Review Article

Open Access
CrossMark

Big data analytics

Abstract

Big data analytics refers to data sets that are too huge in volume generate at high velocity as well as in different varieties. So these data sets are named as big data. They are difficult to handle by traditional methods due to their weak algorithms, high costs and many more. The data is generated by various fields and it has increased from the use of internet. Big data is in structured, unstructured as well as semi-structured form. There are different characteristics famous as "V's" of big data. Then begin the analysis of such huge data. This is necessary for revealing unhidden patterns which may give solutions for various problems and may bring out amazing result which do help the organizations. There are different types of analytics which are choosing according to the type of data they have collected. There are various levels of big data analytics. There are various fields which use big data and its importance is increasing day by data. All most all sectors take advantage of big data. There are certain issues and challenges related to big data and their analysis which are tried to manage. Tool selection is one of the major part as there is no single tool which can handle all things at once.

Keywords: big data, big data analytics, issues, applications, challenges

Volume 2 Issue 5 - 2018

Pritee Chunarkar-Patil, Akshanda Bhosale
Rajiv Gandhi Institute of IT and Biotechnology, Bharati Vidyapeeth Deemed University, India

Correspondence: Pritee Chunarkar-Patil, Rajiv Gandhi Institute of IT and Biotechnology, Bharati Vidyapeeth Deemed University, Pune, Maharashtra, India, Tel 9730038142, Fax +91 20 24365713, Email preeti.chunarkar@bharatividyapeeth.edu, preetichunarkar@gmail.com

Received: September 19, 2018 | **Published:** September 24, 2018

Abbreviations: BI, business intelligence; SQL, structure query language; EHRs, electronic health records; R&D, research and development; BDA, big data analytics; IoT, internet of things; PJ, picojoules; RDBMS, relational database management system; HDFS, hadoop distributed file system

Introduction

Big data term is nowadays used all over the world in every field though it is any forum or organization. Big data is nothing else but data which is in large volume that requires advance technologies to handle as existing traditional technologies cannot manage such enormous datasets, for extracting useful value information. This extraction of value products is the analysis of big data which we call as Big Data Analytics.[1] The revolutionary step in the world of data was the introduction of relational datasets which could be stored in the form of table and which was easily processed whenever needed. Then analysis over the data was done which solved many issues and made life easier. Then came internet world and the big data era came into picture where the problem introduced. This excessive growth was in the beginning of 21st century. Due to excessive use of the World Wide Web we got large volume of data with high velocity and broad variety. Such datasets were difficult to handle and process by traditional techniques and so new data type was introduced known as big data which had all different technologies and concepts.[2] We cannot imagine a world where data is not stored, for example a place where peoples detail otherwise any organization's details, any transactions, any documentation is directly lost after use. Obviously if this would have happened, the ability to produce useful information and to perform exhaustive analytics will be lost by the organization. Even there would be no new opportunities or any significant advantage to any organizations. Any small details of a person whether ranging from there name to the address is a very important aspect to organization which become the building block of it. Now due to the advancement in internet, there are vast details and information provided. Every second data is been created in this world and such enormous data is termed as big data. The need of storing this data is as well a big challenge and analyzing big data is very critical but an important process thus called big data analytics.[3]

Behavioral types of big data

The data is been Categorized into many types according to behavior:

Structured Data

The data stored in relational databases table in the format of row and column. They have fixed structures and these structures are defined by organizations by creating a model. The model allows to store, process as well as gives permission to operate the data. The model defines the characteristics of data including data type and some restriction on the data. Analysis and storing of structured data is very easy. Because of high cost, limited storage space and techniques used for processing, causes RDBMS the only path to store and process the data effectively. Programming language called Structured Query Language (SQL) is used for managing this type of data.

Unstructured data

Data without any specific structure and due to this could not be stored in a row and column format is unstructured data. The data is contradictory to that of structured data. It cannot be stored in a databank. Volume of this data is growing extremely fast which is very tough to manage and analyze it completely. To analyze the unstructured data advanced technology knowledge is needed.

Semi-structured data

Data which is in the form of structured data but it does not fit the data model is semi-structured data. It cannot be stored in the form of data table, but it can be stored in some particular types of files which hold some specific marker or tags. These markers are distinguished by some specific rule and the data is enforced to be stored with a ranking. This form of data increased rapidly after the introduction of the World Wide Web where various form of data need medium for interchanging the information like XML and JSON.[2]

Characteristics of Big data

Different V's are introduced by various publishers some common are

I. Volume- the amount of data created is very huge as compared to ancient time. By some recent estimation, in all sectors there are at least 100 terabytes and many more with more than petabytes.[6]

II. Variety- data is created by machines and individuals that come from different sources so diversity is at great extent. In this some are structured some are unstructured whereas some are semi-structured.

III. Velocity- data formation or generation is very quick nearly unstoppable even if we are in rest

IV. Veracity-data is generated by various different sources so you need to check the veracity i.e. the quality of data. In this the data is entrusted and uncleanness.[26]

V. Complexities- it is the degree of interconnectedness and interdependence of the structures in Big data in such a way that any small or big change in one or few components can lead to drastic change or even no change in the behavior of system.[1]

VI. Visualization- the most crucial part in today's world is visualizing data such large data. We can use chart and graphs, complex spreadsheets and formulas which are effective in conveying meaning.[7]

VII. Value- big data comes in mixture of all structured, unstructured and semi structured out of which we need to extract the data which we need at time. The data which is correlated to what we require is the measure of usefulness and truthfulness of data.[8]

VIII. Validity- accurate and correct data which can be immediately used

IX. Volatility- data shell life and validity of data.[9]

Fields that generates data

All most all fields generate big data. Some major fields where big data plays a major role is

I. Social networking sites: social media that carry information, posts, links etc of different peoples from all over world like Facebook twitter etc.

