International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 2, Issue 5, Nov 2013, 95-100

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

BIG DATA

SUNIL KARAMCHANDANI, BHAVIKA KITAWAT, ABHISHEK SHETTY, VIJAY SHETH & KINJAL MISTRY

D. J. Sanghvi College of Engineering, Mumbai, Maharashtra, India

ABSTRACT

"Big Data" is the current hype in the IT and business world. Companies across almost each industry realize that they need to manage increasingly large data volumes and also analyze the information received from the data in such a way that they can make right decisions fast in order to compete effectively in the market. Technology vendors in the data warehouse spaces say "big data" refers to a traditional data warehousing scene involving data volumes in the single or multi-terabyte range. Newness in technology and larger affordability of digital devices has taken control over today's age of Big Data. Turning Big Data into actionable information requires computational techniques.

New technologies have made it possible not only for scientists and actuaries but also for a wide range of people – including social science academics, marketers, humanities academics governmental organizations, educational institutions, and inspired individuals to construct, share, relate with, and organize data. Enormous data sets that were once abstruse and obvious are being aggregated and made easily accessible. In this paper, we have discussed the various opportunities Big Data brings and the challenges that will be faced in the overall system while implementing it.

KEYWORDS: The Data Deluge, The Age of Big Data

INTRODUCTION

Big data is more than mere hype. What does Big Data mean? Big data simply refers to the massive amounts of data collected over time both structured and unstructured that is so large that it becomes difficult to analyze and handle using traditional database management tools and software techniques. Big data may be as meaningful to business and society just like the internet as it includes business transactions, images, videos, e-mail messages etc. As a result of more data there can be accurate analysis. More the accurate analysis , more the confident decision making process which will in turn lead to better operational efficiencies, reduced risk and cost minimization. IDC defines "big data" as a new generation of technologies and architectures, designed to extract value from very large volumes of a wide variety of data, by allowing high-velocity capture, discovery, and/or analysis. Big data incorporates all types of data (real-time and analytic) managed by upcoming-generation systems to handle increasing volumes of data.

Big data can be petabytes or exabytes of data that consists of billions of records of billions of people from different sources like social networking sites, customer contact center, sales, mobile data etc. The data is loosely structured data which is usually incomplete and inaccessible. The expeditions are due to expansion of data in three dimensions simultaneously which famously are known as the Three "V"s of Big Data. These are:-

Volume: The increase in the volume of data is attributed by many factors. It includes terabytes to petabytes of data. From mega to giga to exa bytes, volume of data is increasing rapidly. Along with volume of data collected, sources of data have also increased significantly. Taking a simple example of a cell phone, the only data it had not long ago was voice calls and texts but with the introduction of smart phones and the astonishing array of sensors, the amount of data they carry is

varying. Present-day organizations are no longer dealing with ponds or lakes of data but are dealing with vast expanses of oceans. And when we are in an ocean, we have a complete different set of challenges to sail our way through the rough seas.

Velocity: This is the least understood of 3 V's. One of the ways to look at velocity as a dimension in which data is expanding is how fast the data is coming in. It's true that the pace of business is inexorably accelerating. Now a day's data is coming faster than ever. It is not just the speed at which data is coming in but also how rapidly it needs to be processed. Business managers need data in real-time since batch processing is not good enough. Now a day's answers like "I will get back to you "is not accepted. Therefore data streaming in at unprecedented speed must be dealt with in an up-to-date manner. Processing data should be faster than the rate at which it is entering the system. Else, we will end up with an ever-growing backlog. Back-logged information will be not useful since there is demand for real time data. Lastly, we need to be prepared not only for consistently high velocity, but also acceleration during intense activity.

Variety: Today data comes in all types of formats such as structured data like numeric data in traditional database and unstructured data like images, audio, video, text documents, email and financial transactions. Every organization collects data and wants to capitalize that data. In the interest of converting this 'dark' unstructured data into information, traditional approaches are no longer sufficient and thus we need new methods. Organizing, incorporating and governing different varieties of data is something that many organizations are still struggling with. Therefore Big Data has gained speeded momentum recently. As a result, organizations need to take urgent steps in order to ensure they are still in this race.

