www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 7 July 2015, Page No. 13438-13446

Challenges of Big Data: Current Analysis

Mr. A. Jamaludeen* Mr. C. Senthil Kumaran** Ms. R. Suriya***

Assistant Professor – Senior Grade
Department of Computer Applications
Christ College of Engineering & Technology
Puducherry, India
E-mail: jam.ahamath@gmail.com

Assistant Professor
Department of Computer Applications
Christ College of Engineering & Technology
Puducherry, India
E-mail: senthilmca81@gmail.com

Research Student
Department of Computer Applications
Christ College of Engineering & Technology
Puducherry, India
E-mail: suriyaragunath@gmail.com

ABSTRACT

In today's technology, the internet has made new sources of vast amount of data available to business executives. Big data is certainly one of the biggest buzz phrases in IT today.It is comprised of datasets too large to be handled by traditional database systems. Big data refers to data volumes in the range of exabytes (1018) and beyond. Such volumes exceed the capacity of current on-line storage systems and processing systems. To remain competitive business executives need to adopt the new technologies and techniques emerging due to big data. We outline some of the challenges of big data in various big sectors. The researcher focuses how Hadoop provides fully functional end to end solutions that address a real world problem. And also this paper indicates the way to success of industries with big data technology.

KEYWORDS: Big data, Database, Hadoop, Internet, Technology.

1. INTRODUCTION

Numerous technological innovations are driving the dramatic increase in data size and data gathering. Data has always been around and there is always been a need for storage, processing, and management of data. However, the amount and type of data captured, stored, processed, and managed in the tools/technologies to gain insights into the data, make decisions, and so on. In ancient days, humans used very primitive ways of capturing/storing data like carving on stones,

metal sheets, wood, etc. Then with new inventions and advancements, humans started capturing the data on paper, cloth, etc.

As time progressed, the means of capturing/storage/management became punching cards followed by magnetic drums, laser disks, floppy disks, magnetic tapes, and finally today we are storing data on various devices like USB Drives, Compact Discs, Hard Drives, etc. In fact, the interest to capture, store, and process the data has enabled human beings to pass on knowledge and research from one generation to the next, so that the next generation does not have to re-invent the wheel. As we can clearly see from this trend, the capacity of data storage has been increasing exponentially, and today with the availability of the cloud infrastructure, potentially one can store unlimited amounts of data. Today Terabytes and Petabytes of data is being generated in our day today life. This is the reason why big data has become a updated area in the IT industries. "Big Data" originally meant the volume of data that could not be processed (efficiently) by traditional database methods and tools.

We define "Big Data" as the amount of data just beyond technology's capability to store, manage and process efficiently. Big Data is a heterogeneous mix of data both structured (traditional datasets –in rows and columns like DBMS tables, CSV's and XLS's) and unstructured data like e-mail attachments, manuals, images, PDF documents, medical records such as x-rays, ECG and MRI images, forms, rich media like graphics, video, audio,animation, contacts, forms and documents. The growth of data will never stop. In the distributed systems world, "Big Data" started to become a major issue in the late 1990's. Due to the impact of the world-wide Web and a resulting need to indexing and querying is rapidly mushrooming content. Big Data is changing the fundamentals of how information is managed and analyzed. Just ten years ago, the largest datasets were in the hundreds of terabytes, but in today's Big Data environment, it is not unusual for Fortune 100 companies to deal with datasets in the dozens or even hundreds of petabytes.Due to its specific nature of Big Data, it is stored in distributed file system architectures.

In contrast, the end to end framework presented in this paper successfully integrates Hadoop provide a fully functional end to end solution that addresses a real world problem. Here the paper presents the overall view of big data concepts and the way to success in industries with big data technology. Then the paper clearly denotes the characteristics, benefits, technologies and implementation of big data. Finally, this paper concludes the challenges of big data in the real world.