II. Search engines: there are lots of data from different databases that retrieve from search engines.

III. Medical history : medical history of patients for various health issues from hospitals

IV. Online shopping: shopping online help to know the preferences of customers on different products.

V. Stock exchange: shares of different companies hold by stock exchange which helps us predict the decision of shareholder.[12]

Traditional data system

Traditional data systems like relational databases were used as major source for storing and analyzing data of business and organizations about 30-40 years back. The systems were designed primarily for handling structured data and the main characteristics of that system was that it was highly organized data. Though there were many other data storage systems but majorly these were used.[10] Traditional data solves large and complex problems in a single computer. It used centralized architecture which is costly and ineffective for large data sets, whereas big data is distributed database architecture. In this architecture large blocks of datasets are divided in small sets and solved, solution to the given problem is calculated by different present in a given computer networks. These computers communicate with each other and find the solution to the given problem. Distributed database is in lower price, improves performance as well as provide better computing. Distributed architecture is based on microprocessors which is economic as compared to centralized which is based on mainframe and distributed has more computational power as compared to traditional. Traditional database systems are based on structured data whereas big data uses semi as well as unstructured data. Traditional database store small amount of data which range from some gigabytes to terabyte however big data can store and analyze data ranging from hundreds of terabytes or petabytes and more. Storing large amount of data reduces the cost which will help the business intelligence (BI).

Data schema (representation of a plan or theory in the form of outline) used by big data is dynamic schema for data storage of both structured and unstructured data. Big data stores data in raw format and schema is applied only when the data is read. Whereas traditional database is based on fixed schema i.e. it cannot be changed once saved. Advantage of dynamic schema is that the information of data is preserved. Data relationship is between data items and system in traditional database system but as big data is massive it cannot find relationship with its items. Traditional database system requires complex and expensive software as well as hardware for managing large amount of data. It requires more numbers of hardware and software resources for moving the data from one system to another which significantly increase cost. In Big data the large data is divided into several systems which decrease the data amount into each system. So big data use is simple, cheap as compared to traditional as well as it make use of commodity hardware and open source. In traditional data base the systems are expensive to store such massive data so all data cannot be stored which decreases the amount of data to be analyzed and decrease the accuracy and confidence. In big data the data is stored and points are correlations are identified which provide high accuracy (Table 1).[11]

Table 1 Comparison of traditional and big data

	Traditional data	Big data	Advantage of big data
Data architecture	Centralized database	Distributed database	Cost effective
Types of data	Structured data	Unstructured and semi-structured	Improve variety
Volume	Small amount of data. Range- Gigabyte - terabytes	Large amount of data. Range- <petabytes.	Cost reduces and help business intelligence
Data schema	Fixed schema	Dynamic schema	Preserves the information in data.
Data Relationship	Relationship with data is explored easily	Difficulty in relationship between data items.	-
Scaling	More than one server for computing	Single server for computing	Cost effective
Accuracy	Less accurate results	High accurate results	Confident results and reliable

Big data history

Big data is a long evolution of capturing and using of data and not a new phenomenon. Big data is the future act that will bring change in the way we run society, just like the other developments in storage of data, processing of data and internet. The ancient history of data is when humans used tally sticks for storing and analysis of data about C 18,000 BCE. The tribal peoples used to mark notches into bones or sticks for calculations, which would make them predict about how long their food would last. One of the earliest prehistoric data storage is Ishango Bone now known as Uganda which was discovered in 1960. Then in C 2400 BCE came the very first device particularly for performing calculations- Abacus. Our first libraries also appeared in this time period which represented our initial step towards mass storage. Then in the period of 300 BC-48 AD the library containing largest collection of data of the historic world which covered pretty much everything which we learned so far was destroyed by Romans accidentally. Then the earliest mechanical computer was discovered by Greek from C 100- 200 AD whose CPU consist of 30 bronze gears. It was designed for astrological purposes and tracking cycle of Olympic Games. After this many small discoveries laid the foundation to emergence of statistics like first recorded experiment in statistical data analysis. In 1880 Hollerith Tabulating Machine was discovered that used punch cards for calculation purposes that completed 10 years of work in 3 months discovered by Herman Hollerith also called as father of automated computation etc. Then started the early stage of modern data storage. In 1928 a German-Austrian engineer Fritz Pfleumer invented a magnetic tape which stored information magnetically. Then came the Business Intelligence and start of large data centers where ideas of relational database and Material Requirement Planning systems were out forward.

In 1989 the first use of the term big data was done by Erik Larson in the Harpers Magazine where he said that "The keepers of big data say they are doing it for the consumer's benefit. But data have a way of being used for purposes other originally intended". The birth of World Wide Web took place that kicked internet into gear in 1991. Google search engine debut in the year 1997. After a couple of years in 1999 big data term appeared in a research paper published by the association for computing Machinery. In that storing large amounts of data and inadequate space for storage as well as analysis difficulties were highlighted. In 2001 characteristics of big data- volume, velocity, variety, was defined by Dough Laney. In 2005 creation of an open source framework- Hadoop took place for storing and analyzing big data sets. Hadoop was famous for its flexibility and management of both structured and unstructured data.[4] Life on earth evolved around 4 billion year ago, in which over last 6 million year human evolution occurred, out of which about 100,000 year ago human language evolution started then about 70,000 year ago cognitive evolution started and then finally was the scientist evolution which is about 500 year ago and fortunately analytics evolution is about just 30 year old but still unfold.It was started in 1990s as Analytics 1.0 also known as 'Business Intelligence'. In this data about production process, interaction of customer etc were collected, combined and analyzed by traditional databases where data used to fit neatly and stored in rows and columns. In analytics 1.0 era, more time was spent on preparing data for analysis than analytics itself by IT & Business Analytic. Then in 2000s came the Analytics 2.0 or well known as big data which had complex queries that had views of both structured as well as unstructured data. To deal with such fast processing across parallel servers software like Hadoop, NoSQL etc.[5]

