Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 875-884

http://www.ripublication.com

A Survey on the Concepts and Challenges of Big Data: Beyond the Hype

Dr.R.Saravanakumar[1] and Dr.C.Nandini[2]

[1]Deparment of Computer science & Engineering, Dayananda Sagar Academy of Tech,& Mang., Bangalore 560082, India.

[2] Department of Computer science & Engineering, Dayananda Sagar Academy of Tech,& Mang., Bangalore 560082, India.

Abstract

Today we are living in the cloud era, were an explosive amount of data being generated every minute of a day. Size is the only dimension that leaps out the big data. The data which has high volume, high velocity, and high varietyof informationassets that demand cost-effective, inventive forms of information processing for enhanced insights and decision making is called as Big Data. Due to huge volume and variety of data, new analysis, decision making algorithms are needed so big data provides opportunities in different places, but it has major research challenges.In recent days big data is one of the hottest research areas, which needs a change in the data processing and storage architecture. This paperattempts to offer the concepts like Data management and Analytics of big data. It also gives the challenges, and aspects of Big Data.

Keywords: Big Data, Analytics, Data Management

1. INTRODUCTION:

Nowadays, the people are living in the epoch, werefiery volume of data being generated every minute of a day. Data from social networking websites (Twitter, Facebook…), sensors, scientific data, medical dataand enterprises, that contribute massiveexplosion of data in large size.From corporate leaders to municipal planners and academics,big data is the main focus, and to some extent fear.The rapid rise of big data has left many unprepared.The impact of Big Data gives not only a huge potential for competition and also growth for the industries and private companies, but the right

use of Big Data increase productivity, innovation, and competitiveness for entire sectors and economies.

The major contribution of this paper is to bring forth the oft-neglected dimensions of big data like velocity, volume and variety etc… The key discourse on the big data,which is influenced and dominated by the advertisingefforts of largesoftware and hardware developers, focuses on predictive analyticsand structured data. But it disregards the largest component of big data,which is unstructured and is available in the form of audio, images, video, andunstructured text. It is estimated that the analytics-ready structured data forms only a small subset of big data. The unstructureddata, especially the data in the form video, is the largest component ofbig data that is archived partially.

The paper organized in such a way it starts with various definitions of big data, big data dimensions, concepts, challenges,and aspects of big data.

2. BIG DATA DEFINITIONS:

'Big data' as a conceptis blossoming and has unreliable origins.Big data definitions have evolved rapidly, which has raised someconfusion between the peoples. This is evident from an online survey of 154 C-suiteglobal executives conducted by Harris Interactive on behalf of SAPin April 2012 ("Small and midsize companies look to make big gainswith big data," 2012). When we ask the question "what is bigdata?" Size is the first characteristic that comes to mind. But there are other characteristics like Volume, Velocity and Variety (3 V's by **Laney, 2001**) of big data that emerged recently. The three V's are the major challenges in data management.

The **"Gartner IT Glossary, n.d."** defines the big data as "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms ofinformation processing for enhanced insight and decision making.''

"John Rauser" states Big data as "Any amount of data that's too big to be handled by one computer".

According to **McKinsey & Co** -Big Data is "the next frontier for innovation, competition and productivity".

"Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data." from **IBM**

"Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning."

3. PHENOMENON OF BIG DATA:

The data has increased into large scale in various fields for the past 20 years. According to the report by International Data Corporation (IDC) in 2011, the overall data created and copied volume in the world was 1.8ZB (~ 10^{21} B), which increased by nine times within five years. The figure will become twice at least every other two years in the future.

