ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2015; 1(12): 791-794
www.allresearchjournal.com
Received: 01-09-2015
Accepted: 02-10-2015

Prof. Navneet Golchha
Department of Computer Science Bhag Singh Khalsa College for Women Abohar, Punjab, India.

Big Data – The information revolution

Navneet Golchha

Abstract

The paper gives an overview of Big Data and the technologies used to build Big Data infrastructure and how Big Data boom can bring about information revolution and assist in development of people, businesses and nation. The paper explores the various possibilities in which Big Data can help in improving decision-making in various streams of development. The paper discusses the outcome that can be achieved by implementing Big Data technology in various sectors. This proposes ideas that can help in achieving our goal of a bright future. It highlights the possibilities that can be achieved in decision making process using Big Data & related technologies.

Keywords: Big Data, Hadoop, HDFS, NoSQL, Cloudera Manager, Map Reduce

Introduction

BIG Data as its name implies is a data that is very big i.e. huge volumes of data which is being generated very fast & is of a wide variety. The modern world is a digital world. Today the speed at which data is generated has increased many folds. This is due to increased use of internet, digital equipments, smart phones, RFID sensors, social networking sites, Satellite imagery, etc. Huge amount of digital data is produced every second. These data can be in the form of text, digital images, videos, social network posts, blogs, call logs, GPS navigation data, sensor generated data etc. We can say that Big Data is characterized by 3 V's i.e. large Volumes, Variety and Velocity. Huge volumes of data which is of a wide variety is produced at very fast speeds. Data we are talking about here is heterogeneous, structured as well as unstructured. This huge amount of data can be of great use if stored & analyzed properly. It can be of great help in decision making, business management, healthcare improvement, trend analysis, weather forecasting, etc. But for this, data has to be stored and analyzed. Due to mostly unstructured nature of data it becomes very difficult for the traditional data management, warehousing & analysis tools to process this data. The data being discussed here is not in gigabytes, terabytes but several petabytes. Infact the global digital content created will increase over the next five years – to about 35 zettabytes. This data can be utilized for the betterment of society, improving business, taking decisions etc. Big Data deals with acquiring data from various platforms and in various forms and storing under unified schemas for analysis. The decision support system can be computerized and digitized by the advancement in the field of Big Data. Lots of research is going on to employ Big Data for improving decision making in business, research, governance, health care. The paper focuses on how Big Data technology can be used in various fields using the digital data present in developing countries like India. Information Technology & Internet has become an integral part of every field like education, transportation, business, health care, etc. The use social networking sites & apps like Google+, Facebook, Twitter, Whatsapp has increased tremendously. Organizations are interested in capturing and analyzing this data because it can add significant value to the decision making process. Using Big Data a lot of improvement can be made in decision making and policy making for the developing countries. It can provide solutions and mechanism for the development of various fields like education, business, law making, weather forecasting, space research, health care etc.

What Is Big Data?

Big Data refers to data that is large enough with respect to Volume, Variety and Velocity [1]. Data is huge in terms of Volume- large quantity of data more than exabyte, Variety-

Correspondence
Prof. Navneet Golchha
Department of Computer Science Bhag Singh Khalsa College for Women Abohar, Punjab, India.

information in the form of text, images, videos, document, GPS data, tweets, blogs, sensor data etc, and Velocity – high speed at which these data are produced. Big Data is sized in peta-, exa-, and soon will be in zetta-bytes. Google receives around 40,000 search queries every second, 112,568 YouTube videos are viewed in 1 second, 3,103 Instagram photos are uploaded in 1 second and Twitter users send roughly 10,710 tweets per second [2]. Around 4883 status updates are done on Facebook every second [3] and more than 50,000 likes take place every second [4]. Additional to these mainly human-generated telecommunication flows, surveillance cameras, health sensors, and the “Internet of things” (including household appliances and cars) are adding a large chunk to ever increasing data streams. The use of social networking sites in developing countries like India is increasing tremendously. Big data refers to storing and capturing this large dataset and mining this data for decision making. The use of Big Data - for analyzing trends and decision making - will become the basis of competition and growth for individual firms, enhancing productivity and creating significant value for the world economy by reducing waste and increasing the quality of products and services. The volume of data generated, stored, and mined for insights, has become crucial to businesses, government, and consumers. Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on" [5]. Every day quintillions bytes of data are produced in the world. Amount of available digital data at the global level grew from 150 exabytes in 2005 to 1200 exabytes in 2012 [6]. It is projected to increase by 40% annually in the next few years [6] which is about 40 times the much-debated growth of the world’s population. Big Data cannot be stored on a single machine; you need parallel system architecture and distributed system architecture to store this data. Since data originates from different sites and in different places, it has to be stored in uniform manner. Then these data is used for analytics purpose as discussed in next section.

