Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.4 No.2, 2015, pp.23-26

DOI: https://doi.org/10.51983/ajcst-2015.4.2.1755

Enormous Possibilities in Big Data: Trends and Applications

Jasmeen Gill[1] and Shaminder Singh[2]
[1]CSE Department, RIMT-IET, Mandi Gobindgarh, India
[2]CSE, Gulzar Group of Institutes, G.T.Road, Khanna, India
E-mail: er.jasmeengill@gmail.com, shamindersinghsohi@gmail.com
(Received 20 August 2015; Revised 4 September 2015; Accepted 25 September 2015; Available online 5 October 2015)

***Abstract* - Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making. In this article, the trends in big data are analyzed with respect to its characteristics. In addition the processing of data is discussed along with various methods and tools like Hadoop. This article also specifies the possibilities of big data in different application areas like IT, Fraud detection, Medical, Retail, Manufacturing, Real estate, Banking etc.**
Keywords: Big Data, Hadoop, 6C System, V3, Fraud Detection.

I. INTRODUCTION

The term "Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze [8]. But as technological advances improve our ability to exploit Big Data, potential privacy concerns could stir a regulatory backlash that would dampen the data economy and stifle innovation [9].

Big data technologies are a set of methodologies that describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis [10].Basically, data is being produced at an ever increasing rate. This growth in data production is being driven by: individuals and their increased use of media; organizations; the switch from analogue to digital technologies; and the proliferation of internet connected devices and systems.

There has also been an acceleration in the proportion of machine-generated and unstructured data (photos , videos, social media feeds and so on) compared to structured data such that 80% or more of all data holdings are now unstructured and new approaches and technologies are required to access, link, manage and gain insight from these data sets[23].Variability. In addition to the speed at which data comes your way, the data flows can be highly variable – with daily, seasonal and event-triggered peak loads that can be challenging to manage [1].

II. BENEFITS OF BIG DATA

Organizations are developing a more complete understanding of their customers than ever before, as they better assemble the data available to them. Public health authorities, for example, have a need for more detailed information in order to better inform policy decisions related to managing their increasingly limited resources [5].

The inability of an organization to benefit from the information it has access to or has generated in the past can result in what has been referred to as 'enterprise amnesia.' Studies, for example, conducted for a major retailer found that out of every 1000 employees hired, two had been previously arrested for stealing from the same store for which they had been rehired[5][10].

Historically, advanced analytics have been used, among other things, to analyze large data sets in order to find patterns that can help isolate key variables to build predictive models for decision-making. Companies use advanced analytics with data mining to optimize their customer relationships [5][12], law enforcement agencies use advanced analytics to combat criminal activity from terrorism to tax evasion to identify theft[13].

III. CHARACTERISTICS OF BIG DATA

Big data can be described by many characteristics of data, namely, volume, variety, velocity, variability, veracity, complexity as shown figure 1 and discussed in [1].

A. Volume

Volume refers to the quantity of generated data. It is important as the size of the data determines the value and potential of the data under consideration and whether it can actually be considered big data or not. The name 'big data' itself means a term related to size, and hence it is the most important characteristic of big data.

B. Variety

The type of content is an essential fact that data analysts must know. This helps people who are associated with and analyze the data to effectively use the data to their advantage and thus uphold its importance.

C. Velocity

Velocity refers to the speed at which the data is generated and processed. In this context, high velocity is also a great challenge in the path of growth and development.

D. Variability

The inconsistency the data show, at times, can hamper the process of handling and managing of data.

E. Veracity

The quality of captured data varies greatly. Accurate analysis depends on the veracity of source data.

