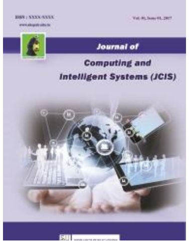
Original Research Article

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in



ISSN: 2456-9496

A Study of Advanced Technologies involving in Big Data Analytics

S. Jayasankari #1, **M. Banu priya**#2

Received on 24th Oct 2017, Accepted on 29th Nov 2017

Abstract — *This paper explains the importance of big data Features, dealing with business effectively. The study and analysis of big data, provides different issues because of the various data types. Heterogeneous mixture data is used to mine different patterns. This paper gives an overview of the advance technologies involving in big data using data mining. The heterogeneous mixture learning techniques is used to analyze the Big data which is helpful to predict the demands.*

Keywords: Big data, Heterogenous Mixture, Hadoop, Volume

1 INTRODUCTION

Today big data create great attention from various people. In the digital computing world, the people from academia and it industry start up to generate and collect the information that exceeds the boundary value. More than 5 billion people connected to the domain have been reviewed. An advance technology in big data using data mining techniques which is explains the advanced tools and technologies of different industries. Big data helps to know the growth in different industries i.e. business application, IT industry, healthcare machines, banking transaction, and social media. Big data either structured or unstructured. Today it works used structured data by basic algorithm. The machine has the capacity to understand and use structured data. Unstructured data is not fit correctly into relational data base like SQL.

A. Big data - Advance Technologies and Challenges

The current population limit exceeds 7.3 billion over 2 billion of people were connected with high speed internet through World Wide and 5 billion of people using Mobile phones. Finally, millions of people have a role to generating large amount of data by increased the speed of use. Big data produce heterogeneous data that are either structured or unstructured. We can characterize by three Aspects: 1. Data are numerous 2. The data cannot be categorized when it used in regular relational databases. 3. Data are generated, updated, captured, and processed very fast. Big data is provided facility to improve such business application and take main role in growth of IT industry. It is creating interest in different area such as health care machines, Banking transaction, Social media and Satellite imaging. Data is placed using highly structured and perfect model to maximize the strength of Information. We explain the characteristics of Big data, what are the different tools are related to big data and various opportunities, issues, challenges are involved in Big data.

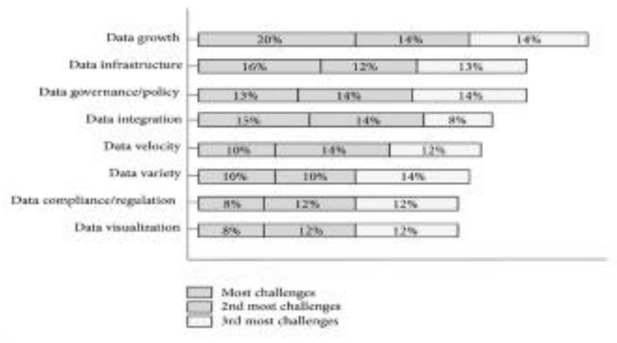


Fig-1 Challenges in Big Data [4]

Today the growth of IT industries mainly depends upon big data because big data has support different data types, which is used to develop number of application, but certain number of issues affected the rapid growth. There are five different issues disturb the rapid development such as volume, velocity complexity, variety, value, the technical group require discussion to represent the serious type of problem. Big data provide opportunities for upcoming researchers. Figure 1 [4] shows what are the issues, how it has affects the growth of organization or different industries. The analysis of challenges in big data has different types of problem, whereas the researchers want to represent efficient and well-constructed pattern. The techniques of data mining provide facility to solve challenges and issues with the help of preprocessing. These preprocessing techniques such as data cleaning, integration, transformation, reduction. This technique helps us to find good solution for inconsistent, noisy data, incompleteness.

* Corresponding author: E-mail: banuchandru146@gmail.com,

[1]Assistant Professor, Dept. of Computer Science, P.K.R Arts College for Women (Autonomous), Gobichettipalayam, Tamilnadu, India.

[2]Research Scholar, Dept. of Computer Science, P.K.R Arts College for Women (Autonomous), Gobichettipalayam, Tamilnadu, India.

Original Research Article

B. *Advanced Technologies to Support Big data processing*

The advance technologies are used to predict the heterogeneous mixture data to create different applications. First, we inherent heterogeneous mixture data by grouping the data. This is the way to improve the accuracy. There are number of methods to grouping the data. Here impossible to verify each and every candidate.

The following three issues used to split the data into several groups. (i) Total numbers of groups (ii) What kind of method for grouping (iii) Select prediction model according to the properties of each group of data.

