IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Big Data: Introduction, Applications, Opportunities & Challenges

Yogesh Kumar,

Assistant Professor, Engineering College Jhalawar Engineering, Rajasthan, India

ABSTRACT

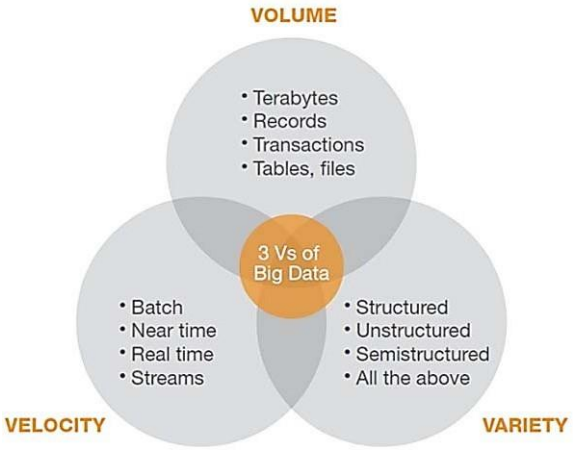
Big data refers to data that is difficult to gather, manage, process, or analyze using standard technologies and techniques due to its great volume (size), variability (complexity), and velocity (pace of expansion). Big Data is a type of data that, due to its size, diversity, and complexity, necessitates the development of novel management strategies, methodologies, algorithms, and analytics. Big data has an ambiguous character and requires extensive processes to locate and transform the data into fresh insights. The types of big data, their prospects, difficulties, and applications are covered in this review study.

KEYWORDS: Big Data, Storage, Volume, Variability and Velocity

INTRODUCTION

Big Data is a term used to describe technique and tools for gathering, managing, distributing, and analyzing datasets with varied structures and sizes up to PB (peta bytes) in size. Huge amounts of data, social media analytics, next-generation data management capabilities, real-time data, and many other topics have all been described using the term "big data". Big data, as defined by Gartner, is characterized by three Vs: volume, variety, and velocity. Gartner first used the terms volume, variety, and velocity to define the components of big data challenges. Big data technologies are described as a "new generation of technologies and architectures, designed to cheaply extract value from very large volumes of a wide variety of data, by enabling the high velocity acquisition, discovery, and/or analysis," according to IDC.

Figure 1: 3Vs of Big Data (Gartner, 2001)



Volume refers to the size of different data generated from different sensors or sources. The benefit of gathering large amounts of data includes the creation of hidden information and patterns through data analysis. Laurila et al. provided a unique collection of longitudinal data from smart mobile devices and made this collection available to the research community. The aforesaid initiative is called mobile data challenge motivated by Nokia. Collecting longitudinal data requires considerable effort and underlying investments. Nevertheless, such mobile data challenge produced an interestingresult similar to that in the examination of the predictability of human behavior patterns or means to share data based on human mobility and visualization techniques for complex data.

Variety refers to the different types of data collected via sensors, smart phones, or social networks. Such data types include video, image, text, audio, and data logs, in either structured or unstructured format. Most of the data generated from mobile applications are in unstructured format. For example, text messages, online games, blogs, and social media generate different types of unstructured data through mobile devices and sensors. Internet users also generate an extremely diverse set of structured and unstructured data.

Velocity refers to the speed of data transfer. The contents of data constantly change because of the absorption of complementary data collections, introduction of previously archived data or legacy collections, and streamed data arriving from multiple sources.

