chemistry, and in other scientific researches.

web based applications encounter big data frequently, for example social computing internet text and documents and internet search indexing. Social computing basically includes things like social network analysis, online communities, recommender system, reputation system, and prediction markets. On the other hand internet search indexing includes ISI,IEEE Xplorer, Scopus, Thompson Reuters etc. If we consider this as an advantage of big data it provides a lot of new opportunities in the knowledge processing task. However we know that one cannot get something good without any hardships or rather challenges.

Now, to handle the challenges we need to have the knowledge of various computational complexities, information security, and computational method to analyze huge amount of data. For example, many statistical methods which work for small amount of data won't work with big data. The various challenges that the health sectors face was already being researched by many researchers.

Here we are classifying the problem into four different categories which are data storage and analysis; knowledge discovery and computational complexities; scalability and visualization of data; and finally information security. These are the issues which are discussed briefly in the following sub sections.

## **DATA STORAGE AND ANALYSIS:**

The amount of data required for various uses, these days, has been growing at a very high rate. Various technological advancements have made it almost a necessity to have resort to high amount of data. But there is always a problem. the loss of data ultimately due to the lack of space. Therefore storage and high speed storage facilities are the main priorities and challenge to be undertaken, nowadays. Another factor to be taken under consideration is the different types of datasets and tasks related to it. This leaves the researchers with a challenge, the challenge of handling the data by reducing, selecting and dealing with high amount of datasets. Structured data storage and retrieval of the same can take place using the methods of data marts, data warehouses and databases. Developing new algorithms to suit the large amount of data and analysis of the same is of utmost importance. So, storage system designs with guaranteed output and machine learning (even with iot) algorithms for the analysis of data for increasing efficiency and reach. SSD or solid state drive and PCM has been introduced but the essential performance for analysing and processing such huge amount of data is not present in the available technologies.

### **Scalability and Visualisation of Data:**

Security and scalability of such huge amounts of data has always been a challenge of utmost importance in front of data analytics researchers. hence, they have adapted to accelerate big data analysis followed by Moore's law(1). Thus, there are shifts in processor technology with the increasing number of cores giving rise to parallel computing. Parallel computing is required for various processes like the financial purposes, social networks, navigation, etc. Now, visualization of data is also important so as to represent the data efficiently by using graphical techniques. Some big companies like amazon and flipkart who have to deal with a huge amount of data use a tool called Tableau for the visualization of huge amounts of data. Thus, as we proceed more into the depths of data analytics we come to see that big data leads to IOT, Machine learning, cloud computing, parallel computing and so much more .....

Big data analytics already have many problems one of the most important problems among them is its scalability and security. previously researchers have paid attentions to accelerate data analysis and its speed up processors which was followed by Moores law. Now, to accelerate data analysis we need to develop sampling, on line, and multiresolution 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, this leads to a dramatic shift in in processor technology being embedded with increasing no. of cores. This shift leads to the development of parallel computing. Parallel computing is basically used in application like navigation, social networks, finance, internet search, timelines etc.

The main objective behind visualizing data is to present them in a better way or to present them adequately using some techniques of graph theory. Graphical representations or rather visualizations provides us with the link between data with proper interpretation.

Many e-commerce sites like flipkart amazon snapdeal etc, have millions of users and billions of goods to be sold each and every month. A lot of data is generated from this. A tool named tableau is used by some companies for big data visualization. This tool has the capability to transform large and complex data into intuitive pictures. This tool really help employees of a company to visualize search relevance, monitor latest customer feedback, and their sentiment analysis. However, current big data visualizations tools mostly have poor performance in all aspects.

After many studies and research we have found that big data have produced many challenges for the developments of both hardware and software which leads to parallel computing , cloud computing , distributed computing, visualisation process, scalability. Now, to overcome this issue, we need to correlate more mathematical models to computer science.

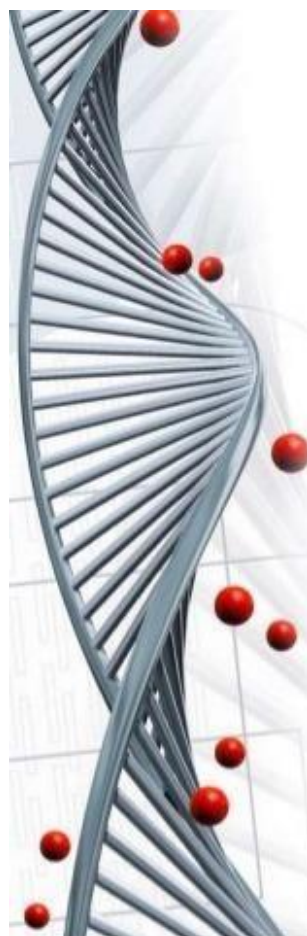
### **INFORMATION SECURITY:**

One of the biggest problems faced in data analytics is the preservation and safeguarding of sensitive information . This literally means that there is always a risk in handling huge amount of data . So , methods like encryption, authorisation or authentication can be involved (1). This has given way to a well known term call INFORMATION SECURITY of big data . So, there should be serious multi-se curity in order to preserve the privacy stored in the big data .

