IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, January 2024

Big Data : Analysis

Ms. Manali Sakpal

Institute of Distance and Open Learning, Mumbai, Maharashtra, India

Abstract*: The amount of data in world is growing day by day. Data is growing because of use of internet, smart phone and social network. Big data is a collection of data sets which is very large in size as well as complex. Generally, size of the data is Petabyte and Exabyte. Traditional database systems is not able to capture, store and analyse this large amount of data. As the internet is growing, amount of big data continues to grow. Big data analytics provide new ways for businesses and government to analyse unstructured data. Now a days, big data is one of the most talked topics in IT industry. It is going to play important role in future. Big data changes the way that data is managed and used. Some of the applications are in areas such as healthcare, traffic management, banking, retail, education and so on. Organizations are becoming more flexible and more open. New types of data will give new challenges as well. The present paper highlights important concepts of Big Data. In this write up we discuss various aspects of big data. We define Big Data and discuss the parameters along which Big Data is defined. This includes the three V's of big data which are velocity, volume and variety. The authors also look at processes involved in data processing and review the security aspects of Big Data and propose a new system for Security of Big Data and finally present the future scope of Big Data.*

Keywords: Petabyte, Zettabytes, Veracity, Valence Rest, Rollback Attack, Sybil Attack, Database, Velocity

I. INTRODUCTION

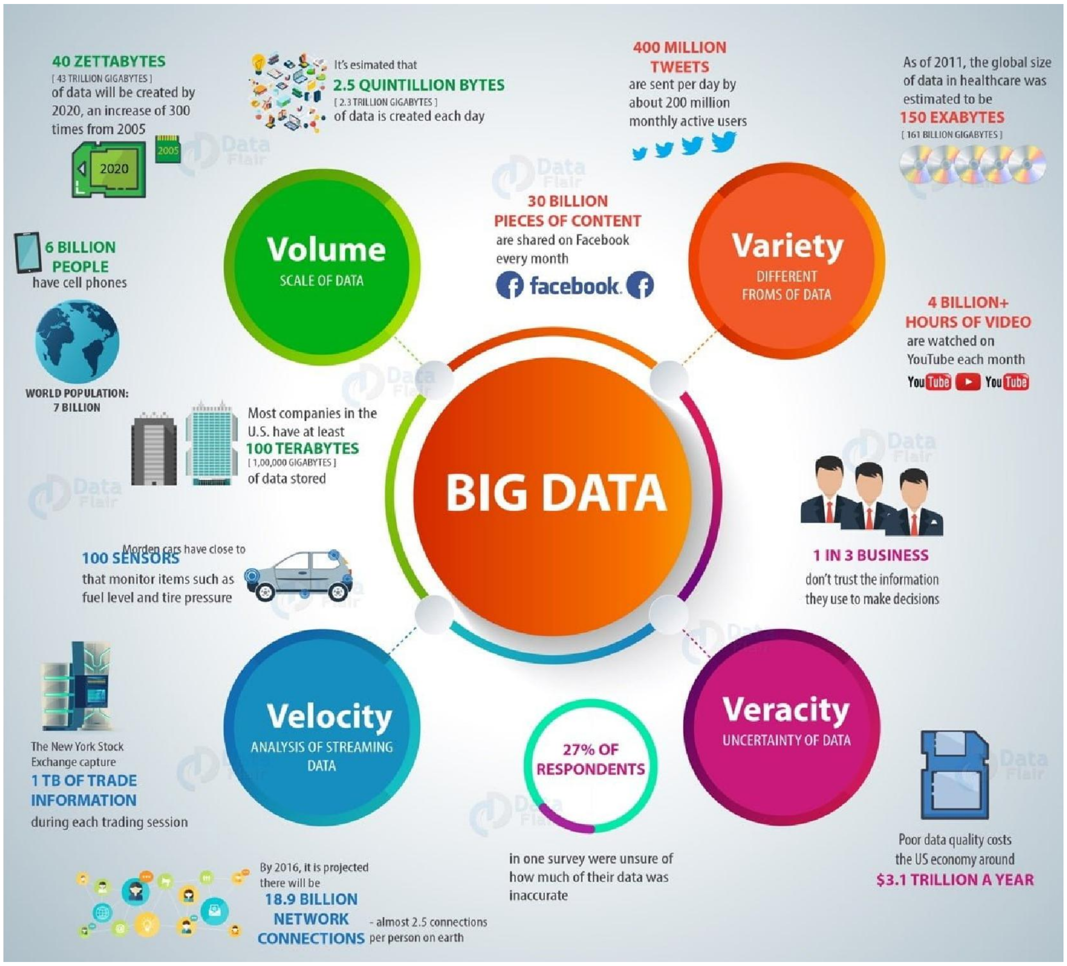
Big data is a collective term referring to data that is so large and complex that it exceeds the processing capability of conventional data management systems and software techniques. However with big data come big values. Data becomes big data when individual data stops mattering and only a large collection of it or analyses derived from it are of value. With many big data analyzing technologies, insights can be derived to enable better decision making for critical development areas such as health care, economic productivity, energy, and natural disaster prediction The term Big Data appeared for the first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey having the title Big Data and the Next Wave of Infra Stress. The first book mentioning Big Data is a data mining book that came to fore in 1998 too by Weiss and Indrukya. The first academic paper having the word Big Data in the title appeared in the year 2000 in a paper by Diebold. The era of Big Data has bought with it a plethora of opportunities for the advancement of science, improvement of health care, promotion of economic growth, enhancement of education system and more ways of social interaction and entertainment. But as is said everything has its flip side as well big data too has its issues. Security and privacy are great issues in big data due to its huge volume, high velocity, large variety like large scale cloud infrastructure, variety in data sources and formats, data acquisition of streaming data, inter cloud migration and others. The use of large scale cloud infrastructure having a varied number of software platforms across large networks of computers increases the region of attack to an all new level of the entire system. The various challenges related to big data and cloud computing and its security and privacy issues and the reasons why they crop up are explained later in details.