Big data analytics

Big data analytics is a method to uncover the hidden designs in large data, to extract useful information that can be divided into two major sub-systems: data management and analysis.[7] Big data analytics is a process of inspecting, differentiating and transforming big data with the goal of identifying useful information, suggesting conclusion and helping to take accurate decisions. Analytics include both data mining and communication or guide decision making. The analytics is concerned with the entire methodology.[15] Big data analytics is used by all most all sectors for increasing productivity and revenue with decrease cost. It helps by optimizing funnel conversion, behavioural analytics, predictive support, market basket analysis, pricing optimization, predict security threat, fraud detection etc.[16] Big data analytics make sense of large volumes of data having variety of data in its raw form that lacks a data model.[17] Organizations collect, store and analyze massive amounts of data which is referred as big data. Collecting and storing such huge amount of data has little value but analyzing gives tremendous value to the data. This analyzed data helps in decision making any many other things.[18] Big data size range from few dozen terabyte to many petabytes in a single data sets. There are obvious difficulties like capturing data, storing, analyzing, visualizing, sharing etc. even the data gained are not in a single format rather they vary tremendously from structured, unstructured and even semi structured. There is a need for exert advanced analytical techniques on big data and this is where big data analytics helps. The analytics process is used to obtain previous unknown, useful hidden patterns, to extract useful unknown relationships. Association rules, clustering, regression etc are the advanced analytical processes used most commonly.[3]

Types of analytics

After collection of data we need to start analyzing it. There are types of analytics which should be used for different types of data. There are 4 types of analytics.

Predictive

Predictive analysis establish previous data patterns and gives list of solutions which may come for given situation predictive analysis study the present as well as past data and predict what may happen in future, give probabilities of what would happen. It is used to your big data to forecast other data which we do not have. This analytical method is one of the most commonly used methods used for sales lead scoring, social media and consumer relationship management data.

Prescriptive

Prescriptive analysis reveals actions and recommend of what step should be taken. It gives answer to the situation in a focused way. Prescriptive data analytics goes one step forward of predictive as it provides multiple actions with likely outcomes for each decision. This method of analytics is not preferred much by organizations, but its data can show impressive result if used correctly.

Diagnostic

Diagnostic analysis looks to the past information and let us know how, what and why happened. It is usually used to uncover any hidden patterns which help for complete root cause as well as identify any factors that are directly or indirectly causing effect. Diagnostic analysis is majorly used in social media for analyzing the number of posts, shares etc.

Descriptive

Descriptive analysis also known as data mining operates what is happening in real-time. It is one of the simplest types of analytics as it converts big data into small bytes. The result is monitored through e-mails or dashboard. It is used by majority of organizations. Descriptive analysis examines historical electricity usage to plan power need and set prices (Figure 1).[13]

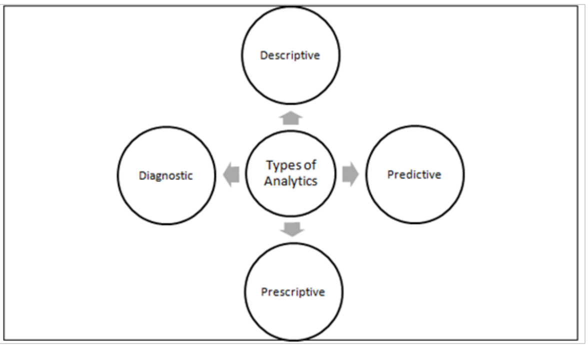

Figure1 Type of analytics.

Levels of big data analytics

Big data analytics developing and implementation is not an easy task, especially when you don't have a data driven culture. Data driven culture is a pre-requisite for big data successful implementation. The right start to big data is to have an understanding of what is it and what can it do to the organization and from there proof of concept with multi-disciplinary team starts developing. This proof of concept is vital to the organization and also for becoming data driven. There are 5 levels of big data maturity within an organization. First level: infancy phase- this is the phase where one starts understanding big data and develops proof of concepts. Second level: Technical adoption: different big data technologies get implemented. This will enable the organization to develop new proof of concepts faster and better. Third level: business adoption- more in deep analysis of structured and unstructured data which results in more sharp, accurate and better decision making of company. Fourth level: Enterprise adoption- the big data adoption across enterprise, which results in united predictive insights of organization. At this level big data analytics has become an integral part of organization. Fifth level: Data & Analytics as a service- at this level the organization operates as a "data service provider". Organization has integrated big data analytics in all levels and now can be seen as 'data companies' no matter what product and service they provide.[14] Levels of big data Commonly there are 4 levels of big data which are considered (Figure 2).[9]

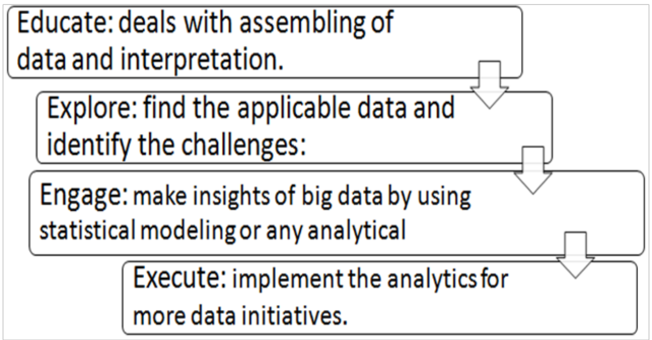
Educate: deals with assembling of data and interpretation.
Explore: find the applicable data and identify the challenges:
Engage: make insights of big data by using statistical modeling or any analytical
Execute: implement the analytics for more data initiatives.

Figure 2 Levels of Big Data.

