BIG DATA: SECURITY ISSUES AND CHALLENGES

[1] **Naveen Rishishwar,** [2] **Vartika** [3]**Mr. Kapil Tomar**
[1,2] Student, Department of Information Technology
[3] Asst. Professor, Department of Information Technology
[1,2] Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, INDIA.
[3] Shri Ram Murti Smarak College of Engineering and Technology, Bareilly, Uttar Pradesh, INDIA.
[1] naveenrishishwar@gmail.com, [2]vartika0112@gmail.com [3]kkapiltomar@gmail.com

***Abstract*— Now a days data is very essential for every company, organization, business and even for any newly startup. Data helps to all to take important decisions related to their growth. When the size of data is increases beyond the limit.When it becomes so large that it is very difficult to manage by the traditional database software. Then we called it Big Data. Big Data deals with unique computational, scalability, storage and processing challenges. Big Data manage the huge amount of data, store it and further process that data in a very efficient manner. So that a valuable information can be retrieved from that huge data to take any decision. Even then many organizations are not using Big Data because there are some issues and challenges which have to be addresses. In this paper, we discussed major available issues and challenges in Big Data. We also discussed some methods to deal with them.**

***Index Terms*—** synthetic polymer.

I. Introduction

Big data is usedto store data just like old method like MYSQL, SQL& many more. It is more fast & useful than previous language. Manipulation rate is fast and easy to manage.

According to Gartner- “Big data is high volume high velocity and high variety (structure, unstructured, semi structure) information assets that requires a new form of processing to enable enhanced decision making,insight discovery and process optimization. Big data is too big, too fast, and too hard for existing technology. Too big means data base of size more than Peta Byte (1000 Tera-Byte). Too fast means quick processing of its requests, too hard means there is no existing tool which is capable of fulfilling all the types of requirement (storage and processing) of big data. [1]

The term “Big Data” is related with managing (manipulating) high amount of data exist in digitalized form that is collected by various companies or organization. For making record so they can easily track their operation. In previous few year growth rate of data is increasing very fast it’s increasing with growth rate of 40% per year that is very high growth rate. We have to do something to for manage it. In 2012 about to zettabyte 1021 byte. In 2020 it will be nearly equal to 45zb. As shown in fig.

Every day we are creating 2.6 Exabyte data

is growing at a 40% compound annual rate nearly45zb in 2020.

.

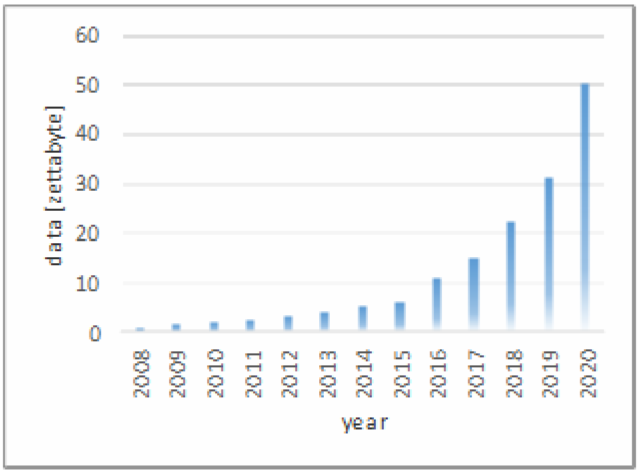

Fig. 1 Data generation in various years

This is not a specific technology. It is a collection of attributes and capabilities. Big data can be described by[2][3]

a. Volume
b. Variety
c. Velocity
d. Veracity
e. Volatility

A. Volume

In today’s scenario rate of the generation data is too much fast. Not only human but also machine is also generating the data. Examples are sensor, machine, network, ecommerce data, human interaction with system like social networking site all these are generating the huge amount of data. Facebook generate data in bulk.