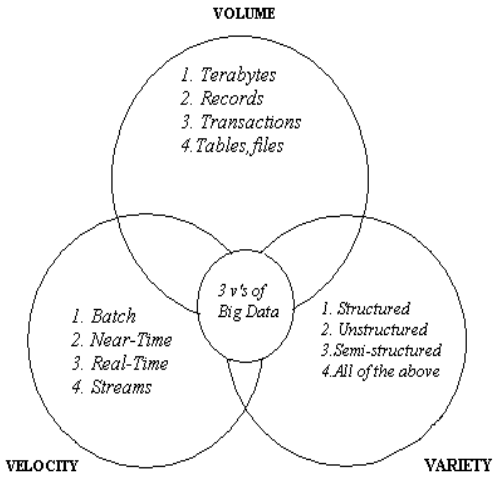

Figure 1

CHALLENGES

While handing a large amount of data, many difficult questions like the ones stated below may arise and it might be difficult to find the most valuable piece of information required at that instant.

- What if the volume of your data gets very and varied that you don't know how to deal with it?
- Is it necessary to store all your data?
- Is it necessary to analyze all the data that is stored in the system?
- How can you find out which data points are really important and how can you make the best advantage of it?

The analysis of large volume of data involves many challenges.

The most salient challenges are stated below

Data Privacy

Privacy is a very sensitive issue and it should not be compromised as it is the backbone of any democracy. We must remain alert in case of privacy being hampered and put in place all necessary safeguards. Privacy, in broader sense is nothing but enclosing a company's information to protect their competitiveness and customers. It can be also understood as a state wishing to preserve their sovereignty and citizens. In both these interpretations, privacy is an encompassing that has a wide range of implications for anyone wishing to explore the use of Big Data for development. Privacy is a fundamental human right that has intrinsic values. It is necessary to ensure an appropriate level of privacy for individuals and societies at large scale. Our basic freedoms, safety, innovation, diversity, etc are at a risk without privacy. A modern society needs privacy to flourish. It is very important to rethink security for information in Big Data uses. Throwing light on individual privacy, it is possible that, in many cases, the users of services and devices generating data (the primary producers) are actually unaware that they are using services that generate data and what it can be used for. It is also unclear that bloggers actually consent to their data being analyzed. People also routinely give their permission to the collection and use the data generated by the web by simply ticking a box. They do not fully realize that their data might be used or misused. To add to it, recent researches have also showed that it is possible to de-identify datasets that were previously anonymised which further raises concerns. Google, Facebook, Twitter, credit card companies and phone companies should emphasize on privacy and not misuse the individual-level information.

Data Access and Sharing

One challenge is the disinclination of private companies and other private institutions to share data about their clients as well as about their own operations. Hindrances like legal or reputational considerations may arise as a need to protect their competitiveness and as a culture of secrecy. In addition, there are technical and institutional challenges as well which includes difficulty in accessing and transferring data when it is stored in places that are inaccessible. There are other technical challenges like inter-compatibility of data and inter-operability of systems. However, these challenges might be less problematic as compared to challenges related to getting formal access or agreement on licensing issues. Any initiative in the field ought to completely recognize the importance of privacy issues and handling data in ways that ensure that privacy is not compromised.

Data Heterogeneity and Incompleteness

When information is consumed by humans, heterogeneity is tolerated and the richness of natural language is appreciated and provides valuable depth in understanding. Whereas, machine analysis algorithm cannot understand nuance and expects homogeneity. As a result, data must be structured as a first step in analysis carefully. Consider for example, a customer who carries out multiple transactions at a bank. We could create one record for every transaction, one record for every transaction carried out in a day, or one record for all lifetime interactions of this customer at the bank. Apart from the first design, the number of transactions per record would vary for each patient. In the example mentioned, the three design choices have successively less structure and greater variety. Many traditional data analysis systems may require greater structure. However, the less structured design is more likely to be effective for various purposes. For example, questions relating to increasing transactions may require a design which is an expensive join operation of the first two designs, but it can be avoided by the third design. Efficient design, access and analysis of moderately structured data require further analysis. Computer systems work efficiently when they can store multiple items that have identical size and structure. Consider a company's record database design which includes fields like the employee's birth date, post, blood type and gender. What is supposed to be done if one the information is not provided by the employee? The record is still placed with

the database. The corresponding attribute values of the unfilled information are set to NULL. If a data analyst decides to classify an employee by his blood type, he must also take into account employees for which this information is not known. In such situations data cleaning and error correction is required. Yet, incompleteness and some errors in data are likely to remain which must be managed during data analysis. Performing this task rightly is a challenge.

Data Analysis and Interpretation

Working with new and varied data from different sources brings about a lot of analytical challenges. The importance and sternness will vary depending on the type of analysis being conducted and the type of information the data might inform. Since the range of data might be very wide, question like "what the data is trying to tell us?" might emerge which is very important in various researches (social science, political, etc). However, there is a general notion that "new" data sources pose specific and acute challenges. Hence, it is essential that these concerns are spelled out in an entirely transparent manner. Also, significant shares of the new digital data sources that make up Big Data are derived from people's own understanding. Perception can be inaccurate and misleading.

