A Survey on Big Data Bio-informatics

R. Samya

Department of Computer Science
Periyar University
Salem, India
samyacs93@gmail.com

***Abstract*- With the advent of Internet of Things (IoT) and Web 2.0 technologies, there has been a tremendous growth in the amount of data generated. This paper emphasizes on the need for big data, technological advancements, tools and techniques used to process big data are discussed. Since, the traditional technologies like Relational Database Management System (RDBMS) have their own limitations to handle big data, new technologies have been developed to handle them and to derive useful insights. The availability of Big Data, coupled with new data analytics, challenges established epistemologies across the sciences, social sciences and humanities, and assesses the extent to which they are engendering paradigm shifts across multiple disciplines. However, the high performance computing devices and software tools to deal with this complex and increased volume of data is still persists as a big challenge among the computer scientists and biologists. The basic objective of this paper is to study various analysis and visualization tools in bioinformatics, bioinformatics big databases and high level bioinformatics system architecture to handle the voluminous data in bioinformatics.**

Keywords: Big Data, Epistemologies, computer scientists, bioinformatics, voluminous data

1. INTRODUCTION

The source of data is generated by fast transition of digital technologies which lead to growth of big data. The collection of large datasets from these devices is difficult to process using traditional database management tools or data processing applications. Generally these data are in pet bytes and beyond. They may be structured, unstructured or semi structured. The Volume, Velocity and Variety which are referred to the 3V's formally define such data. 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 (A. Jacobs, 2009). Variety provides information about the types of data such as structured, unstructured, semi structured etc. The fourth V which also defines it 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. The four V's of Big Data is represented in the figure 1.

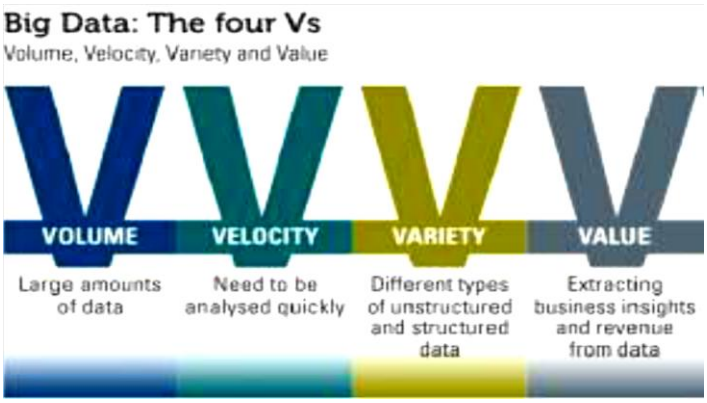

Figure 1. Four V's of Big Data

Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data.

- Data Storage and Analysis.
- Knowledge discovery and Computational.
- Complexity.
- Scalability and Visualization of data Information Security.

It is expected that the growth of big data is estimated to reach 25 billion by 2015. From the perspective of the information and communication technology, big data is a robust motivation to the next generation of information technology industries, which are broadly built on the third platform, mainly referring to big data, cloud computing, internet of things, and social business. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such as data mining and statistical analysis. These challenges generally arise when we wish to perform knowledge discovery and representation for its practical applications. A fundamental problem is how to quantitatively describe the essential characteristics of big data. Additionally, the study on complexity theory of big data will help understand essential characteristics and formation
of complex patterns in big data, simplify its representation, gets better knowledge abstraction, and guide the design of computing models and algorithms on big data. Much research was carried out by various researchers on big data and its trends (Lynch, 2008).

Further, the paper is organized as follows: Section 2 describes Related Study. Section 3 explains the research issues in big data analytics. Section 4 describes Big data Chain Value and finally section 5 concludes the entire research work of this paper.

