International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 05, May 2017
www.ijates.com
ijates
ISSN 2348 - 7550

BIG DATA : AN ANALYSIS OF TOOLS

Amanpreet Kaur

Assistant Professor in Computer Science

Guru Nanak College For Girls,SriMuktsarSahib,Panjab(India)

ABSTRACT

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in area.These type of data set are available in social websites(facebook,twitter,youtube,linkedin and others),Wikipedia, search engines,web sales etc. Examples of such data may include books, journals, documents, metadata, health,attachment in emails and so on .In this paper, I am discussing the dimensions of big data ,its background and analytic tools of big data as traditional tools are not used to abstract knowledge and their problems.

Keywords: Bigdata,Hadoop,NoSql,Textmining,Machine learning

LITERATURE SURVEY

From last many years researcher has completed there work and published numbers of articles and papers in this field.In fact, the amount of digital data that exists is growing at a rapid rate, doubling every two years, and effectively changing the way we live. According to IBM, 2.5 billion gigabytes (GB) of data was generated every day in 2012.An article by Forbes states that Data is growing faster than ever before and by the year 2020, about 1.7 megabytes of new information will be created every second for every human being on this planet.The social media marketing industry report,2015 states that face book users 52%,linkedin 21%,twitter 13%,youtube 4% and others 10 %with this increase in social media the range of data is hiking every second.In 2005 Roger Mougalas from O’Reilly Media coined the term Big Data for the first time, only a year after they created the term Web 2.0. It refers to a large set of data that is almost impossible to manage and process using traditional business intelligence tools.In 2005 the Hadoop was created by Yahoo! built on top of Google’s MapReduce. It’s goal was to index the entire World Wide Web and nowadays the open-source Hadoop is used by a lot organizations to crunch through huge amounts of data.As more and more social networks start appearing and the Web 2.0 takes flight, more and more data is created on a daily basis. Innovative startups slowly start to dig into this massive amount of data and also governments start working on Big Data projects. In 2009 the Indian government decides to take an iris scan, fingerprint and photograph of all of 1.2 billion inhabitants. All this data is stored in the largest biometric database in the world.In 2011 the McKinsey report on Big Data: *The next frontier for innovation, competition, and productivity*, states that in 2018 the USA alone will face a shortage of 140.000 – 190.000 data scientist as well as 1.5 million data managers.

I. INTRODUCTION

The world know is instrumented, interconnected and intelligent as in today's era people and things are interconnected so there is a double digital rates as digital circuits are inexpensive ,we add intelligence to all most everything. As Organization has vital information in the form of structured, unstructured and semi-structured form of data .Big data applies to the information(Structured) that cannot be processed using traditional processes and tools. Structuredreferred to the information with high degree such a information is stored in large databases that is included for internal operation.Structured data is relatively simple to enter, store, query, and analyze, Analysts typically use simple Excel spreadsheets or Structured Query Language (SQL) to perform queries on structured data within relational databases. The data is stored in the data warehouse it must shine with respect to quality as it uses software ETL tools (extract, transformand load) are used to

- Extract data from homogeneous or heterogeneous data sources
- Transform the data for storing it in proper format or structure for querying and analysis purpose
- Load it into the final target (database, more specifically, operational data store, data mart, or data warehouse)Usually in ETL tools, all the three phases execute in parallel

Unstructured data may have its own internal structure, but does not conform into a spreadsheet or database such a data processed to be stored as big data that helps in knowing interest ,plan and decision making for future it act as source for sale,customer interest such as Spotify: an on-demand music service, uses Hadoop big data analytics is not as reliable but used to collect data from its millions of users worldwide and then uses the analyzed data to give informed music recommendations to individual users utilizing big data.Amazon Prime, which is driven to provide a great customer experience by offering, video, music and Kindle books in a one-stop shop also heavily, utilizes big data.Facebook uses big data as

The Flashback memories:Facebook offered its users the option called the 'Flashback', in this video there is a collection of photos and posts that received the most comments and likes by your friends by setting background music.