Fields that use big data

Health sectors

By definition, big data in healthcare refer to electronic health data sets so large and complex which are difficult to manage by traditional software or hardware neither by any traditional tools and methods. Big data analytics plays a vital role in health sector. Benefits of health with related to big data are demonstrated in 3 areas namely to prevent disease, identify risk factor for disease, define intervention for health behavior change. The health care from age has generated voluminous amount of data in the form of records, regulatory requirements, patient care etc. This data is stored in hard copy form most commonly but now everything is rapidly turning to digitization. This reduces the quality of healthcare meanwhile reducing the cost. Big data supports wide range of medical and health care functions to find any previously untapped intelligence. By understanding patterns and trends within the data, big data scientists by the help of big data analytics could improve care, save lives as well as reduce cost. Data in health care sectors include:

I. Clinical data and clinical decision support system like physician written notes and prescription, medical imaging, pharmacy, insurance, other administrative data. About 80% of this health data is unstructured.

II. Patients data in Electronic Health Records (EHRs)

III. Machine generated/ sensor data- monitoring vital signs

IV. Less patient specific information like emergence care data, news feed and article in medical journal.

V. Publication: clinical research and medical reference materials.

VI. Clinical references like the text based practice guidelines and health products data.

VII. Genomic data which represents amount of new gene sequencing data.

VIII. Streamed data- home monitoring, handheld and sensor based wireless or smart devices are the new data sources and types.

IX. Administrative data- this include billing, scheduling and other non health related data.

X. Payer's and Provider's: claims and cost data

XI. Researchers: academic, independent, government.

XII. Developer's: pharmacy and medical device R&D

XIII. Consumer and Marketers: Patient behavior.

Benefits of Using Analytics in Health care sector

i. The large amount of data produced, gives great opportunity to researchers in fields of health informatics, by using tools and techniques for unlocking the hidden patterns.

ii. For individual's/ patients: for deciding any line of treatment for a patient, historical data about the symptoms, drugs, outcomes, responses of different patients is taken into account. The move is towards formulating a (patient on personalized treatment) on the genomic data, locality, area, and lifestyle, response to certain medicines, allergy, and family history. When genome data is known completely, some kinds of relations are established between the DNA and the disease. Then specific treatment is formulated for such small groups of individual. The patient get benefit by various ways like: correct as well as effective line of treatment, better health related decision, preventive steps in time, continuous health monitoring of patients by wireless devices, designing personal line treatment, increase life quality and expectancy.

iii. For hospitals: by various techniques and tools of BDA data available in hospitals gain various benefits: predict patients which like to stay for longer time or get readmit after treatment, identification of patients that are risky for hospitalization, provider could develop pre health plans t prevent hospitalization. Various queries get their answers using these BDA tools, the queries include- will a patient respond positively to the given treatment? If surgery done, will the patent respond to it? Will the patient get prone to catch disease n near future?

iv. The hospital management and administration can take better decisions like- number of patients no getting cured at early stage, number of readmission increasing because of patient getting ill again after treatment, increasing number of staffs on floors for efficient working, plans for frequent post treatment follow ups etc.

v. For insurance companies: government for giving medical claim to patients do large amount of expenditure. By using BDA analysis, prediction and minimizing fraud medical claims is obtained.

vi. For pharmaceuticals: the techniques help R&D to produce drugs, instruments, tools etc in shorter period of time, which are effective in treating specific diseases.

vii. For Government: the demographic data, historical data of disease outbreak, weather data, and data from social media over disease like cholera, flu etc information is used by the government. Government analyzes this massive data to predict epidemics, by finding correlations between weather and disease and accordingly preventive measures are taken. Public health surveillance is improved as well as the response to disease outbreak is quick by using BDA.

viii. Researcher: to improve workflow quality and quantity, like- predictive modeling, statistical tool and algorithms. These improve the outcome of experiment and provide better understanding of developing drugs researchers need new tools. This tool successfully navigates the regulatory approval and marketing process.

ix. Medical device companies: companies are collecting data from hospitals for safety and adverse event prediction. But they wonder what to do with this data and how to integrate the new and old data.

x. By estimations big data can enable more than $300 billion saving per year in US healthcare. Clinical operations and R&D are the largest area of potential savings. Big data could help to reduce waste in the following areas:

Clinical operations

Clinically relevant and cost effective ways to diagnose and treat patients is determined by comparative effective research. There is a big gap in its impact because the Medicaid and Medicare cannot apply this comparative effectiveness. To enhance the efficiency and quality of operations clinical decision support systems are used. This provide real time information to emergency techniques, nurses and doctors to improve triage, diagnosis, treatment choice, prevent iatrogenic infections and readmission, prescription and other medical errors. Transparency about medical data, remote patient monitoring and predictive analysis to identify individuals that would be benefit from proactive care etc are other areas.

R&D

A leaner, more targeted, faster R&D pipeline in drugs and devices is produced by the help of predictive modeling.

Clinical trial designs and patient recruitment to better match treatment to patient could be improved by statistical tools and algorithms. His will help reducing trial failures and speeding cure.

Public health

Analyzing the pattern of disease, tracking disease outbreak, transmission to improve health surveillances and speed response. Accurately targeted vaccines are quickly developed. Torrents of data turning into actionable information which is used to identify needs, service providers and predict as well as prevent crises.[19-23]

Impact of big data on the healthcare system

Discussions of what is right for a patient and right for healthcare ecosystem are getting transformed with the release of big data. Keeping in mind all these changes, patient-centered framework has been created. Five key pathways are considered based on the concept that value is derived from cost spent on healthcare by patient and patient's impacts.

