International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792

Big Data

Ritu Singh Kalsi
Assistant Professor, Dept. of Computer Applications, Post Graduate Govt. College for Girls, Chandigarh, India

***Abstract*: Big data is a buzzword, or catch-phrase, meaning a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity. Ideally, all of this information would be converted into structured data. However, this would be costly and time consuming. Also, not all types of unstructured data can easily be converted into a structured model. For example, an email holds information such as the time sent, subject, and sender (all uniform fields), but the content of the message is not so easily broken down and categorized. Big data is not just for companies and governments but also for all of us individually. We can now benefit from the data generated from wearable devices such as smart watches or smart bracelets. Take the Up band from Jawbone as an example: the armband collects data on our calorie consumption, activity levels, and our sleep patterns. In summary, Big Data is a lot of data produced very quickly in many different forms. In August, 2013 Mark van Rijmenam added "veracity, variability, visualization, and value" to the definition, broadening the realm even further. Rijmenam stated "90% of all data ever created, was created in the past two years. From now on, the amount of data in the world will double every two years."**

***Keywords*: Big data, Buzzword and processing**

1. Introduction

Big data is a buzzword, or catch-phrase, meaning a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity. Ask any Big Data expert to define the subject and they'll quite likely start talking about "The three V's" - "volume, velocity and variety," concepts originally coined by Doug Laney in 2001 to refer to the challenge of data management. He is a pioneer in the field of data warehousing who is considered to have initiated the field of Infonomics (Information Economics). One might understand volume as a huge collection of raw information. Though that is true, it is also about storage costs of data. Important data can be stored on premises as well as on cloud, the latter being flexible option. The speed of processing data is an important factor. Google had already admitted that the speed at which a page loads is essential for better rankings. Apart from the rankings, the speed also provides comfort to users while they shop. The same applies for data being processed for other information. While talking about velocity, it is essential to know that it is beyond just higher bandwidth. It combines readily usable data with different analysis tools. Readily usable data means some homework to create structures of data that are easy to process. The next dimension, variety, spreads further light on this. When it comes to variety, data is of two types, structured and unstructured. Structured data is information, usually text files, displayed in titled columns and rows which can easily be ordered and processed by data mining tools. This could be visualized as a perfectly organized filing cabinet where everything is identified, labeled and easy to access. Unstructured data, usually binary data that is proprietary, is that which has no identifiable internal structure. It's a massive unorganized conglomerate of various objects that are worthless until identified and stored in an organized fashion. Once this organization process has taken place (through the use of specialized software), the items can then be searched through and categorized (to an extent) for obtaining insights. While data mining tools might not be equipped to parse information in email messages (however organized it may be), you may have very good reason to collect and categorize data from this source. This illustrates the importance and plausible breadth of unstructured data. Ideally, all of this information would be converted into structured data. However, this would be costly and time consuming. Also, not all types of unstructured data can easily be converted into a structured model. For example, an email holds information such as the time sent, subject, and sender (all uniform fields), but the content of the message is not so easily broken down and categorized. In summary, Big Data is a lot of data produced very quickly in many different forms. In August, 2013 Mark van Rijmenam added "veracity, variability, visualization, and value" to the definition, broadening the realm even further. Rijmenam stated "90% of all data ever created, was created in the past two years. From now on, the amount of data in the world will double every two years."

2. Big data for organizations

Companies have discovered the competitive advantage to big data analysis. It explains the explosion of behavioral tracking embedded in every piece of technology we own. Collecting, analyzing and storing consumer behavior multiplies the value of a business dollar. When dealing with larger datasets, organizations face difficulties in being able to create, manipulate, and manage big data. This is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyze massive datasets. Leaders

in data analytics face challenges of spreading awareness of technological potential, creating complex and realistic models, and creating the right framework to tie together various types of databases. Despite these problems, big data has the potential to help companies improve operations and make faster, more intelligent decisions. This data, when captured, formatted, manipulated, stored, and analyzed can help a company to gain useful insight to increase revenues, acquire or retain customers, and improve operations. Big Data and analytics are experiencing a dramatic upsurge in innovation and investment. New software and improved approaches in these fields is equipping us with the ability to deal with the challenges of modern-day business. Three prominent ones come to the forefront. First is the emergence of targeted software and services that are helping businesses achieve a more direct, and at times faster, impact on the bottom line. There is a new emerging class of analytics specialists who builds targeted models that have a clear business focus and can be implemented swiftly. We are seeing them successfully applied in a wide range of areas: logistics, risk management, pricing, and personnel management, to name just a few. Because these more specific solutions have been applied across dozens of companies, they can be deployed more readily. Collectively, such targeted applications will help raise management's confidence in investing to gain scale. There's still a need for a shift in culture and for a heavy emphasis on adoption, but the more focused tools represent a big step forward. Second, new self-service tools are building business users' confidence in analytics. One hot term gaining traction in the analytics world is "democratization." Getting analytics out of the exclusive hands of the statistics gurus, and into the hands of a broad base of frontline users, is seen as a key building block for scale. Without needing to know a single line of coding, frontline users of new technology tools can link data from multiple sources (including external ones) and apply predictive analytics. Visualization tools, finally, are putting business users in control of the analytics tools by making it easy to slice and dice data, define the data exploration needed to address the business issues, and support decision making.