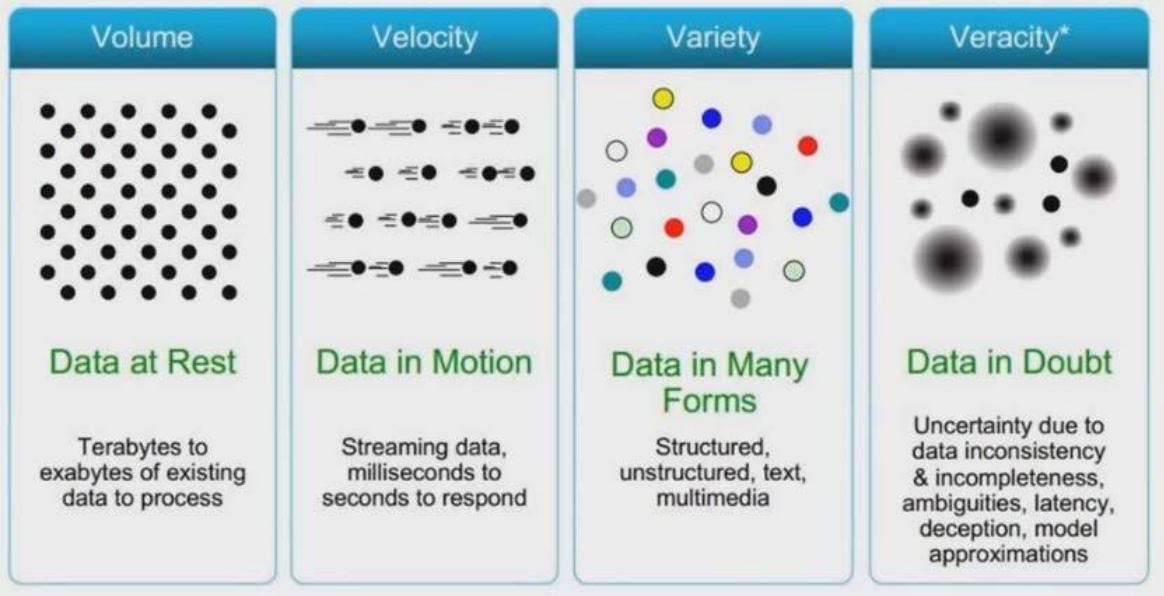


Figure 2. 4Vs

Some authors defined Big Data in four Vs and the 4th V can be veracity, Value, Variability or Virtual. More commonly, big data is a collection of very huge data sets with great diversity of types so that it becomes difficult to process by using traditional data processing platforms.

1.1 Definition of Big Data

Most of the people believe that Big Data means a data in large size of volume, but it is defined by more than just size. The following are popular some definitions

Table 1.1 More definitions of Big Data

Name of Proposer(s)	Definition(s)
Gartner, 2001	”Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”.
Tech America Foundation 2013	“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.”
Manyika, et al., 2011	“Big Data” are datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
Hopkins &Evelson 2012	“Big Data” is techniques and technologies that make handling data at extreme scale affordable.

Paredes 2012	Big Data" is a science of fielding algorithms that enable machines to recognize complex patterns in data. It fuses machine learning with a very deep understanding of computer science and algorithms and that, of course, is key to being able to take machine learning and deploy it in a very scalable way.
Rooney 2012	"Big Data" is the ability to mine and integrate data, extracting new knowledge from it to inform and change the way providers, even patients, think about healthcare.
Shah 2013	"Big Data" is a bubble just filled with hot air – at least for now. Everyone is talking about it bu when you dig a bit deep with a pointed question, very quickly you discover that it has nothing much to do with the Big Data

Data Explosion (where is Big Data exist)

- Over 2.5 EB(Exa Bytes) of data is generated every day.
- A typical, large stock exchange captures more than 1 TB(Tera Bytes) of data every day.
- There are around 5 billion mobile phones (including 1.75 billion smart phones) in the world.
- YouTube users upload more than 48 hours of video every minute.
- Large social networks such as Twitter and Facebook capture more than 10 TB of data daily.
- There are more than 30 million networked sensors in the world.

BIG DATA TYPES

There are three types of Bid Data named as Structured; Semi Structured and Unstructured big data As Big data have variety of data. Here variety mean there are basically three types of data. Structured data, Semi structured data and unstructured data.

Structured data

Data that fit in a fixed field within a record or file is called *structured data*. This includes data contained in relational databases and spreadsheets, and lends itself to that type of processing. It is easy to categorize and analyze and are numbers and words. The sources of structured data are network sensors, smart phones, and global positioning system (GPS) devices,bank transactions data, sales. Etc.