2. LITERATURE REVIEW

"Big Data" derives its name from the fact that the datasets are large enough that typical database systems are unable to capture, save, and analyze these datasets (Manyika et al., 2011). The actual size of big data varies by business sector, software tools available in the sector, and average dataset sizes within the sector (Manyika et al., 2011). Best estimates of size range from a few dozen terabytes to many petabytes (Manyiak et al., 2011).In August 2010, the White House, OMB, and OSTP

proclaimed that Big Data is a national challenge and priority along with healthcare and national security American Institute of Physics (AIP) 2010.In the current marketplace big data analytics hasbecome a business requirement for many organisations looking to gain a competitive advantage as evidenced by IBM's 2011 Global CIO study that places business intelligence and analytics as the main focus for CIOs over the next five years, on top ofvirtualization and cloud computing (IBM (2011)). According to McKinsey,4 Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, store, manage and analyze. There is no explicit definition of how big a dataset should be in order to be considered Big Data. New technology has to be in place to manage this Big Data phenomenon. IDC defines Big Data technologies as a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. According to O'Reilly, "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it."

3. CONCEPT OF BIG DATA

Big data includes structured data, semi-structured and unstructured data. Structured data are those data formatted for use in a database management system. Semi-structured and unstructured data include all types of unformatted data including multimedia and social media content. Big data are also provided by myriad hardware objects, including sensors and actuators embedded in physical objects, which are termed the Internet of Things. The term "Big Data" is believed to be originated from the web search companies who had to query loosely structured very large distributed data.

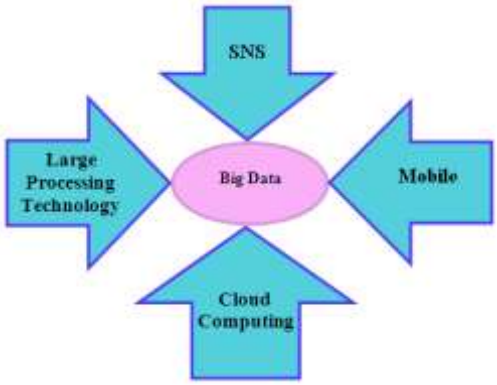

Figure 3.1: Big Data Concept

When making an attempt to understand the concept of Big Data, the word such as "Hadoop" and "Map Reduce" cannot be avoided. Hadoop and HDFS by Apache is widely used for storing and managing Big Data. Analyzing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. MapReduce is widely been used for the efficient analysis of Big Data. Traditional DBMS techniques like Joins and Indexing and other techniques like graph search is used for classification and clustering of Big Data. These techniques are being adopted to be used in Map Reduce.

3.1 Hadoop

Hadoop is a scalable, open source, fault-tolerant Virtual Grid operating system architecture for data storage and processing. It runs on commodity hardware, it uses HDFS which is fault-tolerant

high-bandwidth clustered storage architecture. It runs Map Reduce for distributed data processing and is works with structured and unstructured data.

3.2 Map Reduce

Map Reduce is a programming model for processing large-scale datasets in computer clusters. The MapReduce programming model consists of two functions, map() and reduce(). Users can implement their own processing logic by specifying a customized map() and reduce() function. The map() function takes an input key/value pair and produces a list of intermediate key/value pairs. The Map Reduce runtime system groups together all intermediate pairs based on the intermediate keys and passes them to reduce() function for producing the final results.

4. CHARACTERISTICS OF BIG DATA

Big Data is not just about the size of data but it also includes data variety and data velocity. Together, these three attributes are called as the three V's of Big Data.

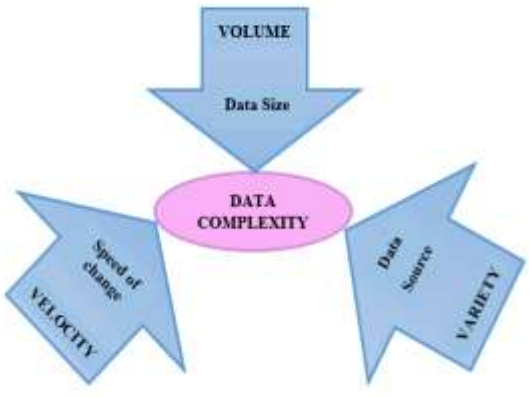

Figure 4.1: Three V's of Big Data

4.1 Volume

Volume refers to the size of data.This data is spread across different places, in different formats, in large volumes ranging from Gigabytes to Terabytes, Petabytes, and even more. Today, the data is not only generated by humans, but large amounts of data is being generated by machines and it surpasses human generated data. This size aspect of data is referred to as Volume in the Big Data world.

