Big Data: A Concept of Managing Huge Data

Gurudatta Verma
Asst. Prof., CSE Dept
GD RCET, Kohka, Bhilai

Giriraj Dey
Student, CSE Dept
GD RCET, Kohka, Bhilai

ABSTRACT
The rapid proliferation of data has lead to the necessity of big data. However the volume of big data is constantly growing. Big data comprises of extremely large data sets which are beyond the scope of management offered by commonly used software. Big Data mining can be described as a set of techniques and technologies used for extracting useful information from these large datasets or streams of data, which due to its volume, variability, and velocity, was not possible earlier. This paper presents an overview of big data, its current status, controversy, future scope and how it can play an essential role in fostering the growth and productivity of a company in today's cut throat competitive world.

KEYWORDS
big data, Hadoop, DBMS.

1. INTRODUCTION
The term 'Big Data' was coined and presented for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the NextWave of InfraStress" [1].

Over the past years there has been an explosion in the amount of data available; primarily because of the diverse methods to gather data and social media which allows users to create contents freely and amplify the already massive Web volume. Furthermore, the advent of smart phones has opened the gateway to get real time data about people from different aspects. The vast amount of data that internet service providers can potentially process, to enhance our daily life has significantly outpaced our past CDR (Call Data Record) based processing for billing purposes.

Big Data has the potential to revolutionize not just research, but also education [2]. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [5]. There is a strong trend for massive Web storage of educational activities over the cloud, and this will create an increasingly enormous amount of detailed data about the students' performance.

Big data offers companies the opportunity to gain an in depth and more accurate insight into customers, partners, and business and grow their competitive edge. Results from the survey indicate that companies are rounding the learning curve for big data. Not only is big data a top strategic priority for large organizations, but most enterprise companies already have a formal big data analytics strategy in place. IT managers are confident in their understanding of big data, and the requests they receive from their constituents indicate that business units have a good grasp of their big data needs.[4]

2. CHALLENGES FOR BIG DATA
The greatest challenges that big data is faces are:

(i) **Volume:** It is the most visible aspect of big data referring to the fact that the amount of data being generated is increasing rapidly. The growth of the internet has boosted the global data production. The solution to this problem was deduced to be the virtualization of storage in data centres, amplified by a significant decrease of the cost of ownership through the generalization of the cloud based solutions. The noSQL database approach is a response to store and query huge volumes of data heavily distributed. [3]

(ii) **Velocity:** This is the rate at which the fast growing data produced is captured. The objective is to collect as much as data possible in short frames of time. The daily additions of connectable devices not only increase the volume but also the velocity. As a result real-time data processing platforms are now considered by global companies as a requirement to gain a competitive advantage. [3]

(iii) **Variety:** This is explained with respect to multiplication of data sources that is the explosion of data formats, ranging from structured text to free text. The requirement to collect and analyse non-structured or semi-structured data goes against the traditional relational data model and query languages. This reality has been a strong motivation to create new kinds of data stores which are able to support flexible data models. [3]

(iv)**Value:** It is a highly subjective aspect referring to the fact that until recently, large volumes of data where recorded but not exploited. Big Data technologies can be visualized as tools to create or mine valuable data from otherwise not fully exploited data; thus converting raw data into potentially useful information that may be used internally, or for intelligent business solutions. [3]

3. APPLICATIONS OF BIG DATA
Big Data finds its use in diverse sectors a few are mentioned below:

Medicine: In the field of medicine Big Data can be used to monitor and improve the health of people moreover it can be used for extremely complex tasks like DNA mining.

Business: Big Data has the ability to lift businesses to greater heights by boosting its efficiency. This is because Big Data will enable companies to offer more diverse and personalized services to their customers. This would lead to healthier relations between the producers and the consumers.

Building Smart Cities: One of the digital technologies that are required to achieve the objectives of a smart city is provided by Big data. Building of a smart city would involve managing voluminous amounts; this data management can be done with the help of Big data.

Improving lives in developing nations: Big data also portrays a key role in bettering lives in developing nations. In order to get a clear picture of how this is done I would like to take the great work Global Pulse [10] is doing as an example.

Global Pulse is a United Nations initiative, launched in 2009, that operates as an innovative lab, focused on mining Big Data for developing countries. They implement a strategy that consists of 1) researching innovative methods and techniques for analyzing real-time digital data to detect early emerging vulnerabilities; 2) assembling free and open source technology toolkit for analyzing real-time data and sharing hypotheses; and 3) establishing an integrated, global network of Pulse Labs, to steer the approach at the country level. Global Pulse describes the key opportunities Big Data offers to developing countries in their White paper "Big Data for Development: Challenges & Opportunities"[11]:

- Early warning: develop prompt response in time of crisis, detecting anomalies in the usage of digital media.
- Real-time awareness: devise programs and policies with a more detailed representation of reality.
- Real-time feedback: note the policies and programs fails, monitoring it in real time, and using this feedback make the needed changes.