Semi structured data

Semi-structured data is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables. XML, NoSql is semi structured data. For example, Web logs, Socialmedia and E-commerce.

Unstructured data

Unstructured data is data that does not follow a specified format for big data and varies in content . . For example, photos, Video, Geological data etc 90% of the generated data is semi-structured or unstructured in other words only 10% of the data can be categorized as structured data. Unstructured text via Twitter Tweets on smart phones, spatial data from tracking devices, Radio Frequency Identification (RFID) devices, and audio and image files updated via smart devices. This unstructured data is Geospatial Data which is received from different devices. Data from mobile devices, sensors, radar, Google Earth is unstructured data or geospatial data.

Processing Tools for Big data

Name	Description	Developed By & Year	OperatingSystem
Hadoop	The Apache distributed data processing software.	Doug Cutting and Mike Cafarella , 2005	Windows &Linux,
MapReduce	A programming model and software framework forwriting applications.	Google, 2004	OS Independent
HPCC	It claims to offer superior performance to Hadoop. Both free community versions and paid enterprise versions are available.	Lexis Nexis Risk Solutions, 2011	Linux
Storm	Storm offers distributed real-time computation capabilities and is often described as the "Hadoop of realtime."	Twitter, 2011	Linux
Cassandra	This NoSQL database is now managed by the Apache Foundation.	Facebook (Now Managed by Apache), 2008	OS Independent
HBase	HBase is the non-relational data store for Hadoop. Features include linear and modular scalability, strictlyconsistent reads and writes, automatic failover supportand much more.	Apache, 2010	OS Independent
MongoDB	MongoDB was designed to support gigantic databases.It's a NoSQL database with document-oriented storage,	10gen (now MongoDB Inc.), 2007	Windows, Linux, Solaris

Why Hadoop?

It is a free, Java-based programming supports the processing of large data sets in a distributed computing environment. Based on Google File System (GFS) and runs a number of applications on distributed systems with thousands of nodes involving petabytes of data. Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model.

Features of Hadoop are Simple and easy programming model, Automatic and linear scalability, Built-in fault tolerance, Cost effective and fast processing in compare to traditional processing.

There are many applications of the Hadoop which includes Log and/or click stream analysis of various kinds, Marketing analytics, Machine learning and/or sophisticated data mining, Image processing, Processing of XML messages, General archiving, including of relational/tabular data, e.g. for compliance and Clinical big data and related medical health informatics fields.

BIG DATA APPLICATIONS

Big data has increased the demand of information management specialists. It is estimated that one third of the globally stored information is in the form of alphanumeric text and still image data, which is the format most useful for most big data applications.

Healthcare Professionals;

Medical Institutions Industry-specific big data challenges The healthcare sector has access to vast amounts of data, but has been plagued by the failure to use data to contain rising healthcare costs and inefficient systems. This is mainly due to unavailability, insufficient or unusable electronic data. Furthermore, health databases containing health-related information have made it difficult to link data that could show useful patterns in the medical field.

Education

Big data is heavily used in higher education. For example, the University of Tasmania. An Australian university with over 26,000 students has a learning and Introduced a management system. Another use case for using big data in education is to measure teacher effectiveness to ensure an enjoyable experience for both students and teachers. Teacher performance can be fine-tuned and measured based on student numbers, subject areas, student demographics, student goals, behavioral categories, and other variables. At the government level, the U.S. Department of Education's Office of Educational Technology is using big data to develop analytics to help students who get lost while taking online big data certification courses. Click patterns are also used to detect boredom. Big data providers in this industry include Knewton, Carnegie Learning and MyFit/Naviance.

Administration

Big Data is advantageous to the use and adoption of governmental procedures and allows for efficiencies in terms of cost, productivity, and innovation. Having stated that, there are several shortcomings in this procedure. To get the desired result, data analysis frequently necessitates the participation of numerous government agencies (both and local).

