International Journal of Advanced Computer Technology (IJACT)
ISSN:2319-7900

AN ANALYSIS OF BIG DATA DISCOVERY AND COLLABORATION

Dr. M.Kumarasamy
Professor in Computer Science,
Sri Venkateswara College of Engineering & Technology, Chennai

Abstract

The term Big Data continues to be a current hot topic around the world, but most of the organizations and institutions still struggle to understand what it really means. The main purpose of this paper is to describe the terms of big data, impact of big data in organizations, Web data, machine data, human generated data, data discovery, data collaboration, Hadoop technologies for big data and zelig model for big data analytics. Big data is a data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it. Big data is comprised of datasets too large to be handled by traditional database systems. To remain competitive business executives need to adopt the new technologies and techniques emerging due to big data. The main purpose of big data is to improve decisions and competitiveness for organizations, institutions and the public administrations, which will create a significant growth of the world economy. In the past, the collection and storage of information was the primary issue. Currently, there are massive amounts of data both structured and unstructured, that need to be analyzed in an iterative, as well as in a time sensitive manner. In response to this need, data analytical tools and services have emerged as a means to solve this problem.

Keywords : *Big Data, Analytics, Structured Data, Web Data, Social Media*

Introduction

Nowadays, many organizations and institutions are collecting, storing, and analyzing massive amounts of data. This data is commonly referred to as "big data" because of its volume, the velocity with which it arrives, and the variety of forms it takes. Big data is creating a new generation of decision support data management. Businesses are recognizing the potential value of this data and are putting the technologies, people, and processes in place to capitalize on the opportunities. A key to deriving value from big data is the use of analytics. Collecting and storing big data creates little value; it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value. Many analytic techniques, such as regression analysis, simulation, and machine learning, have been available for many years. Special attention is given to the technologies, platforms, and approaches for storing and analyzing big data. Security and Privacy concerns about the use of big data are also explored. Big data is changing existing jobs and creating new ones. For example, market researchers must now be skilled in social media analytics. Data management professionals must be able to store massive amounts of data of any structure. Fig.1 represents the architecture of big data.

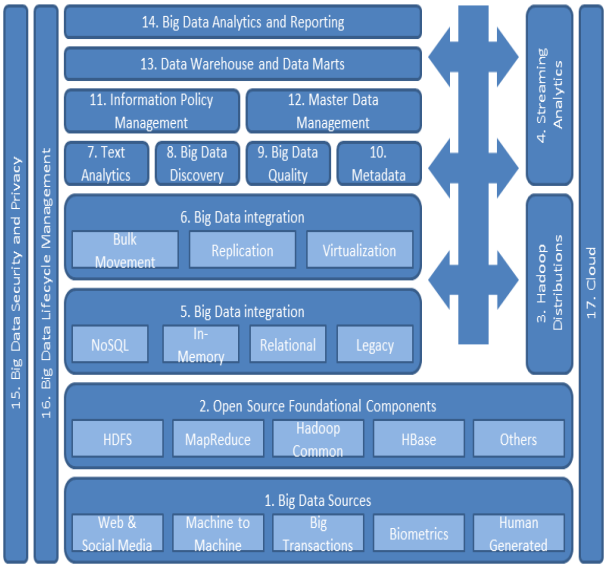

Fig.1 – Big Data Architecture

Technical Drivers of Big Data

Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety;

requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

Volume: covers the size of the data needed for management. There is more data than ever before, its size keep on growing exponentially: 90% of all the data available today were created in the last two years. A short time ago, we were talking about gigabytes, we are talking now relatively about terabytes, petabytes, exabytes and even zettabytes.

Velocity: describes the speed with which the data is generated and processed. We are focused on getting knowledge from the data arriving as streams in real time. More we focus on real time; more we are in big data problem. Gradually, the immediate treatment of data would be the key element of a model big data.

Variety*:* refers to the difference in the type of data we have collected. The data analyzed is not anymore structured as the anterior data, but could be text, pictures, multimedia content, digital traces, sensor data, etc. It is about the ability to give an additional value to the internal traditional data by combining it with a big variety of other external sources of data. Fig.2 represents the technical drivers of big data.

