International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 5 3328 - 3331

Diversified Usages of Big Data : Scope and Challenges

Anu Gupta
Associate Professor, Department of Computer Science and Applications,
Panjab University, Chandigarh
anugupta@pu.ac.in

Abstract— During the last few decades, novel trends in terms of advancement in computational power, increase in storage capacity, growing usage of Web and mass availability of smart devices are leading towards the collection and storage of enormous amount of Data. Big Data Analytics are applied to derive gainful insights and knowledge from this data. The present paper aims to characterize Big Data and then delves in various application domains where Big Data techniques can potentially be applied. The researchers in computer science and allied areas have several research challenges. To address these challenges, new tools and techniques need to be developed and leveraged to solve problems for people, industry and the world at large.

Keywords-*Big Data; Big Data Analytics; Structured Data; Unstructured Data; Hadoop.*

*****

I. INTRODUCTION

Since several decades, data is continuously being gathered on weather, demographics, agriculture, telephone contacts, census, sports, government policies, business transactions etc. In last two decades, growing usage of Internet forced various organizations to collect data about users' preferences for news, music, movies, shopping etc. With the advent of Web 2.0 in the last decade, people started sharing their personal information, photos, videos, opinions etc. over Facebook, Twitter and several other social-media tools. As a result, today we have reached in an era of Data Explosion where large amount of data is being generated by variety of devices surrounding us. Eric Schmidt presented striking statistics about the exponential growth of data i.e. now five Exabyte of data is being produced every two days against same amount of data being generated from the dawn of civilization until 2003 [1]. The data is originating from many sources namely social media sites, sensors, digital photos, business transactions, global positioning systems, Internet of Things (IoT). More than 100 hours of video is being uploaded to YouTube every minute, and even more video is being collected worldwide through the surveillance cameras installed around us [2]. Several mobile-phone apps are also keeping track of our movements [2]. Hence, the data has also been termed as new class of economic asset like currency and gold at World Economic Forum [3].

In the year 2005, the term 'Big Data' was coined by Roger Magoulas from O'Reilly media, that refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size as well as complexity [4]. Big Data can be seen in the finance and business sectors where huge amount of stock trading, banking, online shopping data flows through computerized systems every day and utilized further for inventory monitoring, customer behavior and market trend analysis. In Bio-medical and pharmaceutical sciences, big sets of data such as DNA sequencing, clinical data and patient data are analyzed and used for advanced research. Big data may include surveillance videos, activity logs as well as unstructured text posted on blogs and social media. Big Data contains both structured (e.g. traditional database tables) and unstructured data like e-mail attachments, medical imaging records, tweets, graphics, video and audio [5]. Over 80 percent of enterprise data is unstructured requiring significant storage space and effort to manage [6]. The advancement in computational and storage power is increasingly facilitating the collection, storage and analysis of these Big Data sets.

II. CHARACTERISTICS OF BIG DATA

To characterize Big data, 4 V's have been associated referred as volume, velocity, variety and veracity that are depicted in Fig. 1 [7].

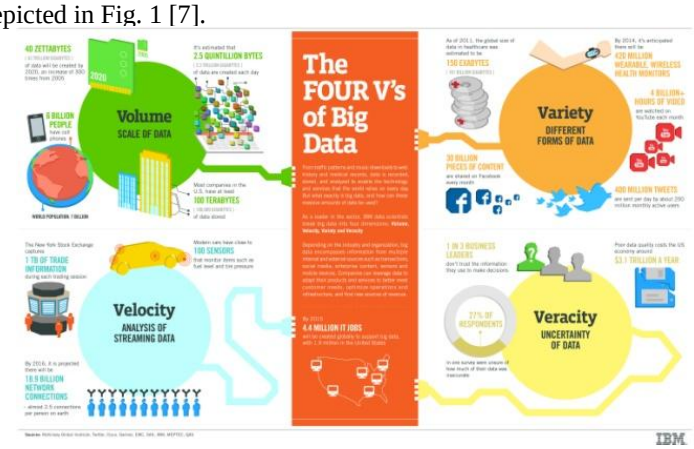

Figure 1. The Four V's of Big Data [7]

Data Volume refers to the vast amount of data that is being generated every second [8]. Facebook ingests about 500 Terabytes of new data every day [9]. Boeing 737 airplane generates 240 Terabytes of flight data during a single flight across the US [10]. The smart phones create and consume large amount of data. Sensors embedded into everyday objects also result in large datasets. Today data is expressed not in Terabytes but in Zettabytes or Brontobytes.

