Big Data: What, Why and Why Not

Surbhi Jain
Assistant Professor, Department of Computer Science, India

Abstract: **Big Data is a term for data that possess larger variety, being received in increasing volumes and with ever higher velocity. Data growth, speed and complexity are being driven by application of billions of intelligent sensors and devices that are transmitting data and by other sources of semi-structured and structured data. The data must be assembled on an ongoing basis, analysed, and then used to provide direction to the business by taking valuable decisions. But, as it is said nothing comes for free; it also increases concerns and problems that must be resolved. There are 3 Big Data concerns that should be of remarkable concern: Data Privacy, Data Security and Data Discrimination. In this paper we will discuss what exactly is Big data along with its evolution as solution for analysing torrents of information of present era. The paper also discusses the prime concerns related to Big data security.**

Key Terms: Big Data, Streaming Data, Breach

1. INTRODUCTION

Big data is a term for **data** sets that are so **large** or complex that primitive **data** processing application software is inadequate to deal with them. Big Data represents a new period in data study and utilization. It is a leveraging open source technology- a robust, secure, highly available, enterprise-class Big Data platform. Challenges include capture, storage, analysis, querying, and updating data safely and securely. While the term "big data" is relatively new, the doing of collecting and storing plethora of information for eventual analysis is ages old. Some of the examples of Big data are:

- The New York stock exchange generates approximately a terabyte of new trade data every day.
- Facebook gets more than 500 terabytes of new data in the form of photos, videos, comments and messages everyday
- Data captured by call centres on everyday basis is yet another example.

2. CHARACTERISTICS OF BIG DATA

The three significant physiognomies of Big Data are volume, variety, and velocity.

Volume. Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. Earlier, it would have been a problem storing such enormous amount of data, but now we have a lot many tools for this.

Velocity. Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with floods of data in near-real time.

Variety. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

Apart from these, two new characteristics are also considered while talking about big data, namely- Variability and Complexity.

3. SIGNIFICANCE OF BIG DATA

Today's data comes from multiple sources, which makes it difficult to be transformed in one common system. This leads to inconsistent data. The amount of data that's being created and stored on a global level is almost unimaginable, and it just keeps growing. That simply means the data which is actually being analysed is just a small fraction of what is actually available to us.

The significance of big data is not based on how much data we have, but how we use that data. We can take data from any source and analyse it to find responses that enable us to produce results in reduced cost and time with smart decision making. When we combine big data with high-powered analytics, we can accomplish professional-allied tasks like:

- Determining core causes of let-downs, issues and defects in near-real time.
- Studying customer's buying habits.
- Recalculating entire risk assortments in minutes.
- Detecting fraudulent behaviour before it affects the organization.
- Planning next course of action judiciously for betterment of organisations, etc.

Getting promotional telephone calls from telecommunication giants is very likely to happen these days with the executive stating here is the mobile plan for you based on your pattern of usage; getting similar calls from Banks, Book sellers, Shopping hubs etc. is also not very unusual. These kind of calls are all based not on analytics but on big data. We can make better predictions and smarter decisions now. The big data movement, like analytics before it, pursues to assemble intelligence from data and translate that into business advantage. However, there are key differences in terms of the three V's of big data mentioned earlier in this paper. As more and more business doings are digitized, new sources of information and ever-cheaper equipment combine to bring us into a new period: one in which large volumes of digital figures exists on almost all areas of interest to a business. Mobile phones, online shopping, social networks, electronic communication, GPS, and instrumented machinery all yield tsunami of data as a by-product of their everyday operations. Each of us is now a moving data producer. The data available are often unstructured—not organized in a database—and clumsy, but there's a large volume of information in it, simply waiting to be released. Analytics brought severe practices to decision making; big data is at once simpler and more powerful.

4. SOURCES OF BIG DATA

Prior to learn how big data can work for the business, we should first know where it comes from. The sources for big data can be broadly compartmentalised in three compartments:

Streaming data- this is the category of data which reaches us from the web of interconnected devices. We analyse this data after receiving and then decide what all to be kept and what to be discarded.

Social media data- this data is gaining popularity day by day. This is generally useful for sales and marketing operations of any business. Since this data is usually very vague and clumsy, it is difficult to organise and structure it for analysis.

