



# INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

## A SURVEY ON BIG DATA ISSUES

AMRINDER KAUR

Assistant Professor, Department of Computer Science & Engineering, University Institute of Engineering and Technology, Kurukshetra University, Haryana

Accepted Date: 29/08/2015; Published Date: 01/09/2015

**Abstract:** - This is data driven era, from many years companies have been building data warehouses to analyze activity and produce insights for decision makers to act on to improve performance. Early analytical systems are based on classic pattern where data from multiple functioning systems is captured, cleaned, transformed & united before loading it into a data warehouse. However, this traditional environment evolves continuously; many new complex types of data have emerged that anyone wants to analyze to enrich what they already know. Customers are creating huge amounts of new data on social networks like tweeter, face book, LinkedIn etc., on-line news matters, competitor online content, climate data etc. Archived data is also being revived for exploration to optimize business operations. The result is richness of new data sources, rapidly increasing data volumes and a outbreak of new data streams that all need to be analyzed.

**Keywords:** Big Data, Volume, Velocity, Variety, Issues

Corresponding Author: ER. AMRINDER KAUR



PAPER-QR CODE

Access Online On:

[www.ijpret.com](http://www.ijpret.com)

How to Cite This Article:

Amrinder Kaur, IJPRET, 2015; Volume 4 (1): 70-75

## INTRODUCTION

Any data that can challenge the current technology in some way can consider as big data. In another words it is data whose scale, variety, and complexity require new architecture, algorithms, techniques and analytics to manage it and extract value and hidden knowledge and patterns from it. Various new sources of big data include structured data i.e. semi structured data and unstructured data i.e. location specific data (from traffic management & from the tracking of personal devices). [1][2] Due to large size of data it becomes very tedious task to perform effective analysis using current techniques. Since Big data is a forthcoming technology in the market which can bring vast benefits to the business organizations and personals. Big data is difficult to work with most relational database management systems and desktop statistics and visualization packages, and requiring massively parallel software consecutively run on tens, hundreds, or even thousands of servers. Due to Big data's important properties like volume, velocity, variety, variability and value it becomes necessary that various challenges and issues associated in adapting this technology are need to be understood [1][5]. In section 2 big data characteristics i.e. volume, velocity, variety, value and veracity will be discussed. In section 3 issues related to characteristics, transportation, storage, data management & processing will be discussed.

## BIG DATA CHARACTERISTICS

### 2.1 Volume:

The word "Big" in big data itself expresses the volume. Currently data existing is in petabytes ( $10^{15}$ ) and is likely to increase to yottabytes ( $10^{24}$ ) in nearby future. Data volume measures the amount of data available to an organization, person etc. Data comes from different sources that may include transactional enterprise data, machine-generated data, social data and data collected from sensors. [6][7]

### 2.2. Velocity

Data is streaming in at unmatched speed and must be tackled with in a timely manner. Velocity is just not being limited to the velocity of incoming data, it also concerned about the speed at which the data flows and accumulated. Social media data streams, such as Twitter, produce a large influx of data at high frequency, which ensures large volumes, over 8 TB per day (Oracle, 2012). Responding quickly enough to deal with data velocity is a challenge for most organizations. For e.g. detection of online fraud in financial activities. [6][7]

### **2.3. Variety:**

Data variety is amount of the richness of the data representation it may include structured, semi structured and unstructured data. Structured data consist of traditional data (machine generated, web log data, point-of-sale data, financial data), semi structured data (web Pages, social media sites, e-mail etc.), whereas unstructured data consists of satellite images, photograph, videos & social media data etc. [6] [7]

### **2.4. Value:**

Data value measures the worth of data in making decisions. User can run queries against the data stored and deduce significant results from the filtered data obtained. These reports help people to find the business trends, patterns and hidden values according to which they can plan future strategies and take decisions.[6] [7]

### **2.5. Veracity**

Big Data veracity refers to the noise and abnormality in data or veracity is about data understandability. The data being stored is useful for the problem being analyzed & solved. It is the biggest challenge in data analysis when compares to other characteristics. Many techniques are used to clean data and processes to avoid 'dirty data' from accumulating in your systems. [6] [7]

## **3. ISSUES IN BIG DATA**

Issues related to Big Data are some of the conceptual points that should be understood by the organization to implement the technology effectively. Big data are important to know and crucial to handle but it is need not to be jumbled

### **3.1. Issues related to the Characteristics**

a. Data Volume: As data volume rises, the value of different data records will decline in proportion to age, type, richness, and amount among other factors. The social networking sites existing are producing data in order of terabytes on each day and this much volume of data is definitely difficult to be handled using the existing traditional systems

b. Data Velocity: Our traditional systems are not capable enough on performing the analytics on the data which is continuous in motion. E-Commerce has fleetly increased the speed and richness of data used for different business transactions. Data velocity issue is much more complicated than a bandwidth issue.