Data Velocity refers to the speed at which new data is being generated [8]. Several messages on social media sites go viral in seconds now a days [8]. On a website, Click streams and ad impressions capture user behavior at millions of events per second. Machine to machine processes exchange data between billions of devices. Infrastructure and sensors generate massive log data in real-time. On-line gaming systems support millions of concurrent users, each producing multiple inputs per second.

Data Variety refers to different types of data that can be used such as geospatial data, 3D data, audio/video clips, log files, social media messages etc. [8]. In the past, structured data that could fit into tables or relational databases was primarily used [8]. Big data technology comprises data of unstructured type also such as messages, social media conversations, photos,

sensor data, video or voice recordings [8]. To extract knowledge, all these types of data need to be linked together and analyzed.

Data Veracity refers to the trustworthiness of the data. With availability of many forms of big data, quality and accuracy are less controllable [8]. The issue of trust and integrity is important for big data just as for traditional relational databases.

III. BIG DATA ANALYTICS

With every passing second, data is becoming more diverse, larger and faster. Big data is not only about collection and storage of data rather deriving gainful insights and knowledge from the data using advanced analytical techniques often termed as Big Data Analytics (BDA) [11]. According to Wikipedia, Big data often refers the use of predictive analytics or certain advanced methods to extract value from data [12]. The true value of data is realized for competitive advantage of a business organization only after it is contextualized and understood in order to deliver insights. BDA is revolutionizing the way we see and process the world. It is the area which can potentially use several advanced analytic techniques such as predictive analytics, data mining, statistical analysis, complex SQL, data visualization, artificial intelligence, natural language processing etc. [13]. Database analytics platforms such as MapReduce, in-database analytics, in-memory analytics and NOSQL data stores are generally used [14-15]. With BDA, attempt is focused to discover new business facts that no one knew before. Advanced analytics is considered to be the best way to discover novel business opportunities, new customer segments, identify the best suppliers, associate products of affinity, understand sales seasonality etc. [13, 16].

The analysts may combine log data with historical data from a data warehouse and may discover new behavior in a subset of the customer base [13]. The discovery would lead to a metric, report, analytic model, or some other product of BI, through which the company could track and predict the new form of customer behavioral change [13]. The basic premise of BDA is to use real-time data for real-time decision-making to become a real-time business. [17].

Hadoop is a framework that supports distributed processing of large datasets across clusters of computers using simple programming models [18]. In BDA, the Hadoop system captures datasets from different sources and then performs functions such as storing, cleansing, distributing, indexing, transforming, searching, accessing, analyzing, and visualizing [19]. As a result, semi-structured and unstructured data are converted into structured data and ready for consumption by end-users in applications, as shown in Fig. 2 [19].

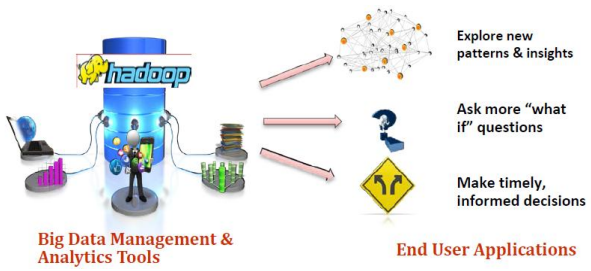

Figure 2. Simplified Big Data Working Process [19]