Other sources- there are a lot of other data sources which are openly available like data portals. This again is challenging to structure this data.

Once all the prospective sources for data have been identified, we can then consider the decisions we will need to make by binding information. The decisions may include its storage, management, structuring, analysis etc. Unlike several years ago, a lot of low-cost options are available now for storing data. Certain organizations do not want to miss analysis of any data they have. This also has become possible these days with today's high-performance technologies such as grid computing or in-memory analytics. Another methodology is to decide beforehand the relevance of data for analysis. Data which is not relevant can be straight way discarded.

The concluding stride in making big data work for the business is to examine the technologies that can support in making the most of big data and big data analytics. Cheap, abundant storage, Faster processors, Affordable open source, distributed big data platforms, Parallel processing, clustering, MPP, virtualization, large grid environments, high connectivity and high throughputs Cloud computing and other flexible resource allocation arrangements can be considered for the same.

5. BIG DATA- WHAT IS DIFFERENT?

A universally known fact is that data-driven decisions are better decisions. Deciding on the basis of evidences rather than intuitions is certainly better for the business managers. This is the reason Big Data has the calibre to revolutionize management. There are companies like Google, Yahoo and Amazon which are kind of born digital, and are already leaders of big data. But the potential to gain competitive advantage from it is even more. The managerial challenges, however, are very real. Senior decision makers have to embrace evidence-based decision making. The companies using Big Data Analytics, need to hire scientists who are capable of doing pattern search on data and translate the results into useful business information. The comprehensive market, however, sees big data quite differently. Rather than the data inferred independently, they see the value anticipated by adding the new data to their present operational or analytical systems. So, it can be said that Big Data defines a complete information management strategy that is a blend of many new types of data and data management with traditional data.

Now the question arises what is different with Big Data which was not available with traditional Data Mining techniques.

Unlike a primitive data warehouse methodology which expects the data to undergo standardized Extraction-Transformation-Loading processes and sooner or later map into predefined schemas, Big data refers data structure and analytics differently. For making changes in the pre-defined schema, the traditional approach follows a very lengthy process while Big Data facilitates capturing of data without requiring a 'defined' data structure. In fact, in most of the cases the structure is derived from the data itself.

Big data count on data scientists and product and process developers more willingly than data analysts which was not so with primitive techniques. Data scientists are the ones who not only understand analytics, but are also well versed in IT. Their skill set includes programming skills, data management skills, mathematical skills and statistical skills. They are thus organized in a manner different to analytical staff of past.

In Big Data analytics, we don't pay attention to the data which is already in stock. Rather we prefer to work on data which is in continuous flow contrary to our traditional mining techniques where the inferences are drawn on the basis of data stocked in warehouses. The ever increasing volume and velocity of data implies that organizations will need to develop endless processes for gathering, analysing and interpreting data.

Big Data introduces new technology, which can separate the new from old. With this new technology, we can speed up the computations. Usually, if we have loads of data increasing and becoming too big for our systems to be managed, we add RAM or remove some processes or maybe we can upgrade to even faster processor. But these kind of quick fixes are not a permanent solution to the growing data. What Big Data does in such scenario is rather than trying to make one system more fancy, it adds more systems to the pool so as to encourage parallelism. Adding more systems to the cluster increases the efficiency and may process many petabytes of information across thousands of nodes. A consequence of increased parallelism is fault tolerance. More the number of nodes in the cluster, more will be the probability of node failures. Big Data systems handles this automatically. Some Big Data systems, for example, has a default 3x duplication of data across nodes, and the system re-directs a computation if some node goes down.

6. BIG DATA HACKS

Big data gives birth to lots of questions for the analytics. A few of them are:

- How will we make use of data? i.e. by selling new products and services or by selling value-added data or by personalizing customer's experience or something else?
- Which business processes can benefit? Operational system or BI system or Reporting?
- Do we need to own and archive the data? If yes, which storage technologies are best for our reservoir?

Apart from these ongoing problems, there are other security related hacks with big data.