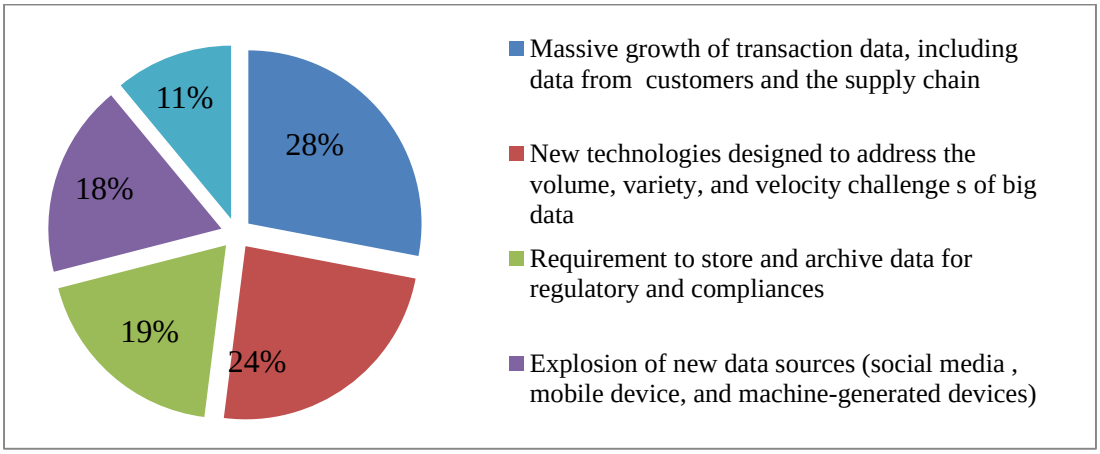
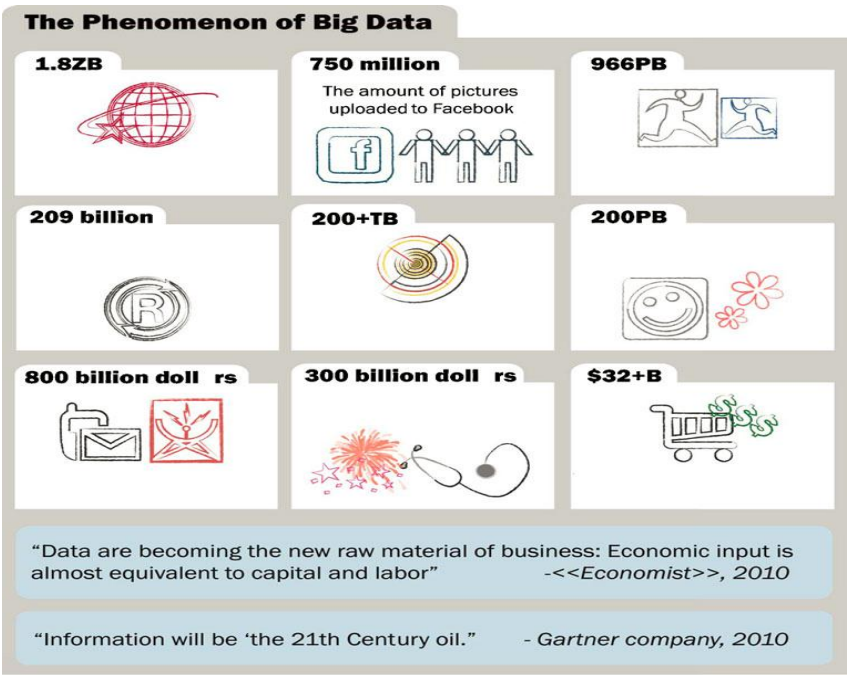


Fig 1. An online survey for big data definitions by global executives in 2014



4. DIMENSIONS OF BIG DATA:

Big data has various dimensions like Volume, Velocity, and Variety each one is described below.

VOLUME: Volume refers to the enormousness of data. The size of Big Data is reported in numerous terabyte to petabyte. **Beaver,Kumar, Li, Sobel, and Vajgel (2010)**report that Facebook processesup to one million photographs per second. One petabyte is equal to 1024 terabytes. Earlier estimates suggest that Facebook stored 260billion photos using storage space of over 20 petabytes.

VELOCITY: Velocity refers to the rate at which data are generated and thespeed at which it should be analyzed and acted upon. Unprecedented rate of data has been created from sensors and smart phones, which are needed for real time analytics and evidence-based planning. The conventional retailers are also generating high frequency data for example Wal-Mart processes more than one million transactions per hour**(Cukier, 2010).**

VARIETY: Variety refers to different type of data like structured, semi-structured and unstructured data. The data available in spread sheets or relational databases in the form of tabular data is structured. Images, audio, video, text (chat messages) is unstructured. Extensible Markup Language (XML), textual language used for exchanging data on web is semi-structured.