IV. KEY APPLICATION AREAS OF BIG DATA

A. *Health Care*

Healthcare is considered to be a highly data intensive industry. A range of health information systems e.g., electronic health records (EHR), picture archiving communications system (PACS), clinical decision support systems (CDSS), and laboratory information systems are used [20]. Along with knowledge of disease, it is equally important to understand the patient [20]. Patient's medical record and current health situation is used in wellness and disease management programs [21]. The computing power of big data analytics may effectively be used to better predict disease patterns and find new cures [11]. Work is underway to gather massive amounts of data regarding behavioral, genetic, environmental, epigenetic, clinical aspects to search for ways to prevent and treat the disease [2]. Satellite and cell phone data is being used to track population movements and predict about the spread of disease. By correlating information about a person's genome, lifestyle habits and location with his medical history, it will become possible to understand the true effectiveness of drugs as well as prescribe personalized medications [2]. This will certainly begin a revolution in health and medicine.

B. *Governance*

Big data analytics is enabling cities to transform into smart cities, where the transport infrastructure and utility processes are joined up [8]. It allows cities to optimize traffic flows based on real time traffic information as well as social media and weather data [8]. Consider a futuristic traffic management system with information regarding thousands of vehicles and local hot spots on roadways [11]. The system may predict potential congestion points along a route chosen by a user and suggest alternatives. It would require evaluating multiple spatial proximity queries working with the trajectories of moving objects [11].

C. *Security and Law enforcement*

Security agencies use big data analytics to make connections and detect patterns so they can prevent and solve crimes [8]. Police forces use large amount of data and puts it through advanced analytics engine to catch criminals and even predict when criminals are likely to strike. Credit card companies use big data analytics to detect fraudulent transactions [8]. Big data has also been used to predict crimes before they happen – a "predictive policing" with three times more accuracy than existing methods of forecasting.

D. *Sports*

In many sports, video analytics is used extensively to track the performance of every player in a football or baseball game [8]. Sensor technology is built into sports equipment such as basket ball or golf club to gather massive data. Many elite sports teams also track athletes outside the sporting environment by using smart technology to track nutrition and sleep as well as social media conversations to monitor emotional wellbeing [11]. The ever-growing world of Big Data has already transformed the sporting arena in surprising ways. Large numbers of cameras are positioned at various angles to record and log the movement of the players as well as the ball. This brings in speed, distance, player separation and ball possession statistics, which can be filtered into user-friendly interfaces so coaches can make better decisions. During training and games, athletes are now equipped with sensors that

monitor every aspect of their performance- from the heart rate and metabolism to reaction time. This data and advanced statistics allow managers and technical staff to determine the factors that influence player performance and consequently, the performance of the team [22].

E. Business and Commerce

Companies can expand their traditional data sets with social media data, browser click stream data and sensor data to get a more complete picture of their customers [8]. The big objective, in many cases, is to create predictive models so that telecom companies can now better predict customer churn; retailers can predict what products will sell, car insurance companies can understand how well their customers actually drive [8,11]. Retailers can optimize stock based on predictive models generated from social media data, web search trends and weather forecasts. They can further improve supply chain or delivery route using data from geographic positioning and radio frequency identification (RFID) sensors. By combining the data about a person's shopping habits with knowledge of his social preferences, health and location; an artificial intelligence based shopping assistant could anticipate what a person wants to wear or to eat [8]. Amazon uses customer data to give us recommendations based on our previous purchases [2]. Google uses search data and other information it collects to sell ads and to provide other services and products.

F. Agriculture

Currently sensors are added to the latest agriculture equipments that help farmers to reduce the downtime of the equipments [23]. The information is combined with historical and real-time data regarding weather prediction, soil conditions, crop features and many other data sets. The Mobile Apps gives access to historical as well as real-time field information; performs soil samples from the field and users can share the information directly with trusted advisors for live remote-advice while on the field [23]. The information can then be used by farmers in decision making regarding cropping, irrigation, ploughing etc. All this will increase the productivity of the crops. Machine optimization uses proactive diagnostics on service issues, to keep the machine running as well as to make it more efficient and productive [23].