Data Privacy: Big data powered apps and services are benefitting us a lot as per our convenience. But do we have any control on how much of our personal information is being shared and used in achieving this? All social networking sites ask to access all data in our mobile phones, be it gallery, contacts or location before granting us any access to them. We cannot use a lot of apps if we deny the permissions they are asking for to withhold our data.

Data Security: The kind of benefits of the product or service always outweighs the concern for data privacy and we usually seek them all the permissions. We just click and agree to our data being used (and finally analysed), but can we trust that the service providers will keep our data safe. It's true that most of this information is not misused but the possibility of sensitive data being misused cannot be denied either. As Big Data is getting voluminous, more and more of our data is exposed to security breaches.

Data Discrimination: We start discriminating between people after knowing the data they possess. Based on person's credentials we judge who is a better candidate for lending money, who will not do any kind of default in his/her insurance policy. Big Data not only helps business managers to become better marketers based on their analysis, but it also allow them to discriminate between people.

There is not an easy solution or a patch to any one of them. The torrent of information collected and the swift change of technology makes solving these concerns even more challenging.

7. BIG DATA BREACHES

- **Yahoo**
  Date:- 2013-14
  Impact:- 1.5 billion user accounts

In September 2016, the once dominant Internet giant, while in negotiations to sell itself to Verizon, announced it had been the victim of the biggest data breach in history, likely by "a state-sponsored actor," in 2014. The attack compromised the real names, email addresses, dates of birth and telephone numbers of 500 million users. The company said the "vast majority" of the passwords involved had been hashed using the robust crypt algorithm.

A couple of months later, in December, it buried that earlier record with the disclosure that a breach in 2013, by a different group of hackers had compromised 1 billion accounts. Not only basic details like names, dates of birth, email addresses and passwords (that are usually not very well protected), security questions and answers were also compromised.

The breaches knocked an estimated $350 million off Yahoo's sale price. Verizon eventually paid $4.48 billion for Yahoo's core Internet business. The agreement called for the two companies to share regulatory and legal liabilities from the breaches. The sale did not include a reported investment in Alibaba Group Holding of $41.3 billion and an ownership interest in Yahoo Japan of $9.3 billion.

Yahoo, founded in 1994, had once been valued at $100 billion. After the sale, the company changed its name to Altaba, Inc.

- **eBay**
  Date:-May 2014
  Impact:- 145 million users compromised

The online auction giant reported a cyber-attack in May 2014 that it said exposed names, addresses, dates of birth and encrypted passwords of all of its 145 million users. The company said hackers got into the company network using the credentials of three corporate employees, and had complete inside access for 229 days, during which time they were able to make their way to the user database.

It asked its customers to change their passwords, but said financial information, such as credit card numbers, was stored separately and was not compromised. The company was criticized at the time for a lack of communication informing its users and poor implementation of the password-renewal process.

- **JP Morgan Chase**
  Date: July 2014
  Impact: 76 million households and 7 million small businesses

The largest bank in the nation was the victim of a hack during the summer of 2014 that compromised the data of more than half of all US households – 76 million – plus 7 million small businesses. The data included contact information – names, addresses, phone numbers and email addresses – as well as internal information about the users, according to a filing with the Securities and Exchange Commission.

The bank said no customer money had been stolen and that there was no evidence that account information for such affected customers was compromised during this attack. Even after this it was reportedly said that the hackers were successful in gaining privilege rights of “root” on more than 90 bank servers. This simply meant that they could do anything and everything with the accounts like withdrawals, transfer of funds, closing of accounts etc. According to the SANS Institute, JP Morgan spends $250 million on security every year.

In November 2015, federal authorities indicted four men, charging them with the JP Morgan hack plus other financial institutions.