On Facebook [5] [6]

Over 4 million post are posted at every minute that is 250 million post every hour,1,00,000 friend request made on Facebook every minute, Facebook support 140 language, 890 people login there Facebook account daily. [4]

1) On Vine

That is not part of Facebook’s empire but still produce huge amount of data, 17,000 videos are played on vine every second also produces high amount of data.

2) Twitter

Twitter is second largest social networking site that also generate the data in bulk like, Approximate 60000 tweets are tweeted each second. [5]

Tinder

Tinder users swipe 10000 matches per second.

3) Pinterest:

Over 10000 pics are pined each second, there are 146 million fashion brands on Pinterest, over 176 million Pinterest accounts have been registered.

4) Instagram:

Instagram Reports 90 Million Active Monthly Users, 8,500 Likes Per Second, 1000 comment per second, 6 million photo per minute,1,000 Comments Per Second, Instagram currently has more than 430 million active users

5) LinkedIn

25 million profiles are viewed every day, more than 3 user signup every second on LinkedIn, there are 300 million people user on LinkedIn, Average users spend 17 minutes on LinkedIn per month.

6) YouTube

YouTube has 1,300,000,000 active users, 5 hour videos are uploaded to YouTube every second,60000 videos are watched every second on YouTube, the average number of mobile YouTube video views per day is 1,000,000,000. [6]

B. Variety

Variety refers for various type of data that is generated by different type of resources like sensor, ecommerce data, photos, audios,videos, email, pdf, that data may be structured or unstructured, we have to store all type so we are going to use big data that is able to do it. Big data is the proper solution of many type of problems that traditional medium have for storing data. [7]

C. Velocity

Velocity refer for the speed of generation of data & its processing to achieve the demand. Velocity deals with data rate at which data processed in various sources like business process, machine, network, social media sites like facebook, twitter, mobiles devices and etc. Data flow is large and continuous in manner. [8]

Researcher& organization management people need this real time data for operating theirorganization, because real time data is very useful for them to operate organization. [9]

D. Veracity

Data should be accurate as much as possible, just because inaccurate data may lead us in direction of our decay. Maintaining accuracy in data is the biggest challenges. Because if that is not accurate that will be very harmful for. Suppose if Flipkart is not having accurate information about the stoke they are still selling that product while that product is out of stock it happensjust because of inaccurate data that can harm company's reputation too. No one wantsto destroy their ownreputation because this may lead them for loss. [10] [11]

E. Volatility

What is the duration of data, it stored at server is known as volatility in Big data. After how much time, data will not be available to use to analysis or to mining. Big Data deals with many real time issues like volume, variety and velocity of the data. There may be so many source of data with different formats. [12]

II. Application of Big Data

1.Understanding and Targeting Customers in E- commerce

E commerce is a very vast and main area big data use today. Big data help in e commerce to understand the customer's needs& there requirements, it also help for targeting the customer for which companies are going to make the products, when we get to know who is going to use their productsand then they start making the products as per their customer's needs to enhance their profit and revenue.

2.Performance Optimization and Improving Health using smart devices

Not only government and companies but we as an individual's now can be benefited from the data that wearable device use and then produce the result.. All these collect the data & help us to improve our health. [13] [14]

Ex. Jawbone is an armband that track our sleeping hours, calorie consumption & send that information to their server & according to your data give you feedback what should you do or not. They are giving you feedback because they have the lots of data & by the help of that data they are able to give you feedback. Each night the collect the data that is equivalent to approximate 60year. For storing such huge data, we use the big data to store all this information. All these data help us to improve our health by previous data record.

3. To make smartCities

Big data is used to improve many aspects of our daily life our surroundings as well as cities and countries where we live.