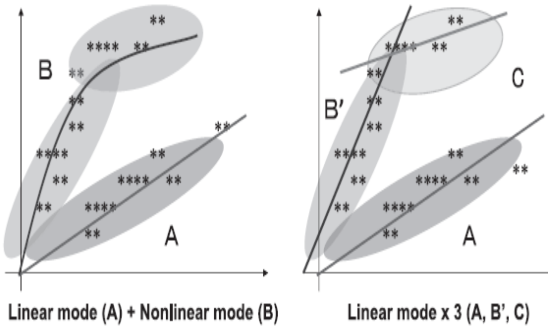


Fig 2 - Illustration of heterogeneous mixture data

These issues solved by following from 1) to 3), with mutual dependences heterogeneous mixture data groups (ellipse A and ellipse B) in (Fig.1 Left) to obtain highly accurate prediction model by grouping data mixture. In right (Fig.1) explains optimum number of group 3, and we used prediction model, but not possible to determine method of group 2) by ignoring 1) and 3). Big data is a large volume of storage to find electricity demand data start with analysis of prediction model by detecting hidden rules.

III TOP 10 IMPORTANT BIG DATA

We discuss ten top most advance technologies these technologies are used to provide standard data.

A. Predictive analytics

The result of hardware and software allow different companies to discover the predictive models by analyzing source of big data which is improve the business task.

B. NOSQL database

Key-value, databases and documents.

C. Search and knowledge discovery

Information (or) pattern is discovered from repositories using tools and technologies. These patterns are structured and unstructured from various data source such as filesystems, databases.

D. Stream analytics

Software (set of instructions) which is used to filter aggregate and analyze a high throughput of data from multiple data source.

E. In memory data fabric

Processing high quantity of data and provide low latency access by distributing data to other sources like DRAM (Dynamic random-access memory), Flash of distributed computer system.

F. Distributed file stores

In a computer network system data is stored on more than one place for redundancy and performance.

G. Data virtualization

There are different data sources such as Hadoop, distributed data in real time which is used to mining information from various data sources.

H. Data Integration

EMR- Amazon Elastic Map Reduce, Apache Pig Map Reduce, Couchbase, Hadoop, Hive Apache and mongo DB these are tools used to combining data from several sources.

I. Data Preparation

Data Preparation or preprocessing which is used for further analysis and processing.

J. Data Quality

Data quality is mainly affect the result of application. Quality of data correctly fit to the operation, when there is no incomplete, noisy and inconsistence. Good and exact result of IT and other organizations only mean the data quality. Which is ensure the growth of health care, social media, IT industries and different social sectors. The quality achieves by complete data and accuracy.

IV BIG DATA AVAILABLE TOOLS AND COMPONENTS

The Big Data and architecture support different organizations because growth and development of organization depends on big data. Main problem in big data is volume, the volume is large, and it is very difficult to handle volume of data. Here two sections are used to detail the volume i. Continues growth of data ii. Rate of hard disk. The usage of unstructured data is increased; Big data not only consider text. Which is gave importance to other data types including video, photo, financial transaction, e-mail, tweeter, mobile phone calls, documents, telemetry, and health care records. The volume of unstructured data is increased by 197% from 2012-2017. Thus, affecting the organization and enterprise environment.

Today Big data has created different new issues. There are five different important issues in big data such as volume, complexity, variety, some additional issues also used to represent problems in technical research managing big data with tool also creating some problem because available tools are not enough to manage data. Today some special tools available to manage data. Big data including Google Big Table Data Stream Management System(DSMS), memcacheDB, voldemort [4], Not Only SQL (NOSQL) AND Simple DB.

Big data not a traditional Data, not only stored in a single machine Increase the efficiency of big data with most repeated tools are Hadoop, BIG Table and MapReduce. It is possible to use Frame works to handle big data.

A.Hadoop

Today one of the latest and hottest tools in big data is Hadoop. Most of the industries 80% of the engineers using this tool. Hadoop has different components. Hadoop is used to manage lot of unstructured logs, events. There is different Component in Hadoop such as Hive mohout Flume, chukwa pig, AVRO, Hcutalog, Zookeeper, Oozie, Fig-3 explains the various components in Hadoop ecosystem.