New pathways: Right Living: Focuses on encouraging patient to make lifestyle choices that make them healthy, like proper diet, appropriate exercises, taking active role in caring of their own if sick etc. Right care: Ensures that patients get most timely and appropriate treatment. It requires coordinated approach instead of relying highly on protocols. Right Provider: In this, patient should be treated by high performing professionals that are best matched to the task and achieve best outcomes. Right Value: in this, providers and payers will continuously enhance healthcare value preserving or improving its quality. It involves multiple measures for ensuring cost effectiveness, eliminating frauds or abuse to system etc. Right Innovation: it involves identification of new therapies, approach to deliver care across all aspects of system and improving innovative engines. This is acquired by advancing medicine and boosting R&D productivity. There should be a better use of prior trial data for capturing this pathway (Figure 3).[27]

Banking

Over a period of time banking systems have undergone some intense process of invention and innovation by which it has allowed bank to diversify their activities, to create new as well as complex products. The banking industry is generating huge amount of data day by day where previously this data was failed to utilize by banks. But nowadays, banks are using this data to reach the main objective of marketing. This data is unlocking secrets of money movements, helping to prevent major disaster and frauds as well as it help to understand customer behavior. This banking and financial industry is one of the biggest adopters of big data technologies. Banks internationally have started to harness the power of data to derive utility across various parts of their functioning. Big data in financial industry is defined as tool that allows an organization to create, manipulate, and manage large sets of data in a given time and the storage that supports such voluminous data. (Meta Group. Application Delivery Strategies; February 2001)

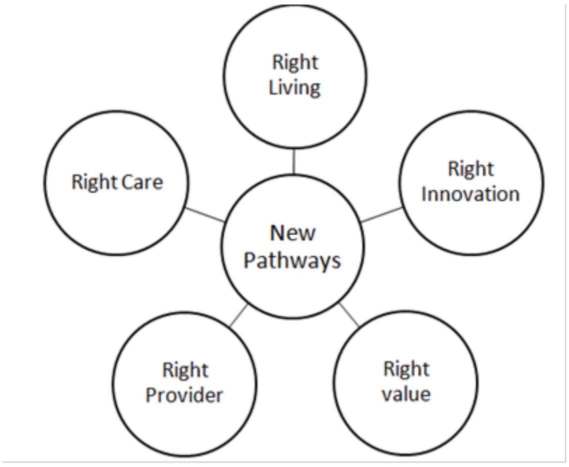

Figure 3 New Pathways for improving healthcare.

Advantages

I. Sentiment analytics: continuously monitoring of customers opinion is needed from banks. Banks need to identify which are their key customers and by their feedback they need to improve their flaws in system. This lead to increase in their productivity and services.

II. Changes in service delivery: whenever a reputation or account range enters into system, it checks through all the information and provide desired information. This allows banks to map work process, save time and prices. Huge information and its proper knowledge allow organization to identify and solve issues before they affect their customers.

III. Fraud detection and prevention: One of the most important obstacles faced by banking sectors is fraud. Big data ensures that no unauthorized transactions are done and provides security as well as safety to the entire system.

IV. Enhanced reporting: getting access to huge amount of data, also contain different needs of different customers. Then Banks offer those needs in a meaningful way. Bank industry provides the exact information required by the customer by using big data.

V. Risk management: early detection of fraud is a massive part of risk management. Large amount of information does the maximum amount o risk management as it will identify fraud. Massive information plays a major role in desegregation of banks needs into a centralized practical platform by which possibilities of losing the information is reduced.

VI. Customer segmentation: by identifying usage of cards by customer, loyalty programs are created. Targeted marketing programs are made as well as relationships are build between valuable customers.

VII. Examine customer feedback: customers sentiments are collected in text form from various social media sites and after collections they are classified into positive and negative. This is used to provide service to customer.

VIII. Detect a customer when about to leave: cost of acquiring new customer is great than retaining old ones. Banks take care of customer's need, by understanding problems and finding solution to it.[28-30]

Big data and bioinformatics

Introduction

First lets us see what bioinformatics is. Bioinformatics is an interdisciplinary field that develops and improves methods for storing, retrieving, organizing and analyzing biological data (Wikipedia). As we enter the digital age, data is produced by various sources not only by people or servers but also from sensor embedded sensors such as mobile phones, MRI scanners, cameras, set-top box etc. due to the digitization of all devices which we use and the information technology, there is quick transition in the information age. Now recently a new trend has emerged to network all the man-made things. These are apart from digitization which communicates with each other through different sensors. These are coined as Internet of Things (IoT). IoT is also growing rapidly. Not all data generated is useful for analysis, only a part of data termed as target-rich data is useful. Like others, the volume of data in bioinformatics is also growing fast. The sources of big data are not limited to experiments of particle physics or search engines. Due to availability of high output of devices at minimal cost and digitization of all processes, data volume is rising in every field considering bioinformatics research also. Decrease computational cost and increased analytical output, supports the trend of rising volume of data. Now in this era, no biologists discover novel biomarker for a disease by using traditional techniques but they depend on vast and continuously growing genomic data. A new era of big data in bioinformatics is rising due to low cost as well as effective technologies for capturing bio data. High volume of data helps for accurate analytics, especially in the field which deals with highly sensitive research like bioinformatics.

Big data types in bioinformatics

There are main five types of big data in bioinformatics research: The human disease network and gene disease association network is also used and are important for research purposes. In analysis of gene expressions, analysis of thouasds of genes over different condition is done. Gene expressions based on microarray is most commonly used for recording expression levels for analysis. There are thrr types of microarray data: gene-sample, gene-time, gene-sample-time. The Gene-samlple profiles record expression levles for various levels of external conditions. The time-sample, record expression levels at different time. Identification of genes that are affected from pathogen or viruses, is helped by gene expression analysis. The results are used for suggesting biomarkers for disease diagnosis and prevention among others. Public resource for microarray databases are: Array Expresses, Gene Expression Omnibus and Stanford Microarray Databases. In DNA, RNA or protein sequence analysis, various analytical methods are process to know their features, functions, structures and evolution. DNA sequencing is used for genomic study, protein association with disease and phenotypes, evolutionary biology etc. RNA sequencing is an alternative for microarray. It is additionally used as for mutation identification, post transcriptional mechanism identification, virus detection and polyadenylation identification. This requires more sophisticated techniques and computing. Important sequence database include DNA Data Bank of Japan, RDP and miRBase. Protein-protein Interaction provides information regarding all biological process. Formation and analysis of this protein-protein network can give a proper knowledge of various functions of protein. PPIs are the basic for Alzheimer's disease and cancer. PPIs are studied in various fields of research like bioinformatics, quantum chemistry, molecular dynamics and biochemistry. Important PPI repositories are DIP, STRING and BioGRID (Figure 4). Analysis of pathway is useful for knowing molecular basics of diseases. Pathway analysis identifies genes and protein associated with the etiology of disease, help to conduct targeted literature searches. It helps to integrate diverse biological information and assign function to genes. Important pathway data sources are KEGG, Reactome, and Pathway Common. Go data base provide species independent gene ontologies for biological process, cellular components and molecular function. The GO data base tool is widely used in bioinformatics research. Most of the tools are third party associated tools, however the GO project itself maintains various tools such as SerbGO for searching accurate GO tools for bioinformatics problem.[33,34]