### **SOLUTIONS FOR PROBLEMS IN DATA ANALYTICS:**

#### **Bio Inspired Computing for Big Data Analytics:**

Bio inspired computing short for biologically inspired computing is a field of study which seeks to solve computer science problems using models of Biology . It relates to connectionism, social behaviour, and emergence. Supercomputing have been made much easier and affordable with the development of virtualization technologies. The huge amount of data that needs computing are generated from variety of resources mainly across the web, this has been happening since the digitization. An intelligent analytics needs to be done by the data scientists and big data professionals, for the analysation of these data s into text, image and video, etc. A rapid increase in technologies is leading to the emerging of things like IOT, cloud computing, bio inspired computing etc. Whereas to bring an equilibrium of data a right platform is very important to be chosen for its analytics.



## Why Bio Computing ??

- Moore's Law states that silicon microprocessor complexity will double in every 18 months.
- One day this will no longer hold true when miniaturization limits are reached.
- Solving complex problems which today's supercomputers are unable to perform in stipulated period of time.
- Require a Successor to Silicon

ABDULLAH FARHAD

Bio-inspired computing techniques the term may seem to be unheard to some but basically it plays a key role in intelligent data analysis and its application to big data. Among the many advantages it has the most advantageous is its simplicity and their rapid convergence to optimal solution mainly when solving service provision problems. Many discussions have already taken place regarding this bio computing and from all these discussions we can conclude or rather our observation is that bio-inspired computing models provide smarter interactions, inevitable data losses and this also helps in handling ambiguities.

### **Quantum computing for big data analytics:**

Quantum computing basically takes advantage of the ability which is strange in subatomic particles that is at a time it can exist in more than one state. Due to this way that these particles of tiny size behave operations that we want to perform can be done in a lesser amount of time and with minimum energy consumption.

Now, in classical computing we all know a bit is a single piece of information that can store either 1 or 0. Whereas, quantum computing uses 'qubits' or quantum bits. But the main difference between normal bit and quantum bit is that it can store much more than normal bit. The main reason behind this is they can exist in any state and at any time.

Now a question may arise in everyone's mind how is this related to big data analytics. From the above paragraph we all are clear that quantum computers will be able to do large calculations in very small amount of time. There are some calculation which today's computers can't do or rather it would take them years to do that. If we use quantum computers it will enable the organizations to deal with large amount of valuable data with ease.

The main features or advantages of quantum computing are it allows for quick detection, data analysis, integration and diagnosis from large amount of scattered data sets.



Quantum computing without any doubt will surely bring in a change in our IT architecture, and even corporate architecture. Quantum computing is already in use in companies like Google and Intel and it is giving them very positive feedback, hence seeing quantum computing used by various other companies is not far away.

### **Iot And Machine Learning For Big Data Analytics:**

Nowadays one of the most spoken about topic in data handling is the IOT. In a nutshell, IOT is the connection of a device to the Internet along with other connected devices. It makes the collection and sharing of data being converted to a giant network(2). IOT is becoming more eye catching due to the use of hand held devices like smart phones, cloud computing and data analytics

