

# Significance of Big Data on Healthcare and DataSecurity

Anish Hemmady

**Abstract-**Nowdays Internet is a great platform for information where user can search what they want but the main problem is some information present on internet is in unstructured format or semi structured format. When user wants any information from retrieval system then he must have data in structured format so that performance efficiency of system gets enhanced. The percentage of unstructured data is growing. This document exemplifies the work carried out towards contribution of big data technology. The facts given below are based on literature survey carried out by myself. Also how Big Data can be used in various fields and its implementation to these fields. Databases of Healthcare sector are huge so this technology plays an important role pertaining to this sector. Big data concept can be very useful for evolution of whole mankind and development of our future. So this Document explores new ideas regarding big data concept. In this paper we have discussed various significance of huge data to healthcare mainly and how unstructured format can be converted to structured format through filtering process I have proposed.

**Index Terms-**implementation, pertaining, evolution, explores, massive data, Big Data, Splunk, Skip Lists.



## 1. INTRODUCTION

Big data is the heart of Modern technologies and sophisticated gadgets. Premier Scientific groups are Intensely focused on it, as evidenced in [1] which is a work of American Statistical association. This whole talk about the term Big Data started way back in 1970 when the Data more than 1 kilobytes was considered Big or Huge, then slowly in 1980 the Data size grew which changed the significance of big data then slowly it grew from kilobytes to megabytes and then to gigabytes. In the last 10 years the data size substantially increased to terabytes which is a lot in terms of Human computing Activities. Society at large is also Interested on it, as documented by major reports in the business and popular press such as [2].

### 1.1 WHY THE TERM 'BIG DATA'

As we discussed the history of big data and how its size grew manifold times throughout these years. The term Big Data was coined by Roger Magoulas as in [3]. Big Data means massive or huge quantity of

data. There are thousands of servers all around our globe and thousands of databases. It's hard to keep a track of all these Databases and we are not able to optimize the usefulness of these databases. Millions of kilobytes of data is wasted through these years as in [4]. So by the year 2020 zettabytes of data will be flowing through the databases so to control such surplus amount of data the term big data is used referencing its amount.

## 2. HOW TO TACKLE BIG DATA

Predominantly there were no problems in tackling huge data collection but now the world has changed a lot since last five to eight years. This is due to changing needs of human beings and expansion of technologies. Nowadays every person has mobiles, personal computers, so data size has increased. In 2005 there was a surge of technologies which would handle huge databases. The Google company first strategized a technique named Map and Reduce. This technology exponentially and drastically changed the world of databases and the world of Internet.

### 2.1. Map and Reduce Technique

Velocity, Volume and Variability are the three things that one should keep in his mind while tackling with massive Data. These three things play an Important role while Judging the concept of big data this technique conceptualized the principle of Master and slave. Also this technique consists of segregation and breaking chunks of data. It consists of two masters and there are n slaves each of the slave consists of Node number which specifies which slave is active. Indexing method is used while data is stored in database also hierarchical tree structure is used for storage of files and documents. Later on this technology was used in Hadoop Systems which now currently manages google, facebook, linkedn and lot more companies.

## 2.2. Hadoop technology

The hadoop system is an apache server as in[5] It consists of map reduce technique and hadoop file systems which are the tools used to manage huge chunks of data. It consists of up the collected information and sends it to output device. The advantages of this technology are

- a. It has inbuilt fault tolerance keeping in mind hardware failures. It stores atleast three copies of Data so there is backup incase there is slave failure. It works by batch processing of input data
- b. It breaks bigger tasks into smaller tasks
- c. It works on distributed systems.

- d. It supports scalability.

This above description is given as per[6]. Hadoop was developed by yahoo but it consists of map and reduce technique developed by google. It is heavily used by Government offices, enterprises.