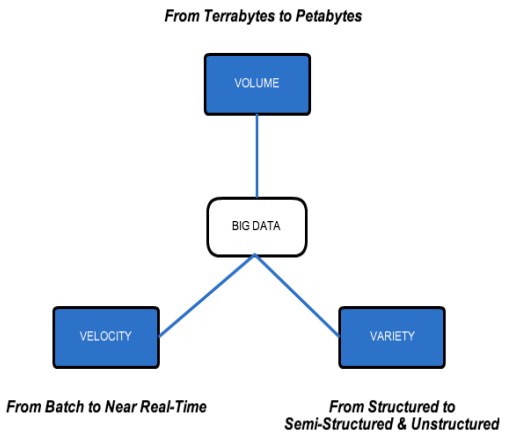

Fig.2 Technical Drivers of Big Data

Web Data & Social Media

The web is a rich, but also very diverse source of data for analytics. For one, there are web sources to directly extract content - knowledge or public opinion - from, which are initially intended for a human audience. These human-readable sources include crawling of web pages, online articles and blogs. The main part of these sources is typically unstructured, including text, videos and images. However, most of these sources have some structure, they are e.g. related to each other through hyperlinks or provide categorization through tag clouds. Next, there is web content and knowledge structured to provide machine-readability. It is intended to enable applications to access the data, understand the data due to semantics, allow them to integrate data from different sources, set them into context and infer new knowledge. Such sources are machine-readable metadata integrated into web pages, initiatives as the linked open data project using data formats from the semantic web standard, but also publicly available web services. This type of data is often graph-shaped and therefore semi-structured.

Machine-to-machine data

Machine to machine communication describes systems communicating with technical devices that are connected via some network. The devices are used to measure a physical phenomenon like movement or temperature and to capture events within this phenomenon. Via the network the devices communicate with an application that makes sense of the measurements and captured events and extracts information from them. One prominent example of machine to machine communication is the idea of the 'internet of things'.

Human-generated data

Human-generated data refers to all data created by humans. It mentions emails, notes, voice recording, paper documents and surveys. This data is mostly unstructured. It is also apparent, that there is a strong overlap with two of the other categories, namely big transaction data and web data. Big transaction data that is categorized as such because it is accompanied by textual data, e.g. call centre agents' notes, have an obvious overlap. The same goes for some web content, e.g. blog entries and social media posts. This shows, that the categorization is not mutally exclusive, but data can be categorized in more than one category.

Big Collaboration and Discovery

ISSN:2319-7900

Big Science generates Big Data that demands Big Collaboration. Algorithms and AI enable the discovery of many of the underlying rules and relationships in biological data. Analytics enable researchers to intersect data sets that are too small to fully understand cancer subtypes and that may have measured expression set on different platforms. Notwithstanding, collaborative expertise—human intelligence and intuition—are required for meaning and interpretation. Big Collaboration extends beyond the borders of traditional research collaboration to include on-demand communication and sharing of protocols, electronic resources, data, and findings among the spectrum of stakeholders in a private or public virtual frontier. At each stage, learning and leveraging information and findings on what hasn't worked is as important as knowing what has worked. Big Collaboration embraces the "from bench to bedside" philosophy and recognizes time, manpower and R&D dollars are scarce resources. Fig.3 represents the data discovery through human-machine collaboration.

