ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

A Study on Big Data: Issues, Challenges and Applications

Sunita Sahu[1], Yugchhaya Dhote[2]

Assistant Professor, Department of Computer Engineering, VESIT, Chembur, Mumbai, India[1,2]

ABSTRACT: With the use of modern application and devices, data is growing with tremendous rate. This data can be audio, video, images, text and transactional data. So Big data is collection of huge amount of data which is complex in nature. This data is nowadays used by the companies or organizations to get deep insights into their businesses. Because of sheer volume of data, it requires the use of high speed parallel and distributed computing. Big data analytics is performing computations and database operations on large distributed data sets to get some meaningful insight from it. The volume, velocity and distributed nature of data poses challenges in data analytics. This paper covers concepts, characteristics and benefits of big data. The various challenges, issues and application areas are also discussed.

KEYWORDS: Big data, Data Analytics, Social media, unstructured data.

I. INTRODUCTION

Big Data is a popular term used to describe a huge volume of data, which is so large that is difficult to store and process with traditional database management systems. Now-a-days terabytes or petabytes of data are pouring into organizations. Big data gives tremendous insight to their business. With the appropriate use of available dataset organizations can formulate the strategies to compete and win. One of the applications of big data is recommendation systems which companies are using to attract their customers by sending them personalized suggestions. Big data technologies are maturing very fast and companies are adopting the big data as a core component to dig the data for exploring the data and results. The big data paradigm is transforming our society and continuously attracting the attention from academia and industry.

Big data analytics is a process of collecting, organizing and examining huge amount of data to identify the hidden pattern, and other useful information which can be used to make better decisions. With the big data analytics techniques, data scientists can analyze huge volume of data that cannot be possible with business intelligence or conventional analytics tools. In [1] authors have categorized the big data analytics in two categories: Stream processing and batch processing. In stream processing, data comes in streams and it is process as soon as possible to generate results. Storm and Apache kafka are popular stream processing models. In batch processing first data is stored and then processed. MapReduce is popular batch processing model.

Google, eBay and LinkedIn were among the first to experiment with big data. Initially they used big data analytics to check if analytical model can be improved and they indeed got positive results. According to [2],Every day, we create almost 2.5 quintillion bytes of data and 90% of the data in the world today has been created in the last two years alone and it is going to double in every two years [3]. Most of the data is created by individual users via social media and other sources of data are sensors data used to gather climate transaction records, digital pictures and videos, transactions records and cell phone GPS signals etc. This data is **big data**.

In today's scenario 75% of the data is generated by social media sites, which are mostly unstructured. According to the recent IDC forecast, Big Data technology and services market represents a fast-growing multibillion-dollar worldwide opportunity and it will grow at a 23.1% compound annual growth rate to $48.6 billion in 2019, or about six times the growth rate of the overall information technology market. Additionally, by 2020 IDC believes that line of business buyers will help drive analytics beyond its historical sweet spot of relational (performance management) to the double-digit growth rates of real-time intelligence and exploration/discovery of the unstructured

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

worlds [15].To reduce the IT cost, a lot of companies are using online big data tools which in turn affects the security and privacy of data, since the entire data is hosted on third party infrastructure [19].

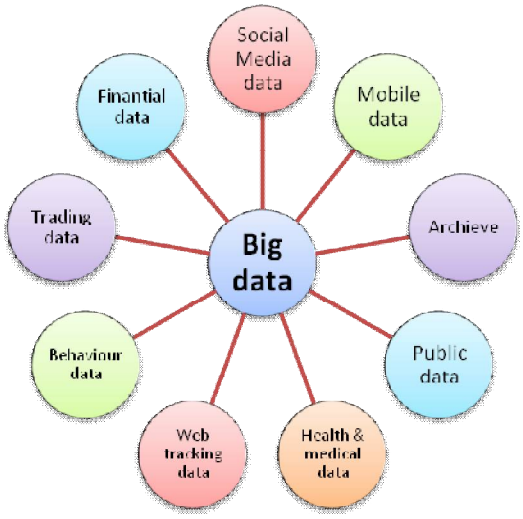

Figure 1: Sources of big data

II. BIG DATA CHARACTERISTICS

Volume: Volume refers to the amount of data generated every second[5]. These data come from emails, tweets, sensor data, photos, and videos that we produce and share. Zettabytes or brontobytes of data is generated every day. According to Sussan Gunelius blog in ACI Facebook users share nearly 2.5 million contents every minute, Twitter users tweet nearly 300,000 times, 220,000 new photos are posted by Instagram users, 72 hours of new video content are uploaded by YouTube users, Amazon generates over $80,000 in online sales[14].Volume introduces challenges because technologies that would work with smaller data sets do not scale up.