Big Data Analytics

Big Data analytics refers to mining knowledge from the information. It turns the unstructured data into actionable information using various machine learning algorithms. Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Analyzing these large data allows us to discover the correlation, facts and other important information that lies in this large data set, which is impossible to be determined by human. The data from various sources are collected; they are refined and stored under uniform schemas. This data is then analyzed into form of reports, graphs, spatial charts, pie diagrams, tables, etc. which can be used for various decisions making.

The primary goal of big data analytics is to help organizations make more informed & efficient decisions by enabling data scientists and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may not be harnessed by traditional data mining and business intelligence software. Big Data analytics refers to tools and methodologies that aim to transform massive quantities of raw data into metadata-for analytical purposes. These analytic tools are based on powerful algorithms that are able to detect patterns, trends, and correlations over various time horizons in the data, but also on advanced visualization techniques as sense making tool [7]. The data is collected from various sources. It is combination of structured, unstructured and semi structured data. The semi-structured and unstructured data may not fit well in traditional data warehouses based on RDBMS. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually. The variety of heterogeneous data is converted into uniform schema and then advanced analysis tools and mining algorithms are deployed for finding out the facts and trends.

Tools and Technologies

This section presents the technologies used to build a big data infrastructure. Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. A 2011 McKinsey report [8] suggests suitable technologies include A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, natural language processing, signal processing, simulation, time series analysis and visualisation. Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data mining, distributed file systems, distributed databases, cloud-based infrastructure (applications, storage and computing resources) and the Internet. Initially we need some storage capabilities which will store huge volumes of data. For that we can use open source tools like.

- Hadoop Distributed File System (HDFS) [9] - An Apache open source distributed file system, which runs on high-performance commodity hardware. Known for highly scalable storage and automatic data replication across three nodes for fault tolerance. Automatic data replication across three nodes eliminates need for backup. Write once, read many times.
- Cloudera Manager [10] - Cloudera Manager is an end-to-end management application for Cloudera’s Distribution of Apache Hadoop. It gives a cluster-wide, real-time view of nodes and services running; provides a single, central place to enact configuration changes across the cluster; and incorporates a full range of reporting and diagnostic tools to help optimize cluster performance and utilization.

Database Capabilities for implementing big data infrastructure can be incorporated by following tools:

- Oracle NoSQL [11]: Not Only SQL. Dynamic and flexible schema design. High performance key-value pair database. Key-value pair is an alternative to a pre-defined schema. Simple pattern queries and custom-developed solutions to access data such as Java APIs.
- Apache HBase [12]: an open source NoSQL database that provides random, real-time read/write access to those large datasets.
- Apache Cassandra [13]: Data model offers column indexes with the performance of log-structured updates, materialized views, and built-in caching
- Apache Hive [14]: Tools to enable easy data extract/transform/load (ETL) from files stored either directly in Apache HDFS or in other data storage systems such as Apache HBase

The big data processing on the sophisticated parallel and distributed system is done by following tools:

- Map Reduce [15] divides a problem into smaller sub-problems, Able to distribute data workloads across thousands of nodes.
- Apache Hadoop [9]: Highly scalable parallel batch processing. Highly customizable infrastructure. Writes multiple copies across cluster for fault tolerance

Big Data as Development Tool

Big Data has revolutionised data analysis & decision making process. It can be adopted in various fields. The advancement of big data can be very beneficial in various fields like education, health care, good government, etc. This section highlights some areas which can be benefited by Big Data Boom.

