



BIG DATA:INTRODUCTION & APPLICATIONS

Computer Science

Santosh Kumar Asst. Prof.,Computer Science & Engg.,Women's Institute of Technology, L.N.M.U, Darbhanga-846004.

Narendra Nath Jha Asst. Prof.,Computer Science & Engg.,Women's Institute of Technology, L.N.M.U, Darbhanga-846004.

ABSTRACT

Big Data is a huge amount of imprecise data in variety of formats generated from variety of sources with rapid speed. It is most buzzed terms among researchers, industries and academia. Big Data is not only limited to data perspective but it has been emerged as a stream that includes associated technologies, tools and real word applications. The objective of this paper is to provide a simple, comprehensive and brief introduction of Big Data. In this paper,. We also discuss some trends and eminent applications in Big Data.

KEYWORDS

Big Data, chare, 4V-Volume, Variety, Velocity, Value,

INTRODUCTION

Nowadays, information technology opens the door through which humans step into a smart society and leads to the development of modern services such as: Internet e-commerce, modern logistics, and e-finance. It also promotes the development of emerging industries, such as Telematics, Smart Grid, New Energy, Intelligent Transportation, Smart City, and High-End Equipment Manufacturing. Modern information technology is becoming the engine of the operation and development of all walks of life. But this engine is facing the huge challenge of big data. Various types of business data are growing by exponential orders of magnitude. Problems such as data collection, storage, retrieval, analysis, and the application of data can no longer be solved by traditional information processing technologies. These issues have become great obstacles to the realization of a digital society, network society, and intelligent society. The New York Stock Exchange produces 1 terabyte (TB) of trading data every day; Twitter generates more than 7 TB of data every day; Facebook produces more than 10 TB of data every day; the Large Hadron Collider located at CERN produces about 15 PB of data every year. According to a study conducted by the well-known consulting firm International Data Corporation (IDC), the total global information volume of 2007 was about 165 exabytes (EB) of data. Even in 2009 when the global financial crisis happened, the global information volume reached 800 EB, which was an increase of 62% over the previous year. In the future, the data volume of the whole world will be doubled every 18 months. The number will reach 35 (zettabytes) ZB in 2020, about 230 times the number in 2007, yet the written record of 5000 years of human history amounts to only 5 EB data. These statistics indicate the eras of TB, PB, and EB are all in the past; global data storage is formally entering the "Zetta era."

Beginning in 2009, "big data" has become a buzzword of the Internet information technology industry. Most applications of big data in the beginning were in the Internet industry: the data on the Internet is increasing by 50% per year, doubling every 2 years. Most global Internet companies are aware of the advent of the "bigdata" era and the great significance of data.

In May 2011, McKinsey Global Institute published a report titled "Big data: The next frontier for innovation, competition, and productivity", and since the report was released, "big data" has become a hot topic within the computer industry. The Obama administration in the United States launched the "Big Data Research and Development Initiative" and allocated \$200 million specifically for big data in April 2012, which set off a wave of big data all over the world. According to the big data report released by Wikibon in 2011, the big data market is on the eve of a growth spurt: the global market value of big data will reach \$50 billion in the next five years. At the beginning of 2012, the total income of large data related software, hardware, and services were around \$5 billion. Greg McDowell, an analyst with JMP Securities, said that the market of big data tools is 11% of overall corporate IT spending, by 2020.

At present the industry does not have a unified definition of big data; big data has been defined in differing ways as follows by various parties:

Big Data refers to datasets whose size is beyond the capability of typical database software tools to capture, store, manage, and analyze. —McKinsey..

Big Data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.

—Gartner.

CHARACTERISTICS:

Big data has four main characteristics: Volume, Velocity, Variety, and Value (referred to as "4V," referencing the huge amount of data volume, fast processing speed, various data types, and low-value density). Following are brief descriptions for each of these characteristics.

Volume: refers to the large amount of data involved with big data. The scale of datasets keeps increasing from gigabytes (GB) to TB, then to the petabyte (PB) level; some even are measured with exabytes (EB) and zettabytes (ZB). For instance, the video surveillance cameras of a medium-sized city in China can produce tens of TB data every day.

Variety: indicates that the types of big data are complex. In the past, the data types that were generated or processed were simpler, and most of the data was structured. But now, with the emerging of new channels and technologies, such as social networking, the Internet of Things, mobile computing, and online advertising, much semi-structured or unstructured data is produced, in the form of text, XML, emails, blogs, and instant messages—as just a few examples—resulting in a surge of new data types. Companies now need to integrate and analyze data from complex traditional and nontraditional sources of information, including the companies' internal and external data. With the explosive growth of sensors, smart devices, and social collaborative technologies, the types of data are uncountable, including text, micro blogs, sensor data, audio, video, click streams, log files, and so on.

Velocity: The velocity of data generation, processing, and analysis continues to accelerate. There are three reasons: the real-time nature of data creation, the demands from combining streaming data with business processes, and decision-making processes. The velocity of data processing needs to be high, and processing capacity shifts from batch processing to stream processing. There is a "one-second rule" in the industry referring to a standard for the processing of big data, which shows the capability of big data processing and the essential difference between it and traditional data mining.

