ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 5, Issue 6, June 2018

Use of Data Mining Technique in Unstructured Data of Big Data

Nilambari Bansod, Harshal Gandhi

Dept. of Computer Engineering, RGCOE, Savitribai phule Pune University, Takali Dhokeshwar, Parner, Ahmednagar, Maharashtra, India

ABSTRACT: Big Data is data of high volume and high variety being produced or generated at high velocity which cannot be stored, managed, processed or analysed using the existing traditional software tools, techniques and architectures. With big data many challenges such as scale, heterogeneity, speed and privacy are associated but there are opportunities as well. Potential information is locked in big data which if properly leveraged will make a huge difference to business. With the help of big data analytics, meaningful insights can be extracted from big data which is heterogeneous in nature comprising of structured, unstructured and semi-structured content. One prime challenge in big data analytics is that nearly 95% data is unstructured. This paper describes what big data and big data analytics is. A review of different techniques and approaches to analyse unstructured data is given. This paper emphasizes the importance of analysis of unstructured data along with structured data in business to extract holistic insights. The need for appropriate and efficient analytical methods for knowledge discovery from huge volumes of heterogeneous data in unstructured formats has been highlighted.

KEYWORDS: Big Data, Unstructured data, Text Analytics, Audio Analytics, Video Analytics, Social Media Analytics.

I. INTRODUCTION

Big data has caught attention of professionals, academicians and researchers since it came into light. This paper explores various aspects of big data and big data analytics. Various definitions have come up during this phase when professionals throughout the world were putting in endeavors to understand and state what big data is. The first thing that comes to mind is its size but it is much more than that.

IDC defines big data as "Big data technologies describe a new generation of technologies and architecture designed to economically extract value from very large volumes of a wide variety of data, enabling high velocity capture, discovery and/or analysis " in 2011 [1]. TechAmerica Foundation defines big data as “Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.”[2].

The traditional data management and analysis systems like Relational Database Management Systems (RDBMS) are not suitable and adequate to process big data as they are based on structured data which is a very small fraction of big data and secondly because they are not scalable to the extremely high rate of generation of big data.

The description of big data is incomplete without mentioning the three V’s of big data which are volume, variety and velocity. These are the fundamental characteristics of big data as given in Fig. 1[3].

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 5, Issue 6, June 2018

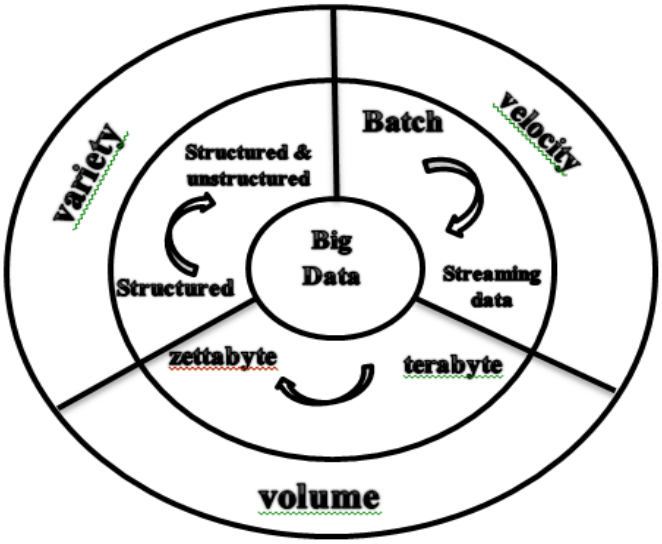

Figure.1 Characteristics of Big Data

1. **Volume –** It refers to the high magnitude of big data which is in the order of terabytes to petabytes and more. For instance, some earlier estimates suggested that 20 petabytes of storage space was used to store 260 billion Facebook photos. In 2010, it was reported that up to one million photographs were processed by Facebook per second [4]. Twitter generates 12terabytesofdatadaily[5].In2012,Facebookstatedthat2.7 billion "likes" and "comments" were registered daily by the users [6].
2. **Variety –** Massive volumes of data of heterogeneous nature is generated by different sources. It consists of structured, semi-structured and unstructured data. Structured data has a fixed format whereas unstructured data has no fixed schema orformat.
3. **Velocity –** Big data is being generated continuously at an exponential rate. 90% of current data is generated in last two years [7, 8]. Social media is one major contributor which is generating data explosively. Sensors, smart phones and internet are leading to huge datafeeds.