G. Weather

Weather influences our decision-making on a daily basis. We are not able to control the weather, but better forecasting will allow us to make more informed plans that can limit financial losses, provide new business opportunities, reduce government spending and even save lives [24]. The volume and diversity of environmental data is increasing exponentially with the usage of new sources of observation (sensors placed on automobiles). Data-driven analytics techniques are also evolving to analyze this tsunami of data, better estimate the long-term state of climate systems and to improve weather forecasting, particularly for severe weather events [24].

H. Education and Training

Education and training can gain benefits from Big Data primarily in form of personalized training just like personalized online shopping. In the current age, several websites (Coursera, edX, Khan Academy etc.) are offering study material online resulting in an explosion of data that can be used to improve educational effectiveness as well as support basic research on learning. Other ways that technology enables learning is through predictive and diagnostic assessments. The predictive assessments seek to evaluate how students will perform on standardized tests, while diagnostic assessments emphasize on finding instructional techniques suitable for individual students. Big data can also help a lot to improve traditional classroom learning environment. In large classes, the teacher does not get to know about the student's habits particularly performance in other classes. A digital tutor could be designed to keep track of a child's progress, teaching-style preferences and intellectual strengths/weaknesses [2]. Using data gathered by digital learning devices, test scores, attendance and habits, the teacher could be informed of which students to focus on, what to emphasize, and how best to teach an individual child [2]. This could change the conventional education system itself.

Above discussed areas are some of the representative areas. But the applications of Big Data are interminable. Currently we are experiencing only the beginning of a revolution into a big data economy.

V. CHALLENGES

Big Data presents many opportunities to influence growth in diverse application domains, but at the same time it brings out many challenges. Few such challenges are being discussed here.

Several innovative data management frameworks have been designed, that are collectively referred as 'NOSQL' frameworks. Some of 'NOSQL' approaches employ hierarchical object representation (XML, JSON, BSON) while others are based on key-value store. Some more approaches are still evolving. For an organization planning for Big Data, selecting a suitable NoSQL tool may be a critical challenge. Another challenge is to import the legacy data which is mostly in structured form and then integrate with new data that is unstructured form comprising emails, text, tweets, images, audio, video accompanied by corresponding metadata. In Big data platform, data is generally migrated from different sources on different schedules that may result in data getting out of synchronisation with the originating system [25]. This may further lead to faulty decision-making. Big Data includes information provided by increasingly diverse sources, of varying reliability. Uncertainty, errors, and missing values need to be managed. Due to large volume, it is impractical to validate every data item. New approaches to data qualification and validation are needed [21].

The one main task of Big Data Analysis is to visualize the results effectively. But bigger volume and larger variety of data make it very difficult to render user-friendly visualizations. Thus new techniques, algorithms and frameworks are needed. Incompatible data formats, non-aligned data structures and inconsistent data semantics represents significant challenges that can lead to analytic sprawl [21]. Along with tools and techniques, there is shortage of good data scientists as well [2]. Instead of traditional method of "bring the data to the code", big data necessitates to process the data "in place" and transmit only the resulting information i.e. "bring the code to the data" [21]. Broad access to huge amounts of data also raises some troubling issues such as privacy, malicious use etc. For electronic health records, there are strict laws governing what data can be revealed in different contexts [11]. In today's complex world, it often requires multiple experts from different domains to really understand the current happenings. A Big Data analysis system must support input from multiple human experts and shared exploration of results [11].

The researchers in computer science and allied areas have several research challenges. There is a dire necessity to improve individual tools that can handle particular type of data effectively and find valuable entities in the data. Historical data-sets need to be integrated with data in motion to provide right context for better data management and analytics. Accordingly, it is also required to adapt and extend existing models to carry out analysis in a convenient and accurate manner. With the rich collection of analytics techniques and tools, a great user experience can be achieved through automation and intelligent guidance. Such newly developed tools and computing environment need to be leveraged to solve important problems for people, industry and the world at large.

VI. CONCLUSION

The progress of Big Data in diverse application domains is not obstructed by ability to capture and collect data. Rather it calls for novel tools, algorithms and frameworks to manage, analyze, summarize, visualize and discover knowledge from huge data in a timely, lucid and scalable fashion. Like any powerful technology, right usage of Big Data has potential to propel towards myriad possibilities.