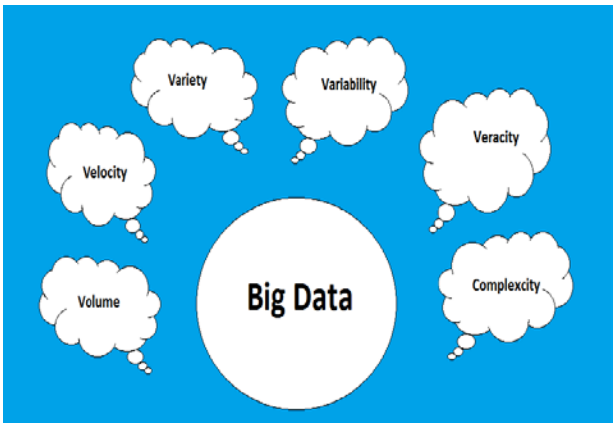

Fig. 1 Characteristics of Big Data

F. Complexity

Data management can be very complex, especially when large volumes of data come from multiple sources. Data must be linked, connected, and correlated so users can grasp the information the data is supposed to convey. Factory work and Cyber-physical systems may have a 6C system:

- Connection (sensor and networks)
- Cloud (computing and data on demand)
- Cyber (model and memory)
- Content/context (meaning and correlation)
- Community (sharing and collaboration)
- Customization (personalization and value)
-

IV. PROCESSING OF BIG DATA

Accelerated processing of huge data sets is made possible by four primary technologies [1]:

A.Grid Computings

A centrally managed grid infrastructure provides dynamic workload balancing, high availability and parallel processing for data management, analytics and reporting. Multiple applications and users can share a grid environment for efficient use of hardware capacity and faster performance, while IT can incrementally add resources as needed.

*B.**Database Processings***

Moving relevant data management, analytics and reporting tasks to where the data resides improves speed to insight, reduces data movement and promotes better data governance. Using the scalable architecture offered by third-party databases, in-database processing reduces the time needed to prepare data and build, deploy and update analytical models.

C.Memory analytics

Quickly solve complex problems using big data and sophisticated analytics in an unfettered manner. Use concurrent, in-memory, multiuse access to data and rapidly run new scenarios or complex analytical computations. Instantly explore and visualize data. Quickly create and deploy analytical models. Solve dedicated, industry-specific business challenges by processing detailed data in-memory within a distributed environment, rather than on a disk.

D.Hadoops

Hadoop is an open source framework for storing and processing large datasets using clusters of commodity hardware. Hadoop is designed to scale up to hundreds and even thousands of nodes and is also highly fault tolerant [6] [14].Distributed File System (HDFS) is a file system that is used to store data across cluster of commodity machines while providing high availability and fault tolerance [6] [15].HDFS is distributed file system for storing and retrieving data for Map Reduce and help to executes jobs for users. Hadoopp form the cluster of data nodes and stored at an on spaceutilization of data nodes on cluster [6] [16[18].Map/Reduce in distributed is currently using by Google [4] [17].

Initially researchers developed a distributed database [4] [20] and for maintaining the work load parallel database system [8] both were successful. Change in data access is major issue in front of distributed and parallel data, to resolve that a new class of system definitions i.e. Key Value. Map Reduce paradigm [4] [22] and its open source implementation platform Hadoop is basically is a solution of this problem of distributed and parallel databases.

Support for Hadoop. You can bring the power of SAS Analytics to the Hadoop framework (which stores and processes large volumes of data on commodity hardware). SAS provides seamless and transparent data access to Hadoop as just another data source, where Hive-based tables

appear native to SAS. You can develop data management processes or analytics using SAS tools – while optimizing run-time execution using Hadoop Distributed Process Capability or SAS environments. With SAS Information Management, you can effectively manage data and processing in the Hadoop environment [1].

We believe the best way to frame why Big Data is important is to share with you a number of our real customer experiences regarding usage patterns they are facing (and problems they are solving) with an IBM Big Data platform. These patterns represent great Big Data opportunities—business problems that weren't easy to solve before—and help you gain an understanding of how big Data can help you (or how it's helping your competitors make you less competitive if you're not paying attention).

V. USE OF BIG DATE

A.IT for IT Log Analytics

Log analytics is a common use case for an inaugural Big Data project. We like to refer to all those logs and trace data that are generated by the operation of your IT solutions as data exhaust. Enterprises have lots of data exhaust, and it's pretty much a pollutant fit's just left around for a couple of hours or days in case of emergency and simply purged. Because we believe data exhaust has concentrated value, and IT shops need to figure out a way to store and extract value from it. Some of the value derived from data ex-trans formed into value-added click-stream data that records every gesture, click, and movement made on a web site [7].

- eBay.com uses two data warehouses at 7.5 petabytes .
- Amazon.com handles millions of back-end operations every day.
- Facebook handles 50 billion photos from its user base.
- Google is handling roughly 100 billion searches per month.