For example, it allows cities to optimize traffic based on real time traffic information as well as social media and weather data. A number of cities are currently piloting big data analytics with the aim of turning themselves into Smart Cities, where all the needful things can be implement to ease our daily life. Where you can check the status of a bus with available seats, route and fare details also. [19]

Another example of the city of Long Beach, California, where they are using smart water meters to detect illegal watering in real time and have been used to help some homeowners cut their water usage by as much as 80 percent. That's vital when the state is going through its worst drought in recorded history and the governor has enacted the first-ever state-wide water restrictions. [17 [18]

4. Improving Security and Law Enforcement
Big data improves the various security features. We all know as our Indian government uses also uses the data to track criminal record & mitigate crime. [22]

For such polices our government start making Adhar card. By that Adhar card they can easily identify the criminal because they have finger print of all Indian & other information so they can control the crime & make our country safer.

5.Improving Science and Research
Big Data bring the various new possibilities by which Science and research is currently transforming. Which is very helpful to discover the new things. Before that it was very tuff to process and analysis the huge amount of data of various formats but now with the help of Big Data it is easy to work on this. [15] for example, CERN, the nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator. Experiments to unlock the secrets of ouruniverse - how it started and works - generate huge amounts of data.
The CERN data center has 65,000 processors to analyze its 30 petabytes of data. [16]

6.Understanding and Optimizing Business
It is very necessary to understand the the need of the business and then optimize the process so that the production and profit both can be increased by using technologies. The process of production how much they need to manufacture and how much they need to cut the production. All these decisions are made on the basis on previous data. They also analyze the supply chain & use the sensor to track goods & delivery vehicles & optimizing the route by using live traffic data. [20]

7.Improving SportsPerformance
Now a days many sports are using big data analytics.There are many smart devices that are track our various activities, get various data from sensors attached to the smart devices and provide feedback , result to us so that we can easily maintain good health and help to improve the game performance. [21]
We have the IBM SlamTracker tool for tennis tournaments; we use video analytics that track the performance of every player in a football or baseball game, and sensor technology in sports equipment such as basket balls or golf clubs allows us to get feedback (via smart phones and cloud servers) on our game and how to improve it.
One of the really cool new things I have come across is a smart yoga mat sensors embedded in the mat will be able to provide feedback on your postures, score your practice, and even guide you through an at-home practice. [23]

8. Optimizing Machine and Device Performance
Big data analytics help machines and devices become smarter and more autonomous. For example, big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings. We can even use big data tools to optimize the performance of computers and data warehouses[7] [8].
Xcel Energy initiated one of the first ever tests of a " smart grid" in Boulder, Colorado, installing smart meters on customers' homes that would allow them to log into a website and see their energy usage in real time. The smart grid would also theoretically allow power companies to predict usage in order to plan for future infrastructure needs and prevent brown out scenarios[8] [9].
In Ireland, grocery chain Tesco’s has its warehouse employees wear armbands that track the goods they take from the shelves, distributes tasks, and even forecasts completion time for a job[11].

III. Reason to use big data

A. It handles a petabyte of data or more
B. It has distributed redundant data storage
C. Can leverage parallel task processing
D. Can provide data processing (MapReduce or equivalent) capabilities
E. Has extremely fast data insertion
F. Has central management and orchestration
G. Is hardware agnostic
H. Is extensible where its basic capabilities can be augmented and altered.

IV. Issues & Challenges using big data

Nothing is perfect each and every thing have their own merit and demerit(pros and cons) so big data also have their own. Some of them are given below with their possible solution

A. Storage issues
Can you imagine how much amount of data is generated daily? Hard to say isn’t it? But according to 2015 big data info graphic contributed by Ben Walker of Voucher Cloud, around 2.5Quintillion (1018byte) [8]
Such high amount of data is able to fill 10 million Blue ray disk. If we stacked all these disc, then its height is 4 time more than Eiffel tower. Generation rate of data is too much fast in 1992 that was 100 Gb per day but in 1997 this become 100Gb per hour. In 2013 it become 28,875 Gb per hour, Experts said that by 2018 rate if generation of data will be 50,000 Gb per second [9] [10]. For storing such high data is not an easy task. [24].