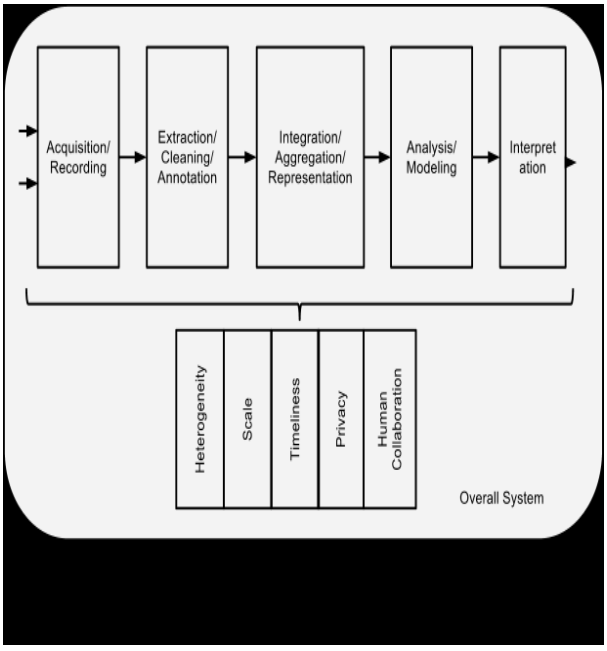

Fig.3 Discovery through Human-Machine Collaboration

Discovery is the desired outcome of an investigative analytical process. Discovery in big data requires the collaboration of man and machine, where the guiding intellect — the ability to posit and infer — is human. In time, artificial intelligence may be able to make suppositions and draw conclusions but, for now, humans still have the advantage. The process of discovery is iterative. The analyst must be able to test a hypothesis against all available data by posing a question that the technology answers in depth and then renders visually, shortening the time between results. This requires the ability to ask questions that were not anticipated by those who built the knowledge base, referred to as ad-hoc queries in the database world. In discovery, you don't know the next question until you get the first answer, and each iteration may require additional datasets for analysis

Zelig Model for Big Data Analytics

The vast promise and broad range of big data applications have steadily begun to be tapped by new tools, algorithms, learning techniques, and statistical methods. The proliferation of tools and methods that have been developed for specific tasks and focused solutions are myriad. But largely, these pioneering tools stand in towering isolation of one another. Often initiated as solutions to specific big data applications, the current open source methods available may each expect different data formats and use different call structures or notations, not to mention different languages. We think the Zelig architecture devised for R can also solve this similar problem for big data science. We propose that a fundamental need in big data science is the proper construction of an *abstraction layer* that allows users to see quantitative problems through their commonality and similar metaphors and attacks, while abstracting away the implementation of any algorithm in any given language on any particular storage device and computational setting. This framework would create an interoperable architecture for big data statistical and machine learning methods. We propose that the architecture developed for Zelig for R can be mirrored in a language agnostic fashion for tools in Scala, Java, Python, and other languages that can scale much more efficiently than R, and can be used to bridge together the growing number of statistics and analytics tools that have been written for analysis of big data on distributed systems (such as Apache Mahout, Weka, MALLET). This will provide easier access for applied researchers, and, going forward, writers of new tools will have the ability to make them more generally available. Critically, such a framework must:

—Allow users to use one call structure, and have access to the range of big data statistical and learning methods written across many different languages.

ISSN:2319-7900

Rather than any user needing to learn new commands, languages, and data structures every time they try a new exploratory model, users will be able to seamlessly explore the set of big data tools applicable to their problems, increasing exploration, code reuse, and discovery.

—Allow any developer of a new tool to easily bridge their method into this architecture.

—Provide common utilities for learning and statistics in big data analytics that can be easily interoperable and available to every model.

There is a large body of general purpose techniques in statistical models and machine learning that are of broad applicability to most any model, but may only be available in a particular open source tool if one of the original authors needed that technique in their own research application. It should not be required of every method author to reinvent each of these wheels, nor should users of tools be constrained to only those techniques of use by the original author of their tool, and our architecture will make all these utilities interoperable across packages.

—Enable interpretation of analytical models in shared and relevant quantities of interest.