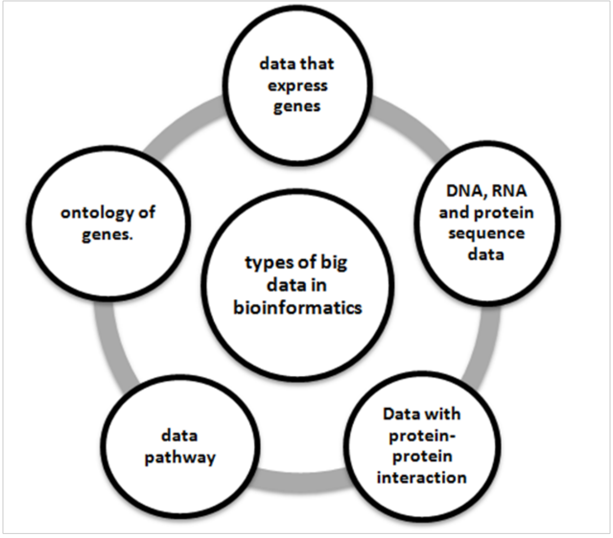
data that express genes
DNA, RNA and protein sequence data
Data with protein-protein interaction
data pathway
ontology of genes.
types of big data in bioinformatics

Figure 4 Types of Big data in Bioinformatics.

Issues

There are some conceptual issues in big data which should be understood by an organization to implement technology effectively. The issues are critical to handle but at the same time important to know.

Issues related to characteristics

Data volume: as the volume of data increase, the value of other data record decrease. The social media itself creates data in tetra byte everyday and this amount of data is definitely difficult to handle by using the existing traditional systems. Data velocity: data is constantly in motion and traditional systems are not capable of performing the analytics on such data. Data velocity is much more than a bandwidth issue. Data variety: data produced is totally different consisting of structured, semi structured as well as unstructured data which traditional data system find difficult to handle. This is the biggest obstacle from the analytic perspective.

Data management issues

Data management is probably the most difficult problem to address with big data. The major obstacles are resolving issue of access, utilization, updating, governance and reference. individuals contribute digital data in various medium in which they are comfortable like documents, drawings, pictures, video, audio, user interface designs etc, with or without adequate knowledge of what, when where, who, why and how it was collected. Digital data is collected in more relax way unlike manual method where rigorous protocols are followed in order to ensure accuracy and validity. By the volume in big data it is impractical to validate every data item so new approach to data qualification and validation is needed. There is no prefect big data management solution yet due to which there is an important gap in research literature on big data is needed to be filled.

Storage and transport

Whenever a new storage medium is invented, each time quantity of data is exploded. Data is being created by everyone and everything not just as here to fore, by professionals such as scientists, journalist, writers etc.[1]

Challenges

Real implementation is usually the challenge in big data. This real implementation hurdles require immediate attention. Without handling of such challenges will definitely lead to technology implementation failure and objectionable results.

Privacy and security

Privacy is one of the most important and sensitive challenge with big data. There are privacy issues in various areas.

Privacy issue in big mobile data

Nowadays mobile is used by every person and they share lot of information on it. Many a time, mobile sends data to service provider without knowing to user. To identify the owner who is using mobile phone and details provide by the service provider is easy so privacy in mobile data is very important. Lee Garber mentioned that a roman security vendor, Bit Defender, found many Android applications, which can easily access and send information without the consent of user. The information includes user's location; contact lists with their number and email ids as well photos. The security vendor also analyzed 836,021 applications on Google Play store and out of which about 33% of application could reveal location related data, about 5% can locate and open photos by user's phones, approximately 3% could reveal their email data, Musolesi has mentioned that in June 2013, there were about 819 million active mobile users on Facebook. This is just about Facebook; there are many other mobile applications too like twitter, Whatsapp from where data is generated at huge amount.

Privacy Issue in Social Media Data

One of the biggest revolutions of past decade is the social media. On social media there is lot of information being shared by user, then sometimes the people close to you share information, which you may not like to disclose on social media. This leads to privacy violation of an individual. But privacy settings are present on Facebook to approve tags. So if someone tagged you in any photo or thoughts, without your approval it will not get disclose on your wall.

Privacy in web Usage Data

Intel wants to makes its internal website energetic, based on web usage data of all the users. Any user can be identified by the browsing information and IP address from the web usage data. Even every activity which the user performs can be detected. By such system user privacy is disobeyed.

Analytical challenge

Big data brings huge amount of analytical challenges along with it. As we know big data brings voluminous data along with it which can be in unstructured, structured as well as semi structured form and to analyze such data requires advanced skills in large number. Type of analysis needed to be done on the data depends on the result obtained. This decision making is done by using any two of these techniques: incorporate voluminous data for analysis or determine upfront by which big data is relevant.