2. AN OVERVIEW OF BIG DATA ENVIRONMENT

2.1 Emerging Technologies for Big Data Analytics

New technologies are emerging to make unstructured data analytics possible and cost-efficient. The new approach redefines the way data is managed and analyzed by leveraging the power of a distributed grid of computing resources. It utilizes easily scalable “shared nothing” architecture, distributed processing frameworks, and non relational and parallel relational databases (L. Wang and J. Shen, 2013).
• Analytics applications architecture:- New data processing systems make the computing grid work by managing and pushing the data out to individual nodes, sending instructions to the networked servers to work in parallel, collecting individual results and then reassembling then to produce meaningful results. Processing the data where it resides is faster and more efficient then first transporting it to a centralized system.
• Data architecture: - To handle the variety and complexity of unstructured data, databases are shifting from relational databases to non relational. Unlike the orderly world of relational databases, which are structured normalized, and densely populated, non relational databases are scalable, network oriented, semi structured, and sparsely populated. NOSQL database solutions do not require fixed table schemas, avoid join operations, and scale horizontally.

2.2 Distributed Frameworks

The Emergence of Apache Hadoop Apache Hadoop is evolving as the best new approach to unstructured data analytics. Hadoop is an open-source framework that uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms and an application layer that manages distributed processing, parallel computation, and workflow configuration management. In addition to offering high availability, Hadoop is more cost efficient for handling large unstructured data sets than conventional approaches, and it offers massive scalability and speed (Kumar, 2103). The entire Apache Hadoop platform is now commonly considered to consist of the Hadoop kernel, MapReduce and Hadoop Distributed File System (HDFS), as well as a number of related projects including Apache Hive, Apache HBase, and others.

2.3 HBase

HBase is an open source, non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File system), providing BigTable-like capabilities for Hadoop. That is, it provides a fault tolerant way of storing large quantities of sparse data. HBase is not a direct replacement for a classic SQL database, although recently its performance has improved, and it is now serving several data-driven websites, including Facebook's Messaging Platform. Use Apache HBase when you need random, real time read/write access to your Big Data (X. Jin, 2015). This project's goal is the hosting of very large tables -- billions of rows X millions of columns -- atop clusters of commodity hardware. Apache HBase is an open-source, distributed, versioned, column-oriented store modeled after Google's Bigtable: A Distributed Storage System for Structured Data by Chang et al. Just as Bigtable leverages the distributed data storage provided by the Google File System, Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.

2..3.1. Features of HBase

- Linear and modular scalability.
- Strictly consistent reads and writes.
- Automatic and configurable shading of tables
- Automatic failover support between Region Servers.
- Convenient base classes for backing Hadoop MapReduce jobs with Apache HBase tables.
- Easy to use Java API for client access.
- Block cache and Bloom Filters for real-time queries.
- Query predicate push down via server side Filters
- Thrift gateway and a REST-ful Web service that supports XML, Protobuf and binary data encoding options
- Extensible jruby-based (JIRB) shell
- Support for exporting metrics via the Hadoop metrics subsystem to files or Ganglia; or via JMX

3. TECHNOLOGIES FOR 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 information science, uncertainty modelling, 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 (R. Kitchin, 2014). Main focus of this section is to discuss open research issues in big data analytics. The challenges in bio-informatics focus on Flood of Data getting visible in genomics, the only one DNA data flood is coming from more than 2500 sequencing tools around the world. Data and Tools need to be more than close. So that data obtained from one tool can be used by another tool easily. Biologists are not computer specialists. Computer experts need to intervene to install and make it functional on behalf of the biologist. A crucial challenge is to combine the capabilities of data resources and tools to create a data exploration and analysis atmosphere that does fairness to the diversity and complexity of biological system data sets. Computing, not sequencing is now the slower and more costly aspect of genomics research. The research issues pertaining to big data analysis are classified into three broad categories namely bio inspired computing, IoT for Big Data Analytics, Cloud Computing and quantum computing.