Big Data using Hadoop / Map Reduce Technology

Of all the platforms and approaches to storing and analyzing big data, none is receiving more attention than Hadoop/MapReduce. Its origins trace back to the early 2000s, when companies such as Google, Yahoo!, and Facebook needed the ability to store and analyze massive amounts of data from the Internet. Because no commercial solutions were available, these and other companies had to develop their own. Important to the development of Hadoop/MapReduce were Doug Cutting and Mike Cafarella who were working on an open-source Web search engine project called Nutch when Google published papers on the Google File System and MapReduce. Impressed with Google's work, Cutting and Cafarella incorporated the concepts into Nutch. Wanting greater opportunities to further his work, Cutting went to work for Yahoo!, which had its own big data projects under way. With Yahoo!'s support, Cutting created Hadoop (named after Cutting's son's stuffed elephant) as an open-source Apache Software Foundation project. *Apache Hadoop* is a software framework for processing large amounts of data across potentially massively parallel clusters of servers. To illustrate, Yahoo has over 42,000 servers in its Hadoop installation. The key component of Hadoop is the Hadoop Distributed File System (HDFS), which manages the data spread across the various servers. It is because of HDFS that so many servers can be managed in parallel. HDFS is file based and does not need a data model to store and process data. It can store data of any structure, but is not a RDBMS. HDFS can manage the storage and access of any type of data as long as the data can be put in a file and copied into HDFS. The Hadoop infrastructure typically runs MapReduce programs (using a programming or scripting language such as Java, Python, C, R, or Perl) in parallel. MapReduce takes large datasets, extracts and transforms useful data, distributes the data to the various servers where processing occurs, and assembles the results into a smaller, easier to analyze file. It does not perform analytics per se; rather, it provides the framework that controls the programs that perform the analytics. Currently, jobs can only be run in batch, which limits the use of Hadoop/MapReduce for near real-time applications. Although Hadoop and MapReduce are discussed and typically used together, they can be used separately. That is, Hadoop can be used without MapReduce and vice versa. This is a simple processing task that could also be done with SQL and a RDBMS, but provides a good example of Hadoop/MapReduce processing. The first step is to *split* the records and distribute them across the clusters of servers (there are only three in this simple example). These splits are then processed by multiple *map* programs (e.g., Java, R) running on the servers. The objective in this example is to group the data by a *split* based on the words. The MapReduce system then merges the *shuffle*/*sort* results for input to the *reduce* program, which then summarizes the number of times each word occurs. This output can then be input to a data warehouse where it may be combined with other data for analysis or accessed directly by various BI tools.

Impacts of Big Data in Organisations

Data are generated in a growing number of ways. Use of traditional transactional databases has been supplemented by multimedia content, social media, and myriad types of sensors. Advances in information technology allow users to capture, communicate, aggregate, store and analyze enormous pools of data, known as "big data". However, the new data collection methodologies pose a dilemma for businesses that have depended upon database technology to store and process data. "Big data" derives its name from the fact that the datasets are large enough that typical database systems are unable to capture, save, and analyze these datasets. The actual size of big data varies by business sector, software tools available in the sector, and

ISSN:2319-7900

average dataset sizes within the sector. Best estimates of size range from a few dozen terabytes to many petabytes. In order to benefit from big data, new storage technologies and analysis methods need to be adopted. Business executives must determine the new technologies and methodologies best suited to their information needs. Business executives ignoring the growing field of big data will eventually become non-competitive.

Challenges and Opportunities with Big Data

In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on painstakingly constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences. Scientific research has been revolutionized by Big Data. In the biological sciences, there is now a well-established tradition of depositing scientific data into a public repository, and also of creating public databases for use by other scientists. In fact, there is an entire discipline of bioinformatics that is largely devoted to the curation and analysis of such data. As technology advances, particularly with the advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially. Big Data has the potential to revolutionize not just research, but also education. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance. It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality , by making care more preventive and personalized and basing it on more extensive continuous monitoring. McKinsey estimates a savings of 300 billion dollars every year in the US alone. In a similar vein, there have been persuasive cases made for the value of Big Data for urban planning, intelligent transportation, environmental modeling, energy saving, smart materials, computational social sciences, financial systemic risk analysis, homeland security, computer security and so on. While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved, there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. 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, and that companies should not focus on just the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability.