The explosive rate of growth of big data presents tremendous opportunity and will yield big economic gains if correctly exploited. The resources needed for exploiting big data are falling short because of the high rate of growth of data. Eric Schmidt, the CEO of Google in the Lake Tahoe Technology Conference held in 2010 [9] quoted that "Between the dawn of civilization through 2003, just 5 exabytes of information was created. That much information is now created every two days and the pace is increasing. People aren't ready for the technology revolution that's going to happen to them." According to a recent study, such amount of data is being generated in every 10 minutes now [10]. It has also been estimated that more than 85% of Fortune 500 organizations will fail to exploit big data for competitive advantage [11]. They will lag behind the 15% organizations that will leverage big data.

II. RELATED WORK

In [1] author has discussed the meaning and importance of big data analysis programming tool use forbig data mining and important of big data , with the example of facebook we can understood that today it isrequired to process large number of data sets ,our traditional data sets are not enough for that ,for exampleinstead of taking large MySql tables we can use caching approach from memcached for n tier elements as Mysqlhas very good performance in read but they are lagging in write ,which leads us to very high reliability but lowpartition tolerance in our CAP model ,another example author has given is Yelp which uses AWS and Hadoopfor data analysis which uses Amazon S3 server to store large datasets which is RAID service.

The author proposed such data analysis using Apache Hadoop and JSON and data stored From Amazon web services using their web services and analyze the data, the analysis showed that this method can analyze the large data from different sources with minimum utilization of resources.

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 5, Issue 6, June 2018

In [2] In this paper author has utilized Nosql database Mongo db to implement the big data analysis asit is advantageous over rigid sql tables which is not useful in today's large scale data for web logs generatedevery day .more over author has compared performance between Mongo db and HDFS frame work using inbuilt map reduce method with mongo db , author has not defined the modern data store technology and integration available with hadoop like Hbase , and HIVE for that experiments and results are shown for large amount of data sets , this is the motive why we choose mongo db data store for Large data sets .With this framework proposed by author the output comparison shows that Figure shows the effect ofthe split size on performance using mongo-hadoop. The number of input records is _ 9:3 million, or 4GB ofinput data. With the default split size of 8MB, Hadoop schedules over 500 mappers; by increasing the split size,we are able to reduce this number to around 40 and achieve a considerable performance improvement. The curvelevels o_ between 128MB and 256MB, so we decided to use 128MB as the split size for the rest of our tests bothfor native Hadoop-HDFS and mongo-hadoop.

In , MS At el.[3] has discussed various security issues and threats available with Big data as data is inzeta byte size it also contains some sensitive and confidential information it is necessary to prevent unauthorizeduse of data so apart from storage retrieve and processing security is also an important concern for data mining ,data application from social web ,consumer oriented work has large impact on big data security according toauthor vast use of smart phones has increased photo uploading and other sensitive information on web it is anissue for that author has proposed metadata analysis in big data which creates an index of each images uploadedon social web and we can identify from link which gives confidentiality over social media, so each images canbe scanned from big data bases of social media and can be apply for future security policies .

In [4] After considering security in analysis we again come with our problem of analysis the big datawith this paper integration of NOSQL with big data analysis author proposed model of unity architecture foranalysis.

In [5] In this paper author has discussed about some very important parameters of mongo db focusingon CAP model and compared various types of data store available with no Nosql and tested them among variousbusiness intelligent system provided , and concluded that Nosql data stores provides huge opportunity where Sqldata bases are not useful basic advantages are their scalability and cross node operation .the intersectionalgorithm for mongo db states the effectiveness of mongo db data store for key value approach to modelize thedata.

In [7] [10] and [11] some practical approaches are shown to interact no sql data stores with varioussystems such as distributed architecture[7] ,Hashdoop[11] and evolution in hadoop [10] are proposed indistributed system data bases are handled by structure system but it fails when data items increase sounstructured data stores are useful for such problems some major industries are capable to develop their ownunstructured data stores for ex. Google's Big table, Yahoo's PNUTS ,Hadoop's Hbase and many more but whatabout small industries ,author stated that there are many open source products are available to handle.

III. UNSTRUCTURED DATA

These data contain complex information such as Email attachments, Images comments on social networking sites. These data cannot be easily analysed. Doug Lancy was the first one talking about 3v's in big data management [3]:
Volume - It describes the amount of data. It refers to mass quantities of data.
Variety - It describes different types of data and sources including structured, semi-structured and unstructured data.
Velocity - It defines the motion of data. Data created rapidly, processed and analysed.