REFERENCES

[1] Eric Schmidt, August 2010, Available: http://techcrunch.com/2010/08/04/schmidt-data/

[2] Vivek Wadhwa, "The rise of big data brings tremendous possibilities and frightening perils", April 2014. Available: http://www.washingtonpost.com/blogs/innovations/wp/2014/04/18/the-rise-of-big-data-brings-tremendous-possibilities-and-frightening-perils/

[3] World Economic Forum, "Personal Data: The Emergence of a New Asset class", Jan. 2011. Available: http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf

[4] Roger Magoulas and Ben Lorica, "Big Data: Technologies and Techniques for Large-Scale Data". Available: http://www.oreilly.com/data/free/files/release2-issue11

[5] Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis: Challenges and Solutions", International Conference on Cloud, Big Data and Trust, Nov. 2013, RGPV, Bhopal.

[6] Seth Grimes, "Unstructured Data and the 80 Percent Rule", Aug. 2008. Available: http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

[7] The Four Vs of Big Data, Available: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[8] Bernard Marr, "Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance", John Wiley & Sons Publishers, Jan. 2015.

[9] Cnet news, Available : http://www.cnet.com/news/facebook-processes-more-than-500-tb-of-data-daily/

[10] Tim Crawford , "Are Enterprises Prepared for the Data Tsunami?", Jan. 2014, Available : http://avoa.com/2014/01/20/are-enterprises-prepared-for-the-data-tsunami/

[11] D. Agrawal et. al, "Challenges and Opportunities with Big Data" , Big Data White Paper- Computing Research Association, Feb 2012. Available: http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

[12] Big Data, Available: http://en.wikipedia.org/wiki/Big_data

[13] Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis: Challenges and Solutions", International Conference on Cloud, Big Data and Trust, Nov. 2013, RGPV, Bhopal.

[14] Dunren Che, Mejdl Safran, Zhiyong Peng, "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities", Database Systems for Advanced Applications, Lecture Notes in Computer Science Volume 7827, 2013, pp 1-15.

[15] Onur Savas, Yalin Sagduyu, Julia Deng, and Jason Li, "Tactical Big Data Analytics: Challenges, Use Cases and Solutions", Big Data Analytics Workshop in conjunction with ACM Sigmetrics 2013, June 21, 2013.

[16] Steve Lohr, "The Age of Big Data", New York Times, Feb. 11, 2012. Available: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[17] Lisa Kelly, "Big Data and Analytics: A Large Challenge Offering Great Opportunities", Available: http://www.computerweekly.com/feature/Big-data-and-analytics-a-large-challenge-offering-great-opportunities

[18] Hadoop, https://hadoop.apache.org/

[19] Jean Yan, "Big Data, Bigger Opportunities", April 2013. Available: http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf

[20] MH Kuo, T Sahama, AW Kushniruk, EM Borycki, DK Grunwell, "Health big data analytics: current perspectives, challenges and potential solutions", International Journal of Big Data Intelligence,Vol. 1, Issue 1, pp.114-126.

[21] S. Kaisler, F. Armour, J. A. Espinosa, W. Money, "Big Data: Issues and Challenges Moving Forward", 46th Hawaii International Conference on System Sciences, Jan. 2013.

[22] Richard Attias, "Big Data: The Next Revolution of Sport", September 2014. Available: http://www.huffingtonpost.com/richard-attias/big-data-the-next-revolut_b_5800342.html

[23] Mark van Rijmenam, "John Deere Is Revolutionizing Farming With Big Data", Available: https://datafloq.com/read/john-deere-revolutionizing-farming-big-data/511

[24] Lorna Garey, "3 Ways Big Data, Supercomputing Change Weather Forecasting", September 2014. Available: http://www.informationweek.com/big-data/big-data-analytics/3-ways-big-data-supercomputing-change-weather-forecasting/a/d-id/1269439

[25] David Loshin, Addressing Five Emerging Challenges of Big Data, June 2014.Available: https://www.progress.com/~/media/Progress/Documents/Papers/Addressing-Five-Emerging-Challenges-of-Big-Data.pdf