Big Data – insights, motivation and challenges

Neelam Singh, Neha Garg, Varsha Mittal

Abstract—Technological innovations, greater affordability of digital devices and an increased ability to collect a wide variety of data from almost every source at an unprecedented scale gave birth to an umbrella term- Big Data‘- for the explosion in the quantity, diversity and heterogeneity of high frequency digital data .Gartner reports that worldwide information volume is growing at a minimum rate of 59% annually. These data hold the potential to allow decision makers to track development progress, improve social protection, to formulate new policies and understand where existing policies and programs require adjustment and improvement.

In this paper we are discussing the various aspects and facets of Big Data , the research challenges that may be addressed to gain deeper insights using various analytical tools and methodologies to improve quality and acceptability of the decision making process.

Index Terms— Big Data, diversity, heterogeneity, relevance, reliability, authority control system, analytical tools.

1 Introduction

Rapid industrialization, greater affordability of devices , the explosion of mobile networks, cloud based infrastructure and new technologies has given rise to incomprehensibly large worlds of information, described as “Big Data”[1]. To take a competitive edge organizations are collecting and storing different types of information about their transactions. Millions of networked sensors are being embedded in the physical world in devices such as mobile phones, RFID, GPS systems, and industrial machines that sense, create, and communicate data in the age of the Internet of Things [2].

According to the 2011 IDC Digital Universe Study, more than a hundred Exabyte of data were created and stored in 2005, the world’s data is reportedly doubling every two years and global annual data creation is set to increase from 1.2 zettabytes in 2012 to 35 zettabytes in 2020 [3].

According to McKinsey [4], Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, store, manage and analyze.

According to O’Reilly, “Big data is data that exceeds the processing capacity of conventional database systems. The data is huge and massive, moves at a very high speed, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it”[5].

2 Motivation

Big Data market is constantly increasing each year. In March 2012, The White House announced a national “Big Data Initiative” that consisted of six Federal departments and agencies committing more than $200 million to big data research projects [6].

Global Pulse which is an innovative lab that is based on the big data mining is also using the Big data to improve the life in developing countries.

In today’s competitive & complex business world the various aspects of business are intermingled. Change in one aspect has direct or indirect effect on the other aspect. Within an organization, this complexity makes it difficult for business leaders to rely solely on experience (or intuition) to make decisions. They need to rely on data - structured, unstructured or semi-structured - to back up their decisions. Existing tools don’t lend themselves to sophisticated data analysis at the scale the user requires. Tools like SAS, R, and Matlab support the decisive analysis but they are not designed for the massive datasets & neither DBMS nor Map Reduce can handle the data that are arrived at high rates. To bridge this gap the “Big Data” came into the scene. Big Data has given the organization a new way to analyze and visualize their data effectively. For example:

Business: Customer Feedback, trends etc.

Health: Health care organizations are leveraging big data technology to capture all the information about a patient to get more complete view for insight into care coordination, health management & outcome. Use of big data helps to build a sustainable healthcare system & increase the access to healthcare.

Energy & utility: Big data can also be the key to actually deploying condition based maintenance program and improve forecasting and scheduling of assets.

3 Big Data Taxonomies

Wherever The term Big Data appeared for the first time in 1998 in a Silicon Graphics (SGI) deck by John Mashey with the title “Big Data and the Next Wave of InfraStress7]. The first academic paper with the words ‘Big Data’ in the title appeared in 2000 in a paper by Diebold [8].

- *Neelam Singh is currently pursuing masters of technology degree program in information technology engineering in Graphic Era University, Dehradun,India, PH-091-9720105097. E-mail:neelamjain.jain@gmail.com*
- *Neha Garg is currently assistant professor in computer science and information technology engineering branch in Graphic Era University, Dehradun, India, PH-091-9897738782. E-mail: nehagarg.february@gmail.com*
- *Varsha Mittal is currently assistant professor in computer application branch in Graphic Era University, Dehradun, India, PH-091-9634435387. E-mail:var.addi@gmail.com*

With the advent of Big Data various terms also originated supporting the retrieval, storage, analysis and computations of this massive data.

Big Data Science: Big data science is the study of techniques covering the acquisition, conditioning, and evaluation of big data. These techniques are a synthesis of both information technology and mathematical approaches.