Natural Resources

In the resources industry, big data enables predictive modeling and supports decision making. It is used to ingest and integrate large amounts of data from geospatial, graphic, textual, and temporal data. Areas of interest where this has been used include seismic interpretation and reservoir characterization. Big data is also being used to solve today's manufacturing challenges and gain a competitive advantage. The chart below shows the current and future potential use of supply chain capabilities from big data, according to Deloitte research.

Public Sectors

There is a Big list of applications, including energy exploration, financial market analysis, fraud detection, health-related research, and environmental protection. Big data is being used in the analysis of large amounts of social disability claims made that arrive in the form of unstructured data. The analytics are used to process medical information rapidly and efficiently for faster decision making and to detect suspicious or fraudulent claims.

The Food and Drug Administration is using Big Data to detect and study patterns of food-related illnesses and diseases. This allows for a faster response, which has led to more rapid treatment and less death. Big Data for several different use cases. Big data is analyzed from various government agencies and is used to protect the country.

Manufacturing

Based on TCS 2013 Global Trend Study, improvements in supply planning and product quality provide the greatest benefit of big data for 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.

Technology

There are many online shopping applications/websites on the market that generate a huge amount of data. For example, eBay.com uses two 7.5 PB (petabyte) and 40 PB data warehouses and a 40 PB Hadoop cluster for search, consumer recommendations, and merchandising. eBay's 90 PB data warehouse.

Amazon.com processes millions of backend operations and requests from over 500,000 third parties every day. The core technology that keeps Amazon running is Linux-based, and as of 2005, it had the world's three largest Linux databases with capacities of 7.8 TB, 18.5 TB, and 24.7 TB. Facebook manages 50 billion photos of him in its user base.

Retail and Wholesale

Retailers and wholesalers continue to collect big data such as customer retention data, point of sale, store inventory, and local demographics. At the 2014 Big Show Retail Conference in New York, companies such as Microsoft, Cisco, and IBM emphasized the need for the retail industry to use big data for analytics and other purposes. Optimized staffing with data from shopping patterns, local events, and more less cheating Timely inventory analysis The use of social media also holds a lot of potential and continues to be slowly but surely adopted, especially in brick-and-mortar stores. Social media is used for customer acquisition, customer retention, product promotion, and more.

Transportation

Government use of big data: traffic control, route planning, advanced traffic systems, congestion management (by predicting traffic conditions) Use of private sector big data in transportation: revenue management, technological advances, logistics and competitive advantage (by consolidating shipments and optimizing freight movements) Personal uses of big data include route planning to save fuel and time, and travel planning in tourism.

OPPORTUNITIES & CHALLENGES

Problems with Big Data Processing

- **Heterogeneity and Incompleteness-** Machine analysis algorithms expect homogeneous data. But Big data is heterogeneous.
- **Scale: Data volume is scaling faster than compute resources, and CPU speeds are static.**
- **Timeliness**: The larger the data set to be processed, the longer it will take to analyze
- **Privacy**: Managing privacy is effectively both a technical and a sociological problem
- **Human Collaboration**: A Big Data analysis system must support input from multiple human experts, and shared exploration of results

More businesses are prepared to pilot and implement big data as a key part of the information management and analytics infrastructure as big data technologies are reaching a degree of maturity. Big data is positioned as the next big step in enabling integrated analytics in many typical business scenarios since it is a collection of cutting-edge disruptive tools and technology. Information technology (IT) professionals and business sponsors will encounter a variety of difficulties as big data makes its inexorable way into the firm. These difficulties must be overcome for any big data initiative to be successful.

There are various hurdles in the processing of big data is that the Data Management Landscape is Uncertain. There are various competing technologies, and there are many competitors within each technical field. Making the best decisions without increasing risk or unknowns associated with the adoption of big data is our first challenge.

