Concepts, Applications and Future Scope of Big Data

T. Karunakar, M. Nagarjuna, K. Hema

Abstract—Big Data refers to the data or sets of records that are too large in volume to be operated using the existing database management tools and techniques. They are produced in many important applications, such as search engines, business informatics, social networks, social media, genomics, meteorology, and weather forecast. Big data presents a big challenge for database and data investigative research. The main objective of this paper is to give a brief introduction of Big Data, its architecture, characteristics and challenges. The hurdles of securing the data and democratizing it have been elaborated amongst several others such as inability in finding sound data professionals in required amounts and software that possess ability to process data at a high velocity. Through the article, the authors intend to decipher the notions in an intelligible manner embodying in text several use-cases and illustrations.

Index Terms—Acquisition, Modeling, Sentiment Analysis, Interpretation.

INTRODUCTION

Recent advancement in technology has led to generation of a great quantity of data from distinctive domains over the past 20 years. Big data is a broad term for data sets so great in volume or complicated that traditional data processing applications are inadequate. Although the big data have large amount of data or volume, it also processes the number of unique characteristics unlike traditional data. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. For example, big data is usually unstructured and requires more time for analysis and processing. This development calls for new system architectures for data acquisition, transmission, storage, and large-scale data processing mechanisms.

Big Data is data that are enormous in size and exceeds the processing capacity of regular or traditional database systems. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The data is so enormous and are generated so fast that it doesn’t fit the structures of normal or regular database architecture. To analyze the data new alternative way must be used to process it.

1 Concepts

“Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few”. such colossal amount of data that is being produced continuously is what can be coined as Big Data. Big Data decodes previously untouched data to derive new insight that gets integrated into business operations. However, as the amounts of data increases exponential, the current techniques are becoming obsolete. Dealing with Big Data requires comprehensive coding skills, domain knowledge and statistics.

Despite being Herculean in nature, Big Data applications are almost ubiquitous from marketing to scientific research to customer interests and so on. We can witness Big Data in action almost everywhere today. From Facebook which handles over 40 billion photos from its user base to CERN’s Large Hadron Collider (LHC) which generates 15PB a year to Walmart which handles more than 1 billion customer transactions in an hour. Over a year ago, the World Bank organized the first WBG Big Data Innovation Challenge which brought forward several unique ideas applying Big Data such as big data to predict poverty and for climate smart agriculture and fore user focused Identification of Road Infrastructure Condition and safety.

ARCHITECTURE OF BIG DATA

Fig 1 Overview of the analytics workflow for Big Data.

One of the most time-consuming and extensive work tasks of
Analytics and investigative approach is preparation of data for analysis and processing; a problem often made worse by Big Data as it already stretches infrastructure to its limits. Performing analytics on huge volumes of data records requires efficient methods to perform operations on the data. Cloud analytics solutions need to consider the multiple Cloud deployment models adopted by enterprises, where clouds can be of certain type.

Private: deployed on a private network, managed by the organization or by an external firm. A private Cloud is suitable for businesses that need the highest level of control of security and data confidentiality.

Public: deployed off-site over the Internet and available to the common people. Public Cloud offers high efficiency and shared resources at cheap rate.

Hybrid: joins both Clouds where additional resources from a public Cloud can be provided as a requirement to a private Cloud. Considering the Cloud deployments, the following scenarios are usually envisioned considering the availability of data and analytics models: (i) data and schema are private; (ii) data is public, structures are private; (iii) data and schema are public; and (iv) data is private, structures are public.

CHARACTERISTICS OF BIG DATA:

![Fig. 2 Characteristics of Big Data](image)

Characteristics of Big Data by what is usually referred to as a multi V model, is shown in Fig. 2. Variety represents the types of records in data, velocity refers to the rate at which the specific amount of data is generated and analyzed, and volume defines the amount or number of records of data. Veracity means how much amount of the data can be trusted given the reliability of its source.

- **Data Volume:** Data volume defines the measures of amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it. As amount of data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among all other factors.

- **Data Velocity:** Data velocity is a mean to measure the speed of data generation, streaming, and arithmetic operations.

- **Data Variety:** Data variety is a measure of the richness of the data representation of the different types of data stored in the database – text, images video, audio, etc.

- **Data Value:** Data value measures the usefulness of data in making decisions. It has been noted that “the purpose of computing is insight, not numbers”. Data science is exploratory and useful in getting to know the data, but “analytic science” encompasses the predictive power of big data.