4.2 Velocity

In terms of velocity, data has gone from being handled in batches and periodically to having to be processed in real time.

4.3 Variety

Variety refers to the different formats in which the data is being generated/stored. Different applications generate/store the data in different formats. Apart from the traditional flat files, spreadsheets, relational databases etc., we have a lot of unstructured data stored in the form of images, audio files, video files, web logs, sensor data, and many others. This aspect of varied data formats is referred to as Variety in the Big Data world.

5. BIG DATA TECHNOLOGIES (HADOOP)

The driving force behind an implementation of big data is the software—both infrastructure and analytics. Primary in the infrastructure is Hadoop. Hadoop is the big data management software infrastructure used to distribute, catalog, manage, and query data across multiple, horizontally scaled server nodes. Yahoo! created it based on an opensource implementation of the data query infrastructure (originated at Google) called Map Reduce. It has a number of commercially

supported distributions from companies such as MapR Technologies and Cloudera. Hadoop is a framework for processing, storing, and analyzing massive amounts of distributed unstructured data. As a distributed file storage subsystem, Hadoop Distributed File System (HDFS) was designed to handle petabytes and exabytes ofdata distributed over multiple nodes in parallel.

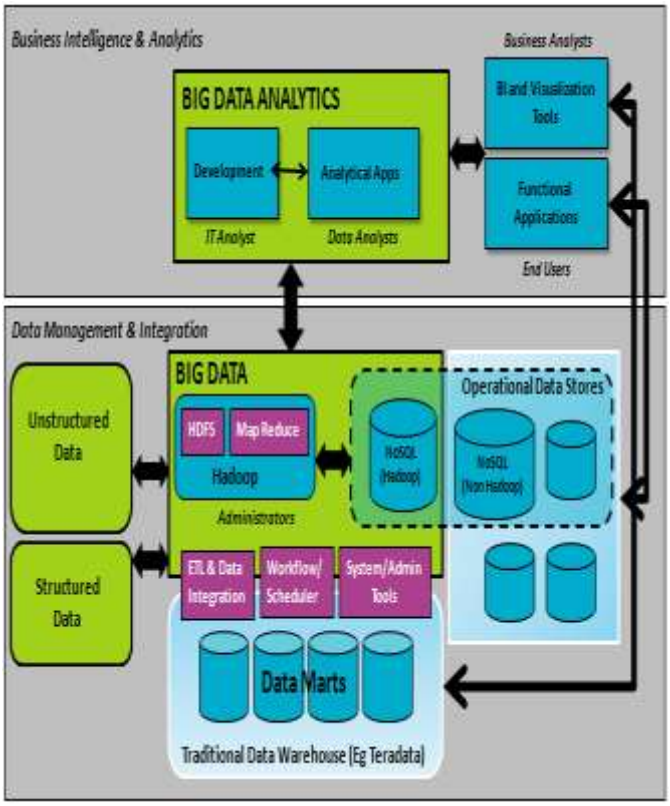

Figure 5.1: High level structure of Hadoop deployment

6. SOURCES OF BIG DATA

Just like the data storage formats have evolved, the sources of data are also evolved. There is a need for storing the data into a wide variety of formats. With the evolution and advancement of technology, the amount of data is being generated. The Sources of Big Data can be broadly classified into six different categories as shown below.

Figure 6.1: Sources of Big Data

6.1 Enterprise Data

There are large volumes of data in enterprises in different formats. Common formats include flat files, emails, Word documents, spreadsheets, presentations, HTML pages/documents, pdf documents, XMLs, legacy formats, etc. This data that is spread across the organization in different formats is referred to as Enterprise Data.

6.2 Transactional Data

Every enterprise has some kind of applications which involve performing different kinds of transactions like Web Applications, Mobile Applications, CRM Systems, and many more. To support the transactions in these applications, there are usually one or more relational databases as a backend infrastructure. This is mostly structured data and is referred to as Transactional Data.

6.3 Social Media

There is a large amount of data getting generated on social networks like Twitter, Facebook, etc. The social networks usually involve mostly unstructured data formats which includes text, images, audio, videos, graphics, animation, etc.

This category of data source is referred to as Social Media.