1) Solution for storing the data
A team of researchers at France’s Institute Charles Sadron and Aix-Marseille University has succeeded in coding binary data into the strand of a synthetic polymer, a minuscule chain of chemical information about 60,000 times thinner than a strand of hair. and capability to store 1 zettabyte of information in just 10 grams of matter.[11] [6][9]
Right now, storing one zettabyte (1 billion terabytes) takes roughly 1000 kilograms of cobalt alloy, the material used in

hard drives. A zettabyte of synthesized polymer would be about 10 grams.

B. Security

One of the biggest challenges with data is data security. Suppose if someone unauthorized get the confidential (credential) data like bank account data, atm pin, nuclear weapon data etc. that will be more devastating consequencesbecause that person& firm with their dangerous intention they can harm us too badly. So privacy & security should we high. [25]

Almost all data security issues are caused by the lack of effective measures provided by antivirus software and firewalls. These systems were developed to protect the limited scope of information stored on the hard disk, but Big Data goes beyond hard disks and isolated systems. Major nine Big Data Security Challenges[26]

- In Big Data Most distributed systems computations have only a single level of protection, which is not recommended.
- Non-relational databases (NoSQL) are actively evolving, making it difficult for security solutions to keep up with demand.
- Automated data transfer requires additional security measures, which are often not available.
- When a system receives a large amount of information, it should be validated to remain trustworthy and accurate; this practice doesn’t always occur, however.
- Unethical IT specialists practicing information mining can gather personal data without asking users for permission or notifying them.
- Access control encryption and connections security can become dated and inaccessible to the IT specialists who rely on it.
- Some organizations cannot – or do not – institute access controls to divide the level of confidentiality within the company.
- Recommended detailed audits are not routinely performed on Big Data due to the huge amount of information involved.
- Due to the size of Big Data, its origins are not consistently monitored and tracked. [8]

C. Processingissue in Bigdata

Data generation rate is too much fast as we earlier discussed by 2018 it will be approximate to 50,000 Gb per Second at that time processing speed will be a major issue, how we will manage such huge data in just 1 second. We really need an advance system that will help us in such fast processes.

D. Privacy in Big Data

One of biggest challenges not only for this particular technology but for all technology is privacy. Data should be as much secure as possible so it not going to be decrypt easily & hence our data will be remaining secure if anyone get our data. [27]

1) e.Redundancy

Data redundancy is also a major issue. In this situation same data is stored in data base more than one time. This increase the size of database and slow down the processing speed also some time create problem like if we update our information at one place and accidently pick redundant data from other place it may lead dangerous result, also generate inconsistency. Redundancy may occur accidently. We have to avoid such type of situation for better use data base. [28].

CONCLUSION

The amount and variety of data is growing very rapidly because of sudden increase of the social sites, search engines, multi media sharing sites, various stock exchange trading sites, online gaming , online survey sites, and various news Sites and so on. Big Data is becoming the very popular research area for scientific data research and for business applications.

Big data analysis helps companies, organizations, government agencies to take better decisions, to predict and to identify new opportunities in their business. In this paper we discussed about the issues and challenges related to big data which helps organizations to better understand their customers and the marketplace and to take better decisions and also helps researchers and scientists to extract useful knowledge out of Big data. In addition to that we introduce some solution for the existing issues and challenges. SO that number of users will increase. That will help also to the research scholars to choose the best upcoming research are for their work.

REFERENCES

[1] L. M. Vaquero L. Rodero-Merino J. Caceres M. Lindner "A break in the clouds: towards a cloud definition" ACM SIGCOMM Comput. Commun. Rev. vol. 39 no. 1 pp. 50-55 2009.

[2] S. Kamara K. Lauter "Cryptographic cloud storage"RLCPS January 2010 Springer.

[3] A. Singhal "Modern information retrieval: A brief overview" IEEE Data Engineering Bulletinvol. 24 no. 4 pp. 35-43 2001.

[4] I. H. Witten A. Moffat T. C. Bell Managing gigabytes: Compressing and indexing documents and images May 1999 Morgan Kaufmann Publishing.