OPPORTUNITIES

Data Revolution: The world is going through data revolution, or flooding of data. The amount of data in our world has been exploding, and analyzing larger data sets will become a key basis of competition. In earlier days, relatively small volume of data was produced, which was handled by analytical expertise, but now-a-days the rate at which data is generated has led to data deluge. According to IDC immense data is growing at 50 percent a year, or more than doubling every two years. The impact of data abundance is beyond business, say scientific research, public health, academia and government. There is no area that is going to be untouched.

Data Acquisition & Recording: Big Data arises from data generating sources for years. Example, our ability to observe everything happening around us, the heart rate of an elderly citizen, weather forecasting, presence of toxins in the air, etc produces nearly 1 million terabytes of raw data per day. In the same way, scientific experiments and research are capable of producing petabytes of data today. The data received should be to remove excessive data. One of the methods is by defining filters in such a way that they should not dispose useful information. Suppose reading of one sensor differs considerably from the rest, there is higher probability of the sensor being faulty, but we cannot be sure if it's an artifact that deserves attention? In such cases, we need intelligent research in data reduction which compresses the size of data, not missing important notification. Also, we are in need of "on-line" analysis techniques which process such streaming data on the fly; we are not in a position to store first and reduce later. The second method is to automatically generate correct data which explains how data is recorded and measured. Especially, in scientific experiments, details of specific experimental conditions and procedures are required to decipher the results correctly, and it is important to record such metadata with observational data

Information Extraction & Cleaning: The set of available data gets younger and younger, i.e. data is getting updated in less than a minute. An increasing percentage of this data is produced and made available real-time. Data is not only becoming available for the computers, but also more understandable. This data contains unstructured data such as words, images and videos which is not in a format ready for analysis. This data cannot be analyzed effectively and hence we require extraction process which pulls out information and expresses it in structured form which is suitable for analysis. Also, artificial intelligence techniques like character and pattern recognition, machine learning and natural-language processing are developed to analyze unstructured data. There is a slight probability of receiving misleading or missing information during survey, historical events or so. In such cases, well-understood error models should be recognized while data cleaning.

Data Integration: Given the variation in the flow of data, it is not sufficient to record it and dump it. As we know there is wide range of data required for scientific experiments. If we dump this data, it is unlikely anyone would ever be able find correction or conduct further research. With adequate metadata, there is some hope for further innovation. Data analysis is considered as the most challenging task when compared to locating, identifying, understanding, or citing data. For effective analysis all of these should be automized which requires differences in data structure and denotation to be expressed in computer understandable language. For this, data integration is required to achieve automated difference resolution with minimal error. Simpler analysis too need database design to store information. Today, these designs are executed in enterprises by highly paid professionals.

Interpretation: To analyze Big Data it is necessary for one to understand the analysis i.e. to interpret these results. This interpretation involves all the assumptions made and retracing the received data or analysis. For every analysis there are possible sources of error: assumptions, experimental conditions, operational system may have bugs which leads misleading results. In such cases, one needs to understand and justify the results produced by the computer which makes the challenge more complex. In short, it is not enough to provide only the results. Rather, supplementary information must be provided that explains the derived results for respective inputs. Such information is known as the provenance of the result. Using techniques to store adequate metadata, we can design a system which provides users the ability to interpret analytical results and to repeat the analysis with different parameters, assumptions, or data sets. Furthermore, user should be able to drill down data into pieces and understand its origin which is a key feature in understanding the data. One way is to enable the users by modifying parameters with every step of analysis and viewing the results of these incremental changes. This makes the user judgemental about the analysis and verifies it with the expected results.

CONCLUSIONS

Even today, there is enough data to be processed by traditional databases, but new systems will accelerate the amount of data collected and access to this huge data. Big data technology of high predictive and mathematical analysis is making it possible to dump out irrelevant information efficiently. The challenge is not of managing big data, but the information received after processing it. There can be many such arguments, but the best way is to start off with small and justify the benefits. After all, big data is leading to higher analysis and opportunities to improve the quality of life.

REFERENCES

1. Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela H. Byers. “Big data: The next frontier for innovation, competition, and productivity.” McKinsey Global Institute (2011): 1-137. May 2011.
2. Lohr, Steve. “The Age of Big Data.” New York Times. 11 Feb, 2012.
3. “The Data Deluge.” The Economist. 25 Feb 2010.
4. www.sas.com
5. adayinbigdata.com
6. www.cra.org