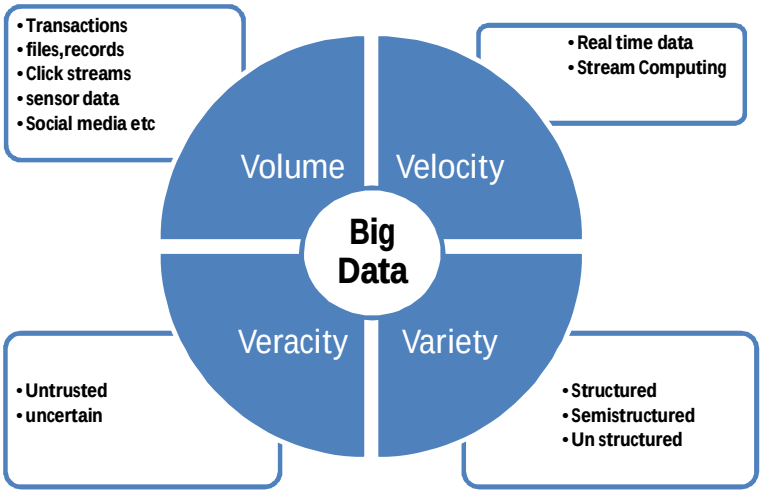

Figure 2: Four V's of Big data

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Velocity: Velocity refers to the speed/rate at which data is generated and processed. Systems should be capable of processing data with variable velocity [13]. Big data technology now allows us to analyze and process the data while it is being generated, without ever putting it into databases.

Variety: Variety refers to the different types of data available to any organization [7,13]. Earlier organizations used to focus only on structured data that can easily fit into tabular structure of relational databases. In today's scenario 80% of data is unstructured or semi-structured. With the help of big data technologies organization can easily handle the variety of data form to take important decisions.

Veracity: Veracity refers to the accuracy or trustworthiness of the data. Accuracy of data is very important because data is processed to take important business decisions.

III. TRAFFIC GENERATED BY VARIOUS POPULAR SITES

The following table details the traffic generated by various popular social media websites.

Popular sites	Active monthly users	Traffic generation
Twitter	310,000,000	• On an average, around 6,000 tweets every second • 350,000 tweets every minute • 500 million tweets per day • Every second tweets are not equal(Irregular traffic)
Facebook	1,100,000,000	• 4,166,667 post every minute. • 250 million post every hour. • 4.5 billion likes generated daily. • 968 million people log onto Facebook daily. • There are 1.31 billion mobile active users. • Photo uploads total 300 million per day. • Every minute 510 comments are posted. • 293,000 statuses are updated every minute.
LinkedIn (World's largest professional network)	255,000,000	• 100 million user visiting LinkedIn per month. • Currently used in over 200 countries and territories. • There are one billion total endorsements on LinkedIn at the moment.
Google+ (Second most popular social media network among US users)	300,000,000	• 30% of the smart phone users google+. • 40% marketers use google+ to market their product.
Tumplr	110,000,000	• 113.6 million Tumplrpost every day • 217 million Tumplrblogs • 5.187 monthly page views • 120,000 daily Tumplrlogin • 99 billion Tumplrpost
Instagram	400,000,000	• 1,736,111 likes every minute on photo • 100 million likes per hour • over 30 billion photos total • share an average of 70 million photos per day

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Flickr (Image hosting and video hosting website)	65,000,000	• 1 million average photo shares on Flickr daily. • 2 million Flickr groups. • 53 million tags have been added on Flickr images.
Vine (Video sharing website)	200,000,000	• 100 million people watch vine videos every month. • 15 billion vine loop played daily. • 8333 videos uploaded by vine user every minute.
YouTube (most widely used video sharing website)	1 billion	• 300 hour video uploaded per minute. • 4 billion video per day. • 1 billion video mobile views.