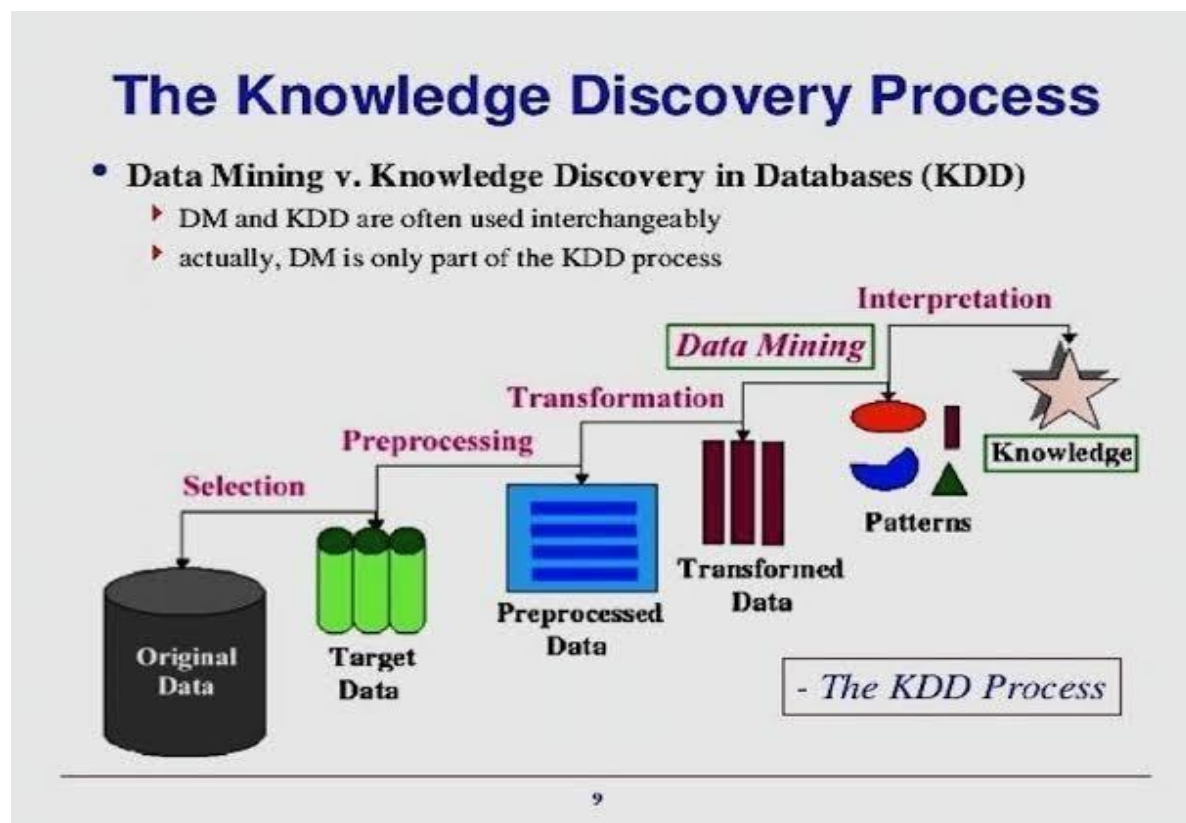
Acquisition of the knowledge from IOT is a big challenge as analysis of IOT data requires the development of infrastructure . Machine learning techniques are needed to be applied to extract a specific data from the continuous data from the IOT device . As the name suggests, machine learning is the tool that can be used to provide a system with the ability to automatically learn and improve without being explicitly programmed . The IOT market is expected to reach a value of around 6.1 billion US dollars by 2024 at a CAGR of 31.8 % (3). Thus , the IOT is in a favourable position to raise benefits to the field of science. One component in the growth of this sector is the IOT data analytics. IOT data analytics is the analysis of great amounts of data generated by devices in the network. As time is progressing , the cost to store data is going down . This creates favourable circumstances for companies to start investing on IOT data analytics . Now, everyday , huge amount of devices are sharing data through sensors on the internet . But unless this data is analysed and stored, it is not effective. Thus IOT and Machine learning go hand in hand with data analytics

### **Cloud Computing For Data Analysis:**

Cloud computing is the recent practice of using a network of servers on the internet to store , handle and operate huge amounts of data and hence can play a huge role in data analytics . This has made supercomputing cheaper and reachable . Cloud platforms make use of virtualisation techniques . The main benefit lies in the fact that one can make available a certain amount of data in the cloud platform and when there is demand of that data they pay for the resources which is required to initiate the completion of the product . This can be used to resolve the problems concerning various domains . The main advantage that cloud computing provides is storage of huge amounts of data which has to be uploaded as well as downloaded in the cloud platform . Many people define cloud computing to be synonymous to Infrastructure-as-a-service where accessing databases, servers and other things are available . Thus, this saves the companies from keeping such huge amounts of data all at once in their respective premises(4). So, data scalability becomes much faster as data sets tend to increase. Moreover , cloud computing and IOT work hand in hand making big data analytics grow in recent years.



