ISSN No. 2395-0900
International Journal of Engineering Studies and Technical Approach
IJESTA

# Comparative Study on Data and Big Data Analytics and its Research Challenges

**Lakshmi . S,**

**Assistant Professor, Department of Computer Science Sri Adi Chunchanagiri Women's College, Cumbum, Tamilnadu - India.**

## ABSTRACT

*This paper aims to provide a comprehensive review of the big data state of the art, conceptual explorations, major benefits, different types of data, the problems in the traditional data analytics process. After that we discussed about what is big data and different parameters of big data. How the problems in traditional data analytics is being resolved by big data analytics and research challenging aspects. In addition to that, several future directions for big data research are highlighted.*

**Keywords:** *Big data; data analytics; data storage procedure, research challenges, big data architecture.*

## I. INTRODUCTION

Big data is a term encompassing different types of complicated and large datasets that is hard to process with the conventional data processing systems. Numerous challenges are in place with big data like storage, transition, visualization, searching, analysis, security and privacy violations and sharing. The exponential growth of data in all fields demands the revolutionary measures required for managing and accessing such data. We have highlighted the need for the research in big data, in order to manage the online bio-logical data avenue. They have foreseen the importance of big data in the biological and biomedical research. It has exploded in such a way that it has marginalized a regulatory schema for personally identifiable information [2]. This is possible by analyzing the meta data and by using the predictive, aggregated findings thereby combining the previous discrete data sets. The significance of big data analytics comes when enterprises choose a technical stack, which dictates the type of data to store and to process. Relational Data Base management Systems are doing fine with structured data and continue to be the choice for many requirements. But for the exponential growth of unstructured data in terabytes or even peta bytes, derived from social networks, sensor networks and other federated data with replications, big data is the answer for handling such data.

Due to the vast use of internet, we got some data having huge-volume, high-velocity, and wide-variety. The relational database could not able to handle and process that data. Hence, a new type of data known as Big Data was introduced with different concepts and different technologies. According to the behaviour, the data can be categorized into the following types.

ISSN No. 2395-0900
International Journal of Engineering Studies and Technical Approach
IJESTA

- Structured Data – the data produced from several research article and generals, business application such as retail, finance, bioinformatics and other such traditional databases in various sources such as RDBMS, OLAP and data warehousing etc.
- Unstructured Data – Data created by the users as social media sites, trading market s data, healthcare data, Web forums etc.
- Semi-structured Data- XML formatted data, HTML,CSV and RDF

## II. TRADITIONAL DATA STORAGE

### PROCEDURE

In traditional data storage procedure the data model defines some properties. The structured data needs to satisfy all the properties defined by the data model in order to be stored in the database. So, the data will only be accepted if and only it satisfies all the properties defined by the data model. If the data doesn't satisfy at least a single property then the data will be rejected. The data basically stored in the row column format in relational databases. SQL is used to handle the data and to process it.

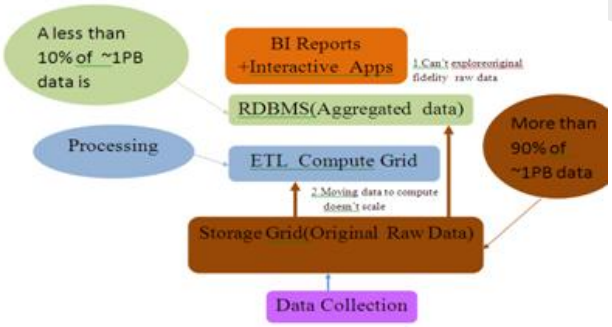


Fig.2.1. Problems in Traditional Data Analytics Procedure

In Traditional data analytics procedure it not so easy to process the data correctly and efficiently. In traditional system, the data is first collected from various sources, then it is stored in the database (Storage grid). The problem arises when large amount of data with high velocity comes to the storage grid. The storage grid can't handle the huge amount of data, even if it stores the data then most of the data gets archived. Suppose 1 PB of data comes to the storage grid in order to analyze, then 90 % of the data gets archived due to shortage of storage space. Now, only 10 % of the data remains for the analysis and the entire 90 % of the data gets archived considering the premature death of the data. The remaining 10 % of the data goes to the ETL compute grid. The ETL (Extraction, Transformation, Load) consists of the following three steps : i. Extraction, ii. Transformation, iii. Load. After that, data goes to the RDBMS where the BI (Business Intelligence) reports are generated. The reports generated by this process are not efficient so unable to provide the effective business solution. So, some advanced tool is required which can analyze the entire data.