Table1:Traffic Generated by Popular Social Media Sites

By observing the above table, we can imagine the amount of data which is generated every day by the various websites and social media users and that too most of the data is unstructured. This data is used to understand the likes and dislikes of the customers and hence, social media is also impacting the e-commerce industries as they are driving the sales.

IV. BENEFITS OF BIG DATA TO ORGANIZATIONS

Organizations are getting many benefits from big data. Some of them are listed below:
Scale Up: With Big data organizations are able to scale very rapidly and elastically across multiple data centers and the cloud, whenever and wherever the need arises.

Cost reduction: Big data technologies like Hadoop and cloud-based analytics can provide substantial cost advantages. All the companies or organizations are employing Big data technologies to augment the existing system.

Faster and better decision making: Large organizations are using Big Data Analytics to improve decision making which is driven by the speed of Hadoop and in-memory analytics.

Performance: In an online world where nanosecond delays can cost heavily on sales, Big data must move at extremely high velocities no matter how much it is scaled or what workloads the databases must perform.

Workload Diversity: Big data comes in all shapes, colors and sizes. Big data uses flexible database design pattern, rather than fixed schema. Organizations/companies want technology to fit their data, not the other way around. Moreover, organizations also want to do much more with their data – perform transactions in real-time, run analytics just as fast and find anything they want in an instant from oceans of data, no matter what form that data may take.

V. CHALLENGES IN BIG DATA ANALYTICS

1. **Speed and data quality:** In today's' competitive environment, companies not only want to mine large amounts of data to take business decisions, they want it quickly and accurate.

2. **Storage and network requirement:** Currently available data storage techniques are not sufficient for storing Big data. Many companies are using cloud as choice to fulfill the storage requirement of Big data [18]. But when it comes to uploading the huge amount of data, network bandwidth becomes the performance bottleneck.

3. **Talent gap:** Enough number of experts are not available for Big data implementation, as it is an emerging technology. At the same time, it does not only require the technology acquaintance but one should have analytical and interpretive skills too[16]. Many companies are hiring Big data consultants to train their employees.

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

4. **Data Integration:** Integrating disparate data from various heterogeneous sources is an open challenge in big data analytics [11].

5. **Selection of right project:** Identifying the right business problem is critical for success and business needs to get involved as quickly as possible.

6. **Budget:** Traditional enterprise data servers are not capable to process Big data [17]. Additional IT investment is required to purchase high performance computing (HPC) servers, GPU units and analytical servers.

7. **IT know-how**: Due to the sheer volume of data, HPC and parallel processing are used in Big data in which several processing threads run concurrently on data. To facilitate parallel processing, storage strategies are also required to be changed.

8. **Data cleanup:** Data cleaning is a process of eliminating incomplete, inaccurate and redundant data from the input data source. It should be the first step of any Big data project. Generally, datasets contain high level of redundancy which should be eliminated to reduce the overall cost of the project [9].

9. **Data confidentiality**: A very lower granularity data which can drive any business such as transactional data which is confidential is used by the big data projects. So it is very important to take proper security measures to protect the data to insure the safety [9].

VI. APPLICATIONS AREAS OF BIG DATA

1. Understanding and Targeting Customers: This is one of the most widely used areas of Big data today. Here, Big data analytics is used to understand customers and their behaviors and preferences. Companies are keen to expand their traditional datasets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers. Recommendation system plays important role here.By analyzing the historical behavior of customers, it sends the personalized recommendations which is win-win condition for both customer and company.

2. Understanding and Optimizing Business Processes: Big data is also widely used to optimize business processes. Retailers are optimizing their stock based on predictions generated from social media data, web search trends and weather forecasts.

3. Improving Healthcare and Public Health: In health care also, we have a large amount of data coming in from various pathological reports, ultrasound and MRIs etc. Nowadays healthcare is using big data technology to predict, understand and avoid various new diseases and improving the quality of life. In the coming future all the individual data from smart watches and wearable devices can be shared with the doctors to analyze our health.

4. Improving and Optimizing Cities/Countries: Many aspects of our cities/countries can be improved by Big data analytics. For example, traffic flows can be optimized based on real time traffic information as well as social media and weather data.