Big Data Frameworks: – Big data frameworks are software libraries along with their associated algorithms that enable distributed processing and analysis of big data problems across clusters of compute units (e.g., servers, CPUs, or GPUs).

Big Data Infrastructure – Big data infrastructure is an instantiation of one or more big data frameworks that includes management interfaces, actual servers (physical or virtual), storage facilities, networking, and possibly back-up systems. Big data infrastructure can be instantiated to solve specific big data problems or to serve as a general purpose analysis and processing engine.

Big Data is not only includes massive size but also data variety, velocity, veracity and value.

Volume- It refers to the scale of the data and processing needs. It calls for scalable storage, and a distributed approach to querying driven by increased data sources and affordability

Velocity- Improved through-put, connectivity and enhanced computing speed of digital devices has not only fastened the production of data but also the retrieval and processing of the data.

Veracity-It refers to the quality and provenance of the information in the face of data uncertainty from many different places. [9]

Variety- Source data has become diverse and complex because it includes not only structured traditional relational data, but also quasi-structured, semi-structured and unstructured data.

Value- The economic value of different data varies significantly. The challenge is identifying what is valuable and then transforming and extracting that data for analysis.

4 Research Trends And Challenges In Handling Big Data

4.1 Research Trends

Major research trends in Big Data can be categorized as follows:

- Storage , Search and Retrieval of Big Data
- Analytics on Big Data
- Computations on Big Data

4.1.1 Storage, Search and Retrieval of Big Data

Storage is very complex and not only does it require managing capacity and finding out the best collection and retrieval methods, it also means to synchronize both the IT and the business teams and paying attention to complex security and privacy issues.

4.1.2 Analytics on Big Data

Big Data analytics comprises of tools, algorithms and architecture that analyze and transform large and massive volumes of data [10].

Big data analytics is a technology-enabled strategy for enabling organization to have a competitive edge over others by analyzing market and customer trends. Analytics on rela-time data, online transactional data gives deeper insights of the trends to make timely and accurate decisions.

4.1.3 Computations on Big Data

Computing is concerned with the processing, transforming, handling and storage of information.

Systems such as Map Reduce, Hadoop have made writing and executing ad hoc big-data analysis and computation easy.

As search engines have transformed information access, other forms of big-data computing can and will transform the activities like medical and scientific research, defense task etc.

4.2 Emerging technologies for Big Data

4.2.1 Column-oriented databases

Column-oriented databases store data in columns, instead of rows, allowing for large data compression and very fast query times.

4.2.2 Schema-less databases, or NoSQL databases

There are several database types that fit into this category, such as key-value stores and document stores, which stores and retrieve large volumes of unstructured, semi-structured, or even structured data.

4.2.3 MapReduce

This is a programming paradigm that allows for massive job execution scalability against thousands of servers or server clusters. Any MapReduce implementation consists of two tasks:

- The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples;
- The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name).

4.2.4 Hadoop[19]

Hadoop is the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data.

4.2.5 Hive

Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster

4.2.6 Storage Technologies

Efficient and effective storage techniques are required for Storing huge volumes of data. The main focus is on data compression and storage virtualization.

4.2.7 Hbase [24]

It is a Scalable Distributed Database which uses HDFS for storage. It supports Structured Data and is Column – Oriented in nature.

4.2.8 Chukwa [25]

It adds the needed semantics for log collection and analysis to monitor Large Distributed System. Chukwa uses an end-to-end delivery model that leverages local on-disk log files which can be easily integrated with legacy systems.

5 CHALLENGES IN HANDLING BIG DATA

5.1 Heterogeneous Data Source

Data required for analytical and computational purpose is strongly heterogeneous which possess typical integration problem (both data and schema integration issues and it also leads to the creation of new tools and architectures for analytics.

5.2 Unstructured Nature of Data Sources

A big challenge to handle Big Data is the transformation of unstructured data into a suitable and structured format in order to design meaningful analytics.

5.3 High Scalability

Data is scaling at an unprecedented rate and it is a challenging issue as data volume is increasing faster than compute resources, and CPU speeds are static.

5.4 Timeliness

With size comes the issue of speed. With increasing volume of data the time to analyze the data will also increase. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data faster.

5.5 Privacy

The privacy of data is another major and important concern in the context of Big Data. There is a great public fear regarding the inappropriate use of private data, particularly through relating of data from multiple sources. Handling privacy is effectively both a technical and a sociological problem, which must be realized to take advantage of Big Data.