I Voted:Facebook successfully tied political activity to user engagement when they came out with a social experiment by creating a sticker allowing its users to declare "I Voted" on their profiles. Tracking cookies: Facebook tracks its users across the Web by tracking cookies of the user. If a user is logged into Facebook and simultaneously browses the Web, Facebook can track the sites they are visiting. It shows you the related topics,advertisement,groups and so on.

Abstract data from LIKES: A recent study conducted showed that is viable to predict data accurately on a range of personal attributes that are highly sensitive just by analyzing the 'Likes' that have been clicked by a user on Facebook.The work has been conducted by researchers at Cambridge University and Microsoft Research shows how the patterns of Facebook 'Likes' can very accurately predict your sexual orientation, satisfaction with life, intelligence, emotional stability, religion, alcohol use and drug use, relationship status, age, gender, race and political views among many other. So, it is important to manage the information and analyze the data there are numbers of big data analysis tools used to handle this.

II. BIG DATA AND PARAMETERS

Big Data: Big Data refers to humongous volumes of data that cannot be processed effectively with the traditional applications that exist. The processing of Big Data begins with the raw data that isn't structured and is impossible to store in the memory of a single computer. It is immense volumes of data, both unstructured and structures. Big Data is something that can be used to analyze insights which can lead to better decision and strategic business moves.

In a 2001 research report and related lectures, META Group (now Gartner) defined data growth as Vscharacteristics of big data are defined as Vs

2.1 Volume

It refers to just about size in Gigabyte, Terabyte, Petabytes to zettabytes .Big data size is a constantly moving target, from a few dozen terabytes to many petabytes of data. AsFacebook is generating one terabyte of data in sec.Its a large amount of load to handle. Volume of data being stored is exploding as in year 2000 it was 80,000 petabytes of data and in year 2020 will be 35 zettabytes of data.The data is increasing day to day but percent of data to be analyzed is on decline.

2.2 Variety

It represent all types (formats,text,database,pictures,excel,pdf) of data may be in relational or non-relational. In the companies only 20% of data is structuredand rest 80% is unstructured.The data is in different formats so difficulty to store in structured format, it is challenging feature in big data.

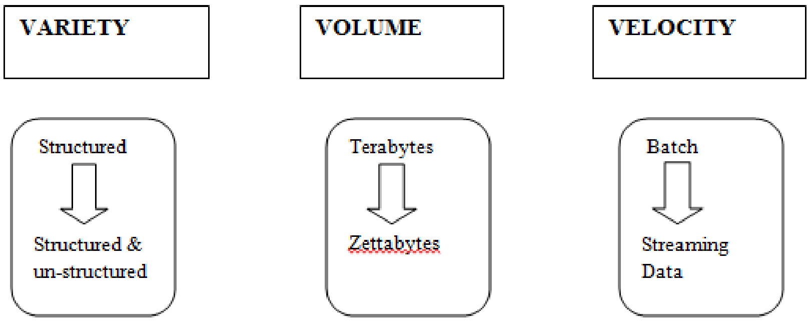

FIGURE I

Characteristics of Big Data

2.3 **Velocity:**It considers how quickly the data is arriving and stored .The speed at which the data is flowing.As the Speed in twitter is 14, 300 tweets per second, 2 000 pictures on Instagram per second,Facebook processes 750TB/day of data.IBM has revealed the new concept of stream computing. In which you does n't only used running queries against static data but also get continuous updated result from GPS data referred in real time.

In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization such a high amount of Vs are challenging factor in the big data

III. ANALYSIS TOOLS IN BIG DATA

Traditional tools such as relational databasesystem, Data mining and Data warehousing are not capable of handling large amount of unstructured data. There are some specialized tools used for this purpose as data sources are unpredictable,multi-structured and massive.

3.1 Hadoop

Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. Hadoop is important as its ability to store and process huge amounts of any kind of data, quickly. With data volumes and varieties constantly increasing, especially from social media and the Internet of Things (IoT), that's a key consideration.Hadoop's distributed computing model processes big data fast. The more computing nodes you use the more processing power you have. Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.Unlike traditional relational databases, you don’t have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.The open-sourceframework is free and uses commodity hardware to store large quantities of data. You can easily grow your system to handle more data simply by adding nodes. Little administration is required.