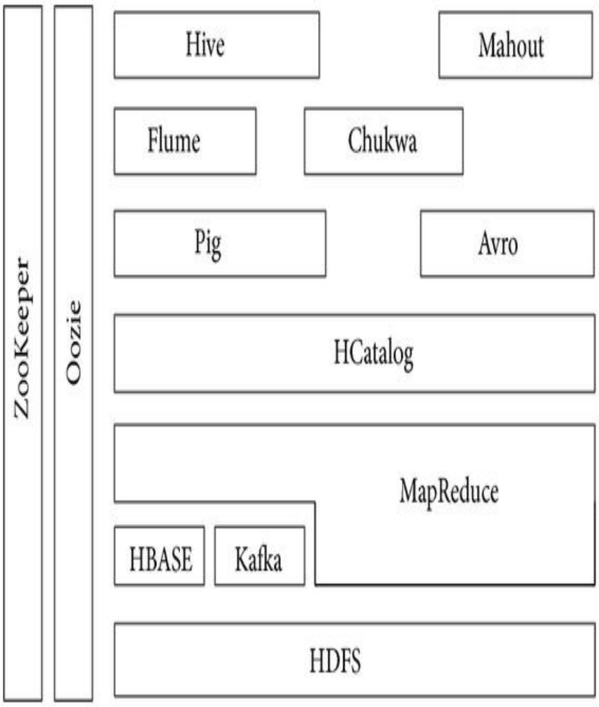


Fig-3 Components of Hadoop

HDFS - There HDFS has two different types of nodes name-node which is acts like master, second node data node acts like slave node. HDFS is the component of Hadoop . It is more complex and given complexities and uncertainties , the problem in this component too much of data focus the single system. HDFS have secondary name node to facilitate to handle large amount of data.

HBASE - is a open source, the result of the operation produce large data sets When it is involved in read and write operation . API-(Application programming interface) is used interface with HBASE.These component allow the user to perform read and write operation in all rows and small subset of all colume used.

Zookeeper - This component contains master and slave node. It is called as distributed service provider. Zookeeper involve to configures, and maintain huge amounts of data. This component has facility to provide 1. synchronization which means different clock Paulse used to data transfer between input output devices CPU. 2. group services another facility of Zookeeper. This component is used activate the distributed processes to control and manage a name space of the data registers.

Two different master and slave node used in distributed system for split the file system based on hierarchy.

Hcatalog - is mainly used to create no of tables for storage purpose.These component combine with other components such as pig, MapReduce and Hive. Hcatalog used to expand expand HBase, which is establish the user communication using HDFS data at the same time enable the data sharing between tools and execution platforms.

*Hive*QL - is own query language of Hive, Hadoop ecosystem has no of platform , Hive is the subplatform in the Hadoop ecosystem.This query language executed with the help of MapReduce then activate the user-defined functions (UDFs). These component contain has three data structures: partitions, bucket and tables. These related structures used to store different information.

Pig - This is component of Hadoop ecosystem. Pig generates High level scripting language, the own data type of Pig used to represent semistructured data. Which is operates at runtime.

Mahout - This component is used to support data mining and machine learning algorithms. There are four different types of groups are used to mining patterns from data warehouse such as collective, filtering, categorization, and clustering.

Oozie - used for execution and job flow. Oozie used to complete various task. Which is based on Apache Hadoop framework. tasks are completed with the help of directed acyclic graph (DAG).

Chukwa - is a framework which if used for data collection and analysis both are important to relate with MapReduce and HDFS. Chukwa framework mainly from its development stage to processing. This component collects, perform certain operation using distributed systems and stores them into Hadoop ecosystem. It is not support other module Chukwa is took part in distribution of Apache Hadoop.

Avro - passes data from one stage to another stage. The qualities of Avro are suited to scripting languages such as Pig.

Flume - This component mainly used to send large amounts of data in and out of Hadoop system. There are two different types of nodes are 1. sources and 2. sinks. Sources include Avro, files system and logs, whereas sinks refer to HDFS. Flume transforms data before it shuttled into the sink

Table 1: Functionalities and component of Hadoop

Hadoop component	Functions
(1) HDFS	➢ Storage ➢ Replication
(2) MapReduce	➢ Distributed processing ➢ Fault tolerance
(3) HBASE	➢ Increased speed of read/write access
(4) Hcatalog	➢ Metadata
(5) Pig	➢ Support for Scripting languages
(6) Hive	➢ SQL (Structure Query Language)
(7) Oozie	➢ Workflow ➢ Scheduling
(8) ZooKeeper	➢ Coordination
(9) Kafka	➢ Messaging ➢ Data integration
(10) Mahout	➢ Machine learning ➢ Artificial intelligence

V BIG DATA HEALTHCARE

Today number of research work in big data based on healthcare. Big data for healthcare explain what the part of engineers and researchers is. We know that engineers and researchers have same mind set, only they focus to define the problem and what type of machine learning algorithm used to develop big data technologies

They problem solving is different from knowledge discovery. VPH approach is applied in personalized healthcare VPH approach is used to overcome the limitation of biology founded observational data affected by issues. We will explain how big data methods strength and technologies support and empower VPH approaches. The fundamental difference between researchers and engineers, engineers deal with problem when they find fragility and mistrust, the medical researchers have well established body of knowledge. Clinical researcher's mechanistic models "too simple ".