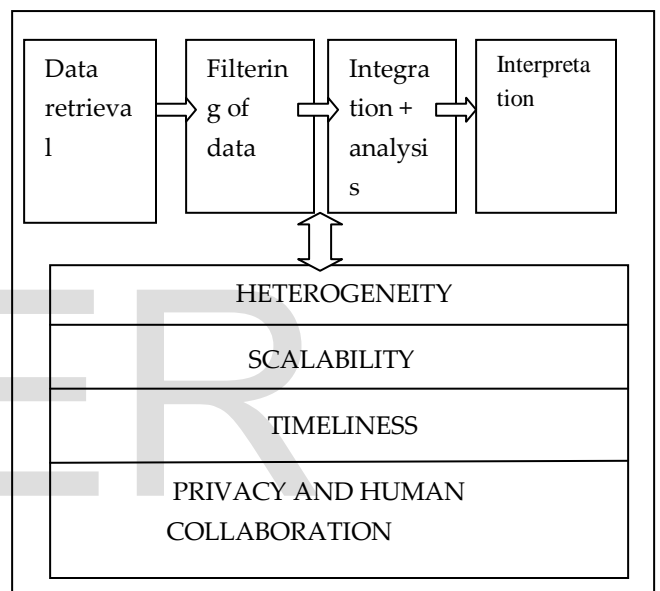


Fig. 1. The above figure illustrates pipelining in analysis of big data

## 3. USES OF BIG DATA IN VARIOUS

### FIELDS

The sumptuous amount of data collected through various databases around the world have be brought in proper use of mankind.Major amount of data are obtained from hospitals and medical care centres.

### 3.1. Healthcare

I shall discuss in this document about the technological advancement in healthcare systems and the use of massive data this technologies operate on.According to fred trotter in his article as in[7] due to more and more technological advancements sometimes errors are caused in healthcare like overtreatment,undertreatment of patients which he refers as hacking of healthcare.We can reduce these errors by analyzing biomedical technologies and patients response to these systems by storing their data into a database just like storage of wine for many years we have improved its storage quality and quantity.We can have e-patients which are solely depended on advancement of technology.These patients are minute to minute informed about their choices they have to make while eating food this is possible only through large data collected through various systems and refined to its proprer use.

Big Data can be used in bioinformatics and systems biology.Its the goal for medical research according to Ranjan perera who is scientific director of analytical genomics and bioinformatics at Sanford burnham medical centre according to[8].The different fields are genomics,proteomics,metabolomics,lipidomics.The numbers behind these fields involved are huge approximately 20,300 coding genes.As we learn more about how genes respond to the various treatment therapies can be decided based on ones gene structure according to Adam ghozdik,so by collecting large amount of data around the world we can improve our strategies towards various diseases.We can use series of layers to collect huge amount of data since at molecular level large amount of data is generated.According to Sumit chanda as in[9] who studies cellular proteins involved in influenza A and retrovirus in HIV implements series of system level approaches to study large amount of data at molecular level.storage problems are caused due to large amount of incoming data due to which storage capacity of processors available have become less and data has become more causing disproportion between data and available memory of processor.

By studying various data around the world we can devise a method for curing grave disorders and

diseases like asthma,aids and many more which are incurable till today.Sanford burnham scientists are facing problems for storing such large flow of incoming data eventhough the capacity for storing data at that centre is 10.4 petabyte as in[10].

In future we may need yottabytes of data storage capacity which means humungous quantity of data is flowing through current databases and servers.Previously in olden times if a person staying in village had any ailments there were no doctors available for him but now the times have changed doctors can stay connected to them or even cure them by staying millions of miles away just by using conference call this rapid development in technology has drastically changed the lives of many villages across the globe.

### 3.2 Data security

Data security means keeping data stored in databases secure which means there should not be any leakage of data.Several techniques are used to secure the data like loosless decomposition technique in database.proper authentication and credentials should be provided while constructing a database.proper pl sql queries should be implemented while securing databases also view techniques should be used while showing employees and customers their information.Different techniques are implemented by large database storing companies like hadoop so that there is no private information leaked.Nowdays people are hacking in to databases of stock market also.The biggest hacking into systems include sonys playstation network in 2011 as in[11].so how does hadoop security work .hadoop security follows some protocols which are necessary while keeping data secured.These include

- a. Enforcement of hadoopfile system file permission
- b. Delegation of tokens for subsequent authentication check.
- c. Block access tokens for access control to data blocks
- d. Pluggable authentication for http web consoles and network encryption

This means data security plays an important role in conserving the databases hence saving so many peoples private lives.

## 4. SOLUTIONS FOR 'BIG DATA' CRISIS

The solution for huge data crisis is unknown till now. In this document I shall show some of the ways we can solve this problem. First we have to understand why this crisis is caused its because various formats of files and documents are available in today's times because of it there is variable form of data. Do we should first converge all these formats into single format which we can easily filter out. Steps we should take while making database according to me are

1. proper disintegration of data and at the end proper integration of data with proper and verified output.
2. Redundancy of data should be strictly avoided as it fill in the void spaces of databases which leads to wastage of space
3. proper tracking of data should be implemented by various algorithms. Tree hierarchy and indexed structures should be used.