6.4 Activity Generated

The Activity Generated sources of big data include data from medical devices, censor data, surveillance videos, satellites, cell phone towers, industrial machinery, and other data generated mostly by machines. These types of data are referred to as Activity Generated data.

6.5 Public Data

This data includes data that is publicly available like data published by governments, research data published by research institutes, data from weather and meteorological departments, census data, Wikipedia, sample open source data feeds, and other data which is freely available to the public. This type of publicly accessible data is referred to as Public Data.

6.6 Archives

Organizations archive a lot of data which is either not required anymore or is very rarely required. In today's world, with hardware getting cheaper, no organization wants to discard any data, they want to capture and store as much data as possible. Other data that is archived includes scanned documents, scanned copies of agreements, records of ex-employees/completed projects, banking transactions older than the compliance regulations. This type of data, which is less frequently accessed, is referred to as Archive Data.

7. BIG DATA ADVANTAGES

In Big data, the software packages provide a rich set of tools and options where an individual could map the entire data landscape across the company, thus allowing the individual to analyze the threats he/she faces internally. This is considered as one of the main advantages as big data keeps the data safe. With this an individual can be able to detect the potentially sensitive information that is not protected in an appropriate manner and makes sure it is stored according to the regulatory requirements. There are some common characteristics of big data, such as

- Big data integrates both structured and unstructured data.
- Addresses speed and scalability, mobility and security, flexibility and stability.
- In big data the realization time to information is critical to extract value from various data sources, including mobile devices, radio frequency identification, the web and a growing list of automated sensory technologies.

8. BENEFITS OF BIG DATA

Now–a-days the big data is emerging technologies in modern world. Follow are just few benefits which are very much known to all of us:

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies

and retail organizations are planning their production.

- Using these data regarding the previous medical history of patients, hospitals are providing better and quick service.
- Financial services providers are adopting big data analytics infrastructure to improve their analysis of customers to help determine eligibility for bank loans, equity capital, insurance, mortgage, or credit.
- Airlines and trucking companies are using big data to track fuel consumption and traffic patterns across their fleets in realtime to improve efficiencies and save costs.
- Healthcare providers are managing and sharing patient electronic health records from multiple sources—imagery, treatments, and demographics and across multiple practitioners. In addition, pharmaceutical companies and regulatory agencies are creating big data solutions to track drug efficiency and provide more efficient and shorter drug development processes.
- Telecommunications are using big data solutions to analyze user behaviors and demand patterns for a better and more efficient power grid. They are also storing and analyzing environmental sensor data to provide insight into infrastructure weaknesses and provide better risk management intelligence.
- Media and entertainment companies are utilizing big data infrastructure to assist with decision making around customer life cycle retention and predictive analysis of their user base, and to provide more focused marketing and customer analytics.

9. ISSUES OF BIG DATA

There are three fundamental issue that need to be addressed in dealing with big data:

9.1 Storage issues

Current disk technology limits are about 4terabytes per disk. So, 1 exabyte would require 25,000

disks. Even if an exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an exabyte would takes about 2800 hours. Two solutions manifest themselves. First, process the data in place and transmit only the resulting information. In other words, "bring the code to the data", vs. the traditional method of "bring the data to the code". Second, it performs on data and transmits only that data which is critical to downstream analysis. In either case, integrity of data should be transmitted along with the actual data.

9.2 Management issues

Management is the most difficult problem to address with big data. This problem first surfaced a decade ago in the UK eScience initiatives where data was distributed geographically and "owned and managed" by multiple entities. Resolving

issues of access, metadata, utilization, updating, and governance have proven to be major stumbling blocks.

9.3 Processing issues

Assume that an exabyte of data needs to be processed in its entirety. For simplicity, assume the data is chunked into blocks of 8 words, so 1 exabyte =1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for end-to-end processing would be 20 nanoseconds. To process 1K petabytes would require a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing.