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 5, Issue 6, June 2018

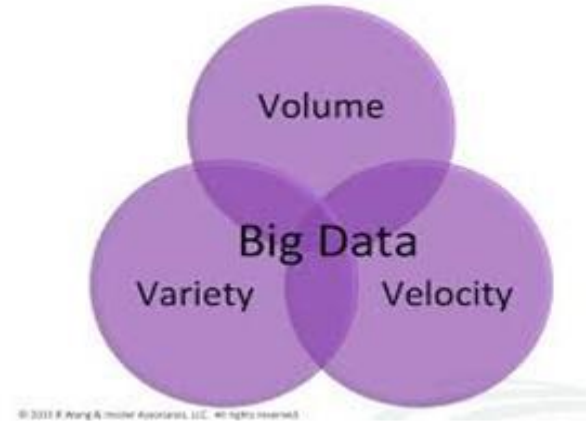

Figure 2. Unstructured 3V Big Data Management

When the capacity of Data Warehouse grew from 50 GB to 1 TB – 100TB. Data was in structured form when it creates from many organizations. Data goes from three properties like volume, Variety and velocity. Many companies were facing the problem on how to expand the capacity of data warehouse to accept the new requirement.

IV. BIG DATA ANALYTICS

Big data analytics enables organizations to analyze a mix of structured, semi structured and unstructured data in search of valuable business information. Makinsey's internal Think- Tank, the Mckinsey Global Institute, published a major study in June 2011 on Big Data [12]. Its overloading conclusion: Big Data is "a key basis of competition and growth". The term Analytics (including its Big Data form) is often used broadly to cover any data-driven decision making [8]. The term analytics divided into two groups: Corporate analytics and Academic research Analytics. In Corporate Analytics, Team uses their expertise in statistics and Data mining. In Academic Analytics, Researchers analyze data to test Hypotheses and form theories [8].
In Big Data Analytics, Researchers found that the generated data divided into various Big Data application such as follows [2].
1 Structured Analytics-In structured analytics, large quantity of data is generated from business and scientific research fields. These data is managed by RDBMS, Data warehousing, OLAP and BPM. Data grown by various research area like Privacy preserving data mining, E-commerce.
2 Text Analytics-In Text analytics, Text is one of the most common forms of storing the information and it includes Email communication, documents, and Social media contents. Text analytics also known as Text mining, refers to the process of extracting useful information from large text. Text mining system is based on text representation and Natural Language Processing (NLP) with emphasis on the latter [2].
3 Web Analytics-The aim of Web analytics is to retrieve, extract the information from Web Pages. Web Analytics also called Web mining
4 Multimedia Analytics Recently multimedia data, including images, audio, and video has grown at a tremendous rate. Multimedia analytics refers to extract interesting knowledge and semantics captured in multimedia data. Multimedia analytics covers many subjects like Audio Summarization, Multimedia annotation, Multimedia indexing and retrieval.
5 Mobile Analytics Mobile data traffic increased 885PBs Per Month at the end of 2012. Vast volume of application and data leads to mobile analytics.

V. SIMULATION RESULTS

In 21stcentury Big Data is the modern kind of electricity power that transforms everything it touches in business, government, and private life. Every day, we generate more than 2.5 quintillion bytes of data and 85% of the data in the world today has been created in the last two years only. In which 80% of data captured today is unstructured such as climate information, data post to social media sites, digital pictures and video, purchase transaction records and getting GPS signals from cell phone. All of this is example of unstructured data is Big data.In 2010, Google estimated that

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 5, Issue 6, June 2018

every two days at that time the world generated as much data as the sum it generated up to 2003. In spite of the very recent "Big Data Executive Survey 2013" by New Vantage Partners [15] that states "It's about variety, not volume", lots of people with author would still believe the prime issue with Big Data is scale or volume.Big data has a great variety of data forms: text, images, videos, sounds, and data which have extreme scale. Big data frequently comes in the form of streams of a variety of types.The growth of data will never stop. According to the 2011 IDC Digital Universe Study, in 2005 there were created and stored 130 exabytes of data. The amount grew to 1,227 exabytes in 2010 and is projected to grow at 45.2% to 7,910 exabytes in 2015[14].In addition to being the hottest new trend in business and government, Big Data is fast becoming a persistent force in modern science. American president Barack Obama administration started a $200M Big Data in Science scheme with the goals of improving economic growth which creates lots of jobs in various sector such as education, health, energy, environmental, public safety, and global development.