5. In agriculture: In the coming decades agriculture will be transformed by the use of big data analytics. Sensors can be deployed on the forms and collected data is used to detect the reactions of crop on different environmental condition, soil conditions, water level etc.

6. Financial Trading: Big data is used widely today in high-frequency trading. Here big data analytics is used to make trading decisions.

ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

VII. CONCLUSION

Traditional RDBMS system are inappropriate for handling big data challenges because of lack of support for semi structured or unstructured data, and it requires expensive hardware to scale out. This paper describes the new emerging technology, Big Data, its characteristics and benefits. Various challenges associated with it are high cost of initial infrastructure, massive processing power and providing proper security mechanisms. The paper also highlights the some of the application areas of Big data.

REFERENCES

1. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: a technology tutorial. *Access, IEEE, 2*, 652-687
2. www-01.ibm.com/software/data/bigdata/what-is-big-data.html
3. J. Gantz and D. Reinsel, ``The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,'' in *Proc. IDC iView,IDC Anal. Future*, 2012.
4. Ward, Jonathan Stuart, and Adam Barker. "Undefined by data: a survey of big data definitions." *arXiv preprint arXiv:1309.5821* (2013).
5. Ammu, Nrusimham, and MohdIrfanuddin. "Big Data Challenges."*International Journal of Advanced Trends in Computer Science and Engineering* 2.1: 613.
6. Michael, Katina, and Keith Miller. "Big data: New opportunities and new challenges [guest editors' introduction]." *Computer* 46.6 (2013): 22-24.
7. Sagiroglu, Seref, and DuyguSinanc. "Big data: A review." *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013.
8. Kaisler, Stephen, et al. "Big data: Issues and challenges moving forward."*System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, 2013.
9. Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: A survey." *Mobile Networks and Applications* 19.2 (2014): 171-209.
10. Labrinidis, Alexandros, and H. V. Jagadish. "Challenges and opportunities with big data." *Proceedings of the VLDB Endowment* 5.12 (2012): 2032-2033.
11. Cuzzocrea, Alfredo, Il-Yeol Song, and Karen C. Davis. "Analytics over large-scale multidimensional data: the big data revolution!." *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*. ACM, 2011.
12. Martin, Kirsten E. "Ethical issues in the Big Data industry." *MIS Quarterly Executive, Forthcoming* (2015).
13. Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.
14. SUSAN GUNELIUS, *"data Explosion in 2014 Minute by Minute – Info graphic" JULY 12, 2014,ACI blog*
15. *http://www.idc.com/getdoc.jsp?containerId=prUS40560115*
16. *Katal, Avita, Mohammad Wazid, and R. H. Goudar. "Big data: issues, challenges, tools and good practices." Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013.*
17. *Mary shacklett "10 roadblocks to implementing Big Data analytics" http://www.techrepublic.com/blog/10-things/10-roadblocks-to-implementing-big-data-analytics.*
18. Katal, Avita, Mohammad Wazid, and R. H. Goudar. "Big data: issues, challenges, tools and good practices. " *Contemporary Computing (IC3), 2013 Sixth International Conference on*. IEEE, 2013.
19. Ji, Changqing, et al. "Big data processing in cloud computing environments."*Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on*. IEEE, 2012.
20. Cuzzocrea, Alfredo, Il-Yeol Song, and Karen C. Davis. "Analytics over large-scale multidimensional data: the big data revolution!." *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*. ACM, 2011.
21. Constantine, C. (2014). Big data: an information security context. *Network Security, 2014*(1), 18-19.
22. Bizer, Christian, et al. "The meaningful use of big data: four perspectives--four challenges." *ACM SIGMOD Record* 40.4 (2012): 56-60.

BIOGRAPHY

Sunita Sahu is an Assistant Professor in Computer Engineering Department, VESIT, Chembur, Mumbai, Mumbai University. She received Master of Technology (M.Tech) in Computer Science in the year 2010 from Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV) Bhopal, India. Her research interests are Big data Analytics, Mobile Ad Hoc Networks, Cloud Computing, Network security etc.

Yugchhaya Dhote is an Assistant Professor in Computer Engineering Department, VESIT, Chembur, Mumbai, Mumbai University. She received Master of Technology (M.Tech) in Information Technology in the year 2013 from Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV) Bhopal, India. Her research interests are Social media analysis, Big data analytics, Cloud Computing etc.