B.Fraud Detection Pattern

Fraud detection comes up a lot in the financial services vertical ,but if you look around, you will find it in any sort of claims –or transaction-based environment (online auctions, insurance claims, under writing entities , and so on).Pretty much any where some sort of financial transaction is involved presents a potential for misuse and the ubiquitous specter of fraud. If you Leverage Big Data platform, you have the opportunities to do more than you have ever done before to identify it or, better yet, stop it. Several challenges in the fraud detection pattern are directly attributable to solely utilizing conventional technologies .The most common, and recurring, theme you will see across all Big Data patterns is limits on what can be stored as well as available computer sources to process your intentions. Without Big Data technologies, these factors limit what can be modeled .Less data equals constrained modeling. What's more, highly dynamic environments commonly have cyclical fraud patterns that come and go in hours, days, or weeks. If the data used to identify or bolster new fraud detection models is not available with low latency, by the time you discover these new patterns, it's too late and some damage has already been done [7].

C.Medical Field

The healthcare industry historically has generated large amounts of data, driven by record keeping, compliance & regulatory requirements, and patient care [3] [25]. While most data is stored in hard copy form, the current trend is toward rapid digitization of these large amounts of data. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery meanwhile reducing the costs, these massive quantities of data (known as 'big data') hold the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management [3] [26][27][28][29]. Reports say data from the U.S. healthcare system alone reached, in 2011, 150 exabytes. At this rate of growth, big data for U.S. healthcare will soon reach the zettabyte (1021 gigabytes) scale and, not long after, the yottabyte (1024 gigabytes) [3] [30]. Existing analytical techniques can be applied to the vast amount of existing (but currently unanalyzed) patient-related health and medical data to reach a deeper understanding of outcomes, which then can be applied at the point of care. Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient.

D.Manufacturing

Big data provides an infrastructure for transparency in manufacturing industry, which is the ability to unravel uncertainties such as inconsistent component performance and availability. Big Data Analytics for Manufacturing Applications can be based on a 5C architecture (connection, conversion, cyber, cognition, and configuration).

E.Retail

Walmart handles more than 1 million customer transactions every hour.

F.Banking

FICO Card Detection System protects accounts worldwide. The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.

G. Real estate

Windermere Real Estate uses anonymous GPS signals from nearly 100 million drivers to help new home buyers determine their typical drive times to and from work throughout various times of the day.

H. Science

The Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per

second. There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.99995% of these streams, there are 100 collisions of interest per second.

V. CONCLUSION

Big data refers to structured, unstructured and semi structured data which means data is having variety. Big Data is also referred to huge data set having really huge magnitude, (really a huge volume).

This article explores that big data analytics are asset of advanced technologies designed to work with large volumes of heterogeneous data leading to improved performance via data driven decision making. They also help in improving the outcomes by reducing the cost in different fields like IT, Fraud detection, Medical, Retail, Banking, Manufacturing, Real estate etc. Moreover, the analytical tools have made it possible to analyze and manage huge amount of data resolving scalability issues. So Big data technologies are foreseen as to not only support the ability to collect large amounts, but more importantly, the ability to understand and take advantage of its full value.

REFERENCES

[1] Troester, M., "Big Data Meets Big Data Analytics", SAS, white paper, SAS Institute Inc., pp.1-13.

[2] Arun Thomas George et al., Big Data Spectrum by Infosys, pp.1-61.

[3] Raghupathi, W., and Raghupathi, V., "Big data analytics in healthcare: promise and Potential", Health Information Science and Systems, 2, 3, 2014.

[4] Manekar, A., and Pradeepini, G., "A Review on cloud based big data analytics", ICSES- Journal on computer networks and communication, 1, 1, pp.6-9, 2015.

[5] Ann Cavoukian and, Jeff Jonas, Privacy by Design in the Age of Big Data, pp. 1-17, 2012.

[6] Singh, D., and Reddy, C., K., "A survey on platforms for big data analytics", journal of big data: a springer open journal, 1, 8, pp. 1-20, 2014.

[7] Zikopoulos, P., C., Eotan C., Deroos, D., and Lapis, G., "Under - standing Big Data", McGraw-Hil, 2012.