- **Complexity:** Complexity measures the amount of inter-connectedness and interdependence and overlapping of data in big data structures such that even a slight change in one or a few elements can affect very large changes or a small change that ripple across or cascade through the system and substantially affect its behavior, or no change at all.

Considering data velocity, it is considered that, to complicate matters further, arrival of data and processing or analyzing data are performed at different speeds, as illustrated in Fig. 4. Whilst for some applications, the arrival and processing of data can be performed in a block, other analytics applications require continuous and real-time analyses sometimes require immediate action upon processing of incoming data streams i.e. the action is taken when the data flows continuously.

![Fig. 3 Velocity of Big Data](image)

![Fig. 4 Variety of Big Data](image)

From an analytic perspective, it is probably the biggest obstacle to effectively use large volumes of data. Incompatible data formats, incomplete data, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic spread out over a large area in an untidy or irregular way.
Big Data is slowly becoming ubiquitous. Every arena of business is examining all the assumptions made and retracing the analysis, has to interpret these results. This interpretation does not occur out of nowhere in a vacuum. Usually, it involves limits value if users cannot understand or interpret the analysis correctly and completely is a continuously a technical challenge faced.

- **Data Acquisition and Recording:**
  Big Data does not fall out of sky or neither is it grown on tree: the generated data is recorded from some data generating source. Another biggest challenge is to automatically produce the correct metadata to describe what data is recorded and how it is recorded and measured. Another important issue here is data provenance. Recording information about the data at its generation is not usefulness. This information can be understood completely and flowed along through the data analysis pipeline.

- **Information Extraction and Cleaning:**
  Frequently, the data generated and stored will not support the format which needs to be ready for analysis. We cannot use the unsupported data in this form and still effectively analyze it. To do the analysis, we need a process to extract the information that selects the required information from the underlying unstructured data and expresses it in a structured form suitable for analysis and process. Analyzing and processing the data correctly and completely is a continuously a technical challenge faced.

- **Data Integration, Aggregation, and Representation:**
  Data stored in databases are not similar, they are heterogeneous data i.e. Variety of data. It is not enough simply to record the data and store it into a repository. Data analyzing and processing is considerably more challenging than simply locating, identifying, understanding, and citing data.

- **Query Processing, Data Modeling, and Analysis:**
  Methods for querying and mining the records of Big Data are fundamentally different from that of traditional statistical analysis on small samples like tables, views, cursor etc. Big Data is often noisy, dynamic, heterogeneous, inter-related interdependent, overlapping and untrustworthy.

- **Interpretation:**
  Having the ability to analyze or process Big Data is off the limits value if users cannot understand or interpret the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. This interpretation does not occur out of nowhere in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis.

### 2. APPLICATIONS OF BIG DATA.

Big Data is slowly becoming ubiquitous. Every arena of business, health or general living standards now can implement big data analytics. To put simply, Big Data is a field which can be used in any zone whatsoever given that this large quantity of data can be harnessed to one’s advantage. The major applications of Big Data have been listed below.

- **The Third Eye- Data Visualization**
  Organizations worldwide are slowly and perpetually recognizing the importance of big data analytics. From predicting customer purchasing behavior patterns to influencing them to make purchases to detecting fraud and misuse which until very recently used to be an incomprehensible task for most companies big data analytics is a one-stop solution. Business experts should have the opportunity to question and interpret data according to their business requirements irrespective of the complexity and volume of the data. In order to achieve this requirement, data scientists need to efficiently visualize and present this data in a comprehensible manner. Giants like Google, Facebook, Twitter, EBay, Wal-Mart etc., adopted data visualization to ease complexity of handling data. Data visualization has shown immense positive outcomes in such business organizations. Implementing data analytics and data visualization, enterprises can finally begin to tap into the immense potential that Big Data possesses and ensure greater return on investments and business stability.

- **Integration- An exigency of the 21st century**
  Integrating digital capabilities in decision-making of an organization is transforming enterprises. By transforming the processes, such companies are developing agility, flexibility and precision that enables new growth. Gartner described the confluence of mobile devices, social networks, cloud services and big data analytics as the as nexus of forces. Using social and mobile technologies to alter the way people connect and interact with the organizations and incorporating big data analytics in this process is proving to be a boon for organizations implementing it. Using this concept, enterprises are finding ways to leverage the data better either to increase revenues or to cut costs even if most of it is still focused on customer-centric outcomes. Such customer-centric objectives may still be the primary concern of most companies, a gradual shift to integrating big data technologies into the background operations and internal processes.