## III. INTRODUCTION TO BIG DATA

In a general meaning Big data is the huge amount of data. But the only parameter i.e., amount can't express the definition of big data completely. So, basically it is identified according to the large-volume, high-velocity and wide-variety of information.

ISSN No. 2395-0900
International Journal of Engineering Studies and Technical Approach
IJESTA

### 3.1 Five v's of Big Data

There are many properties associated with big data. The prominent aspects are Volume, Variety, Velocity, Variability and Value.

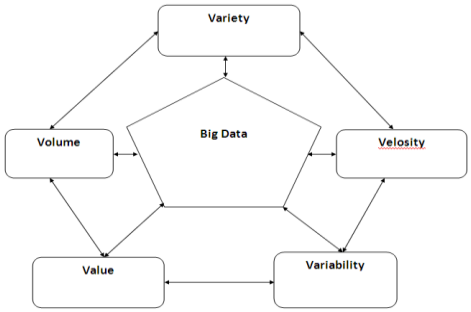


Fig.3.1 Five V's of Big Data

There are many properties associated with big data. The prominent aspects are Volume, Variety, Velocity, Variability and Value.

**Volume:** The volume of big data is exploding exponentially day to day. The data accumulated through social websites and sensor networks going to cross from petabytes to Zetabytes.

**Variety:** Data produced are from different categories, consists of unstructured, standard, semi-structured and raw data which are very difficult to be handled by traditional systems.

**Velocity:** This is a concept which indicates the speed at which the data generated and become historical. Big data is able to handle the incoming and outgoing data rapidly.

**Variability:** It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set.

**Value:** All enterprises and e-commerce systems are keen in improving the customer relationship by providing value added services. For that, study on customer attitudes and trends in the market are to be analyzed.

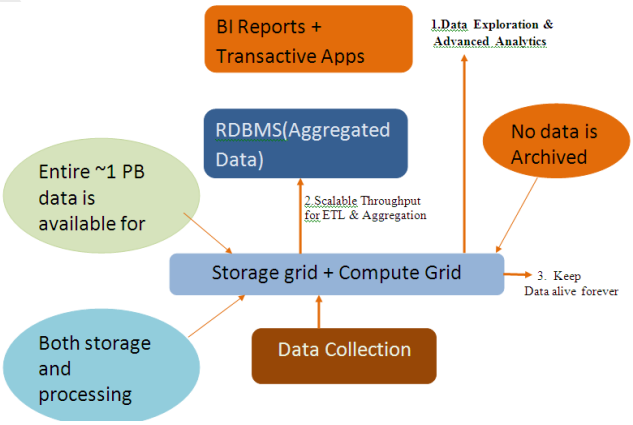


Fig.3.2.Solving the problems of Traditional Data Analytics with Big Data

If we can replace traditional storage grid with an advanced storage grid which itself used as a compute grid then the problem can be easily resolved. In place of traditional data storage grid if we will use some big data tools like Hadoop, then our problem will be solved. Hadoop stores the data In a distributed file system known as HDFS( Hadoop Distributed File System) due to which the storage and processing become easier and faster. Hive (DW system) is the ETL tool in Hadoop. After analysis the entire data stored in the RDBMS where BI reports are generated which is the most accurate and provide the effective business solution

## IV. RESEARCH AREAS AND THE CHALLENGES

There are six major research areas identified as shown in Figure-3. They are:

a) Applied ontology
b) Security
c) Storage and Transport
d ) Mobility

### 4.1 Applied ontology

Applied Ontology is the way of applying the ontological resources to the domains like Geography as well as Bio-medicine. These works can be done within the semantic web framework. Applied ontology involves in looking the relationship between a person's world and his actions. The 9thOntology Summit held on Jan'16-2014 under the theme "Big Data and Semantic Web Meet Applied Ontology" [4] has concluded the following;

- ♣ To build bridges between Semantic web, big Data, Linked Data and Applied Ontology.
- ♣ Performance issues, Challenges in scalability while combining big data and Semantic web. Technically, automated reasoning tools to be developed to make use of ontologies. Large common reusable ontologies and ontological analytical techniques to overcome the engineering bottlenecks are the need of the hour.

### 4.2 Security

Big data involved with many use cases like Staging, Pre-processing, Processing, Meta data storage and to store short term as well as long term fact data. For serving each use case multi-facets of infrastructure required. Safe and private transactions are the two major concerns of IT. But the safety and privacy becomes a question mark as the data volume of big data fast grows. When we consider the safety aspect, the existing cryptography standards can not meet the demands of big data [5]. Hence; effective mechanisms to handle structured, semi-structured, unstructured data have to be investigated and to be developed. Data acquisitions of users like habits, personal interests and the like through websites may happen with the permissions of the users or maybe when the users are not aware of. But the same may be leaked while storing, transmitting or handling. According to report [6], researcher Ron acquired 2.8 GB of Facebook user's data and made it available to download on the internet. Hence, Privacy protection is another challenging problem in big data.

### 4.3 Storage and transport

Big data stores and handles data in different way from traditional data warehouses. Big data comprises massive sensor data, raw and semi-structured log data of IT industries and the

exploded quantity of data from social media. As per the examples given in [7], current disk technologies are limited to store 4 TB per disk. For storing 1 exa byte, it requires 25, 000 disks which will overwhelm the existing communication technologies. That means such phenomenon demands for a revolution in storage and communication technologies.

### 4.4 Mobility

Enterprises are poised to extend more investment in applications which support mobile devices. Very great potential to increase productivity is on the way for businesses when they combine enterprise process automation and mobile computing technologies. Increasing location based datasets, influx of data from mobile applications, their size and variety exceeds the capacity of spatial and mobile computing technologies. Mobile users contribute a lot to big data analytics through their online activities. The convergence of traditional routing services (including GPS and spatial data) into the big data paradigm has to face major challenges. First, it increases the computational cost because it magnifies the impact of routing queries to mobile devices. Second, it uses geographical reasoning in remote sensing and inference over time and space. The built-in motion detectors in mobile phones derive a huge amount of data from every user's life. How efficiently utilize these data and how to carry and share through limited bandwidth mobile stations, are the other challenges.

## V. TECHNICAL CHALLENGES IN BIG DATA

Whenever new technologies evolve, they meet with new challenges in all the aspects. Once the functional challenges are in place, the next kin is the technical challenges. Big data faces many technical challenges which are on the roadway of the research.

### A) Failure handling

Devising 100% reliable systems on the go is not an easy task. Systems can be devised in such a way that the probability of failure must fall within the permitted threshold. Fault tolerance is a technical challenge in big data. When a process started it may involve with numerous network nodes and the whole computation process becomes cumbersome. Retaining check points and fixing the threshold level for process restart in case of failure, are greater concerns.

### B) Data heterogeneity

Big data deals with unstructured, semi-structured and structured data. Linking unstructured data with structured data, converting data from one form into another required form needs a lot of research.

### C) Data quality

Huge amount of data pertaining to a problem is undoubtedly a big asset for both Business as well as IT leaders. For predictive analysis or for better decision making amount of relevant data helps a lot. But the quality of such data is based on the source through which they are derived. Though big data stores large relevant data, the accuracy of data is completely dependent on the source domains. Hence, there is a question of how far the data can be trusted and it definitely requires appropriate trust agent filters.

ISSN No. 2395-0900
International Journal of Engineering Studies and Technical Approach
IJESTA

## VI. APPLICATIONS OF BIG DATA

The several areas of Big Data Computing are described below.

### A) Scientific Explorations:

The data collected from various sensors is analyzed to extract the useful information for societal benefits. E.g. physics and astronomical experiments-a large number of scientists collaborating for designing, operating and analyzing the products of sensor networks and detectors for scientific studies. Earth Observation Systems (EOS) -

Information gathering and analytical approaches about earth's physical, chemical and biological systems via remote sensing technologies, to improve social and economic well-being and its applications for weather forecasting, monitoring and responding to natural disasters, and climate change predictions etc.

### B) Healthcare:

Healthcare organizations would like to predict the locations from where the diseases are spreading so as to prevent further spreading. However, to predict exactly the origin of the disease would not be possible, until there is statistical data from several locations. In 2009, when a new flu virus similar to H1N1 was spreading, Google has predicted this and published a paper in the scientific journal Nature, by looking at what people were searching for, on the internet.

### C) Governance:

Surveillance system analyzing and classifying streaming acoustic signals, transportation departments using real-time traffic data to predict traffic patterns, update public transportation schedules.Security departments analyzing images from aerial cameras, news feeds, and social networks or items of interest. Social program agencies gain a clearer understanding of beneficiaries and proper payments.Tax agencies identifying fraudsters and support investigation by analyzing complex identity information and tax returns.
Sensor applications such stream air, water and temperature data to support cleanup, fire prevention and other programs.

### D) Financial and Business Analytics:

Retaining customers and satisfying consumer expectations are among the most serious challenges facing financial institutions. Sentiment analysis and predictive analysis would play a key role in several fields like travel industry-for optimal cost estimations, retail industry-products targeted for potential customers, Forecast analysis –estimating the best price estimations etc.

### E) Web Analytics:

Several web sites are experiencing millions of unique visitors per day, in turn creating a large range of content. Increasingly, companies want to be able mine this data to understand limitations of their sites, improve response time, offer more targeted ads and so on. This requires tools to perform complicated analytics on data that far exceeds the memory of a single machine or even in cluster of machines.

## VII. CONCLUSIONS

In this paper, we have done an elaborated study on Big Data and its research challenges. We have presented the research opportunities. We propose the technical view of big data comprising the various classes. Several areas of Big Data Computing are described in this paper. After this survey we got that data is growing rapidly irrespective of the type and size. So, we can conclude that in the near future we may deal with some new data definition having more advanced characteristics and there will be some more advanced tools which can solve the problems caused by that new type of data. In Future, Detecting the problems in Big data analytics and Solutions of the detected problem.

## REFERENCES

i. S.Sruthika, N. Tajunisha, A Study On Evolution Of Data Analytics To Big Data Ana lytics and Its Research Scope. In: Paper presented in IEEE Sponsored International Conference on Innovations in Information Embedded and Communication Systems ICIIECS, 2015.

ii. Y.Demchenko, P.Membrey (2014), Defining Architecture Components of the Big Data Ecosystem. in: Paper published in Collaboration Technologies and systems (CTS), pp-104 -112. Doi: 10.1109/CTS.2014.6867550.

iii. Julie M. David, Kannan Balakrishnan, (2011), Prediction of Learning Disabilities in School-Age Children using SVM and Decision Tree, Int. J. of Computer Science and Information Technology, ISSN 0975-9646, 2(2), pp829-835.

iv. Albert Bifet, (2013), "Mining Big data in Real time", Informatics 37, pp15-20.

v. Richa Gupta, (2014), "Journey from data mining to Web Mining to Big Data", IJCTT, 10(1), pp18-20.

vi. Howe AD, Costanzo M, Fey P, et al. 2008.

vii. Big data: The future of biocuration, Nature. 455(7209): 47-50. Doi: 10.1038/455047a.

viii. Crawford Kate and Jason Schultz. 2014. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, Boston College Law Review. 55(93): 93-128.

ix. http://www.techrepublic.com/blog/the-enterprise-cloud/cloud-computing-and-the-rise-of-big-data/.

x. http://ebiquity.umbc.edu/blogger/2014/01/14/2014-ontology-summit-big-data-and-semantic-web meet-applied-ontology/.

xi. Min Chen, Shiwen Mao, Yunhao Liu. 2014. Big Data: A Survey, Mobile Networks and Applications. 19(2): 171-209.

xii. Tasevski P. 2011. Password attacks and generation strategies, Tartu University: Faculty of Mathematics and Computer Sciences.

xiii. Ibrahim Abaker Targio Hashema, Ibrar Yaqooba, Nor Badrul Anuara, Salimah Mokhtara, Abdullah Gania, Samee Ullah Khanb. 2015. The rise of "big data" on cloud computing: Review and open research issues, Information Systems, Elsevier. 47: 98-115.