This open source software platform managed by the Apache Software Foundation has proven to be very helpful in storing and managing vast amounts of data cheaply and efficiently. Basically, it’s a way of storing enormous data sets across distributed clusters of servers and then running “distributed” analysis applications in each cluster.Big Data applications will continue to run even when individual servers or clusters fail. And it’s also designed to be efficient, because it doesn’t require your applications to shuttle huge volumes of data across your network.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly available service on top of a cluster of computers, each of which may be prone to failures.

3.2 NoSQL

A NoSQL referring to "non SQL", "non relational" database provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases. NoSQL encompasses a wide variety of different database technologies that were developed in response to the demands presented in building modern applications. Developers are working with applications that create massive volumes of new, rapidly changing data types: structured, semi-structured, unstructured and polymorphic data. The NoSQL Database Types used are

3.2.1Document databases: pair each key with a complex data structure known as a document. Documents can contain many different key-value pairs, or key-array pairs, or even nested documents.

3.2.2.Graph stores : are used to store information about networks of data, such as social connections.

3.2.3.Key-value stores : are the simplest NoSQL databases. Every single item in the database is stored as an attribute name (or 'key'), together with its value.

3.2.3.Wide-column stores : It store columns of data together, instead of rows.

3.3 R (PROGRAMING LANGUAGE)

Is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis sets a limit on the most memory it will allocate from the operating system

memory.limit()

?memory.limit

memory.size()

We use memory.size() to change R's allocation limit. Memory limits are dependent on your configuration. If you are running 32-bit R on any OS, it will be 2 or 3Gb .If you're running 64-bit R on a 64-bit OS, the upper limit is effectively infinite, you still shouldn't load huge datasets into memory – Virtual memory, swapping, etc.Under any circumstances, you cannot have more than 2,147,483,647 rows or column. So only R is not enough to handle the large database.

3.4 Text Mining

Text Mining is a process established to obtain information from unstructured texts. With the help of linguistic, statistical and mathematical processes, patterns and structures are selectively sought and information extracted by Text Mining. The text mining is used as digital information is increasing highly professional Text Mining software solutions provide support here, helping everywhere where data and information are found in text documents and not in databases.

3.5 Machine Learning

Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed. The evolution of machine learning is from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. They learn from previous computations to produce reliable, repeatable decisions and results .While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data.

VI. CONCLUSION

I have described how big data is massively increasing know a days and helping organizations to understand Big Data and more and more companies are slowly adopting and moving towards it. The concept of big data is generated from the multiple sources forming the complexity such as volume, velocity and variety. I have described various analysis tools in big data, its uses in face book, music, and amazon.Big data leads towards intelligence and lead towards intelligence decision making and prediction towards personal interest from unstructured data.

REFERENCES:

[1].Chen,MIN,Shiwer Mao and YunhaoLiv “Big Data:ASurvey”Mobile Network And Applicationsn19.2(2ing 014):171-209

[2.] NielRaden,Hired Brains INC:Big Data Analysis Architecture

[3.] Puneet Singh Duggal,SanchitPaul,Department of Computer Science&EngneeringBirlaInstitute of Technology Mesra,Ranchi,India,” Big Data Analysis: Challenging and Solutions”,International conference on Cloud, Big Data and Trust 2013,Nov 13-15,RGVP.

[4.] AdityaB.Patel,ManashviBirla,Ushma Nair,(6-8 Dec 2012), “Adressing Big Data Problem using Hadoop and Map Reduce”.

[5] Paul C.Zikopoulos,ChrisEaton,DirkDeroos,ThomasDeulseh,GeorgeLapis:Understanding Big data-analystics for Enterprise Class Hodoop and Streaming Data.

[6.] https://en.wikipedia.org/wiki/Programming_with_Big_Data_in_R

[7.] https://www.sas.com/en_us/insights/analytics/machine-learning.html