Value: Because of the enlarging scale, big data's value density per unit of data is constantly reducing, however, the overall value of the data is

increasing. Big data is even compared to gold and oil, indicating big data contains unlimited commercial value. According to a prediction from IDC research reports, the big data technology and services market will rise from \$3.2 billion in 2010 to \$16.9 billion in 2015, will achieve an annual growth rate of 40%, and will be seven times the growth rate of the entire IT and communication industry. By processing big data and discovering its potential commercial value, enormous commercial profits can be made. In specific applications, big data processing technologies can provide technical and platform support for pillar industries of the nation by analyzing, processing, and mining data for enterprises; extracting important information and knowledge; and then transforming it into useful models and applying them to the processes of research, production, operations, and sales. Meanwhile, many countries are strongly advocating the development of the "smart city" in the context of urbanization and information integration, focusing on improving people's livelihoods, enhancing the competitiveness of enterprises, and promoting the sustainable development of cities. For developing into a "smart city," a city would need to utilize the Internet of Things, Cloud computing, and other information technology tools comprehensively; integrate the city's existing information bases; integrate advanced service concepts from urban operations; establish a widely deployed and deeply linked information network; comprehensively perceive many factors, such as resources, environment, infrastructures, and industries of the city; build a synergistic and shared urban information platform; process and utilize information intelligently, so as to provide intelligent response and control for city operation and resource allocation; provide the intelligent basis and methods for the decision making in social management and public services; and offer intelligent information resources and open information platforms to enterprises and individuals.

BIG DATA APPLICATIONS:

Big data technologies have wide and long list of their applications. It is used for Search Engine, Log Processing, Recommender System, Data Warehousing, Video and Image Analysis, Banking & Financial, Telecom, Retail, Manufacturing, Web & Social Media, Medicine, Healthcare, Science & Research and Social Life. We are discussing some of the eminent applications here

Politics

Big Data analytics help Mr. Barack Obama to win the US presidential election in 2012. His campaign was built of 100-strong analytics staff to shake dozens of terabyte scale data. They used a combination of the HP Vertica massively parallel processing analytical database and predictive models with R and Stata tools.

National Security

Babak Akhgar et. al authored a book on Application of Big Data for National Security. Authors relate the Big Data technologies to national security and crime detection and prevention. They present strategic approaches to deploy Big Data technologies for preventing terrorism and reducing crime.

Health Care and Medicine

Big Data technologies can be used for storing and processing medical records. Streaming data can be captured from sensors or machines attached to patients, stored in HDFS and analyzed quickly [1]. With Big Data tools and human genome mapping, there may be a commonplace for people to have their genes mapped as the part of their medical record. Genetic determinants that cause a disease will be easy to find, which help in the development of personalized medicine.

Science and Research

Science and research are now driven by technologies. BigData adds new possibilities to them. CERN, the European Organization for Nuclear Research have started the world's largest and most powerful particle accelerator, Large Hadron Collider (LHC). The experiment generated excessive amount of data. Data center at CERN has 65,000 processors, which analyzed 30 petabytes of data. Its computing powers of thousands of computers are distributed across 150 data centers worldwide.

Social Media Analysis

IBM provides a social media analytics, a powerful SaaS solution to discover hidden insights from millions of web sources. It is used by businesses to gain a better understanding of their customers, market and competition.

It captures consumer data from social media, predicts customer behavior and creates customized campaign.

CONCLUSIONS:

Big Data is the hot frontier of today's information technology development. The Internet of Things, the Internet, and the rapid development of mobile communication networks have spawned big data problems and have created problems of speed, structure, volume, cost, value, security privacy, and interoperability. Big data technology is constantly growing with the surge of data volume and processing requirements, and it is affecting our daily habits and lifestyles.

Big Data is not only concerned to data big in volume but also data with big velocity, big variety and big veracity. Big Data has introduced a new attitude in data processing and analysis and new opportunities to provide solutions of real world problems, which are considered infeasible as before.

REFERENCES

- [1] Zikopoulos PC, Eaton C, DeRoos D, Deutsch T, Lapis G. Understanding big data. New York, NY: McGraw-Hill; 2012.
- [2] Bell G, Hey T, Szalay A. Beyond the data deluge. *Science* 2009;323(5919):1297-8.
- [3] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: the next frontier for innovation, competition, and productivity. MacKinsey Global Institute; 2011.
- [4] Big Data Research and Development Initiative, http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.
- [5] Gupta R, Gupta H, Mohania M. Cloud computing and big data analytics: what is new from databases perspective? Big data analytics. Berlin, Heidelberg: Springer; 2012. p. 42-61.
- [6] Li GJ, Cheng XQ. Research status and scientific thinking of big data. *Bull Chin Acad Sci* 2012;27(6):647-57 (In Chinese).
- [7] Apache Hadoop, <http://hadoop.apache.org/>.
- [8] Khetrapal A, Ganesh V. HBase and hypertable for large scale distributed storage systems. Department of Computer Science, Purdue University; 2006.
- [10] Available from: <http://cassandra.apache.org/>.