10. ILLUSTRATION OF BIG DATA

This table shows the example for big data

Data Set /Domain	Description
Large HadronCollider/Particle Physics (CERN)	13-15 petabytes in 2010
InternetCommunications(Cisco)	667 exabytes in 2013
Social Media	12+ Tbytes of tweetsevery day and growing. Average retweets are 144per tweet.
Human DigitalUniverse	1.7 Zbytes (2011) -> 7.9Zbytes in 2015 (Gantz and Reinsel 2011)
British Library UKWebsite Crawl	~ 110 TBytes per domaincrawl to be archived
Other	RFIDS, smart electricmeters, 4.6 billioncamera phones w/ GPS

11. CONCLUSION

Big data is the "new" business and social science frontier. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to massage and process data. Moreover, it has become clear that "more data is not just more data", but that "more data is different". "Big data" is just the beginning of the problem.With careful planning and predetermined expectations, creating an optimized big data deployment is relatively straightforward. Keep in mind only three or four years ago, broad commercial appeal for big data implementations was not a key requirement in data center design. Previously untapped sources of data are able to be stored and processed. Unstructured data previously available, such as invoice data, can be stored in a new, more convenient and meaningful format, and can employ text searching techniques. Big data poses opportunities and challenges for businesses. Business owners need to follow trends in big data carefully to make the decision that fits their businesses. We must support and encourage fundamental research towards addressing these technical challenges to achieve the promised benefits of Big Data.

REFERENCES

Agneeswaran VS, Tonpay P, Tiwary J (2013) Paradigms for realizing machine learning algorithms. Big Data 1(4):207–214.

American Institute of Physics (AIP). 2010. CollegePark, MD, (http://www.aip.org/fyi/2010/)

Borthakur D (2008) HDFS architecture guide. HADOOP APACHE PROJECT. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.pdf

Brad Brown, Michael Chui, and James Manyika, Are you ready for the era of „big data"?,McKinseyQuaterly,Mckinsey Global Institute, October 2011.

Carlos Ordonez, Algorithms and Optimizations for Big Data Analytics: Cubes, Tech Talks,University of Houston, USA.

Conte, R., Gilbert, N., Bonelli, G., &Helbing, D. (2011). FuturICT and social sciences: Big Data, big thinking. ZeitschriftfürSoziologie, 40, 412–413.

Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in USENIXSymposium on Operating Systems Design and Implementation, San Francisco, CA, Dec. 2004, pp. 137–150.

Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(1):107–113

Digital Universe Study (on behalf of EMC Corporation) (2012) Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. In: http://idcdocserv.com/1414

DunrenChe, MejdlSafran, and ZhiyongPeng, From Big Data to Big Data Mining: Challenges, Issues, and Opportunities, DASFAA Workshops 2013, LNCS 7827, pp. 1–15, 2013.

Fan J, Han F, Liu H (2014) Challenges of big data analysis. NatlSci Rev. Oxford University Press

Fox, B. 2011. "Leveraging Big Data for Big Impact", Health Management Technology, http://www.healthmgttech.com/

Gartner Group. 2012. "Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big Data By 2015," http://www.gartner.com/newsroom/id/2207915.

Hadoop,"PoweredbyHadoop,"http://wiki.apache.org/hadoop/PoweredBy.

Hadoop. http://hadoop.apache.org/

IBM (2011) The Essential CIO. In: http://www-935.ibm.com/services/uk/cio/pdf/CIE03073-GBEN-01.pdf

IBM. 2013. "What Is Big Data?," http://www.ibm.com/big-data/us/en/.

Jacobs, A. 2009. "Pathologies of Big Data", Communications of the ACM, 52(8):36-44.

Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013.

Lee K-H, Lee Y-J, Choi H, Chung YD, Moon B (2012) Parallel data processing with MapReduce: a survey. ACM SIGMOD Record 40(4):11–20

Manyika, J., Chui,. M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011, June). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Retrieved from

http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_ innovation

MarcinJedyk, MAKING BIG DATA, SMALL, Using distributed systems for processing, analysing and managing large huge data.

McKinsey Global Institute (2011) Big data: The next frontier for innovation, competition, and productivity. In:

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Mervis, J. 2012. "Agencies Rally to Tackle Big Data", Science, 336(4):22, June 6, 2012.

O'Reilly," Big Data Glossary", September 2011.

The Age of Big Data. Steve Lohr. New York Times, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

VinayakBorkar, Michael J. Carey, Chen Li, Inside "Big Data Management":Ogres, Onions, or Parfaits?, EDBT/ICDT 2012 Joint Conference Berlin, Germany,2012 ACM 2012, pp 3-14.