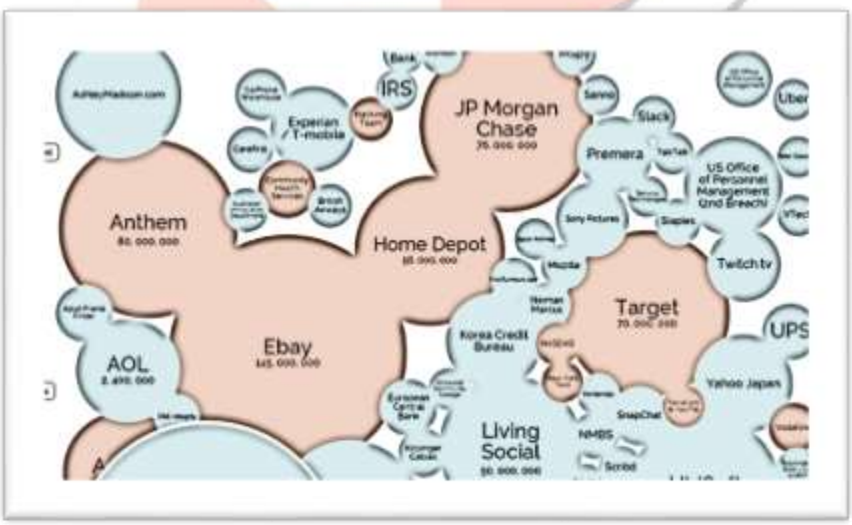
- **Adobe**
  Date: October 2013
  Impact: 38 million user records

Originally reported in early October by security blogger Brian Krebs, it took weeks to figure out the scale of the breach and what it included. The company originally reported that hackers had stolen nearly 3 million encrypted customer credit card records, plus login data for an undetermined number of user accounts.

Later in the month, Adobe said the attackers had accessed IDs and encrypted passwords for 38 million “active users.” But Krebs reported that a file posted just days earlier, “appears to include more than 150 million username and hashed password pairs taken from Adobe.” After weeks of research, it eventually turned out, the hack had also exposed customer names, IDs, passwords and debit and credit card information.

In August 2015, an agreement called for Adobe to pay a $1.1 million in legal fees and an undisclosed amount to users to settle claims of violating the Customer Records Act and unfair business practices. Later in November 2016, an amount of $1 million was paid to customers.

There are a lot of such data breaches including zomato, sony etc. The cases mentioned here are just a few of them to give a hint of the problems associated with very popular big data.



Picture taken from public source

8. FUTURE OF BIG DATA

Big Data is everywhere. We do not want to miss out on anything important and that is the reason there is a vital need to collect and preserve whatever data is being produced. There is a huge amount of data floating around. What we do with that data is all we are concerned about. Data is useless if we can’t analyse it. Big data analytics has thus become crucial. We will be able to improve business and decision making capabilities and gain competitive advantage over the rivals. The demand for professionals who are trained in analytical skills is going up day by day. This will eventually give job opportunities to lot many people. Big data is a journey not a destination. Many market front-runners are already using big data and big data analytics, making their competitors lag behind. The future of big data will come earlier and faster for some organisations than for others. Thinking about the future of big data, thought leaders anticipate an all new kind of experience, where information is shared in a more manageable way, with a path from selling products to selling an outcome.

9. CONCLUSION

Today is the information age. As data is increasing day by day on an ever faster rate, there is a huge need of tools and technologies which can handle this flow. It is revolutionizing almost all spheres of our life, ranging from social networking to banking, from science to government. Potentials of Big Data take account of origination, growth and long term sustainability. Threats include privacy breaches, security hacks, and breach of data integrity. Though, Big data needs to be exploited in an open and transparent manner, but we should be able to ensure having enough skills to make effective and safe use of technology at our disposal.

10. REFERENCES

[1] Nathan Marz, James Warren, “Big Data: Principles and Best Practices of Scalable Real-Time Data Systems”, dreamtech publications, 2015
[2] https://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[3] Viktor Mayer-Schönberger, Kenneth Cukier, “Big Data: A Revolution That Will Transform How We Live, Work, and Think”, 2013
[4] Willem Vermeend, “The Impact of the Internet: How the Internet is changing the way we think, learn, work, do business and make money”.
[5] J Manyika, M Chui, B Brown, J Bughin, R Dobbs, Charles Roxburgh, A.H. Byers, “ Big data: The next frontier for innovation, competition, and productivity”, May 2011
[6] Paul Zikopoulos, Chris Eaton, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data”, McGraw-Hill Osborne Media, 2011
[7] Andrew McAfee and Erik Brynjolfsson, “Big Data: The Management Revolution”, Harvard Business Review, October 2012