International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

Big Data Analytics: A Brief Survey

M. Praveen Kumar[1], SP. Santhosh Kumar[2], G. Ramya[1]
[1, 2]Assistant Professor
[1]Department of IT, [2]Department of CSE,
Rathinam Technical Campus, Coimbatore, Tamilnadu, India

ABSTRACT
In recent days, the size of the informations generated from modern information systems and digital technologies like IoT and Cloud computing is huge (ie. In TB). With this huge sized data, it is quite difficult to analysis and it is in the need of more effects at multiple levels to extract data. Big data analysis is the technique and used both for research and development. The idea of this paper is to give the brief about the big data concepts. Additionally, it will support for the researchers who is doing their research in the area of big data.

Keywords: *Big data analytics; Massive data; Structured data; Unstructured Data*

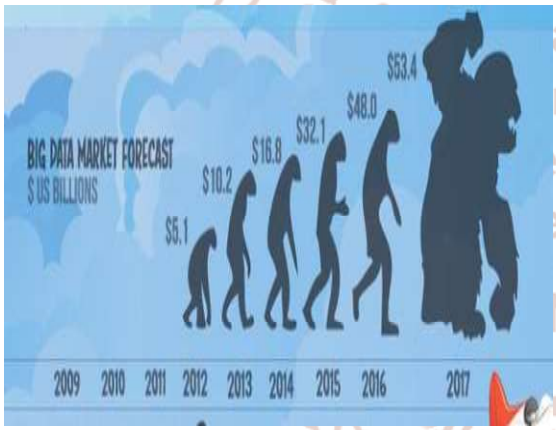
I. INTRODUCTION

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. It provides evolutionary breakthroughs in many fields with collection of large datasets. In general, it refers to the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing applications. These are available in structured, semi-structured, and unstructured format in pet bytes and beyond. Formally, it is defined from 3Vs to 4Vs. 3Vs refers to volume, velocity, and variety. Volume refers to the huge amount of data that are being generated everyday whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi structured etc. The fourth V refers to veracity that includes availability and accountability. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques [1]. Some of these extraction methods for obtaining helpful information was discussed by Gandomi and Haider [2]. The following Figure 1 refers to the definition of big data. However exact definition for big data is not defined and there is a believe that it is problem specific. This will help us in obtaining enhanced decision making, insight discovery and optimization while being innovative and cost-effective. It is expected that the growth of big data is estimated to reach 25 billion by 2015 [3]. From the perspective of the information and communication technology, big data is a robust impetus to the next generation of information technology industries [4], which are broadly built on the third platform, mainly referring to big data, cloud computing, internet of things, and social business. Generally, Data warehouses have been used to manage the large dataset. In this case extracting the precise knowledge from the available big data is a foremost issue. Most of the presented approaches in data mining are not usually able to handle the large datasets successfully. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such as data mining and statistical analysis. These challenges generally arise when we wish to perform knowledge discovery and representation for its practical applications. A fundamental problem is how to quantitatively describe the essential characteristics of big data. There is a need for epistemological implications in describing data revolution [5]. Additionally, the study on complexity theory of big data will help understand essential characteristics and formation of complex patterns in big data, simplify its representation, gets better knowledge abstraction, and guide the design of

computing models and algorithms on big data [4]. Much research was carried out by various researchers on big data and its trends [6], [7], [8].

II. What is BIG DATA?

- 'Big Data' is similar to 'small data', but bigger in size
- but having data bigger it requires different approaches:
  –Techniques, tools and architecture
- an aim to solve new problems or old problems in a better way
- Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.
- Walmart handles more than 1 million customer transactions every hour.
- Facebook handles 40 billion photos from its user base.
- Decoding the human genome originally took 10years to process; now it can be achieved in one week.



III. Four Characteristics of Big Data V3s

1st Character of Big Data – Volume

- A typical PC might have had 10 gigabytes of storage in 2000.
- Today, Facebook ingests 500 terabytes of new data every day.
- Boeing 737 will generate 240 terabytes of flight data during a single flight across the US.
- The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds containing environmental, location, and other information, including video.