- Healthcare Sector: In health care historical data of all patients, various images form scanning devices like X-rays, ECG and MRI images, etc can be used to find out some symptomatic patterns, prescription trends etc. The outbreak of any epidemics can be traced using Google Trends data. For example, if the Google trends' data shows high number of search for a word/phrase 'Swine Flu' in a particular region then it can be predicted that there is an outbreak of swine flu in that region, so appropriate preventive measures can be taken before it spreads massively. Mining vast quantities of health-related online data can give us various information that can be useful in Medical Field. For this some applications can be created that capture the search patterns on various search engines. Examples of such applications are 'Google Flu Trends', 'Google Dengue Trends' and Influenza detection [16]. Computer scientists at John Hopkins University analysed over 1.6 million health-related tweets (out of over 2 billion) posted in the US between May 2009 and October 2010 using a sophisticated proprietary algorithm and found a 0.958 correlation between the flu rates, they modelled based on their data, and the official flu rate [17].
- Good Governance, Policy Making and Security: Based on the data collected from various sources like census, income tax records, social behaviour of people on social networking sites and apps, historical data etc. government can predict the needs and problems of the citizen. So government can make policies that are beneficial for the citizens. For example, an application can be designed which can analyse and give data about areas where per capita income is very low, such areas can be identified and proper schemes can be implemented for improving the conditions of folks. The govt. will have the detailed report of population and their distribution socially and economically which will help them to enact different laws and policy. Government can decide how much monetary funds need to be allocated for Food Security based on the statistics of the people in the country. Based on the large data captured from call logs, SMS logs, CCTV footages, internet search patterns etc. government agencies can track the behaviour of persons which can assist in finding suspicious behaviours, terrorist activities, criminal activities etc. Applications can be developed based on big data that will alarm such anti-social activities.
- Environment & Sustainability: Developing countries face major threat from nature i.e. Flood, drought, earthquakes or any other natural disaster. Any such natural calamity dents the economic status of the country. Big Data can prove fruitful in such cases to tackle the situation. Applications can be developed which can store data generated by satellite imagery, sensors & other digital equipments, along with the historical data & then analyse the big data to provide accurate and timely information and predictions regarding rain fall, humidity, seismic activity, snow fall etc. The big data analytics can alarm us in advance so that we can take some corrective and preventive actions to safeguard from the impact of the calamities. Applications can also be developed that can keep track of activities going on in an ecosystem and provide useful stats regarding endangered species, depleting water table, depleting resources etc. Big Data can be useful in finding new sources of natural resources like coal, gold, iron ore, petrochemicals etc.
- Business Development: Big Data Analytics can be used in promotion and sales of products according to personalised needs. Online shopping sites and apps keep track of their customers activities and capture data like the products purchased, viewed, brands shopped etc. and on the basis of analysis of that data, customers are presented with more personalized view of products while they are shopping and also sending them personalized mails, messages and promotional offers. This proves to be helpful to businesses as well as customers. Data from social networking sites, RFID codes, credit/debit card statements etc can be used to discover the spending pattern of customers. Big Data analytics can be used in decision support system of many companies to target new market or launch new product. Big Data also offers opportunities to improve conventional business processes, such as customer interaction through sales and service websites and call centre functions, factory automation and quality assurance, and many more

Future Research

Big Data is here to revolutionize the world and impact our lives in a positive way. Techniques and models can be developed that can integrate data from different online as well as offline sources. Cheap techniques can be searched out to store this massive data. Open source tools discussed above can be used for mining and managing these data. If systems, as discussed above, can be developed and deployed, they can be used by governments and other agencies for the betterment of society and country. The businesses can also benefit from big data systems and thus boost the economy of the country. A cost efficient decision support system application comprising of sophisticated hardware and open source tools can be developed in future which will help the society a great deal.

Conclusion

Big Data will definitely improve the decision support system and in result the decision making in every area. It will bring a revolution in the society. Application needs to be developed to implement the above mentioned changes. Big Data will prove to be beneficial in the development of developing countries like India, where decision support system and decision making process has to be up to the mark. Big Data technology based applications though are not as simple to build. As it involves data, there are privacy issues. Privacy of the individuals who own the data or are associated with the data is very critical factor. The academic, public sector and private sector collaboration is required also. Since big data has to be stored and implemented using sophisticated hardware which is costly, government should assist and fund the research work. We can hope for such initiatives by government which will go a long way in development of our country.

References

1. www.technologytransfer.eu/article/98/2012/1/What_Is_Big_Data_and_Why_Do_We_Need_It_.html
2. http://www.internetlivestats.com/one-second/
3. https://zephoria.com/top-15-valuable-facebook-statistics/
4. http://onesecond.designly.com/
5. Data, data everywhere. The Economist. 25 February, 2010
6. Helbing Dirk. Stefano Balietti. From Social Data Mining to Forecasting Socio-Economic Crises.
7. Bollier Davi. The Promise and Peril of Big Data. The Aspen Institute, 2010.
8. Manyika James, Chui Michael, Bughin Jaques, Brown Brad, Dobbs Richard, Roxburgh Charles, Byers Angela Hung Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May, 2011.
9. Michael Franklin, Alon Halevy, David Maier, From Databases to Dataspaces: A New Abstraction for Information Management ACM SIGMOD Record December, 2005.
10. White Paper. Using Cloudera to Improve data Processing Cloudera, Inc 1-888-789-1488 or 1-650-362-0488
11. White Paper. Oracle NoSQL Database an Oracle White Paper September, 2011.
12. Shoji Nishimura, Sudipto Das, Divyakant Agrawal, Amr El Abbadi. HBase: A Scalable Multi-dimensional Data Infrastructure for Location Aware Services.
13. White Paper. Introduction to Apache Cassandra" By datastax Corporation July, 2013.
14. W Jung Hyun Kim, Xilun Chen, Maria Luisa Sapino. Hive Open Research Network, Action research in the UK construction industry - the B-Hive Project. August, 2001.
15. Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplied Data Processing on Large Clusters Google Research Paper.
16. Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, Larry Brilliant. Detecting Influenza epidemics Using Search Engine Query Data. Nature 457.7232, 2008, 1012-1014.
17. Paul MJ, M Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. Rep. Center for Language and Speech Processing at Johns Hopkins University, 2011.