3.1 Bio inspired computing for Big Data Analytics

Bio inspired computing is a technique inspired my nature to address complex real world problems. Biological systems are self organized without a central control. A bio-inspired cost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance (Z. Huang, 1997). 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 data and

categorizing into text, image and video, etc. will require lot of intelligent analytics from data scientists and big data professionals. Proliferations of technologies are emerging like big data, IoT, cloud computing, bio-inspired computing etc. Whereas equilibrium of data can be done only by selecting right platform to analyze large and furnish cost effective results. Bio-inspired computing techniques serve as a key role in intelligent data analysis and its application to big data. These algorithms help in performing data mining for large datasets algorithms help in performing data mining for large datasets due to its optimization application. The most advantage is its simplicity and their rapid convergence to optimal solution while solving service provision problems.

3.2 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. The new regulation of future will be eventually; everything will be connected and intelligently controlled (K. Kambatla, 2014). The concept of IoT is becoming more pertinent to 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. In a broader sense, just like the internet, Internet of Things enables the devices to exist in a myriad of places and facilitates applications ranging from trivial to the crucial. Several diversified technologies such as computational intelligence, and big data can be incorporated together to improve the data management and knowledge discovery of large scale automation applications. Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing. Therefore, it is essential to develop infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and the re-searchers can develop tools to extract meaningful information from these data using machine learning techniques. Under-standing these streams of data generated from IoT devices and analyzing them to get meaningful information is a challenging issue and it leads to big data analytics. Machine learning algorithms and computational intelligence techniques is the only solution to handle big data from IoT prospective. Key technologies that are associated with IoT are also discussed in many research papers. Figure 2 depicts an overview of IoT big data and knowledge discovery process.