2nd Character of Big Data – Velocity

- Click streams and ad impressions capture user behaviour at millions of events per second
- high-frequency stock trading algorithms reflect market changes within microseconds
- machine to machine processes exchange data between billions of devices
- infrastructure and sensors generate massive log data in real-time
- Online gaming systems support millions of concurrent users, each producing multiple inputs per second.

3rd Character of Big Data – Variety

- Big Data isn't just numbers, dates, and strings. Big Data is also geospatial data, 3D data, audio and video, and unstructured text, including log files and social media.
- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.
- Big Data analysis includes different types of data

4th Character of Big Data – Veracity

- Refers to the biases, noise and abnormality in data.
- Data that is being stored, and mined meaningful to the problem being analyzed.
- Inderpal feel veracity in data analysis is the biggest challenge when compares to things like volume and velocity.

IV. FEATURES OF BIG DATA

1. Storing Big Data

- Analyzing your data characteristics
  - Selecting data sources for analysis
  - Eliminating redundant data
  - Establishing the role of NoSQL
- Overview of Big Data stores
  - Data models: key value, graph, document, column-family
  - Hadoop Distributed File System
  - HBase
  - Hive

2. Selecting Big Data stores

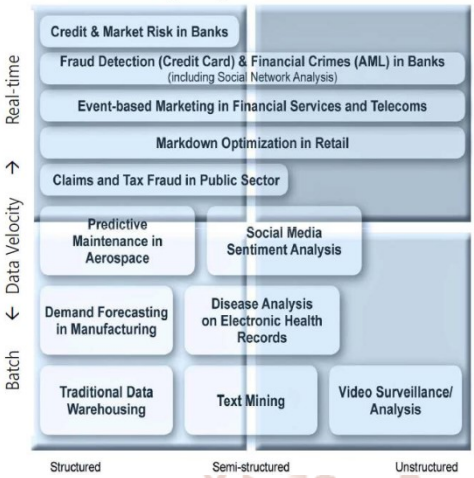
- Choosing the correct data stores based on your data characteristics
- Moving code to data
- Implementing polyglot data store solutions

- Aligning business goals to the appropriate data store

3. Processing Big Data

- Integrating disparate data stores
  - Mapping data to the programming framework
  - Connecting and extracting data from storage
  - Transforming data for processing
  - Subdividing data in preparation for Hadoop Map Reduce
- Employing Hadoop Map Reduce
  - Creating the components of Hadoop Map Reduce jobs
  - Distributing data processing across server farms
  - Executing Hadoop Map Reduce jobs
  - Monitoring the progress of job flows

4. The Structure of Big Data

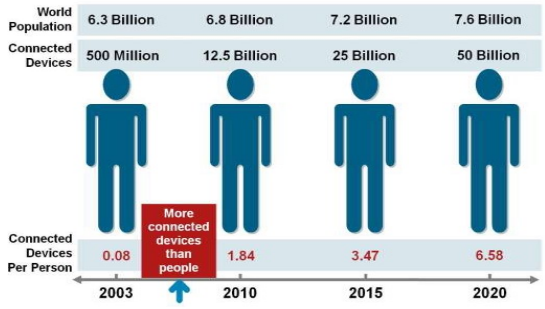


- Structured
  - Most traditional data sources
- Semi-structured
  - Many sources of big data
- Unstructured
  - Video data, audio data

5. Why Big Data?

Growth of Big Data is needed:

- Increase of storage capacities
- Increase of processing power
- Availability of data(different data types)
- Every day we create 2.5 quintillion bytes of data;
- 90% of the data in the world today has been created in the last two years alone
- FB generates 10TB daily
- Twitter generates 7TB of data Daily
- IBM claims 90% of today's stored data was generated in just the last two years.

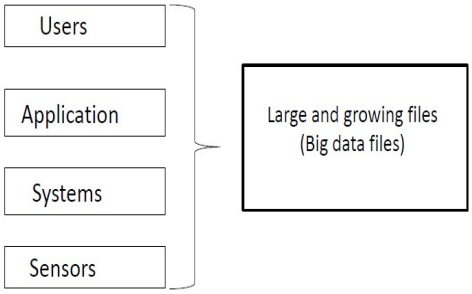


6. How Is Big Data Different?

- Automatically generated by a machine (e. g. Sensor embedded in an engine)
- Typically an entirely new source of data (e. g. Use of the internet)
- Not designed to be friendly (e. g. Text streams)
- May not have much values
  - Need to focus on the important part