c. Data Variety: Big data is totally different consisting of structured (machine generated, web log data, point-of-sale data, financial data), semi structured (web Pages, social media sites, e-mail etc.) and even unstructured data (satellite data, video, audio etc.) which is hard to be tackled by the existing analytic systems. From an analytic viewpoint, it is probably the biggest hurdle to effectively using large dimensions of data. Significant challenges include incompatible data formats, non-aligned data structures, and unpredictable data semantics that can lead to analytic sprawl.

d. Data Value: Stored data is being used by different organizations for data analytics. Data value is usefulness of data for the organization. It will reveal insights of data. Business leaders and the IT professionals use their organization data just to increase value of their business and getting more and more profit

e. Data Complexity: It concerned with the effect of changes on data i.e. minute change in one or more elements will lead to large changes, small changes or no changes in big data structures. One current difficulty of big data is working with it using relational databases and statistics/visualization packages, requiring software consecutively running on many servers. It is essential to link, match, clean and transform data across systems coming from numerous sources. To analyze the data effectively it is essential to connect and correlate relationships, hierarchies and multiple data linkages otherwise data can quickly spiral out of control.[3][4]

### 3.2. Storage and Transport Issues

The quantity of data has exploded each time when a new storage medium invented. Recent data explosion is mainly due to social media. Furthermore, data is being created by everyone and everything, (from ubiquitous devices to Super Computers, by professionals such as researcher, reporters, writers, students etc.).

Current disk technology is limited near about terabytes ( $10^{12}$ ) per disk. So, 2 Exabyte ( $10^{18}$ ) would need 50,000 disks. If an Exabyte of data could be processed on a single system, it would be incapable to directly attach the required number of disks. Access to that data would overwhelm present communication networks. [7]

### 3.3. Data Management Issues

Data Management is the most difficult problem to address with big data. Major stumbling blocks in data management is to resolve different problems like access, utilization; updating, governance, and reference. The sources of the data are distinguished by size, by format, and by method by which data is collected. Individuals contribute into digital data in the form of

documents, drawings, pictures, sound and video recordings etc., with or without including sufficient metadata unfolding what, when, where, who, why and how it was gathered and its source. During the collection of data by manual methods, where severe protocols are often followed in order to ensure accuracy and validity where as digital data collection is not constrained at all. According to volume which is generated every day, it is not possible to validate each and every data item. Need of new approaches to data qualification and validation are desirable. The abundance of digital data representation forbids a personalized methodology for data collection. Conclusively, there is no complete & perfect big data management solution yet. [8]

### **3.4. Processing Issues**

Suppose that an Exabyte of data needs to be processed. For ease, assume the data is chunked into blocks of 8 words, so 1 Exabyte = 1K petabytes. Assuming a processor uses 100 instructions on one block at 5 gigahertz, the required time for end to end processing would be 20 ns. For processing of 1K petabytes would require 635 years roughly as a processing time.

Thus, effective processing of Exabyte of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable [8]

### **CONCLUSION**

As Big Data is current technology, it need new algorithms, new techniques to tackle its volume, variety and velocity. Big data is tremendous amount of data which has been evolved from multiple sources like weather forecasting, social sites etc. With the help of this data analysis companies and business leaders will identify new opportunities. This paper will help researcher to identify various issues concerned with big data.

### **REFERENCES:**

1. Agrawal Divyakant etal., “ White Paper: Challenges and Opportunities with Big Data”, February 2012
2. Ferguson Mike.,” White Paper: Architecting a Big Data Platform for Analytics”, Intelligent Business Strategies, October, 2012
3. Jaseena K. U. etal. , “ Issues, Challenges & Solutions: Big Data Mining” , Academy & Industry Research Collaboration Center, Vol 4, 2014

4. Katal,A etal.," Big Data: Issues, challenges, Tools and Good Practices", IEEE, pp 404-409, August 2013
5. Kattiman Suvarna. L. etal.," A survey on handling Big Data issues by Improving Performance on Hybrid Cloud Management Through Agent Based Cloud Computing Framework", International Journal of Advanced Technology in Engineering and Science, Vol 3, Issue 1, April 2015
6. Satyanarayana Lenka Venkata, "A Survey on Challenges & Advantages in Big Data", International Journal of Computer Science & Technology, Vol 6, Issue 2, June 2015.
7. Stephen Kaisler etal.," Big Data: Issues and Challenges Moving Forward", 46th Hawaii International Conference on System Sciences, 2013
8. Yadav Chanchal etal., "Algorithm and approaches to handle large Data-A Survey", International Journal of Computer Science and Network, Vol 2, Issue 3, 2013