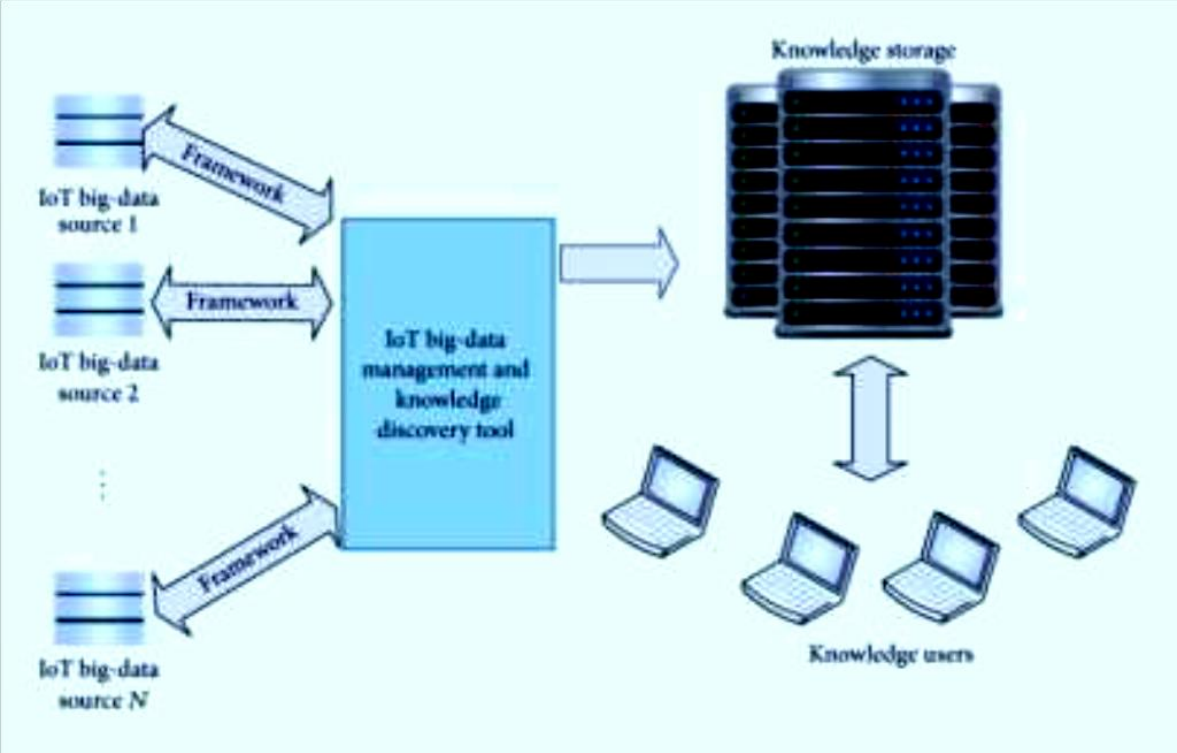

Figure 2. IoT Big Data Knowledge Discovery

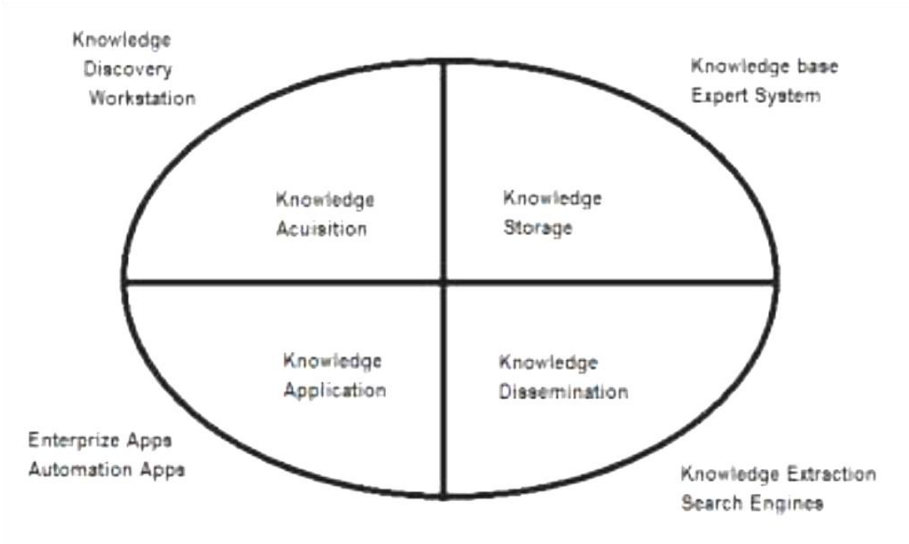

Figure 3. IoT Knowledge Exploration Syste*m*

3.3 Cloud Computing for Big Data Analytics

The development of virtualization technologies has 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 flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been one of the most robust big data techniques. 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 data and cloud computing are discussed in detail by many researchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, and resource management . So Cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools. Big data application using cloud computing should support data analytic and development.Big data forms a framework for discussing cloud computing options. Depending on special need, user can go to the Market place and buy infrastructure services from cloud service providers such as Google, Amazon, IBM, software as a service (SaaS) from a who le crew of companies such as NetSuite, Cloud9, Job science, etc. Another advantage of cloud computing is cloud storage which provides a possible way for storing big data. The obvious one is the time and cost that are needed to upload and download big data in the cloud environment (X. Jin, 2015). Else, it becomes difficult to control the distribution of computation and the underlying hardware. But, the major issues are privacy concerns relating to the hosting of data on public servers, and the storage of data from human studies. All these issues will take big data and cloud computing to a high level of development.

3.4 Quantum Computing for Big Data Analysis

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously. This exponential improvement in computer systems might be possible. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computers and 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. For example, 100 qubits in quantum systems require 2100 complex values to be stored in a classic computer system. It means that many big data problems can be solved much faster by larger scale quantum computers compared with classical computers. Hence it is a challenge for this generation to build a quantum computer and facilitate quantum computing to solve big data problems.

4. BIG DATA VALUE CHAIN

Value Chain, the concept introduced by Porter (1980), refers to a set of activities performed by a firm to add value at each step of delivering a product/service to its customers. In a similar way, data value chain refers to the framework that deals with a set of activities to create value from available data (L. Wang and J. Shen, 2013). It can be divided into seven phases: data generation, data collection, data transmission, data pre-processing, data storage, data analysis and decision making.

