ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.3, Special Issue 4, April 2015

National Conference On Emerging Trends in Information, Digital & Embedded Systems (NC'e-TIDES -15)

Organized by

Dept. of ECE, Annamacharya Institute Of Technology & Sciences, Rajampet, Andhra Pradesh-516126, India held on 28th February 2015

# Big Data Analytics: Issues and Challenges

P.S.Nagendra Babu[1], R.V.Ramana Kumar[2], K.Phaneendra Kumar[3]

Lecturer, Department of Computer Engineering,, Government Polytechnic, Obulavaripalli, Kadapa(dist), Andhra Pradesh, India[1]

Principal, Government Polytechnic, Obulavaripalli, Kadapa(dist), Andhra Pradesh, India[2]

Lecturer, Department of ECE, Government Polytechnic, Obulavaripalli, Kadapa(dist), Andhra Pradesh, India[3]

**ABSTRACT:** Big data, true to its name, deals with large volume of data characterized by velocity, variety and value. The data is being collected and stored at unprecedented rates. There is a great challenge not only to store and manage the large volume of data, but also to analyze and extract meaningful information from it. There are several approaches to collecting, storing, processing, and analyzing big data. The main focus of this paper is to discuss about various challenges related to processing and analysis of Big data.

**KEYWORDS:** Big data, Big data Analytics, Challenges

## I. INTRODUCTION

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences. The concept of big data has been endemic within computer science since the earliest days of computing. "Big Data" originally meant the volume of data that could not be processed (efficiently) by traditional database methods and tools. The original definition focused on structured data, but most researchers and practitioners have come to realize that most of the world's information resides in massive, unstructured information, largely in the form of text and imagery. We define "Big Data" as the amount of data just beyond technology's capability to store, manage and process efficiently. These imitations are only discovered by a robust analysis of the data itself, explicit processing needs, and the capabilities of the tools (hardware, software, and methods) used to analyze it.

The current growth rate in the amount of data collected is staggering. A major challenge for IT researchers and practitioners is that this growth rate is fast exceeding our ability to both: (1) design appropriate systems to handle the data effectively (2) analyze it to extract relevant meaning for decision making. In this paper we discuss about the big data analysis phases and also challenges related to its analysis.

## II. WHAT IS BIG DATA

Big Data means Data sets whose volume and /or variety is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time that is relevant to business. The difficulty can be related to data capture, storage, search, sharing, analytics and visualization etc. It means data that's too big, too fast, or too hard for existing tools to process. Here, "Too big" means that organizations increasingly must deal with petabyte-scale collections of data that come from click streams, transaction histories, sensors, and elsewhere. "Too fast" means that not only is data big, but it must be processed quickly — for example, to perform fraud detection at a point of sale or determine which ad to show to a user on a webpage. "Too hard" means data that doesn't fit into an existing processing tool or that needs some kind of analysis that existing tools can't readily provide.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.3, Special Issue 4, April 2015

National Conference On Emerging Trends in Information, Digital & Embedded Systems (NC'e-TIDES -15)

Organized by

Dept. of ECE, Annamacharya Institute Of Technology & Sciences, Rajampet, Andhra Pradesh-516126, India held on 28th February 2015

Big Data is characterized by the following 5 **Vs**:

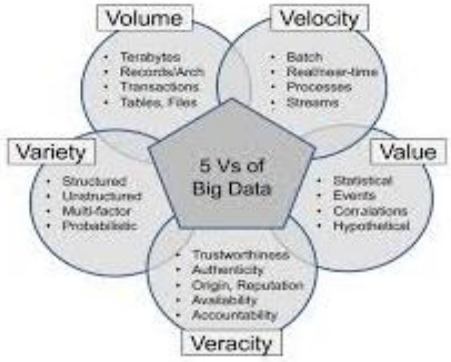


Figure 1: The 5 V's of Big data

• **Volume** - the vast amount of data generated every second that are larger than what the conventional relational database infrastructures can cope with.
• **Velocity** - the frequency at which new data is generated, captured, and shared.
• **Variety** - the increasingly different types of data (from financial data to social media feeds, from photos to sensor data, from video capture to voice recordings) that no longer fits into neat, easy to consume structures.
• **Veracity** - the disarrayed data (Facebook posts with hash tags, abbreviations, typos, and colloquial speech)
• **Value** - Data value measures the usefulness of data in making decisions. It has been noted that "the purpose of computing is insight, not numbers". Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data.

## III. BIG DATA ANALYTICS

Big Data analytics – the process of analyzing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools.

The technological advances in storage, processing, and analysis of Big Data include (a) the rapidly decreasing cost of storage and CPU power in recent years; (b) the flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage; and (c) the development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing. These advances have created several differences between traditional analytics and Big Data analytics (Figure 2).


Figure 2. Technical factors driving Big Data adoption

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.3, Special Issue 4, April 2015

National Conference On Emerging Trends in Information, Digital & Embedded Systems (NC'e-TIDES -15)

Organized by

Dept. of ECE, Annamacharya Institute Of Technology & Sciences, Rajampet, Andhra Pradesh-516126, India held on 28th February 2015

1. Storage cost has dramatically decreased in the last few years. Therefore, while traditional data warehouse operations retained data for a specific time interval, Big Data applications retain data indefinitely to understand long historical trends.
2. Big Data tools such as the Hadoop ecosystem and NoSQL databases provide the technology to increase the processing speed of complex queries and analytics.
3. Extract, Transform, and Load (ETL) in traditional data warehouses is rigid because users have to define schemas ahead of time. As a result, after a data warehouse has been deployed, incorporating a new schema might be difficult. With Big Data tools, users do not have to use predefined formats. They can load structured and unstructured data in a variety of formats and can choose how best to use the data.