A. Big data technologies and skills

Interpretation of Big Data can bring about insights which might not be immediately visible or which would be impossible to find using traditional methods. This process focuses on finding hidden threads, trends, or patterns which may be invisible to the naked eye. This requires new technologies and skills to analyze the flow of material and draw conclusions. Apache Hadoop is one such technology, and it is generally the software most commonly associated with Big Data. Apache calls it "a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models." Just as Big Data can be both a noun and a verb, Hadoop involves something that is and something that does - specifically, data storage and data processing. Both of these occur in a distributed fashion to improve efficiency and results. Hadoop is open-source and there are variants produced by many different vendors such as Cloudera, Hortonworks, MapR and Amazon. There also other products such HPCC and cloud-based services such as Google Big Query. Skills are also brought to the table by Big Data scientists who obtain business value from a plethora of information by analyzing it for meaning and trends. This requires mathematical and statistical expertise as well as creative, communicative, problem-solving, and business skills, making it a very complex but incredibly valuable role. New fields have developed to train for this expanding career path, and there is a wealth of advice for those aspiring to enter the Big Data industry - which is expected to see a 500 percent job increase from January 2014 to January 2016.

B. Big data applications

Big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers. The big objective, in many cases, is to create predictive models. Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts. Big data is not just for companies and governments but also for all of us individually. We can now benefit from the data generated from wearable devices such as smart watches or smart bracelets. Take the Up band from Jawbone as an example: the armband collects data on our calorie consumption, activity levels, and our sleep patterns. The computing power of big data analytics enables us to decode entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns. What's more, big data analytics allow us to monitor and predict the developments of epidemics and disease outbreaks. Integrating data from medical records with social media analytics enables us to monitor flu outbreaks in real-time. Science and research is currently being transformed by the new possibilities big data brings. Take, for example, CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator. The CERN data center has 65,000 processors to analyze its 30 petabytes of data. However, it uses the computing powers of thousands of computers distributed across 150 data centers worldwide to analyze the data. Such computing powers can be leveraged to transform so many other areas of science and research. High-Frequency Trading (HFT) is an area where big data finds a lot of use today. Here, big data algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make, buy and sell decisions in split seconds. Big data is used to improve many aspects of our cities and countries. For example, it allows cities to optimize traffic flows based on real time traffic information as well as social media and weather

data. A number of cities are currently piloting big data analytics with the aim of turning themselves into Smart Cities, where the transport infrastructure and utility processes are all joined up. Where a bus would wait for a delayed train and where traffic signals predict traffic volumes and operate to minimize jams.

3. Future big data and conclusion

Talking about the future of big data is somewhat beside the point, because it's very much a "here and now" phenomenon. There are only two certainties in big data today: It won't look like yesterday's data infrastructure, and it'll be very, very fast. And the possibilities will be astounding, blurring the lines of industries and fundamentally altering the way businesses interact with their customers and each other. This latter trend is evident in the rise of Apache Spark and real-time analytics engines, but it's also clear from the parallel rise of real-time transactional databases (NoSQL). The former is all about lightning-fast data processing, while the latter takes care of equally fast data storage and updates. Market research firm IDC forecasts that the market for Big Data is expected to grow from $3.2 billion in 2010 to $16.9 billion in 2015 in its report, Worldwide Big Data Technology and Services 2012-2015. Job growth will also be significant. McKinsey says that in 2014 the U.S. alone faces a shortfall of 140,000 to 190,000 people to fill big data jobs, with an additional shortage of 1.9 million analysts and managers. They say that by 2018, the U.S. won't be able to fill 50 percent to 60 percent of these roles. Darin Stewart of InformationWeek said in a recent article about big data, "The age of information overload is slowly drawing to a close. Enterprises are finally getting comfortable with managing massive amounts of data, content and information. The pace of information creation continues to accelerate, but the ability of infrastructure and information management to keep pace is coming within sight. Big Data is now considered a blessing rather than a curse."

References

[1] http://www.powershow.com/
[2] https://en.wikipedia.org/wiki/Big_data
[3] https://www.mongodb.com/big-data-explained
[4] http://www.ibm.com/developerworks/library/bd-archpatterns1/