---

**Fig.6 Analysis as generated by IBM Institute of Business Value 2014 Analytics study.**

**Big Data in Healthcare:**

Healthcare is one of those arenas in which Big Data ought to
have the maximum social impact. Right from the diagnosis of potential health hazards in an individual to complex medical research, big data is present in all aspects of it. Devices such as the Fitbit, Jawbone and the Samsung Gear Fit allow the user to track and upload data. Soon enough such data will be compiled and made available to doctors, which will aid them in the diagnosis. Several partnerships like the Pittsburgh Health Data Alliance have been established. The Pittsburgh Health Data Alliance is a collaboration of the Carnegie Mellon University, University of Pittsburgh and the UPMC. In their website, they state, “The health care field generates an enormous amount of data every day. There is a need, and opportunity, to mine this data and provide it to the medical researchers and practitioners who can put it to work in real life, to benefit real people…….The solutions we develop will be focused on practitioners who can put it to work in real life, to benefit real patients diagnosis will be analyzed and compared with the symptoms of others to discover patterns and ensure better treatment. IBM has taken initiative in a large scale to implement big data in healthcare systems be in its collaboration with healthcare giant Fletcher Allen or with the Premier healthcare alliance to change the way unstructured but useful clinical data is made available to more medical practitioners so as to improve population health. Big Data can also be used in major clinical trials like cure for various forms of cancer and developing tailor-made medicines for individual patients according to their genetic makeup. To summarize, Sundar Ram of Oracle stated, “Big Data solutions can help the industry acquire organize & analyze this data to optimize resource allocation, plug inefficiencies, reduce cost of treatment, improve access to healthcare & advance medicinal research.

- **Big Data and the World of Finance:**

Big Data can be a very useful tool in analyzing the incredibly complex stock market moves and aid in making global financial decisions. For example, intelligent and extensive analysis of the big data available on Google Trends can aid in forecasting the stock market. Though this is not a fool-proof method, it definitely is an advancement in the field. A research study by the Warwick Business School drew on records from Google, Wikipedia and Amazon Mechanical Turk in the time period of 2004-2012 and analyzed the link between Internet searches on politics or business and stock market moves. In the paper, the author states, “We draw on data from Google and Wikipedia, as well as Amazon Mechanical Turk. Our results are in line with the intriguing possibility that changes in online informationgathering behavior relating to both politics and business were historically linked to subsequent stock Market moves…Our results provide evidence that for complex events such as large financial market moves, valuable information may be contained in search engine data for keywords with less-obvious semantic connections to the event in question.

Overall, we find that increases in searches for information about political issues and business tended to be followed by stock market falls. Big Data is also being implemented in a field called Quantitative Investing where data scientists with negligible financial training are trying to incorporate computing power into predicting securities prices by drawing ideas from sources like newswires, earning reports, weather bulletins, Facebook and Twitter. One very interesting avenue of using Big Data in finance is the sentiment extraction from news articles. Market sentiment refers to the irrational belief in investors about cash-flow returns. The Heston-Sinha’s Application of the Machine Learning algorithm provides us with the probability of an article being ‘positive’, ‘negative’ and ‘neutral’ using two other popular methods, one being with the use of the Harvard IV Dictionary.

In general, big data is set to revolutionize the landscape of Finance and Economy. Several financial institutions are adopting big data policies in order to gain a competitive edge. Complex algorithms are being developed to execute trades through all the structured and unstructured data gained from the sources. The methods adopted so far has not been completely adept, however, extensive research ensures growing dependence of the stock markets, financial organizations and economies on big data analytics.

- **Big Data in Fraud Detection:**

Forensic Data Analytics or FDA has been an intriguing area of interest in the past decade. However, very few companies are actually using FDA to mine big data. The reasons for this unfortunate situation vary from the deficit of expertise and awareness, developing the right tools to mine big data to lack of appropriate technology and inability to handle such humungous quantities of data. Ernst & Young undertook the Global forensic data analytics survey in 2014 and found that,
Today, Big Data is influencing IT industry like few technologies have done before. The massive data generated from sensor-enabled machines, mobile devices, cloud computing, social media, satellites help different organizations improve their decision making and take their business to another level.

"Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives," - Susan Hauser, corporate vice president of Microsoft.

Data is the biggest thing to hit the industry since PC was invented by Steve Jobs. As mentioned earlier in this paper, every day data is generated in such a rapid manner that, traditional database and other data storing system will gradually give up in storing, retrieving, and finding relationships among data. Big data technologies have addressed the problems related to this new big data revolution through the use of commodity hardware and distribution. Companies like Google, Yahoo!, General Electric, Cornerstone, Microsoft, Kaggle, Facebook, Amazon that are investing a lot in Big Data research and projects. IDC estimated the value of Big Data market to be -about $ 6.8 billion in 2012 growing almost 40 percent every year to $17 billion by 2015.

By 2017, Wikibon’s Jeff Kelly predicts the Big Data market will top $50 billion.

"Demand is so hot for solutions that all companies are exploring big data strategies. The problem is that the companies lack internal expertise and best practices… the side effect is that there is a services and consulting boom in big data. It’s a perfect storm of product and services” says Wikibon’s Jeff Kelly. Recently it was announced that, Indian Prime Minister’s office is using Big Data analytics to understand Indian citizen’s sentiments and ideas through crowd sourcing platform www.mygov.in and social media to get a picture of common people’s thought and opinion on government actions. Google is launching the Google Cloud Platform, which provides developers to develop a range of products from simple websites to complex applications. It enables users to launch virtual machines, store huge amount of data online, and plenty of other things. Basically, it will be an one stop platform for cloud based applications, online gaming, mobile applications, etc. All these required huge amount of data processing where Big Data plays an immense role in data processing.

The predictions from the IDC Future Scope for Big Data and Analytics are:

1. Visual data discovery tools will be growing 2.5 times faster than rest of the Business Intelligence (BI) market. By 2018, investing in this enabler of end-user self-service will become a requirement for all enterprises.
2. Over the next five years spending on cloud-based Big Data and analytics (BDA) solutions will grow three times faster than spending for on-premise solutions. Hybrid on/off premise deployments will become a requirement.
3. Shortage of skilled staff will persist. In the U.S. alone there will be 181,000 deep analytics roles in 2018 and five times that many positions requiring related skills in data management and interpretation.
4. By 2017 unified data platform architecture will become the foundation of BDA strategy. The unification will oc-

---

---
cur across information management, analysis, and search technology.

5. Growth in applications incorporating advanced and predictive analytics, including machine learning, will accelerate in 2015. These apps will grow 65% faster than apps without predictive functionality.

6. 70% of large organizations already purchase external data and 100% will do so by 2019. In parallel more organizations will begin to monetize their data by selling them or providing value-added content.

7. Adoption of technology to continuously analyze streams of events will accelerate in 2015 as it is applied to Internet of Things (IoT) analytics, which is expected to grow at a five-year compound annual growth rate (CAGR) of 30%.

8. Decision management platforms will expand at a CAGR of 60% through 2019 in response to the need for greater consistency in decision making and decision making process knowledge retention.

9. Rich media (video, audio, image) analytics will at least triple in 2015 and emerge as the key driver for BDA technology investment.

10. By 2018 half of all consumers will interact with services based on cognitive computing on a regular basis.

Big data isn't new, but now has reached critical mass as people digitize their lives. "People are walking sensors," said Nicholas Skytland, project manager at NASA within the Human Adaptation and Countermeasures Division of the Space Life Sciences Directorate.

Taking an average of all the figures suggested by leading big data market analyst and research firms, it can be concluded that approximately 15 percent of all IT organizations will move to cloud-based service platforms, and between 2015 and 2021, this service market is expected to grow about 35 percent.

CONCLUSION:
Big data is able to process and store that data and probably in bulk of amount in soon future. Hopefully, technology will get better. New technologies and tools that have ability to record, monitor measure and merge all kinds of data surrounding us, needs to be introduced very soon. Industries need new technologies and tools for anonymizing data, analysis, tracking and inspecting information, sharing and maintaining, private data in future. So many aspects of life which generates the big data on daily basis that manages big data world need to be shined as possible.

There are too much of future important challenges in Big Data management and analytics that arise from the nature of data: large, diverse, and evolving.

REFERENCES: