The Study of Big Data: The Emerging Step of Innovations

Ms. Megha Shinde
Student, Dept. of Information Tech.
Shah & Anchor Kutchhi Polytechnic,
Mumbai, India
meghashinde135@gmail.com

Mr. Ritesh Rathi
Lecturer, Dept. of Information Tech.
Shah &Anchor Kutchhi Polytechnic,
Mumbai, India
ritesh.rathi@sakp.ac.in

Mr. Abhijeet Pasi
Lecturer, Dept. of Computer Tech.
Shah & Anchor Kutchhi Polytechnic,
Mumbai, India
abhijeet.pasi@sakp.ac.in

Abstract: The data collected from people, devices, networks and other sources was stored at large scales but by some or the means data used to get scrambled and sometimes used to be lost. People were unable to search their documents and files on their stored area. So as a solution to this mess, Big data came into existence. The idea behind Big data was to segregate the data into different data sets were the data had to go from an activity of data analyses. This will lead to handling and analyzing the data properly and get appropriate data on time with accuracy.

*****

I. INTRODUCTION

When the word "Big Data" comes ahead, various questions comes in mind:

- What?
- Why?
- How?
- Where?
- When?

This paper may help to find some relevant answers to these questions. Many organizations use several data processing applications to manage their data. But at times the data is not properly processed and due to this there is always a risk of losing data or not accurately getting it at the time of retrieval. In the way of getting a solution for these problem, Big data has occurred. The agenda behind Big data is that to separate the data into data sets that are more large and complex in size as per their classification. This classification is created after a process which involves Acquisition, Extraction, Integration, Analysis, Interpretation and Decision.

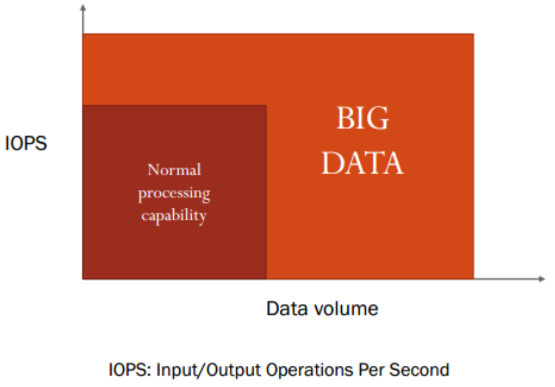

IOPS
BIG DATA
Normal processing capability
Data volume
IOPS: Input/Output Operations Per Second

Figure 1. When data becomes "Big"[1]

II. WHAT IS BIG DATA?

Big data applies when normal processing capability of any traditional processing application exceeds and still the data generated requires more space to be stored even if it is of no use. Big data is a term utilized to refer to the increase in the volume of data that are difficult to store, process, and analyze through traditional database technologies. The nature of big data is indistinct and involves considerable processes to identify and translate the data into new insights. The term “big data” is relatively new in IT and business[2]. At times it becomes awkward for organizations to work with using standard statistical software that are not so capable for processing the information.

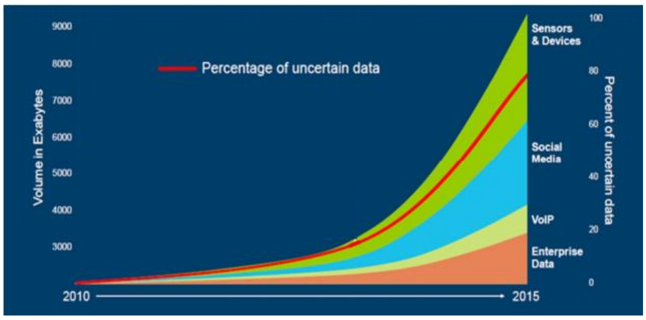


Figure 2. Graph of Percentage of uncertain data [1]

1.1 Characteristics of Big Data

The characteristics of Big data can be described as:

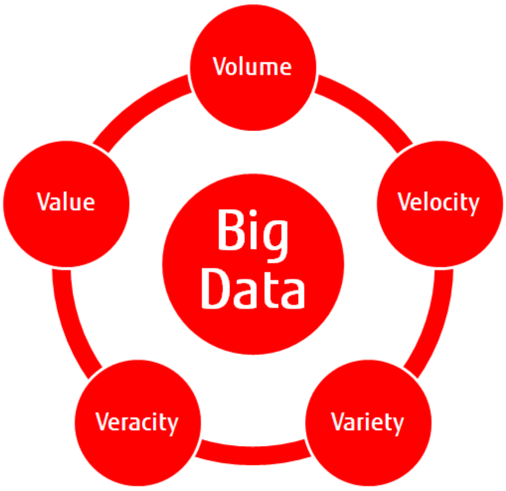


Figure 3. 5 V's of Big Data [3]

- Volume

The volume is the most challenging aspect when Big data comes to mind. The companies already have stored their information in the forms of logs but have shortage of capacity for processing it. The ability to process large amounts of information defines the benefits of big data. Big data provides volumes such as: Terabyte (10^{12}), Petabyte (10^{15}), Exabyte (10^{18}), Zettabyte (10^{21}).

- Velocity

Velocity refers to the increasing speed at which this data is created, so the increasing speed at which the data can be processed, stored and analyzed by relational databases. Velocity refers to the speed at which new data is generated and the speed at which data moves around. About social media messages going to viral in seconds.[4]

- Variety

Variety refers to the various types of data generated with sensors, smart phones or social networks. The data can be in two forms: structured and unstructured. The data mostly generated is in unstructured form. The next aspect of big data is its variety. The data needs to be analyzed and then categorized them which is suitable to Big data.

- Veracity

Quality plays an important role when we provide something of use. So in case of Big data, veracity plays this role. When we are dealing with a high volume, velocity and variety of data, it is not possible that all of the data is going to be 100% correct there will be dirty data. The quality of the data being captured can vary greatly. The data accuracy of analysis depends on the veracity of the source data [4].

- Value

Value is a vital aspect of Big data. It becomes useless when the data retrieved is of no value. If the data stored is in proper way and the retrieved data is in well format and provides some value when used is called as value given by Big data.

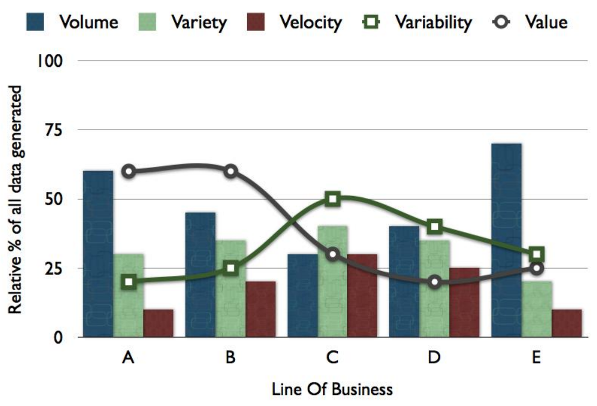


Figure 4. Statistics of 5 V's of Big Data [5]

1.2 Big Data Workflow

The following figure shows the life cycle of data:

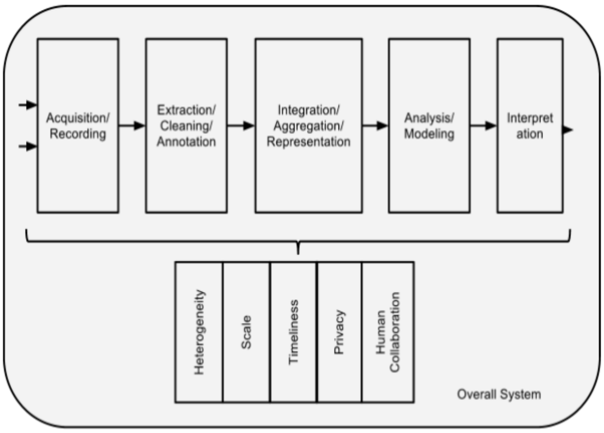


Figure 5. The Big Data Analysis Pipeline.Major steps in analysis of big data in the flow at top. Below it is big data need that makes these tasks challenging [6].

- Acquisition

Big Data does not arise out of a vacuum: it is recorded from some data generating source. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information. The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. Metadata acquisition systems can minimize the human burden in recording metadata. Another important issue here is data provenance. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline.

- Extraction

Frequently, the information collected will not be in a format ready for analysis. We cannot leave the data in this form and still effectively analyze it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge.

- Integration

Data analysis is considerably more challenging than simply locating, identifying, understanding and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then “robotically” resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution.

- Analysis

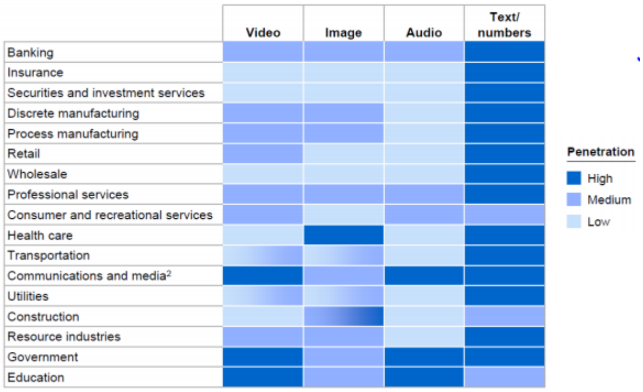
Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. Big Data is also enabling the next generation of interactive data analysis with real-time answers. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses.

- Interpretation

Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis.

1.3 Types of Available Data

The data is generated in various types of sectors that are as follows :



	Video	Image	Audio	Text/ numbers
Banking	▨	▨	▨	■
Insurance	□	□	□	■
Securities and investment services	□	□	□	■
Discrete manufacturing	▨	▨	□	■
Process manufacturing	▨	▨	□	■
Retail	▨	□	□	■
Wholesale	□	□	□	■
Professional services	▨	▨	▨	■
Consumer and recreational services	▨	□	▨	▨
Health care	□	■	□	■
Transportation	□→▨	□→▨	□	■
Communications and media[2]	■	▨	■	■
Utilities	□→▨	□→▨	□	■
Construction	□	▨→■	□	▨
Resource industries	▨	▨	□	■
Government	■	▨	■	■
Education	■	▨	■	▨

Penetration
■ High
▨ Medium
□ Low

Figure 6. Types of Data generated in various sectors [1]

III. WHY BIG DATA IS IMPORTANT?

Big Data is important as there is increases of storage capacities, increase of processing power and availability of data. When big data is effectively and efficiently captured, processed and analyzed, When big data is effectively and efficiently captured, processed, and analyzed, companies are able to gain a more complete understanding of their business, customers, products, competitors, etc. which can lead to efficiency improvements, increased sales, lower costs, better customer service, and/or improved products and services [7]. Technological growth and easy access to sophisticated gadgets have led to a digital data explosion. Complex data generated by network traffic and collected from applications and process logs, outputs from numerous digital devices, interactions on the web and social media sites, digital photographs, satellites, are common examples of Big Data. Technological growth and easy access to sophisticated gadgets have led to a digital data explosion. Complex data generated by network traffic and collected from applications and process logs, outputs from numerous digital devices, interactions on the web and social media sites, digital photographs, satellites, are common examples of Big Data [8].

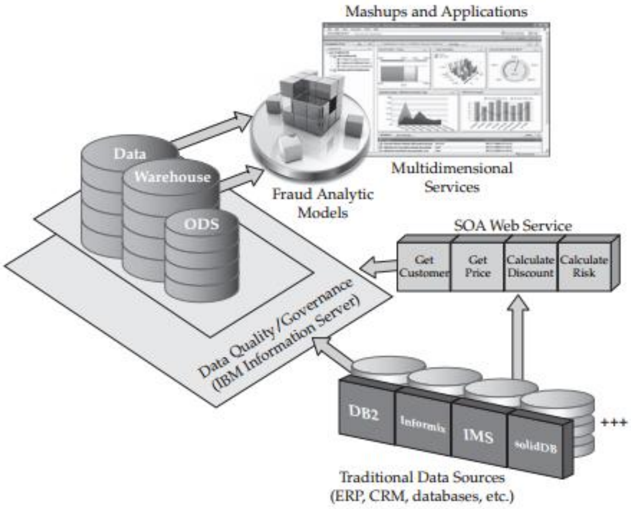


Figure 7. Traditional way of finding the Fraud Data [9]