Technical challenge

Fault tolerance: with the invention of new technologies it is always predetermined that whenever failure occurs the damage should be in acceptable origin. As it is nearly impossible to invent a foolproof 100% reliable, fault tolerant machine the main task is reduce the chances of failure to an acceptable level. But the harder we attempt to reduce the chance of failure the higher is the cost. Two methods which reduce the chance of failure in big data:

I. Divide the whole computation to be done in task and assign these tasks to different nodes for computation.

II. One node is assigned to observe that whether the nodes are working efficiently.

If something gets wrong that particular task gets restart. But sometime the whole computation is not possible to be divided into task thus restarting the whole computation becomes an unmanageable process. This is usually avoided by applying check points which checks the system after certain intervals of time and if any failure occurs the computation can restart from the last checkpoint.

Scalability

This issue of big data has lead towards cloud computing which is now compiled with dissimilar work with varying performance goals into a large cluster. A high level of sharing is required which is expensive and brings various challenges along with it. These challenges are like how to execute various jobs and run them so that each workload is cost effective. Along with high level sharing it requires dealing with system failure in effective manner. System failure occurs very frequently as operations are on large clusters. There is large change in the technologies begin used like hard disk drives (HDD) are replaced by solid state drives.

Quality of data

Collection of huge amount of data and its storage comes at a cost so if more data is used it will definitely leads to appreciable results. Business leaders will always take more and more data but in case of IT leaders, they will first prefer technical aspects before storing of all data. Thus big data basically deals with quality of data storage than irrelevant high amount of data. This leads to better results and conclusion to be drawn. This later leads to origin of various queries like that how can we get sure about the relevant data, how much data should be enough for decision making, data stored is accurate or not to draw conclusion.

Heterogeneous data

Working with unstructured data is almost difficult as well as costly. Unstructured data include almost every kind of data that is been produced like social network interaction, recorded meeting, handling PDF documents etc. converting all such unstructured data to structured is not at all easy. Along with that structured data is highly organized, managed and integrates with databases easily but in the case of unstructured it is completely raw and unorganized.

Human collaboration

Input from multiple human consultants and shared exploration of result should be supported by data analysis system. There should be a separate space and time for these multiple consultants when it gets too costly to assemble the whole team in one room. This distributed professional input should be accepted by the system at the same time the system should support the collaboration.

Hardware

We all know the economy progress in the last 20 year is pushed by the exponential growth in the Information Communication technology. By the exponential growth world, in every 18 month computer chip capacity gets double (this in accordance to Moore's law), every 12 months data storage gets doubles and every 9 months communication bandwidth doubles. It results in reduced component size and less energy consumption. In 2013 the energy consumption for CPU Floating Point Unit (FPU) was 100 PicoJoules (pJ) and it is assumed to be reduced at least ten times in 2018. Lastly, by power consumption, computer processing power for big data is limited. CPU processing speed in gigahertz has reached a certain limit due to overheating issue. To reduce the power consumption multi-core with low frequency clock is used. Many cores are not elixir as data movements also requires energy.[24-31]

Tool selection

Big data involves various tools and techniques that are used to manage the data. There is no single technology that can cope up with all the characteristics of big data at once, which is a major challenge in today's time. Application of advanced analytical technique to large data sets i.e. to big data is big data analytics. These advanced analytical techniques are collection of various types of tools and techniques that support analytics. Types of tools include: predictive analytics, statistical analytics, artificial intelligence, complex SQL etc. Analytics requires massive performance and scalability. For big data processing and analytics, there are many open source tools available. One of the well known open source tool is Hadoop.[24-32]

Hadoop

Hadoop is an open source framework created by the creator of Apache Lucene, Doug Cutting. It is written Java. By using simple programming models, it allows distributed processing of large data sets across clusters. Hadoop provides distributed storage and computation across clusters of computers as well as it is designed to scale up from single server to thousands of machine.

Traditional approach

Traditionally, enterprises had computers for big data storing and processing and then for analysis purpose it used to be present. Data was stored in Relational Database Management System (RDBMS) like Oracle database, MS SQL Server etc. For proper interaction with databanks, complex software's are written. This traditional approach goes for smaller amount of data that could be easily held by standard servers and can be processed in an optimum way by processor. But when data volume increase it consumes time as well as the processing task becomes tedious. To tackle such voluminous based problems, goggle picked a new approach by using an algorithm called MapReduce. In MapReduce, the algorithm divides the input in smaller parts and assigns those parts to many computers connected over network, and then the result is collected to form a result with final output. Hadoop runs applications where data is processed in parallel on different computer nodes using MapReduce algorithm.

Hadoop architecture

Hadoop framework includes 4 models

I. Hadoop Common: they contain java libraries and utilities that are required by other Hadoop modules. The java libraries provide file system and OS level abstraction. It contains necessary Java files and scripts that are required to start Hadoop.

II. Hadoop Yarn:

III. HDFS (Hadoop Distributed File System): it is a distributed file system that provides high throughput computing access to application data.

IV. Hadoop MapReduce: for large scale data processing this is programming model.

HDFS

It is a distributed file system in Hadoop and is an extended version of Google File System. It holds very large amount of data on cluster of machines and provides easy access. Such huge data files are stored across multiple machines in a redundant fashion to rescue system in possible loss in case of failure. It uses a master slave architecture, in which the master node assigns a task and controls the slave nodes. Slave node has an empty slot. The Map and Reduce functions run in parallel on these nodes. Master node is single in number whereas slave nodes are multiple. Hadoop has a set of identical nodes manages the processing as well as the storage part by Hardware point of view. By using cheap commodity HDFS provides replicated storage for data and by default this replication factor is 3. HDFS is fault tolerant and designed by using low cost hardware as a result cost effective. For accessing small number of large files HDFS gives best performance.