### 4.1 Skip lists Technique

According to me we can devise a faster searching, retrieving and storing algorithms which can insert and delete even if infinite data is incoming. One method is to use SkipLists data structure which are like linklists but better and faster than linklists as they have different levels of nodes. So each block of data can be inserted quickly by checking these levels rather than searching the entire line of numbers also we can visit maximum number of nodes using skiplists. Given below is a small algorithm of skiplists for generating random level structure

Algorithm of skiplists:

Steps:

1. Make all the nodes level 1

$J=1$

2. While the number of nodes at level  $j > 1$  do
3. For each  $i$ th node at level  $j$  do

If  $I$  is odd

If  $I$  is not the last node at level  $j$

Randomly or devise a code to promote it to level  $j$

4. Else do not promote
5. Else if  $I$  is even and node  $i-1$  was not promoted then promote it to level  $j+1$

### 4.2 Splunk technology alongwith Map Reduce

Splunk is general purpose search, analysis and reporting engine. Splunk is deployed for managing various functions, applications and so on. Splunkmapreduce can be combinedly used for collecting, storing and processing data. A lot of processing and compiling time will be saved thus helping big data unstructured format easily convertible to structured format.

MASSIVE DATA INSERTED

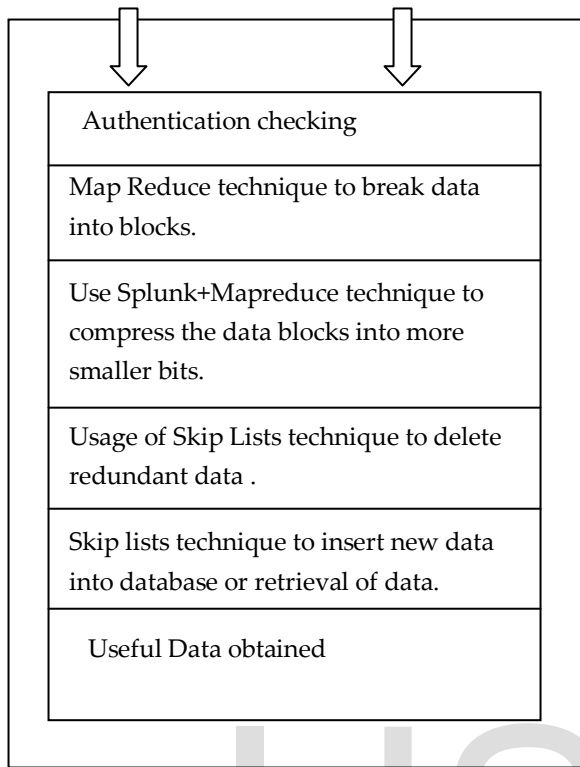


Fig. 2. The above mentioned could be filtering process for huge data

## 5. CONCLUSION

In this document I have highlighted the facts about Big data, its terminology, significance to our healthcare and data security systems and so on I have illustrated a diagram for filtering process of huge data. Proper sense about BIG DATA is brought up in this document. This document consists of different inferences to technology we have known and what advancements we can make according to me. The massive data filtering process helps unstructured format of data to be converted to structured format so that information retrieval system gives efficient and effective output.

Thus BIG DATA filtering process can be a success in near future helping towards bright future of our world.

## 6. REFERENCES

- [1] SallieAnnKeller, Steven E koonin, Stephanie shipp ,“Big Data and city living-what can it do for us”Wileys Significance magazine ISSN-1740-9713/ vol.no 9 issue 4 ,page.no 4-7,2012.
- [2] Steve Lohr,“How Big Data became so Big”, New York Times,Aug 2012.
- [3] Roger Magoulas,Ben Loriga, “Technologies and techniques for large scale data” Oreilly Radar ISSN-1935-9446/issue 11,page.no 32,feb 2009.
- [4] Fluidcreativity “Data Wastage Report”, 15 Apr 2011.
- [5] Lam,Chuck “Hadoop in action” Manning publications ISBN-1-935182-19-6/first edition,page .no 325,22 June 2009.
- [6] HasanMir “zeroprotraining.co” Apr-2013
- [7] FreddTrotter,David Ulham,“Hacking Healthcare:A guide to standards,workflows and meaningful use” Oreilly Media ISBN-978-1449305024 /page.no 248, 5 aug 2011.
- [8] communications staff “The impact of Big data on medical research” SanfordBurnhamScience Blog, 25 june 2013
- [9] Sumit Chanda “The impact of Big data on medical research” SanfordBurnhamScience Blog, 25 june 2013
- [10] communications staff “The impact of Big data on medical research” SanfordBurnhamScience Blog, 25 june 2013
- [11] Kevin T. Smith “Big Data Security-The evolution of hadoops security model” infoq.com,14 Aug 2013.

Author :Anish Hemmady

Bachelor Of Engineering Student

Stream:Computer Engineering

Currently studying in Senior Year

College:Pillais Institute Of Information Technology

Location:New Mumbai