The excitement surrounding big data applications seems to imply that there is a sizable community of professionals ready to aid in deployment. This is known as the "Big Data Talent Gap." The talent gap, however, presents our second difficulty because this is not the situation at this time. Our third difficulty is data accessibility and integration. The volume and variety of data that must be absorbed into a big data environment might be overwhelming for an untrained data practitioner. The potential for time gaps to affect data currency and consistency increases when more data sets from various sources are added to an analytical platform. This is our fourth difficulty.

Making Sense of the Big Data Platform for Information Finally, if the information cannot be adequately provisioned back within the other components of the enterprise information architecture, using big data for various purposes, from storage augmentation to enabling high-performance analytics, is hindered, making big data syndication our fifth challenge.

CONCLUSION

We have entered an era of Big Data. This paper describes the concept of Big Data along with Vs. and focuses on Big Data processing challenges. Billions of devices are generating unstructured data every day.

Big data is a set of techniques and technologies that require new forms of integration in order to discover great hidden value from diverse, complex and large-scale data sets. Big data requires great technology to efficiently process large amounts of data in an acceptable amount of time. It is hoped that big data technology will enable to provide the most relevant and accurate social sensing feedback to better understand our society in real time. Methodologies (Hadoop/Map reduce) adopted for Big Data have kept up with the increasing data but the scope for complexities associated with the real-time analysis of such data resulting in quality information is still open.

Furthermore, due to ever-growing data sets, the need arises for technological advancements towards information management

REFERENCES

1. OGC-OpenGIS Consortium et al. The opengis abstract specification-topic 7: The earth imagery case, 1999
2. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers A.H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 1-137.
3. Giachetta R, A framework for processing large scale geospatial and remote sensing data in MapReduce environment, 2015
4. Krämera M , Sennera I A modular software architecture for processing of big geospatial data in the cloud,2015
5. Lee J.-G., Han J., Li X., Trajectory outlier detection: a partition-and-detect framework, in: Proceedings of 24th International Conference on Data Engineer-ing, Cancun, Mexico, 2008, pp.140–149.
6. Lee J.-G., Han J., X. Gonzalez Li, H., TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering, Proc. VLDB Endow. 1(1) (2008) 1081–1094.
7. L.Chang, R.Ranjan, Z.Xuyun, Y.Chi,D.Georgakopoulos, C.Jinjun, Public Auditing for Big Data Storage in Cloud Computing – a Survey, Computational Science and Engineering(CSE), 2013 IEEE 16th International Conference on, 2013, pp.1128–1135.
8. M. Cox, D. Ellsworth, Managing Big Data For Scientific Visualization, ACM Siggraph, 1997.
9. O. Kwon, N. Lee, B.Shin, Data quality Q7 management, data usage experience and acquisition intention of big data analytics, Int.J.Inf. Manag. (2014).
10. K. Singh, S. C. Guntuku, A. Thakur, C.Hota, Big data analytics framework for peer-to-peer botnet detection using random forests, Inf. Sci.(2014).
11. Meeker M., 2012 KPCB internet trends year-end update, http://www.slideshare.net/kleinerperkins/2012- kpcb-internet-trends-yearend-update, Dec. 2012.
12. Mohamed Eldawy A, Mokbel F: A Spatial MapReduce Language, 2014
13. J.L.Schnase, D.Q.Duffy, G.S.Tamkin, D.Nadeau, J.H.Thompson, C.M.Grieg, M.A.McInerney, W.P.Webster, MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud- enabled Climate Analytics-as-a-Service, Computers, Environment and Urban Systems, (2014).
14. B.K.Tannahill, M.Jamshidi, System of systems and bigdata analytics – bridging the gap, Comput.Electr.Eng.4 (2014)2–15.
15. J.Abawajy, Symbioses of Big Data and Cloud Computing: Opportunities & Challenges, (2013).
16. S. Aluru, Y.Simmhan, A special issue of journal of parallel and distributed computing: scalable systems for big data management and analytics, J.Parallel Distrib.Comput.73 (2013)896.
17. S.Hipgrave, Smarter fraud investigations with big data analytics, Netw.Secur. 2013(2013)7–9.