Distributed & Heterogeneous Big Data

Big data refers to huge, heterogeneous, distributed and often unstructured digital content that is difficult to process using traditional data management tools and techniques. The term encompasses the complexity and range of data and data types, real-time data collection and processing needs, and the value that can be obtained by smart analytics. Big data also has new sources, like machine generation (e.g., log files or sensor networks), mobile devices (video, photographs, and text messaging), and machine-to-machine. Characteristics of big data are volume, variety, velocity,

DOI: 10.48175/IJARSCT-15074

veracity and value. Volume, It estimates that 2,500,000,000,000,000,000 bytes of data are created now each day. Velocity, increasing data rates because of network bandwidth. Variety additional unstructured data types. Veracity of data, large set of decisions and analysis. Finally, the Value can be extracted and analysed for useful data findings.


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's esimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data is created each day
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
Volume
SCALE OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
facebook
Variety
DIFFERENT FROMS OF DATA
6 BILLION PEOPLE
have cell phones
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
YouTube
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[1,00,000 GIGABYTES]
of data stored
BIG DATA
Morden cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
1 IN 3 BUSINESS
don't trust the information they use to make decisions
Velocity
ANALYSIS OF STREAMING DATA
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
The New York Stock Exchange capture
1 TB OF TRADE INFORMATION
during each trading session
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
- almost 2.5 connections per person on earth
Data Flair

When compared to relational database records big data is less structured. Highly developed data mining techniques and associated tools can help take out information from huge, difficult datasets and create new insights in big data mining techniques in a limited time [15]. When dealing out a query in big data, speed is an important claim [16]. However, the process may take time because mostly it cannot traverse all the related data in the whole database in a short time. In this case, index will be an optimal choice. At present, indices in big data are only aiming at simple type of data, while big data is becoming more complicated. The combination of appropriate index for big data and up-to-date preprocessing technology will be a desirable solution when we encountered this kind of problems. There is several application paradigms used for big data problems. The traditional serial algorithm is inefficient for the big data. Cloud‟s reduced cost model to use hundreds of computers for a short time costs of the big data application. By using online big data application, a lot of companies can greatly reduce their IT cost. However, security and privacy affect the entire big data storage and processing, since there is a massive use of third-party services and infrastructures that are used to host important data or to carry out critical operations. The scale of data and applications grow exponentially, and bring huge challenges of dynamic data monitoring and security protection. Unlike traditional security method, security in big data is mainly in the form of how to process data mining without exposing sensitive information of users. Besides, current technologies of privacy protection are mainly based on static data set, while data is always dynamically changed, including data pattern, variation of attribute and addition of new data. Thus, it is a challenge to implement effective privacy protection in this complex circumstance.

DOI: 10.48175/IJARSCT-15074

II. BIG DATA GOALS

Big data helps to achieve various goals, which are the following: 1. Cost Reduction Hadoop is a framework for storing huge amount of data on distributed clusters. In Hadoop cluster, one year storage cost for one terabyte is $2,000. That is 800 times less than the traditional relational databases. 2. Time Reduction Macy's merchandise pricing optimization application calculates data sets in seconds or in minutes which actually can take hours for calculation. 3. Support in Internal Business Decisions The main idea of big data is to assist in the interior company decisions like, what kind of new products should be offered to people? How much stock should be detained? And what must be the cost of our item? 4. Developing New Big Data-Based Offerings Big data must be used to create new products and offerings. LinkedIn is the top example, which has used big data to develop products and offerings, including jobs you may be interested in, who have viewed my profile, people you may know, and numerous others. These ideas have pulled people to LinkedIn.