The Big Data mining revolution is not restricted to the developed nations, as mobiles are spreading in developing nations as well. It is anticipated that there are over five billion mobile phones, out of which 80% are located in developing countries.

4. CURRENT AND EXPECTED USE OF BIG DATA

Today, big data is being used to gather multiple types of insight. Among the respondents, generating competitive intelligence and determining staffing levels and productivity are the most frequent uses of big data analytics. By 2016, the expectation is that big data will be used most often to help improve operational efficiency (30 percent) and identify new revenue sources (28 percent), as well as provide insight to other areas of business, such as reducing IT costs, improving business agility, and bettering prices. [4]

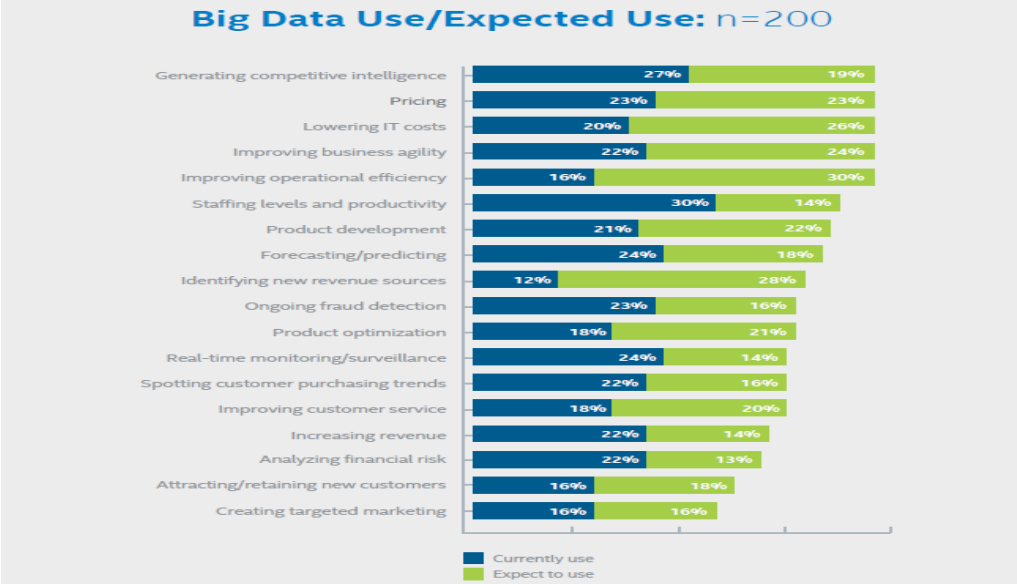

Fig 1. Current and expected use of big data by 2016[4]

5. BIG DATA SOLUTIONS

5.1 BIG DATA ANALYSIS PLATFORMS AND TOOLS

1. Hadoop

The Apache distributed data processing software is so pervasive that often the terms "Hadoop" and "big data" are used together. The Apache Foundation also sponsors various related projects that extend the capabilities of Hadoop.

2. MapReduce

Described as "a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes " in its website; MapReduce was developed by Google. It's used by Hadoop, as well as many other data processing applications.

3. Storm

Currently owned by Twitter, Storm offers distributed real-time computation capabilities and is described as the "Hadoop of realtime." It's highly scalable, robust and works with almost all programming languages.

5.2 DATABASES/DATA WAREHOUSES

1. MongoDB

MongoDB is designed to support voluminous databases. It's a NoSQL database with document-oriented storage, full index support, replication, high availability and more advanced features. Commercial support is available through 10gen.

2. FlockDB

It is better known as Twitter's database, it was designed with an objectve to store social graphs (i.e., who is following whom and who is blocking whom). It offers horizontal scaling and fast reading and writing.

3. Cassandra

This was originally developed by Facebook, but now this NoSQL database is managed by the Apache Foundation. It's used by many organizations with huge, active datasets, like Twitter, Constant Contact, Reddit, Urban Airship Netflix, Cisco and Digg.

5.3 OPEN SOURCE SOLUTIONS VS COMMERCIAL SOLUTIONS

Roughly a quarter (26 percent) of respondents has deployed open-source Hadoop software, double the number of those who have deployed a commercial distribution of Hadoop framework (12 percent). [4]