Map reduce

MapReduce programming model for parallel processing of large data sets. This YARN based system was proposed by Goggle to support data-intensive applications running on parallel computers. With programming model it is a software framework for writing applications that speedily immense amounts of data in parallel on massive clusters of calculate nodes. MapReduce consist of two important functional programming primitives: Map and Reduce. The Map function perform task of filtering and sorting. It takes an input data to generate list of intermediate value pairs. Reduce function perform the task of summarizing the result which is applied to the set of intermediate pairs with same value. Multiple reducers are used to parallelize the aggregation. MapReduce work is to process the data through data node. Data node always communicates with Name node continuously through heart beat message. At specific interval data node sends the heart beat to name node indicating that it is alive and working properly. If the data nodes fail to beat and the name node could not get the beat after a particular time period, it assumes the data node to be dead and passes the work of that data node to somewhere else.. There is a system that teams along all intermediate pairs based on intermediate value and passes them to reduce function for producing results; this system is termed as MapReduce runtime system. MapReduce specialty lies in its simplicity because the programmers just need to focus on data processing function rather than on parallelism details.[20-25]

Conclusion

The journey of this review was not easy at all. Many difficulties came and once I even thought that my work would not be completed on time. The research papers helped me in understanding my topic easily. I have put my sincere effort to complete my review. There are some flow chart diagrams which help to understand the topic at a glance. There are almost all topic covered which deals with big data. There is a brief history of the topic too. Detailed information of the sectors that utilize such data is given along with the advantages.

Acknowledgments

Undertaking a review on “Big Data Analytics” and submitting detailed dissertation would not have been possible without the remarkable guidance and support of my review guide Mrs. Preeti Chunarkar, Department of Bioinformatics, Rajiv Gandhi Institute of IT and Biotechnology, whose valuable contributions paved my way towards completing of my review. I am highly grateful and obliged to her for her dedicative support.I wish to express my profound sense to gratitude to Dr. G.D. Sharma, Principal, Rajiv Gandhi Institute of IT and Biotechnology, for kind blessing and encouragement. I would like to thank Dr. S. Shaikh, HOD, Department of biotechnology, for her great support. I would like to thank my parents for their motivation and encouragement during completion of my review.

Conflict of Interests

The author declares there is no conflict of interest.

References

1. Lenka Venkata Satyanarayana. A survey on challenges and advantages in big data. *IJCST*. 2016;6(2):115–119.
2. Ananta Chandra Das, Santosh Kumar Pani, Sachi Nandan Mohanty. A comparative study on data analytics and big data analytics. 2016;4:67–75.
3. Nada Elgendy, Ahmed Elragal. Big data analytics: A literature review paper. *Springer International Publishing*. 2014;8:214–227.
4. Bernard Marr. A brief history of big data. 2015;2.
5. Shankar megnatha. The evolution of analytics. 2016;10.
6. Sas white paper. 2016.
7. Naresh Babau, Suneetha Mane. A comprehensive survey of big data analytics and techniques. *IJCST*. 2016;14(10):177–184.
8. Pooja Vajpayee. *Big data analytics, its privacy and security challenge*. 2016;4:12–15.
9. M Kavita. *Journey through big data analytics*. 2016;4(9):73–74.
10. George Trujillo, Rommel Garcia. *Understanding the big data world*. 2015;8:2.
11. Deepali Agarwal. *Difference between traditional data and big data*. 2016;6.
12. Nirali Honest, Atul Patel. A survey of big data analytics. *IJIST*. 2016;6(3):35–43.
13. Nidar R. *Big data and analytics: which type of analytics does your business need?* 2015;9.

14. Van Rijmenam. *Five levels of Big Data Maturity in an organization*. 2014.

15. Irwin King, Michael R. Lyu and Haiqin Yang, online learning for big data analytics. 2013. p. 1–116.

16. Why Big Data Analytics?

17. Hugh Watson. *Tutorial:big data analytics: concepts, Technologies, and Applications*. 2014;34(4):65.

18. Jasleen Kaur Bains. Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2016;6(4):430–435.

19. Benjamin Allen. Big Data Analytics and the Current Generation. *IJCSI*. 2015;3(9):8171–8177.

20. Priyanka k, Nagarathna Kulennavar. A Survey On Big Data Analytics In Health Care. *IJCSIT*. 2014;5(4):5865–5868.

21. Jasleen kaur Bains. Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges. *IJARCSSE*. 2016;6(4):430–435.

22. Bonnie Feldman, Ellen M, Martin. *Big Data in Healthcare Hype and Hope*. 2012;8:1–56.

23. Brijesh Mehta, Udai Pratap Rao. Privacy preserving unstructured big data analytics: issues and challenges. *Procedia Computer Science*. 2016;78:120–124.

24. Mikin dagli, Brijesh B, Mehta. *Big data and Hadoop: A Review*. 2014;2(2):192–196.

25. Chaitanya kadam, Yasoda Thapa. *Big Data: Features, Challenges & Solutions*. 2016;5(5):1649–1652.

26. Peter Groves, Basel Kayyali, David Knott, et al. The big data revolution in health care. *Mckinsey& Company*. 2013;1:1–22.

27. Krish Khambadk. *Big Data technology use cases for banking and financial services*. 2013.

28. Abhinav Kathuria. Impact of Big Data analytics on banking sector. *IJSETR*. 2016;5(11):3138–3141.

29. Alina pop. The New Banking Sector. Towards Reforming the Too Big to Fail Banks. *Procedia Economics and Finance*. 2015;23:1485–1491.

30. Simon C, lin, Eric Yen. Challenges of Big Data Analytics. *IJCA*. 2013;3:17–22.

31. Sahil kalra, Aarati Mahajan. Big challenges, big data. *IJCA*. 2015;131(11):14–18.

32. Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, et al. Big Data Analytics in Bioinformatics: A Machine Learning Perspective. Journal of latex class files. 2014;13(9):1–20.

33. Prakash Nemade, Heena Kharche. Big data in bioinformatics & the era of cloud computing. *Journal of latex class files*. 2013;4(2):53–56.