III. BIG DATA APPLICATIONS

1. Understanding and Targeting Customers

This is one of the biggest and most publicized areas of big data use today.

Here, big data is used to better understand customers and their behaviors and preferences.

Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers.

The big objective, in many cases, is to create predictive models. Using big data, Telecom companies can now better predict customer churn; Wal-Mart can predict what products will sell, and car insurance companies understand how well their customers actually drive. Even government election campaigns can be optimized using big data analytics. Some believe, Obama's win after the 2012 presidential election campaign was due to his team's superior ability to use big data analytics.

2. Understanding and Optimizing Business Processes

Big data is also increasingly used to optimize business processes.

Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts.

One particular business process that is seeing a lot of big data analytics is supply chain or delivery route optimization. Here, geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic data, etc.

3. Improving Healthcare and Public Health

The clinical trials of the future won't be limited by small sample sizes but could potentially include everyone.

Big data techniques are already being used to monitor babies in a specialist premature and sick baby unit.

By recording and analyzing every heart beat and breathing pattern of every baby, the unit was able to develop algorithms that can now predict infections 24 hours before any physical symptoms appear.

4. Improving Sports Performance

Most elite sports have now embraced big data analytics.

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.

Many elite sports teams also track athletes outside of the sporting environment – using smart technology to track nutrition and sleep, as well as social media conversations to monitor emotional wellbeing.

5. 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.

6. Improving Security and Law Enforcement

Big data is applied heavily in improving security and enabling law enforcement.

The National Security Agency (NSA) in the U.S. uses big data analytics to foil terrorist plots (and maybe spy on us).

Others use big data techniques to detect and prevent cyber attacks.

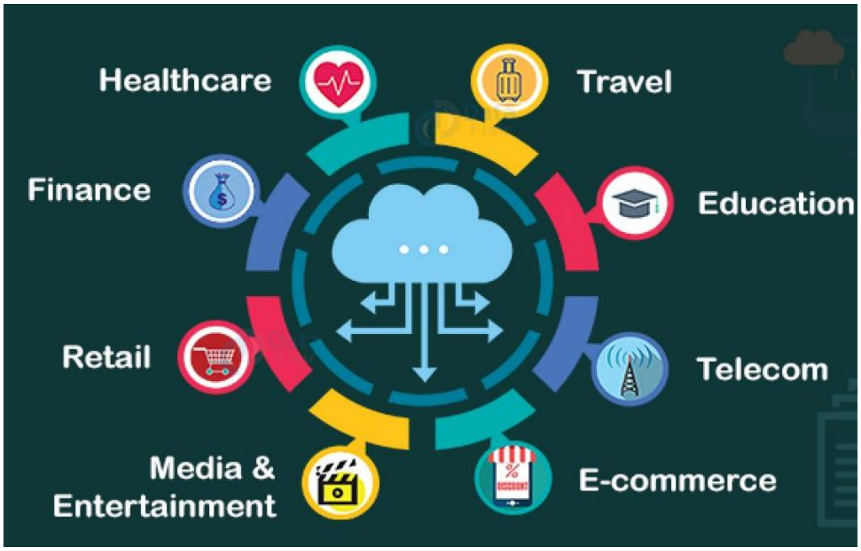
Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data use it to detect fraudulent transactions.

7. Improving and Optimizing Cities and Countries

Big data is used to improve many aspects of our cities and countries.

For example, it allows cities to optimize traffic flows 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 the transport infrastructure and utility processes are all joined up, where a bus would wait for a delayed train and where traffic signals predict traffic volumes and operate to minimize jams.



IV. CHALLENGES IN BIG DATA

Meeting the challenges presented by big data will be difficult. The volume of data is already enormous and increasing every day.

The velocity of its generation and growth is increasing, driven in part by the proliferation of internet connected devices.

Furthermore, the variety of data being generated is also expanding, and organization"s capability to capture and process this data is limited.

Current technology, architecture, management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data.

V. SECURITY AND PRIVACY

Cloud security alliance big data working group identify top security and privacy problems that need to capture for making the big data computing and infrastructure more secure.

Most of these issues are related to the big data storage and computation.

Some of the challenges are secure data storage. Various security challenges related to data security and privacy are discussed in which include data breaches, data integrity, data availability and data backup.

VI. DYNAMIC PROVISIONING

A service of the cloud computing is infrastructure as service in which it provides computation resources on demand, many cloud related companies are implementing this concept and to making it easy for customers to access these services.

Current frameworks do not have the property of the dynamic provisioning.