In addition to the 3 V's, we also have other dimensions like Veracity, Value and Variability which are described below.

VERACITY: Veracity represents the unreliability of data, which is inherent from different sources. It devised as fourth V by IBM. For example Social media has sentimental data of customers which are uncertain in nature. However they are valuable information that can be used for analytics which is another facet of Big Data.

VALUE: Value defines the attribute of big data. Oracle introduces Value and defines that big data are often characterized byrelatively "low value density". That is, the data received in theoriginal form usually has a low value relative to its volume. By analyzing large Volume of data High Value can be obtained.

VARIABILITY: Variability refers to the variation in data flow rates. The Velocity of big data is inconsistent and has periodic troughs and peaks. So Variability is in needed, introduced by SAS.SAS also introduces another term Complexity, which denotes that big data generated from myriad resources. Complexity arises after collecting data from different sources, it has to connect, clean, match and transform the data.

5. CONCEPTS IN BIG DATA:

The concept of big data focuses in two main sub processes: Data Management and Analytics. Data management should store, acquire, retrieve and prepare the data for

analytics by using various technologies like acquisition, annotation, aggregation etc... Analytics is used to analyze and acquire intelligence from the large amount of data. Now different analytical methods for structured, semi-structured and unstructured big data are given below.

The different big data analytics are Text Analytics, Audio Analytics, Video Analytics, Social Media Analytics and Predictive Analytics. In the following sections every technique is briefly reviewed.

Text Analytics

The term text analytics or text mining refers that extraction high quality information from textual data by using statistical pattern learning. It also encompasses statistical analysis, machine learning and computational linguistics which supports evidence-based decision making. The text analytics methods are Information Extraction (IE), Text Summarization, Question Answering and Sentiment Analysis. Tools for text analytics are SAS Text Analytics, IBM Text Analytics, SAP Text Analytics etc…

Audio Analytics

Audio analytics is used for unstructured audio data to analyze and extract the information from human spoken language which is also referred as speech analytics. It mainly used in call centers for analyzing million hours of recorded calls efficiently, which improve customer experience, monitor compliance with private policies and enhance sales turnover rates. It has two approaches transcript based approach and phonetic based approach. Tools for audio analytics are Marsyas, Vamp, SoundRuler and WaveSurfer.

Video Analytics

Video analytics is to monitor, detect, analyze, and determine meaningful data (temporal and spatial events) from video streams which is mainly used in health care, retail, transport, security and safety. Video analytics is also called as Video Content Analysis because in recent years this technology is used CCTV and Surveillance camera's for detecting breaches in restricted areas, recognizing suspicious activities, detecting loitering in precisezones etc… in effective and efficient way. Tools used are Ooyala, Vidyard, Vimeo Analytics etc…

Social Media Analytics

Social media analytics is a process of gathering information from social media websites (Facebook, Twitter**)**, blogs (Buffer Social, Grow) and analyze the data for decision making in business. It extracts the user opinion which is in the form of structured and unstructured data from social media. Content based analytics and Structure based analytics were used to perform social media analytics. Tools used are ViralWoot, Collecto, SumAll, Tailwind, Beevolve etc…

Predictive Analytics

Based on historical and present data predicting the future is predictive analytics. It does not tell what will happen in future, it articulates what might happen in future with reliability. Predictive analytics is based on statistical methods. Tools to perform predictive analytics are splunk, medalogux etc…

6. CHALLENGES OF BIG DATA

In the big data and analytics due to different techniques there are more challenges which are listed below.

1. Size –Volume
2. Awareness in using analytics
3. No relation towards user data
4. Data Visualization
5. Performance and Scalability
6. Distributed Storage
7. Content Validation

Size

In present era, new technologies have been hosted to reduce human burden and workload which enable us to store and query large data sets. Now it is difficult to use the complete data set because of volume (size) that enables us to new techniques.

- a new algorithm
- a new technology platform
- an ability to understand the data structure and business values
- an essential asset is intellectual property and patent portfolios

To overcome "Data Scientists" with multidisciplinary ideas are needed to face the competitive edge.