5.6 Data Integration

Big Data Integration is multidimensional and multidisciplinary and requires multi-technology method which poses a big challenge.

5.7 Finding and relating Anomalies in Human Ecosystems

A challenge when attempting to measure or detect anomalies in human ecosystems is the characterization of (ab) normality.

6 CONCLUSION

We are in an era of Big Data. Proper and effective analysis of large volumes of data will lead to faster advances in many scientific disciplines and improving the profitability and success of many enterprises.

While the potential benefits of Big Data are real and significant, and some initial successes have been achieved in some of the projects, there remain many technical challenges that must be addressed to fully realize the hidden potential of Big data. The large size of the data is a major challenge however, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity [Gar2011], and that companies should not focus on just the first of these.

The challenges include issue of large volume, but also heterogeneity, undefined structure, error-handling, privacy, timeliness, security provenance, integration and visualization. These technical challenges are found in large variety of application domains, and therefore impose a huge cost. Furthermore, these challenges will require transformative solutions, and will require a wide range of tools, methodologies and applications to deal with. In order to attain the promised benefits of Big Data these things has to be taken under strong consideration so that full potential can be derived to gain a competitive edge.

References

[1] The Economist. A special report – Managing Information. Data, data everywhere. February 25, 2010. http://www.economist.com/node/15557443

[2] McKinsey Quarterly. March 2010. Article. The Internet of Things. http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_things

[3] John Gantz and David Reinsel.. Extracting Value from Chaos. Sponsored by EMC Corporation. IDC iView. June 2011. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[4] James Manyika, et al. Big data: The next frontier for innovation, competition, and productivity. May 2011. Report. McKinsey Global Institute. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

[5] Edd Dumbill. What is big data?. January, 2012. http://strata.oreilly.com/2012/01/what-is-big-data.html

[6] Oscar Renalias. Unlocking Value in (Big) Data. 2012. http://www.slideshare.net/oscarrenalias/unlocking-value-in-your-big-data

[7] Jackie Fenn, Emeritus Hung LeHong. Emerging Technologies: What's Hot for 2012 to 2013. September 19, 2012. http://public.brighttalk.com/resource/core/3297/september_19_hype_cycle_2012-fen_-lehong_6009.pdf

[8] F. Diebold. Big Data - Dynamic Factor Models for Macroeconomic Measurement and Forecasting. Discussion Read to the Eighth World Congress of the Econometric Society, 2000.

[9] T.Morgan. IBM Global Technology Outlook 2012. In Technology Innovation Exchange, IBM Warwick, 2012. http://anlenterprises.com/2012/10/30/ibms-4th-v-for-big-data-veracity/

[10] Michael Schroeck, Rebecca Shockley, Dr. Janet Smart, Professor Dolores Romero-Morales and Professor Peter Tufano. Analytics: The real-world use of big data. 2012 Executive Report. IBM Institute for Business Value. http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf

[11] King, Gary N. and Eleanor Powell. How Not to Lie Without Statistics. Working Paper. Harvard University. 22 Aug. 2008. http://gking.harvard.edu/gking/files/nolie.pdf

[12] Dan Vesset, et al. Worldwide Big Data Technology and Services 2012-2015 Forecast.
Market Analysis. IDC. http://www.idc.com/getdoc.jsp?containerId=233485

[13] Cisco. The Internet of Things: How the Next Evolution of the Internet is Changing Everything. http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

[14] IDC. IDC's Worldwide Big Data Taxonomy. 2011. http://www.idc.com/getdoc.jsp?containerId=231099

[15] Thusoo, et al. R. Hive – A Petabyte Scale Data Warehouse Using Hadoop. 2010. Proceedings of ICDE, 2010.

[16] Alan Gates. Programming Pig. Dataflow Scripting with Hadoop.O'Reilly.2011. pp 1-2.

[17] Greg Emmerich. Demystifying Big Data: Skytree Brings Machine Learning to the Masses .2013. UW Madison M.S. Biotechnology Program. Advanced Biotechnology: Global Perspectives. Thesis Paper. April 16th, 2013.

http://gregemmerich.wordpress.com/2013/04/17/demystifying-big-data/

[18] Ariel Rabkin, Randy H. Katz.2010. Chukwa: A system for reliable large-scale log Collection. University of California at Berkeley, March 5, 2010. http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-25.pdf