Here is an issue that Compute resources can be insufficient for the submitted job, some process may requires more resources.

Another issue is scheduling and protection algorithm, current algorithms does not consider these aspects.

VII. ALGORITHMS

Organizations were granting the papers by capturing key words from the abstract and titles.

Analyzing the science with hand was a difficult task.

After that, work was done by the program analyst. They use algorithms to do this work.

These algorithms can be varying from each other. This difference can reduce the effectiveness and reliability of the final result.

Improvement in the data management will result in better technology but it will face many issues.

Misuse of Big Data

Challenges including potential misuse of big data are here, because information is power.

Types of the data which people will produce in the future are unknown. To overcome these challenges we have to strengthen and increase our intent and capacity.

Data Management Data Big data comes in all shapes, colors and sizes. Rigid schemas have no place here; instead you need a more flexible design. You want your technology to fit your data, not the other way around. And you want to be able to do more with all of that data – perform transactions in realtime, run analytics just as fast and find anything you want in an instant from oceans of data, no matter what from that data may take.

VIII. CONCLUSION

To handle big data and to work with it and obtaining benefits from it a branch of science has come up and is evolving, called Data Science. Data Science is the branch of science that deals with discovering knowledge from huge sets of data, mostly unstructured and semi structured, by virtue of data inference and exploration. It's a revolution that's changing the world and finds application across various industries like finance, retail, healthcare, manufacturing, sports and communication. Search engine and digital marketing companies like Google, Yahoo and Bing, social networking companies like Facebook, Twitter and finance and e commerce companies like Amazon and EBay are requiring and will require a lots of data scientists. As far as security is concerned the existing technologies are promising to evolve as newer vulnerabilities to big data arise and the need for securing them increases.

REFERENCES

[1] www.coursera.org, Introduction to Big Data, University of California, San Diego. https://www.coursera.org/learn/big-data-introduction

[2] http://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report. Published: 4th January 2014 in Technology, Education Harsh Kishore Mishra. Center for Computer Science and Technology. School of Engineering and Technology, Central University of Punjab, Bhatinda

[3] Schmitt, C., Shoffner, M., Owen P., Wang, X., Lamm, B., Mostafa, J., Barker, M., Krishnamurthy, A., Wilhelmsen, K., Ahalt, S., &Fecho, K. (2013): Security and Privacy in the Era of Big Data: The SMW, a Technological Solution to the Challenge of Data Leakage. RENCI, University of North Carolina at Chapel Hill. Text: http://dx.doi.org/10.7921/G0WD3XHT Vol. 1, No. 2 in the RENCI White Paper Series, November 2013.Created in collaboration with the National Consortium for Data Science. (www.data2discovery.org)

[4]. Big Data Meets Big Data Analytics, www.sas.com/offices, 2012.

[5]. Shvachko, K., Kuang, H., et al, "The Hadoop distributed file system. In Mass Storage Systems and Technologies (MSST)", 2010, IEEE, pp.1-10.

[6]. Dean, Jeffrey, and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters", Communications of the ACM, Vol.51, No.1, pp.107-113, 2008.

[7] C. Mbohwa and A. K. Sahu, "Performance assessment of companies under IIoT architectures: application of grey re- lational analysis technique," in Proceedings of the 2018 In- ternational Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1350–1354, Coimbatore, India, July 2018.

[8] J. Park, H. Park, and Y. Choi, "Data compression and pre- diction using machine learning for industrial IoT," in Pro- ceedings of the 2018 International Conference on Information Networking (ICOIN), pp. 818–820, Chiang Mai, Thailand, January 2018.Y. Son and K. Lee, "Cloud of things based on linked data," in Proceedings of the 2018 International Conference on Infor- mation Networking (ICOIN), pp. 447–449, Chiang Mai, Thailand, January 2018.

[9] Y. Wu, "Research on depth estimation method of light field imaging based on big data in internet of things from camera array," IEEE Access, vol. 6, pp. 52308–52320, 2018.

[10] C. Yin, J. Xi, R. Sun, and J. Wang, "Location privacy pro- tection based on differential privacy strategy for big data in industrial internet of things," IEEE Transactions on Industrial Informatics, vol. 14, no. 8, pp. 3628–3636, 2018.

BIBLIOGRAPHY

Ms. MANALI KASHINATH SAKPAL has completed Bachelors in Computer Application from S.N.D.T College Shirgaon Ratnagiri, affiliated to S.N.D.T. University Mumbai 2011. Presently she is pursuing MCA from Institute of Distance and Open Learning. Along with she is having Teaching professional experience of overall 6 years of experience. Currently working as Craft-Instructor In Government Industrial Training Institute(Women), Ratnagiri.

DOI: 10.48175/IJARSCT-15074