Awareness in using analytics

There is lack of understanding on how to use analytics to reduce the size, to improve the business value. This occurs because the objects which have to be modeled are huge, complex, and distributed. To overcome new modeling and simulation software are needed which should be simple, robust, distributed and parallel computing.

No relation towards user data

A major challenge in taking spatial data into account is it has no relation to the context about user's history, location, tasks, habits, etc… Here the goal is to take the context of user information that is not related to user and present the right information to right people. To perform this Context Awareness is the best approach which also increases the quality of big data. In big data context, Contextualization is used to combine heterogeneous form of data that improves that quality of classifier.

Data Visualization

Nowadays more peoples are using big data, which leads to a vibrant problem data visualization. In the internet when the users accessing the complex information, performing the associated tasks, using social networks (Facebook, twitter, etc…), conceptual networks the design issues multiply rapidly. To overcome this visual clutter which degrade the system performance has to be properly used.

Performance and Scalability

In Big Data, performance and scalability are two main issues that lead to store and process huge volume of data in big data systems and technologies. Process analysis can be used improve the performance and scalability of big data.

Distributed Storage

Obviously it is known that Big Data Analytics accessing distributed data sets has been increased. In recent days the organizations store and access their data in distributed environment for ease. So the volume and velocity of distributed storage increases which is a major challenge. Also security is the most complex problem in distributed data sets to solve. One solution is Data Marts can be used.

Content Validation

There is large number and type of different sources like blogs, social networking, news sites, and different type of content such as tweets, comments, articles, etc… has vast information which is difficult to validate and is major challenge. Machine Learning algorithms are used to extract information from web documents and validation can be performed.

7. BIG DATA TOOLS USED IN ACADEMIA

Many emerging methods have been developed to analyze, understand big data. Big data analytics like web-based, mobile health, behavioral and administrative data focused by genome. The large scale data produced in the procedure of sequencing,mapping, and analyzing the human genome placesgenomics in the realm of big data. Hundreds of petabytes ofdata may easily be generated by sequencing multiple humangenomes, and the analysis of the gene will further create moredata. There are platforms for genomics analysis. For example,The ENCODE consortium is an encyclopedia of functionalDNA elements to be used by the scientific community.NextBio offers a platform that sits on top of existing systemsto aggregate and analyze genomic data. Bina Technologies has built a system to enable users to access and analyzegenomic sequencing data, and its hybrid architecture keepsparts of data on the premises and parts in the cloud, which isused to speed up sequencing time and facilitate data transfer.

The Portable Genomics uses a mobile visualization platformfor genomics. The visualization concept has brought genomicsto professionals and consumers to

understand and use personalizedand preventive medicine. The analytics in this area aremostly based on either Frequentist or Bayesian approaches, although classical data mining techniques are increasinglybeing utilized. The joint work of computational biostatisticians,geneticists, computer scientists, and engineers is neededto advance this big genetic data area.Web-based and mobile health intervention studies have theadvantages of combining tailored approaches of face-to-faceinterventions with the scalability of public health interventionsvia the Internet with lower cost. It is a promising solutionfor healthcare due to its accessibility and time and cost savings.

They have been developed for the following clinical areas:

• Chronic conditions, such as heart disease, arthritis, and asthma.

• Health promotion, such as alcohol reduction, smoking cessation, diet, and exercise.

• Mental health, such as anxiety and depression.

The following table lists some tools used in various industries to manage big data.

COMPANY	SYSTEM
IBM	**Apache Hadoop, InfoSphere**
Cloudera	**CDH, Cloudera Standard, Cloudera Enterprise**
Oracle	**Oracle Big Data Appilance**
Google	**BigTable**
Yahoo!	**Sherpa**
Amazon	**SimpleDB**
Microsoft	**Dryad**
Facebook	**Apache Cassandra**
Hypertable	**HyperTable**
ASF	**Apache CouchDB**

8. STRENGTH AND WEAKNESS OF BIG DATA

Strengths

- Products are renovated.
- Risk Analysis performed.

Weakness

- Accuracy and uncertainty are difficult to assess.
- No privacy and confidentiality.