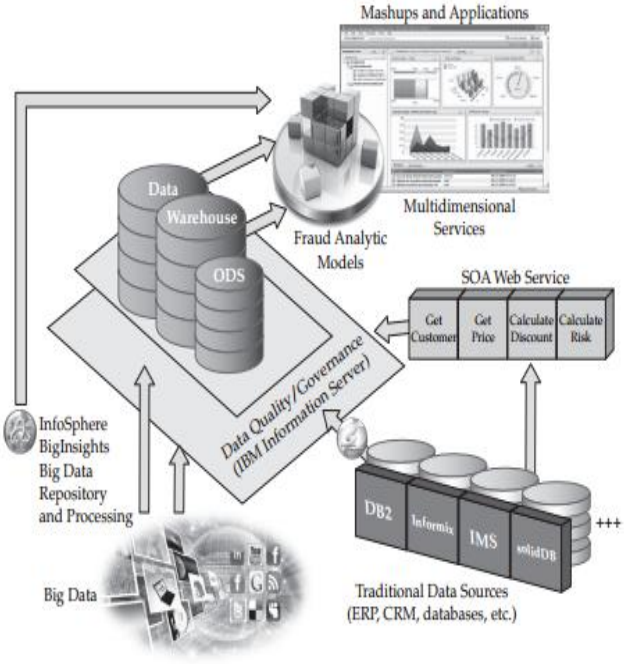

Mashups and Applications
Data
Warehouse
ODS
Fraud Analytic Models
Multidimensional Services
SOA Web Service
Get Customer
Get Price
Calculate Discount
Calculate Risk
Data Quality/Governance (IBM Information Server)
InfoSphere BigInsights Big Data Repository and Processing
Big Data
DB2
Informix
IMS
solidDB
+++
Traditional Data Sources (ERP, CRM, databases, etc.)

Figure 8. Modern way of finding the Fraud Data[9]

The Big data sets a new of finding the fraud data as the big data storage is the huge storage of information and so any event can be generated to scam the data so its need to be kept safe.

IV. RISKS AND CHALLENGES OF BIG DATA

- Performance, performance, performance!
- Data grows faster than energy on chip
- Efficiency
- Scalability
- Effectiveness
- Heterogeneity
- Flexibility
- Privacy
- Costs

V. CONCLUSION

- We live in the era of Big Data.
- Wide range of availability in different areas.
- Big opportunities to solve big problems.
- They can create value.
- The challenge is how to manage and use them.
- New technologies are needed.
- Methodological aspects are important.
- A rapidly evolving area.
- Data scientists: the current hottest profession in IT.

REFERENCES

[1] Riccardo Torlone. "Big data: an introduction". Universita` Roma Tre

[2] Ibrahim; Targio Hashem, Abaker; Yaqoob, Ibrar; Badrul Anuar, Nor; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "big data" on cloud computing: Review and open research issues". Information Systems 47: 98–115. doi:*10.1016/j.is.2014.07.006*

[3] http://infiniment-it-fujitsu.com/les-5-v-du-big-data-et-le-fujitsu-m10/

[4] Ishwarappa , J. Anuradha." A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology ".

[5] Christian Reilly. "The 5V’s of The Data Landscape" on November 9, 2012

[6] "Challenges and Opportunities with Big Data" A community white paper developed by leading researchers across the United States.

[7] "Why is BIG Data Important?"A Navint Partners White Paper May 2012

[8] Sangita Garg. "The New Frontier for the Pharmaceutical and Life Sciences Industry: Real Big Value from Big Data".

[9] Chris Eaton, Dirk Deroos, Tom Deutsch, George Lapis, Paul Zikopoulos. "Understanding Big Data"