[5] D. Song D. Wagner A. Perrig "Practical techniques for searches on encrypted data" Proc. of S&P2000.

[6] E.-J. Goh "Secure indexes" <em>Cryptology ePrintArchive2003: http://eprint.iacr.org/2003/216.

[7] Y.-C. Chang M. Mitzenmacher "Privacy preserving keyword searches on remote encrypted data" Proc. of ACNS2005.

[8] R. Curtmola J. A. Garay S. Kamara R. Ostrovsky "Searchable symmetric encryption: improved definitions and efficient constructions" Proc. of ACM CCS2006.

[9] D. Boneh G. D. Crescenzo R. Ostrovsky G. Persiano "Public key encryption with keyword search" Proc. of EUROCRYPT2004.

[10] M. Bellare A. Boldyreva A. ONeill "Deterministic and efficiently searchable encryption" Proc. of CRYPTO2007.

[11] M. Abdalla M. Bellare D. Catalano E. Kiltz T. Kohno T. Lange J. Malone-Lee G. Neven P. Paillier H. Shi "Searchable encryption revisited: Consistency properties relation to anonymous ibe and extensions" J. Cryptol.vol. 21 no. 3 pp. 350-391 2008.

[12] J. Li Q. Wang C. Wang N. Cao K. Ren W. Lou "Fuzzy keyword search over encrypted data in cloud computing" <em>Proc. of IEEE INFOCOM'10 Mini-ConferenceMarch 2010.

[13] D. Boneh E. Kushilevitz R. Ostrovsky W. E. S. III "Public key encryption that allows pir queries" Proc. of CRYPTO2007.

[14] P. Golle J. Staddon B. Waters "Secure conjunctive keyword search over encrypted data" Proc. of ACNSpp. 31-45 2004.

[15] L. Ballard S. Kamara F. Monrose "Achieving efficient conjunctive keyword searches over encrypted data" Proc. of ICICS2005.

[16] D. Boneh B. Waters "Conjunctive subset and range queries on encrypted data" Proc. of TCCpp. 535-554 2007.

[17] R. Brinkman "Searching in encrypted data" <em>PhD thesis2007.

[18] Y. Hwang P. Lee "Public key encryption with conjunctive keyword search and its extension to a multi-user system" Pairing2007.

[19] J. Katz A. Sahai B. Waters "Predicate encryption supporting disjunctions polynomial equations and inner products" Proc. of EUROCRYPT2008.

[20] A. Lewko T. Okamoto A. Sahai K. Takashima B. Waters "Fully secure functional encryption: Attribute-based encryption and (hierarchical) inner product encryption" Proc. of EUROCRYPT2010.

[21] E. Shen E. Shi B. Waters "Predicate privacy in encryption systems" Proc. of TCC2009.

[22] C. Wang N. Cao J. Li K. Ren W. Lou "Secure ranked keyword search over encrypted cloud data" Proc. of ICDCS' 102010.

[23] S. Yu C. Wang K. Ren W. Lou "Achieving secure scalable and fine-grained data access control in cloud computing" Proc. of INFOCOM2010.

[24] C. Wang Q. Wang K. Ren W. Lou "Privacy-preserving public auditing for data storage security in cloud computing" Proc. of INFOCOM2010.

[25] S. Zerr E. Demidova D. Olmedilla W. Nejdl M. Winslett S. Mitra "Zerber: r-confidential indexing for distributed documents" Proc. of EDBTpp. 287-298 2008.

[26] S. Zerr D. Olmedilla W. Nejdl W. Siberski "Zcrberr: Top-k retrieval from a confidential index" Proc. of EDBTpp. 439-449 2009.

[27] Y. Ishai E. Kushilevitz R. Ostrovsky A. Sahai "Cryptography from anonymity" Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Sciencepp. 239-248 2006.

[28] W. K. Wong D. W. Cheung B. Kao N. Mamoulis "Secure knn computation on encrypted databases" Proceedings of the 35th SIGMOD international conference on Management of datapp. 139-152 2009.