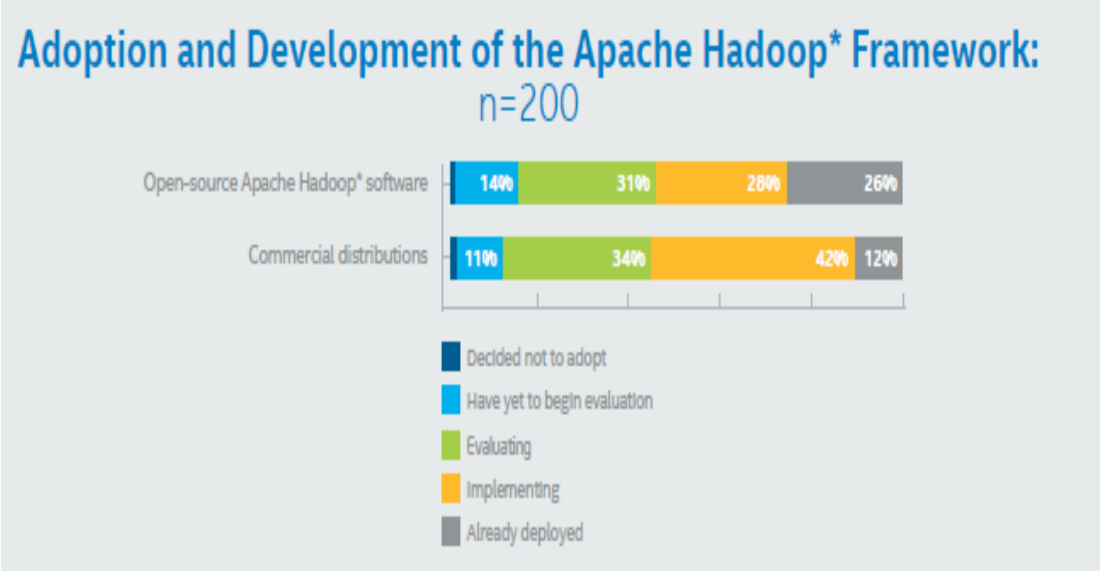

Fig 2. Adoption and Development of the Apache Hadoop Framework: n=200 [4]

6. CONTROVERSIES ABOUT BIG DATA

Like with every technology or concept there are many controversies about Big Data. Some are mentioned

Big Data may be a hype to promote Hadoop based computing systems. Hadoop is not always the best tool [7]. It appears that data management system vendors try to promote and sell systems Hadoop based systems.

Is there really a need to distinguish Big Data analytics from data analytics? This has also been a recurring question, as data will always continue grow and it will never be small.

The main objective is to retrieve desired significant results; this objective may be lost while dealing with such large data sets.

It is not mandatory for bigger data to be always better data. It also depends on whether the data is noisy or not, and whether it is really a reflection of the required data.

Information and Communication Technology enabled organization to work with a greater deal of efficiency on one hand and on the other it created a digital divide. Similarly big data has also created a digital divide. This digital divide will be between the organizations or people who can use big data to manage data and the ones who cannot. This digital divide will be mainly be seen between the large organizations and small organizations. Thus the gap between the wealthy organizations and poor organizations will be widened.

7. FUTURE SCOPE

Data is likely to be updated or changed over time, hence it is essential that Big Data mining techniques are able to adapt and in some case detect this change.

Visualization of data is another factor of consideration. One of the major tasks of Big Data analysis is to chalk out an appropriate way to visualize the results. Since the data taken into account is extremely large, it is very difficult to make user-friendly visualizations.

When dealing with Big Data, the amount of storage space required is an obvious concern. To resolve this issue there are two main approaches:

- **Compression:** In this approach we don't lose any data.
- **Sampling:** In this approach more relevant data is selected.

Both the methods have their own pros and cons. Compression, is a time consuming process. In other words space is saved in the bargain of time. It may also be considered as a transformation from time to space. In case of sampling, we lose information in the bargain of saving space.

8. CONCLUSION

After a thorough consideration of the controversies and the advantages that Big data offers, it is clear that Big data has more to give to the community than it has to take from the community. With the growth in population and its demand our data is bound to increase and the Big data technology is perfect to cater of this need. Big data will give organization the wings it needs to soar to greater heights.

9. REFERENCES

[1] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.

[2] Advancing Personalized Education. Computing Community Consortium. Spring 2011,jan1.

[3] Big Data A new World of Opportunities ,Nessi White Paper , December 2012.

[4] “Big Data Analytics” Intel’s 2013 IT Manager Survey on How Organizations Are Using Big Data.

[5] Getting Beneath the Veil of Effective Schools: Evidence from New York City. Will Dobbie, Roland G. Fryer, Jr. NBER Working Paper No. 17632. Issued Dec. 2011.

[6] Challenges and Opportunities with Big Data, Divyakant Agrawal, Philip Bernstein, Elisa Bertino ,Susan Davidson, Umeshwas Dayal, 1-1-2011

[7] J. Lin. MapReduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not a Nail! CoRR, abs/1209.2191, 2012

[8] J. Gama. Knowledge Discovery from Data Streams. Chapman & Hall/Crc Data Mining and Knowledge Discovery. Taylor & Francis Group, 2010.

[9] R. Smolan and J. Erwitt. The Human Face of Big Data. Sterling Publishing Company Incorporated, 2012.

[10] UN Global Pulse, http://www.unglobalpulse.org.

[11] E. Letouz_e. Big Data for Development: Opportunities & Challenges. May 2011.