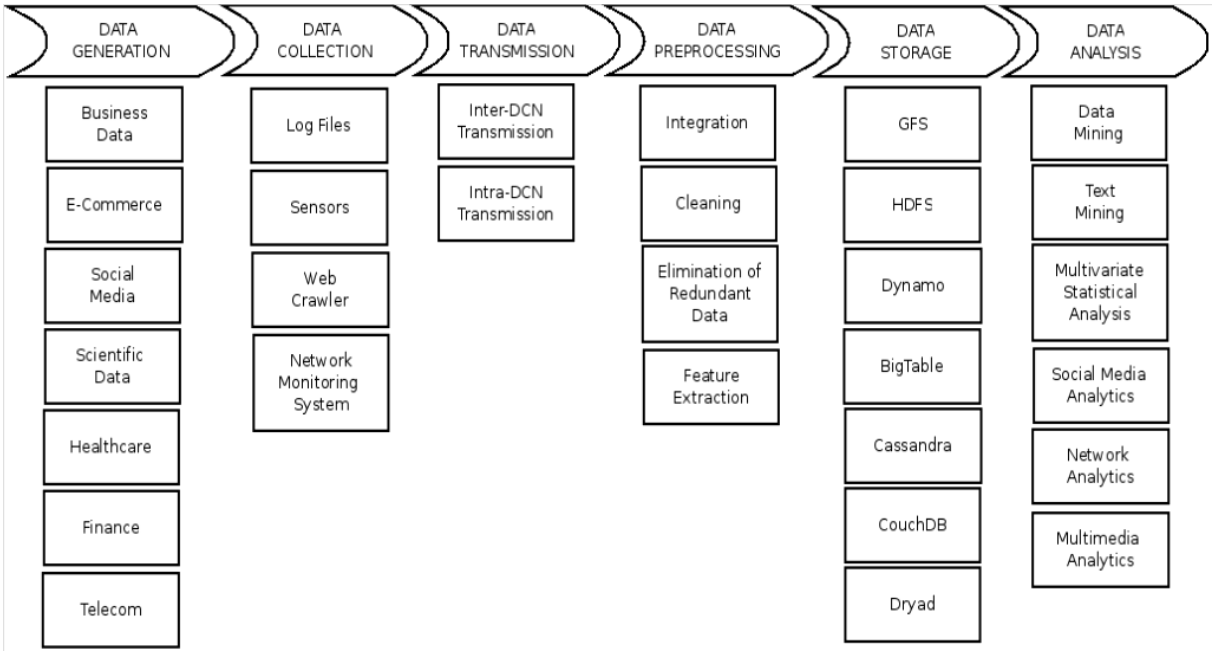

Figure 4. Big Data Value Chain

4. 1. Data Generation

The first and foremost step the big data value chain is the generation of data. As discussed in the previous section, data is generated from various sources that include data from Call Detail Records (CDR), blogs, Tweets and Face book Page.

4. 2. Data Collection

In this phase, the data is obtained from all possible data sources. For instance, in order to predict the customer churn in Telecom, data can be obtained from CDRs and opinions/complaints of the customers on Social Networking Sites such as Twitter (in the form of tweets) and Face book (opinions shared on the company's Face book page). The most commonly used methods are log files, sensors, web crawlers and network monitoring software.

4. 3. Data Transmission

Once the data is collected, it is transferred to data storage and processing infrastructure for further processing and analysis. It can be carried out in two phases: Inter-Dynamic Circuit Network (DCN) transmission and Intra-DCN transmissions. Inter-DCN transmission deals with the transfer of data from the data source to the data

centre while the latter helps in the transfer within the data centre. Apart from storage of data, data centre helps in collecting, organizing and managing data.