V. APPLICATIONS AND TOOLS

1. Big Data sources



2. Big Data Analytics

- Examining large amount of data
- Appropriate information
- Identification of hidden patterns, unknown correlations
- Competitive advantage
- Better business decisions: strategic and operational
- Effective marketing, customer satisfaction, increased revenue

3. Types of tools used in Big-Data

- Where processing is hosted?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is stored?
  - Distributed Storage (e.g. Amazon S3)
- What is the programming model?
  - Distributed Processing (e.g. Map Reduce)
- How data is stored & indexed?
  - High-performance schema-free databases (e.g. MongoDB)

- What operations are performed on data?
  - Analytic / Semantic Processing

4. RISKS OF BIG DATA

- Will be so overwhelmed
- Need the right people and solve the right problems
- Costs escalate too fast
- Isn't necessary to capture 100%
- Many sources of big data is privacy
- self-regulation
- Legal regulation

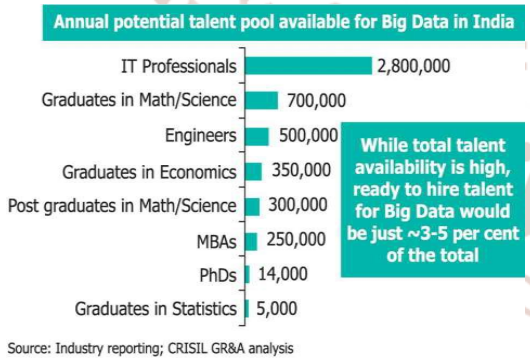
VI. BENEFITS AND FUTURE

1. How Big data impacts on IT

- Big data is a troublesome force presenting opportunities with challenges to IT organizations.
- By 2015 4.4 million IT jobs in Big Data ; 1.9 million is in US itself
- India will require a minimum of 1 lakh data scientists in the next couple of years in addition to data analysts and data managers to support the Big Data space.

2. Potential Value of Big Data

- $300 billion potential annual value to US health care.
- $600 billion potential annual consumer surplus from using personal location data.
- 60% potential in retailers' operating margins.



3. Benefits of Big Data

- Real-time big data isn't just a process for storing pet bytes or exabytes of data in a data warehouse, It's about the ability to make better decisions and take meaningful actions at the right time.
- Fast forward to the present and technologies like Hadoop give you the scale and flexibility to store data before you know how you are going to process it.
- Technologies such as Map Reduce, Hive and Impala enable you to run queries without changing the data structures underneath.
- Our newest research finds that organizations are using big data to target customer-centric outcomes, tap into internal data and build a better information ecosystem.
- Big Data is already an important part of the $64 billion database and data analytics market
- It offers commercial opportunities of a comparable scale to enterprise software in the late 1980s
- And the Internet boom of the 1990s, and the social media explosion of today.

4. Future of Big Data

- $15 billion on software firms only specializing in data management and analytics.
- This industry on its own is worth more than $100 billion and growing at almost 10% a year which is roughly twice as fast as the software business as a whole.
- In February 2012, the open source analyst firm Wikib on released the first market forecast for Big Data , listing $5.1B revenue in 2012 with growth to $53.4B in 2017
- The McKinsey Global Institute estimates that data volume is growing 40% per year, and will grow 44x between 2009 and 2020.

VII. CONCLUSION

In recent years the amount of data has been increased drastically. To analyse those data became a challenging task for humans. In this paper, we discussed about big data and its various challenges and tools to analyse huge data. From the brief concepts, it is easily understood that every big data tool has specific functions. We belive that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

REFERENCES

1. D. P. Acharjya, Kauser Ahmed P, A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools,(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.
2. M. K. Kakhani, S. Kakhani and S. R. Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.

3. A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015), pp.137-144.
4. C. Lynch, Big data: How do your data grow?, Nature, 455 (2008), pp.28-29.
5. X. Jin, B. W. Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research, 2(2) (2015), pp.59-64.
6. R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1) (2014), pp.1-12.
7. C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275 (2014), pp.314-347.
8. K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.
9. S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of map reduce for imbalanced big data using random forest, Information Sciences, 285 (2014), pp.112-137.
10. MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.
11. R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, 2013, pp.17-22.