- Users Data kept in safe.
- Maintenance cost reduced.
- Data quality is low.
- Unknown Population Representation.

9. ASPECTS OF BIG DATA

1. Ability to align the data analysis algorithm towards the constraints generated like size, modularity, heterogeneous etc… of massive data collection.
2. Need more visual representation of different types of data at temporal & spatial level.
3. Simple rules have to be derived for content validation.
4. Quality Constraints should be defined for both storage and processing of big data.
5. Analysis tool has to be developed which should support interaction methods and also comparison among different scales and aggregations.

6. CONCLUSION:

This objective of this paper is to reflect the definition, dimension, concepts, challenges, strengths, weakness and some aspects of big data. The paper first offers what is big data from different author's perspective. Then highlights various V's – dimension of big data and tells that Veracity & Velocity are equally important. The paper main focus is on the concepts of big data which gives brief note about the data management and analytics. The paper also highlights the major challenges of big data that affects the social and business networks. Going beyond samples, additional valuable insights could beobtained from the massive volumes of less 'trustworthy' data.

REFERENCES

[1]. Laney, D. (2001, February 6). 3-D data management: Controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[2]. Gartner IT Glossary (n.d.). Retrieved from http://www.gartner.com/it-glossary/big-data/

[3]. McKinsey Global Institute, *Big Data: The next frontier for innovation, competition and productivity* (June 2011)

[4]. Beaver, D., Kumar, S., Li, H. C., Sobel, J., &Vajgel, P. (2010). Finding a needle inhaystack: Facebook's photo storage.In Proceedings of the nineth USENIX conference on operating systems design and implementation (pp. 1–8). Berkeley, CA,USA: USENIX Association.

[5]. Cukier K., The Economist, Data, data everywhere: A special report on managing information, 2010, February 25, Retrieved from http://www.economist.com/node/15557443
[6]. http://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics
[7]. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al.(2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.
[8]. Aggarwal, C. C. (2011). An introduction to social network data analytics.In C. C.Aggarwal (Ed.), Social network data analytics (pp. 1–15). United States: Springer.
[9]. Arbesman, S. (2013), "Five Myths about Big Data," Washington Post, August 16, 2013. (http://articles.washingtonpost.com/2013-08-16/opinions/41416362_1_big-data-data-crunching-marketing-analytics).
[10]. NESSI – Big Data White Paper - A New World of Opportunities, December 2012
[11]. A. Gandomi, M. Haider / International Journal of Information Management 35 (2015) 137–144
[12]. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis.National ScienceReview, 1(2), 293–314.
[13]. http://www.cio.com/article/2385497/data-management/6-practical-predictive-analytics-tools.html.
[14]. http://docs.splunk.com/Documentation/Splunk/6.2.3/Search/Aboutpredictiveanalytics
[15]. Gundecha, P., & Liu, H. (2012). Mining social media: A brief introduction. Tutorials inOperations Research, 1(4).
[16]. https://blog.bufferapp.com/social-media-analytics-tools
[17]. http://www.switchvideo.com/2012/07/10/top-tools-to-simplify-video-analytics/
[18]. Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., &Tufano, P. (2012).Analytics: The real-world use of big data. How innovative enterprises extractvalue from uncertain data. IBM Institute for Business Value. Retrievedfrom http://www-03.ibm.com/systems/hu/resources/the real word use ofbig data.pdf
[19]. Jiang, J. (2012). Information extraction from text. In C. C. Aggarwal, & C. Zhai (Eds.),Mining text data (pp. 11–41). United States: Springer.
[20]. YouTube Statistics (n.d.). Retrieved from http://www.youtube.com/yt/press/statistics.html
[21]. Min Chen · Shiwen Mao · Yunhao LiuMobile NetwAppl (2014)Big Data: A Survey, Springer Science+Business Media New York 2014
[22]. Hua Fang, Zhaoyang Zhang, Chanpaul Jin Wang, Mahmoud Daneshmand, Chonggang Wang, and Honggang Wang. "A Survey of Big Data Research", 0890-8044/15/$25.00 © 2015 IEEE, IEEE Network • September/October 2015