## IV. CHALLENGES IN BIG DATA ANALYTICS

Having described the multiple phases in the Big Data analysis pipeline, we now turn to some common challenges that underlie many, and sometimes all, of these phases.

### A. *Heterogeneity and Incompleteness*

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

Consider an electronic health record database design that has fields for birth date, occupation, and blood type for each patient. What do we do if one or more of these pieces of information is not provided by a patient? Obviously, the health record is still placed in the database, but with the corresponding attribute values being set to NULL. A data analysis that looks to classify patients by, say, occupation, must take into account patients for which this information is not known. Worse, these patients with unknown occupations can be ignored in the analysis only if we have reason to believe that they are otherwise statistically similar to the patients with known occupation for the analysis performed. For example, if unemployed patients are more likely to hide their employment status, analysis results may be skewed in that it considers a more employed population mix than exists, and hence potentially one that has differences in occupation-related health-profiles.

Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge.

### B. *Scale*

Of course, the first thing anyone thinks of with Big Data is its size. After all, the word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.3, Special Issue 4, April 2015

National Conference On Emerging Trends in Information, Digital & Embedded Systems (NC'e-TIDES -15)

Organized by

Dept. of ECE, Annamacharya Institute Of Technology & Sciences, Rajampet, Andhra Pradesh-516126, India held on 28th February 2015

First, over the last five years the processor technology has made a dramatic shift - rather than processors doubling their clock cycle frequency every 18-24 months, now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. In the past, large data processing systems had to worry about parallelism across nodes in a cluster; now, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that were applied in the past for processing data across nodes don't directly apply for intra-node parallelism, since the architecture looks very different; for example, there are many more hardware resources such as processor caches and processor memory channels that are shared across cores in a single node.

The second dramatic shift that is underway is the move towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals (e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap) into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger clusters (that are required to deal with the rapid growth in data volumes). This places a premium on declarative approaches to expressing programs, even those doing complex machine learning tasks, since global optimization across multiple users' programs is necessary for good overall performance. Reliance on user-driven program optimizations is likely to lead to poor cluster utilization, since users are unaware of other users' programs. System-driven holistic optimization requires programs to be sufficiently transparent, e.g., as in relational database systems, where declarative query languages are designed with this in mind.

C. *Timeliness*

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge as described earlier, and a timeliness challenge described next.

There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria. For example, consider a traffic management system with information regarding thousands of vehicles and local hot spots on roadways. The system may need to predict potential congestion points along a route chosen by a user, and suggest alternatives. Doing so requires evaluating multiple spatial proximity queries working with the trajectories of moving objects. New index structures are required to support such queries. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

D. *Privacy*

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

An ISO 3297: 2007 Certified Organization Vol.3, Special Issue 4, April 2015

National Conference On Emerging Trends in Information, Digital & Embedded Systems (NC'e-TIDES -15)

Organized by

Dept. of ECE, Annamacharya Institute Of Technology & Sciences, Rajampet, Andhra Pradesh-516126, India held on 28th February 2015

Consider, for example, data gleaned from location-based services. These new architectures require a user to share his/her location with the service provider, resulting in obvious privacy concerns. Note that hiding the user's identity alone without hiding her location would not properly address these privacy concerns. An attacker or a (potentially malicious) location-based server can infer the identity of the query source from its (subsequent) location information. Several other types of surprisingly private information such as health issues (e.g., presence in a cancer treatment center) or religious preferences (e.g., presence in a church) can also be revealed by just observing anonymous users' movement and usage pattern over time. Note that hiding a user location is much more challenging than hiding his/her identity. This is because with location-based services, the location of the user is needed for a successful data access or data collection, while the identity of the user is not necessary.

There are many additional challenging research problems. For example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases. In addition, real data is not static but gets larger and changes over time; none of the prevailing techniques results in any useful content being released in this scenario. Yet another very important direction is to rethink security for information sharing in Big Data use cases. Many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.

## V. CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

## REFERENCES

[1] Douglas and Laney, "The importance of 'big data': A definition," 2008.
[2] JASON. 2008. "Data Analysis Challenges", The
Mitre Corporation, McLean, VA, JSR-08-142
[3] "Big Data Analytics" by Sachidanand Singh – Paper published in 2012 International Conference on Communication Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India
[4]. Prof. Roberto V. Zicari (2013), "The challenges and opportunities of big data".
[5] "Big Data: Issues and Challenges Moving Forward" *46th Hawaii International Conference on System Sciences*, 2013.
[6]. Hansen, C. (2013), "Big Data: A Scientific Visualization Perspective", SCI Institute Professor of Computer Science, University of Utah
[7] " Big Data Visualization tool with Advancement of Challenges" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, issue 3, March 2014.
[8]. Tien, J.M. (17-19 July, 2013), "Big Data: Unleashing information".
[9].Big data analytics. Adoption and employment trends 2012-2017. e-skills UK,January 2013. Available from http://ec.europa.eu/digital-agenda/en/news/bigdata-analytics-assessment-demand-labour-and-skills-2012-2017 Accessed 25 June.
[10] X. Zhou, J. Lu, C. Li, and X. Du, "Big data challenge in the management perspective," *Communications of the CCF*,vol. 8, pp. 16–20, 2012.23 2014