In the following we discuss how VPH works.VPH approach which is used to prove mechanistic models.VPH provide accurate prediction Big data can help to transform biomedicine competition into collaboration, significantly VPH approach accepted for clinical practice. Modern big data technologies involved to analyze a large data set of thousand patients in short time using clusters and correlations, modern predictive model developed by using statistical and machine learning techniques [2], [3], Here new approaches are called FRAX, these methods not only count new patients, but also collected different information correlation related to patients.

The main problem in this FRAX is mechanistic knowledge is quite incomplete when we identify the by empirical modeling to elaborate mechanistic theory. Today big data for healthcare used "grey-box" model to deal with patients problem. These methods combine mechanistic knowledge with phenomenological model and provide solution for physiology, biomechanics, biochemistry and biophysics. This model switch to overcome problem in FRAX large amount of validated knowledge used to identify patient with specific data. In many times mechanistic models very expensive so we move to reduced-order model (or) also called Meta models. This model used to store input-output in a data repository this model accurately replaces expensive model cheaper /faster. Another approach is used for Nonlinear Auto Regression moving average model for nonlinear system identification.

VI TECHNIQUES OF BIG DATA

Analysis of techniques in Big Data explain clear concept with the help fishbone diagram. It is a big data technique consist of Data base, Hadoop platform, management., Data Set, Integrated Development Environment (IDE), Computing, Algorithm and File System. IDE is a tool which is related to software application and these tool consists of default tools like build automation ,debugger, source code editor, IDE help us for intelligent code compilation, these are the different technique in Big Data , to identify cause effect is help to design product or identify the factors in fish bone diagram, this technologies are used to solve the problem in techniques .the cause effect method fits perfectly should support technical analysis of project management in the IPTECH platform. IPTECH is digital technology related company which is support the clients to deliver e-business solutions.

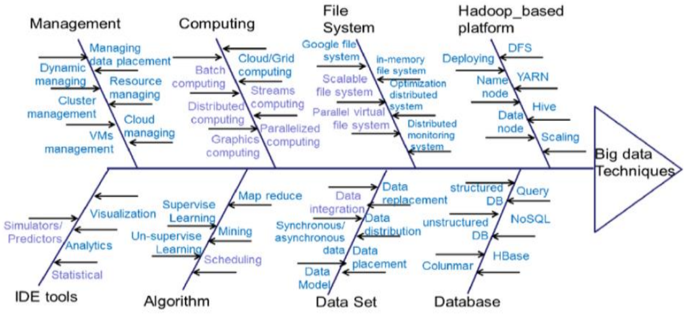


Fig-4 Fishbone diagram of Big Data techniques

VII CONCLUSION

Discussed what is the meaning of big data, what are the different components in big data, how to handle problem with big data tools and advantages of heterogeneous data types. The problem solved using variety of techniques Here big data create chance for upcoming researchers to reduce the problem. Big data help us to create innovative ideas in different field such as health care machines, Banking transaction, Social media and Satellite imaging. The researchers don't want to repeat mistake. Big data create chance for upcoming researchers, because this platform is very interesting, and provide facility, invoke to continue research in specific applications

REFERENCES

[1] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," Nature, vol. 457, no. 7232, pp. 1012–1014, 2009.

[2] A. Wright, "Big data meets big science," Common. ACM, vol. 57, no. 7, pp. 13–15, 2014.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh,

[4] H. Byers, "Big data: The next frontier for innovation, competition, and productivity," Mckinsey Global Inst., 2011.

[5] D. Noble "Genes and causation" Philos. Trans. A Math. Phys. Eng. Sci. vol. 366 no. 1878 pp. 3001-3015 Sep. 2008.

[6] A.Martin computational fluid dynamics simulations of cerebrospinal fluid motion in the cervical spine" PLoS One vol. 7 no. 12 pp. e52284 . I. Yiallourou J. R. Kroger N. Stergiopulos D. Maintz B. A. Martin A. C. Bunck "Comparison of 4D phase-contrast MRI flow measurements 2012.

[7] E fracture risk calculation: The FRAX algorithm," *Curr. Osteoporos. Rep.*, vol. 7, no. 3, pp. 77. V. McCloskey, H. Johansson, A. Oden, and J. A. Kanis, "From relative risk to absolute –83, Sep. 2009.

[8] D. Noble, "Genes and causation," *Philos. Trans. A Math. Phys. Eng. Sci.*, vol. 366, no. 1878, pp. 3001–3015, Sep. 2008.

[9] O. Johnell and J. A. Kanis, "An estimate of the worldwide prevalence and disability associated with osteoporotic fractures," *Osteoporos Int.*, vol. 17, no. 12, pp. 1726–1733, Dec. 2006.

[10] A. Wright, "Big data meets big science," *Commun. ACM*, vol. 57, no. 7, pp. 13–15, 2014.