## Knowledge discovery



in big data the main issue is knowledge discovery and representation which contains several sub-topics like authentication, archiving, management, preservation of data, recovery of data.

there are several techniques for knowledge discovery and representation like soft set, near set, formal concept analysis, rough set, fuzzy set etc which are used to solve different problems. Each problem has its own technique. and sometimes hybrid techniques are created to solve the problem, but in case of large data sets some time the techniques fail. the size of the data is constantly increasing and the techniques that are available to us might fail to process them the most common approach in case of large data is data warehouses and data marts. data warehouse is mainly used for the storing the data and data mart is based on data warehouse and facilitates analysis. The checking of the huge data sets require complex methods. The analysis of large amount of data leads to computational complexities due to several as there are several issues like inconsistency and uncertainty, generally the systematic modeling of the comprehensive mathematical system is done, it may face some difficulties to compute such large amount of data the main concept is to minimize the cost of processing the data. as we have tools of low specifications the computation

of large amount of data is hard, as it faces many challenges. which in turn can help us to develop a technique that can solve the computational complexities, uncertainty, inconsistency and many more in an effective way



## TOOLS FOR BIG DATA

Here to control the huge amount of data some emerging tools or techniques are there known as MapReduce, Apache Spark and Storm. Majority of the tools follow the process of batch processing which is based on the Apache Hadoop infrastructure like Mahout and Dryad. In case of the stream data application the process that is used known as real time analytics. The examples of large scale streaming platform are Storm and Splunk. When the users are allowed to interact with their analysis directly that process is known as interactive analysis.

### Apache Hadoop And Mapreduce :

The most successful process are Apache Hadoop and MapReduce in case of big data analysis, it consists of Hadoop kernel, MapReduce, Hadoop distributed file system (HDFS) and Apache Hive etc. To process large amount of data sets which is based on divide and conquer method. MapReduce is used map and reduce steps are programmed into divide and conquer method. The Hadoop works on two types of nodes master and worker node. The worker nodes distribute the smaller sub problems in the map sets, and the master nodes divide the input into smaller sub problems and it also combines the outputs for all subproblems in reduce step.

### Apache Mahout

[5] Apache Mahout is one of the main tools it is a scalable machine learning algorithms which is mainly focuses on the linear algebra to analyse the data. It is a project which is made by the Apache Software Foundation. The core algorithms of Apache Mahout is the pattern mining, regression, dimensionality reduction, evolutionary algorithms and many more. The main goal is to solve every problem and challenges, the companies who are using scalable machine learning are Google, Yahoo, IBM, Flipkart, etc

### Dryad

Another famous programming model which is used for handling large context based on dataflow graph. The programming model consists of a cluster of computing nodes and an user who will use the resource of that cluster to run the program in a distributed way. The machine consists of multiple cores or processors and the user need to use several of those machines. But here the user does not need to know or have any knowledge about the concurrent programming, Dryad has lots of functions like job graphs, scheduling of machines for available processes and transition failure handling in cluster and many more. Dryad was originally built upon open source DSpace repository software, it was developed in Massachusetts Institute of Technology.

### Apache Drill

Apache Drill is an interactive analysis of big data which is flexible as it supports many types of query language, data format, data source. Its speciality is to exploit nested data. The Apache Drill

has the capability to process petabytes and go through trillions of records in seconds, the map reduce process is used by it to do the batch analysis and for storage it uses HDFS.

## **CONCLUSION:**

In the recent years data is being produced in huge amounts and at a fast pace. Analyzing these huge amounts of data is too tough for a normal human being. In this paper we survey the various problems faced during research, the challenges faced and the various methods and tools used to analyze the big data. From our research, we understand that every big data platform has its individual focus. Each big data platform has a specific function and task to perform. Different methods 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 the coming days researchers will pay more attention to solve problems of big data effectively and efficiently. Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. In the future we will see the rapid application of big data analytics in the healthcare industry. As big data analytics becomes more mainstream, issues such as safeguarding security, establishing standards and continually improving their tools will gather attention. Big data analytics are still in the process of development but too many rapid advances in platforms and tools can accelerate their maturing process.

## **REFERENCES:**

1. A survey on big data analytics by Debi Prasanna Acharya
2. IBM.com
3. www.itransition.com-Sandra Khvoynitskaya
4. Quora-Mike Chan
5. Grant Ingersoll- Introducing Apache Mahout.
6. Raghupathi W: Data Mining in Health Care. Healthcare Informatics: Improving Efficiency and Productivity. Edited by: Kudyba S.2010, Taylor and Francis, 211-223.
7. Burghard C: Big Data and Analytics Key to Accountable Care Success.2012, IDC Health Insights
8. Large Scale Distributed Data Science using Apache Spark. Authors: James G.Shanahan, Laing Dal
9. Acharya D.P., Kauser Ahmed P., A Survey On Big Data Analytics: Challenges, Open Research Issues And Tools.