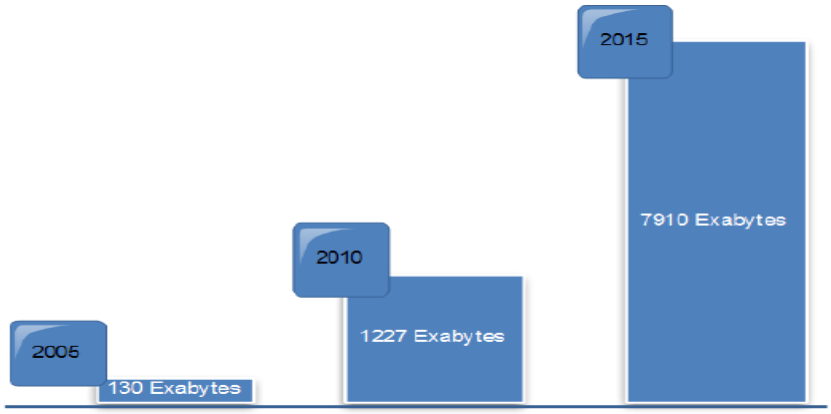

Figure 3. A Decade of Digital Universe Growth: Storage in Exabyte

VI. CONCLUSION AND FUTURE WORK

Big data has invaluable information which if extracted completely can make a big difference to business gains. In this paper, various aspects of big data analytics along with the role of unstructured data have been discussed. Various application areas have been described. The analytical techniques and approaches of unstructured data such as text, audio and video have been reviewed. These techniques include information extraction, summarization, question-answering and sentiment analysis for text analytics, the transcript-based and phonetic- based approach for audio analytics, automatic video indexing and retrieval for video analytics and content based and structure based approaches, community detection, social influence analysis and link prediction for social media analytics. It has also been emphasized that there is a need of knowledge discovery from unstructured data along with structured data in businesses to gain competitive edge over others. This fact has been supported by giving various real-life examples. Unstructured data has a wealth of information and is in a very large fraction in comparison to structured data. In predictive analytics, both structured and unstructured data hold significance for extracting knowledge and making predictions in business. Cost effective and efficient tools and techniques are needed to analyse structured and unstructured data in real time. With data increasing over time, opportunities are arising in every field whether it is businesses, medicines, research, weather forecasting, etc. Data is a natural resource these days and if utilized effectively, the future will yield cost- effective and critical insights to all mankind problems. We intend to work in this area in future and come up with better approaches to big data analytics using unstructured data.

REFERENCES

1. Jyoti Nandimath , Ankur Patil , Ekata Banerjee , PratimaKakade :"Big Data Analysis using Apache Hadoop " In SKNCOE Pune India,2013
2. E. Dede, M. Govindaraju ,D. Gunter, R. Canon, L. Ramakrishnan "Performance Evaluation of a MongoDB and HadoopPlatform for Scientific Data Analysis" In Lawrence Berekely National Lab Berkeley, CA 94720

ISSN: 2395-7639

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

(A Monthly, Peer Reviewed Online Journal)

Visit: www.ijmrsetm.com

Volume 5, Issue 6, June 2018

3. Matthew Smith, Christian Szongott,BenjaminHenne, Gabriele von Voigt" Big Data Privacy Issues in Public Social Media",2013
4. Ramon Lawrence :"Integration and Virtualization of Relational SQL and NoSQL Systems including MySQL and MongoDB"At 2014 International Conference on Computational Science and Computational Intelligence,2014
5. Laurent Bonne, Anne Laurent, Michel Sala,BenedicteLaurent,NicolasSicard:" REDUCE, YOU SAY: What NoSQL can do for Data Aggregation and BI in Large Repositories "In 2011 22nd International Workshop onDatabase and Expert Systems Applications,2011
6. AlexandruBoicea, Florin Radulescu, Laura IoanaAgapin" MongoDBvs Oracle - database comparison "2012 Third International Conference on Emerging Intelligent Data and Web Technologies,2012
7. Suyog S. Nyati, ShivanandPawar ,Rajesh Ingle: " Performance Evaluation of Unstructured NoSQL data over distributed framework" 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI),2013
8. LiorOkman,Nurit Gal-Oz,, YaronGonen,, Ehud Gudes ,Jenny Abramov:" Security Issues in NoSQL Databases" 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11,2011
9. ChanchalYadav, ShuliangWang ,Manoj Kumar: Algorithm and approaches to handle large Data- A Survey. IJCSN International Journal of Computer Science and Network, Vol 2, Issue 3, 2013
10. RuxandraBurtica, Eleonora Maria Mocanu, MugurelIonutAndreica, NicolaeTapus: Practical application and evaluation of no-SQL databases in Cloud Computing , ©2012 IEEE
11. RomainFontugne,,Johan Mazel, Kensuke FukudaHashdoop: A MapReduce Framework for NetworkAnomalyDetectionCERN – European organization for nuclear Research 2014 IEEE INFOCOMWorkshops: 2014 IEEE INFOCOM Workshop on Security and Privacy in Big Data,2014