4. 4. Data Pre-processing

The data collected from various data sources may be redundant, noisy and inconsistent, hence, in this phase; the data is pre-processed to improve the data quality required for analysis. This also helps to improve the accuracy of the analysis and reduce the storage expenses. The data can be pre-processed with the help of following steps:

a. Integration: The data from various sources are combined to provide a unified and uniform view of the available data. Data federation and data warehousing are the two commonly used traditional methods. Data warehousing executes the Extract, Transform, and Load (ETL) process. During extract process, the data is selected, collected, processed and analyzed. The process of converting the extracted data to a standard format is called Transformation process. In Loading, the extracted and transformed data is imported into a storage infrastructure. In order to make data integration dynamic, data can be aggregated from various data sources using a virtual database. It does not contain any data but the details regarding the information related to original data or metadata can be obtained.
b. Cleaning: The data is checked for accuracy, completeness and consistency. During this process, the data may be deleted and modified to improve the data quality. The general process followed includes following five processes: error types are defined and determined, errors are identified from the data, errors are corrected, error types and corresponding examples are documented, and data entry procedure may be modified to avoid future errors.
c. Elimination of Redundant Data: Many datasets have surplus data or data repetitions and are known as data redundancy. This increases the storage cost, leads to data inconsistency and affects the quality of data. In order to overcome this, various data reduction methods such as data filtering and compression, are used. The limitation of these data reduction techniques is that they increase the computational cost. Hence, a cost-benefit analysis should be carried before using data reduction techniques.

4. 5. Data Storage

The big data storage systems should provide reliable storage space and powerful access to the data. The distributed storage systems for big data should consider factors like consistency (C), availability (A) and partition tolerance (P). According to the CAP theory proposed by Brewer (2000), the distributed storage systems could meet two requirements simultaneously, that is, either consistency and availability or availability and partition tolerance or consistency and partition tolerance but not all requirements simultaneously. Considerable research is still going on in the area of big data storage mechanism. Little advancement in this respect is Google File System (GFS), Dynamo, BigTable, Cassandra, CouchDB, and Dryad.

4. 6. Data Analysis

Once the data is collected, transformed and stored, the next process is data exploitation or data analysis, which is enumerated using the following steps:

a. Define Metrics: Based on the collected and transformed data, a set of metrics is defined for a particular problem. For instance, to identify a potential customer who is going to churn out, a number of times he/she contacted (be it through a voice call, tweets or complaints on Face book page) can be considered.
b. Select architecture based on analysis type: Based on the timeliness of analysis to be carried out, suitable architecture is selected. Real-time analysis is used in the domain where the data keeps on changing constantly and there is a need for rapid analysis to take actions. Memory-based computations and parallel processing systems are the existing architectures. Fraud detection in retail sectors and telecom fraud are the examples of real-time analysis. The applications that do not require high response time is carried out using offline analysis. The data can be extracted, stored and analyzed relatively later in time. Generally used architecture is Hadoop platform.
c. Selection of appropriate algorithms and tools: One of the most important steps of data analysis is selection of appropriate techniques for data analysis. Few traditional data analysis techniques like cluster analysis,

regression analysis and data mining algorithms, still hold good for big data analytics. Cluster analysis is an unsupervised technique that group's objects based on some features. Data mining techniques help to extract unknown, hidden and useful information from a huge data set. The 10 most powerful data mining algorithms were shortlisted and discussed. Various tools are available for data analysis including open source softwares and commercial softwares. Few examples of open source softwares are R for data mining and visualization, Weka/Pentaho for machine learning and RapidMiner for machine learning and predictive analysis.

d. Data Visualization: The need for inspecting details at multiple scales and minute details gave rise to data visualization. Visual interfaces along with statistical analyzes and related context help to identify patterns in large data over time. Visual Analytics (VA) is defined as "the science of analytical reasoning facilitated by visual interactive interfaces". Few visualization tools are Tableau, QlikView, Spotfire, JMP, Jaspersoft, Visual Analytics, Centrifuge, Visual Mining and Board. A comparison of visualization tools based on their data handling functionality, analysis methods and visualization techniques.

4. 7. Decision Making

Based on the analysis and the visualized results, the decision makers can decide whether and how to reward a positive behavior and change a negative one. The details of a particular problem can be analyzed to understand the causes of the problems take informed decisions and plan for necessary actions. Having discussed about how value can be extracted from big data, an industry regardless of sector should consider three criteria before implementing big data analytics: can useful information be obtained in addition to those obtained from the existing systems, will there be any improvement in the accuracy of information obtained using big data analytics and finally, will implementation of big data analytics help in improving the timeliness of response.