[8] Manyika, J.,et. al., "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute. , 2011,Online:http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

[9] Tene, O., and Polonetsky J., "Privacy in the age of big data: A time for big decisions", Stanford Law Review 64, 63, 2012.

[10] Gantz. J., and Reinsel. D., "Extracting value from chaos", IDC, 2011, Online: http://www.emc.com/collateral/analyst-reports/idc - extracting-value-from-chaos-ar.pdf.

[11] Jonas, J. , "On how data makes corporations dumb. GigaOm", Online: http://gigaom.com/2010/10/11/jeff-jonas-big-data/,2010.

[12] Marsella, A., and Banks, M. , "Making customer analytics work for you!" , Journal of Targeting, Measurement and Analysis for Marketing, vol.13, no. 4, pp. 299-303,2005.

[13] Jonas, J., and Harper, J., "Effective counterterrorism and the limited role of predictive data mining", Policy Analysis, CATO Institute, Washington, DC, 584, pp.1-11, 2006.

[14] Hadoop. http://hadoop.apache.org/

[15] Borthakur D., HDFS architecture guide, Hadoop Apache Project , 2008, http://hadoop.apache.org/docs/r1.2.1/hdfs _design.pdf.

[16] Vavilapalli VK, Murthy AC, Douglas C, Agarwal S, Konar M, Evans R, Graves T, Lowe J, Shah H, and Seth S "Apache hadoop yarn: Yet another resource negotiator", In: Proceedings of the 4th annual Symposium on Cloud Computing., pp. 5,2013.

[17] Dean J., and Ghemawat S., "MapReduce: simplified data processing on large clusters ", Commun. ACM, vol.51, no.1, pp.107–113, 2008.

[18] Xin RS, Gonzalez JE, Franklin MJ, and Stoica I Graphx, "A resilient distributed graph system on spark". First International Workshop on Graph Data Management Experiences and Systems , pp. 2, 2013.

[19] Divyakant Agrawal, Sudipto Dasand Amr E l Abbadi "Big Data and Cloud Computing: Current State and Future Opportunities", EDBT2011, March22–24, 2011, Uppsala, Sweden ACM 978-1-4503-0528-0/11/0003.

[20] J. B. Rothnie Jr., P. A. Bernstein, S. Fox, N. Goodman, M. Hammer, T. A. Landers, C. L. Reeve, D. W. Shipman, and E. Wong, "Introduction to a System for Distributed Databases (SDD-1)", ACM Trans. Database Syst., vol. 5, issue 1, pp.1–17, 1980.

[21] D. J. Dewitt, S. Ghandeharizadeh, D. A. Schneider, A. Bricker, H.I.Hsiao, and R. Rasmussen., " The Gamma Database Machine Project", IEEE Trans. On Knowl. and Data Eng., 2(1): 44–62, 1990.

[22] HadoopDistributedFileSystem:ArchitectureandDesignhttp://hadoop.apache.org/common/docs/r0.18.2

[23] "Big Data Strategy- issues papers", pp.1-12, March 2013.

[24] Martin Hilbert, "Big Data for Development: A Review of Promises and Challenges" , Journal Development Policy Review of the Overseas Development Institute ,pp 1-41,2014.http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-7679.

[25] Raghupathi W., "Data Mining in Health Care ". In Healthcare Informatics, Improving Efficiency and Productivity, Edited by Kudyba S. Taylor & Francis, pp.211–223, 2010.

[26] Burghard C, "Big Data and Analytics Key to Accountable Care Success". IDC Health Insights, 2012.

[27] Dembosky A., "Data Prescription for Better Healthcare." Financial Times, pp. 19, December 12, 2012. Available from: http://www.ft.com/intl/cms/s/2/55cbca5a-4333-11e2-aa8f-00144feabdc0.html#axzz2W9cuwajK.

[28] Feldman B, Martin EM, Skotnes T: "Big Data in Healthcare Hype and Hope." Dr. Bonnie 360,2012. http://www.west-info.eu /files/big-data-inhealthcare.pdf.

[29] Fernandes L., O'Connor M., and Weaver V., "Big data, bigger outcomes", Journal AHIMA, pp. 38–42, 2012.

[30] IHTT ,Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/