5. TOOLS AND APPLICATION OF BIG DATA BIO-INFORMATICS

Computer Science and application is a versatile discipline which has crept into almost every other discipline and now research in specific area has turned into interdisciplinary research. There are variety of core and derived areas in which the bioinformatics are being applied for research and development. These are as follows

Table 1: List of Bio-informatics Tools, Applications and URL

Tool	**Application**	**URL**
Hydra	Framework to support mass spectroscopy data Analysis	http://people.ucalgary.ca/~dschriem/ (Vinod Kumar, 2016)
Chaos game	Genome sequence	https://github.com/usm/usm.github.com/wiki
Representation based tool	alignments were performed using parallelised approach to obtain longest similar Sequence	
BioPig	Hadoop toolkit to analyse large-scale sequence data	https://github.com/JGI-Bioinformatics/biopig
BlueSNP	Hadoop-based R package to support genome-wide association studies	https://github.com/ibm-bioinformatics/BlueSNP
Rainbow	Used for clustering and assembling RAD-seq data	http://www.mybiosoftware.com/rainbow-v2-0-3-clustering-assembling-short-reads-rad.html
PaRFR - parallel random forest regression on Hadoop	PaRFR supports the identification of longitudinal changes in human brain structure developed as a result of Alzheimer's disease using genome-wide studies	http://wwwf.imperial.ac.uk/~gmontana/parfr.htm

Tool	Application	URL
Mercury	Genome sequence analyser deployed in Cloud	https://www.hgsc.bcm.edu/software/mercury
ADAM	Big data platform for genomic data processing	http://bdgenomics.org/projects/adam/
The University of Tokyo DPC Project	A user-friendly tool to manage health services and clinical research data	https://github.com/hiromasah/charsiu
Cloudwave	Hadoop infrastructure for distributed processing from electrophysiological recordings for epilepsy clinical research	http://prism.case.edu/prism/index.php/Cloudwave
BINDSURF	GPU framework to detect protein binding sites in Ligand	http://omictools.com/bindsurf-s5620.html
Corbi	A software for simulation of biomolecular networks in R	http://doc.aporc.org/wiki/Corbi
cGRNB	A web server for generating combinatory gene	https://omictools.com/combinatorial-gene-regulatory-networks-builder-tool

Table 2: Fields of Big Data Bio-Informatics

Application Field	Description
Biotechnology	A broad discipline for developing new technologies, New tools and products using biological processes, organisms, cells or cellular components.
Molecular Medicine	The understanding of the molecular mechanisms of disease, infections empowers better handlings, treatments and even defensive tests to be developed. (Kalyan Nagaraj, 2018)
Bio-Weapon	Biologists have been successful in creating virus poliomyelitis entirely with the artificial means.
Alternative Energy Sources	Genome of the microbe Chlorobium Tedium is capable of generating the energy form the light. Biologists very much interested and studying about this microbe as a source of alternate energy.
Development of Drought Resistance Varieties	To create crop assortments equipped for enduring decreased water conditions, more noteworthy resistance against soil alkalinity, free from aluminium and iron poisonousness.
Waste Cleanup	Biologists are very intelligently making effort to find the organism of Deinococcus radiodurans because these organisms are very potential in washing out the radiation infected and toxic chemical locations.
Vetinary Science	Knowledge of the biology of the living beings dairy animals, sheep, pigs will have tremendous effects for enhancing the production and health of animals and eventually have advantages for human sustenance.
Gene Therapy	Gene therapy is the methodology used to treat, cure or even avoid diseases by changing the genes expression of any human.
Antibiotic Resistance	The investigation of antibiotic resistance helps in finding harmfulness area made up of various antibiotic resistant genes that may add to the bacterium's change from an innocuous gut microbes to a threatening intruder.
Comparative Studies	Bioinformatics tools may be utilized to perform comparative assessments between the facts and figures, sites and biochemical activities of genes in diverse organisms
Drug Development	Bioinformatics is very helpful in depth understanding of disease mechanism that

Application Field	Description
	helps in drug development , i.e. drugs that are effective in treatment with fewer side effects in lower costs.
Microbial Genome Applications	The entry of the complete genome sequence and their capability to give a more noteworthy knowledge into the microbial world and its abilities could have expansive and sweeping ramifications for environment, well-being, vitality and industry applications.
Preventative Medicine	Safety measure, for example, change of way of life or having treatment at the most punctual conceivable stages when they will probably be effective, could bring about tremendous advances in our battle to vanquish diseases.
Personalized Medicine	Personalized medicine is developed based on the individual's genetic inheritance knowledge of drug response body.
Crop Improvement	Findings of comparative genetics of the plant genomes suggest that information obtained from the model crop system can be used to suggest improvements to other food crops. Currently, the complete genomes of Arabidopsis thaliana (water cress) and Oryza sativa (rice) are available.
Insect Resistance Plants	Insect resistance plants which can resist insect attack will prove a boon to produce more nutritional quality crops due reduced use of insecticides.
Forensic Analysis of Microbes	Experts use the genome tools to help in the forensic investigation in criminal cases.

6. CONCLUSION

In recent years data are generated at dramatic pace analyzing these data is challenging for a general man. From this survey, it is understood that every big data platform has its individual focus some of them are designed for batch processing whereas some are good at real-time analytic. Each big data platform also has specific functionality. The capability for organizations to collect and process Big Data about individuals or groups causes various privacy concerns. Further study could address the ethical issues that may arise from Big Data and the measures that society can take to mitigate such concerns. This survey provides a comprehensive summary about a number of recent studies and techniques. Future applications of big data analytics must focus on development of high-end integrated technologies to process massive biological data at minimal cost, increased speed and robust security measures to accelerate bioinformatics research. Bioinformatics and analytics together can generate new wave of career opportunities in research and development (R&D) sectors in pharmaceutical industries as well as information technology (IT) sectors in coming future.

REFERENCES

A. Jacobs. (2009). The pathologies of big data. Communications of the ACM, 36-44.

C. L. Philip, Q. C. Data-intensive applications,challenges, techniques and technologies: A survey on big data, Information Science.

Haider, A. G. (2015). Beyond the hype: Big data concepts, methods and analytics,. International Journal of Information Management, 137-144.

K. Kambatla, G. K. (2014). Trends in big data analytics. Journal of Parallel and Distributed Computing , .2561-2573.

Kalyan Nagaraj, G. S. (2018). Emerging trend of Big Data Analytics in Bioinformatics: a literature review. International Journal of Bioinformatics Research and Applications, 12.

Kumar, T. K. (2103). Big data analytics: A framework for unstructured data analysis, 153-156.

L. Wang and J. Shen. (2013). Bioinspired cost-effective access to big data. International Symposium for Next Generation Infrastructure, 1-7.

Lynch, C. (2008). Big data: How do your data grow?,Nature. 28-29.

M. K.Kakhani, S. K. (2015). Research issues in big. International Journal of Application or Innovation in Engineering & Management, 228-232.

MH. Kuo, T. S. (2013). Health big data analytics: current perspectives, challenges. International Journal of Big Data Intelligence , 17-22.

R. Kitchin. (2014). Big Data, new epistemologies and paradigm shifts,. Big Data Society , 1-12.

R. Nambiar, A. S. (2014). A Look at challenges and opportunities of big data analytics in healthcare, 114-126.

Vinod Kumar, R. M. (2016). Big Data Analytics : Bioinformatics Perspective . International Jouranal of Innovations & Advancement in Computerscience, 2347-8616.

X. Jin, B. W. (2015). Significance and challenges of big data research. Big Data Research , 59-64.

Z. Huang. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining,. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.