Conclusion

In this paper, we discussed a brief overview of the Big data topic, including the main concerns and the main challenges for the future. Big Data will allow us to extract insights that no one has extracted before. However, it is still under development and current approaches and tools are very limited to deal with the new real Big Data requirements. Further work will be focused on this corner. This paper also presents the introduction of big data, web data, human data, machine data, data discovery and collaboration. The use of recent advances in different fields of large-scale data analysis is promoted in the heuristics framework, focusing on applications in medicine, biology and technology. Organizations want more business value from big data, and analytics is an important route to value.

Many of the most beneficial applications of big data involve discovery. However, as dataset sizes grow, a collaborative human-machine approach to discovery is required to enable humans to cope with the size and complexity of the datasets. Discovery in big data typically involves fusing information from many different sources and then testing hypotheses, expressed as complex queries, against the entire dataset. Big data can be viewed as the latest generation in the evolution of decision support data management. The need for data to support computer-based decision making has existed at least since the early 1970s with *DSS*. Big data is the

ISSN:2319-7900

fourth generation decision support data management. The ability to capture, store, and analyze high-volume, high-velocity, and high-variety data is allowing decisions to be supported in new ways. It is also creating new data management challenges. Technical challenges are common across a large variety of application domains. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

References

[1] Big Data Now: 2012 Edition. Techreport, O'Reilly Media, Inc., 2012.

[2] Bollier D. (2010) "The Promise and Peril of Big Data." Washington, D.C.: The Aspen Institute.

[3] Cooper, B.L., H.J. Watson, B.H. Wixom, and D.L. Goodhue (2000) "Data Warehousing Supports Corporate Strategy at First American Corporation," *MIS Quarterly*, (24)4, pp. 547-567.

[4] Davenport, T.H. and J.G. Harris (2007) *Competing on Analytics: The New Science of Winning*, Boston: Harvard Business School Press.

[5] Gautham Vemuganti, "challenges and opportunities", Infosys Labs Briefings, VOL 11, NO 1, 2013.

[6] Hadoop 1.2.1 Documentation. Online documentation, Apache Software Foundation, 2013. URL http://hadoop.apache.org/docs/r1.2.1/index.html. Accessed: 27-03-2013.

[7] King, Gary. 2007. An introduction to the Dataverse network as an infrastructure for data sharing. *Sociological Methods and Research* 36:173–99.

[8] Lakshman, Avinash, and Prashant Malik. 2010. Cassandra: A decentralized structured storage system. *ACM SIGOPS Operating Systems Review* 44 (2): 35–40.

[9] Latamore B. (2011) "Big data is where problem becomes opportunity says EMC's Chuck Hollis." Available at: http://wikibon.org/wiki/v/Big_Data_is_Where_Problem_Becomes_Opportunity_Say s_EM C's_Chuck_Hollis Accessed on October 5, 2011.

[10] Nauman Sheikh, "Big Data, Hadoop, and Cloud Computing, Implementing Analytics", Morgan Kaufmann, 2013.

[11] Neubauer, J., Vesely, V. (2011). Change point detection by sparse parameter estimation. *Informatica*, 22(1), 149–164.

[12] Nunkesser, R, Morell, O. (2010). An evolutionary algorithms for robust regression. *Computational Statisticsand Data Analysis*, 54(12), 3242–3248.

[13] Pederson, S (2012), Exploiting Big Data from the DeepWeb: The new frontier for creating intelligence. BrightPlanet, Sioux Falls, South Dakota.White paper available (http://www.brightplanet.com/2012/07/creating-intelligence-from-big-data-whitepaper/)

[14] Sakalauskas, L., Zavadskas, E. (2009). Optimization and intelligent decisions. *Technological and EconomicDevelopment of Economy*, 15(2), 189–196.

[15] Saul, L.K, Roweis, S.T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4, 1 9–155.

[16] The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. *National Public Radio*, Nov. 30, 2011. http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data.

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

[18] Thibaud Chardonnens, "Big Data analytics on high velocity streams: specific use cases with Storm", Software Engineering Group, Department of Informatics, University of Fribourg, Switzerland, 2013.

[19] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.

[20] Watson, H. J. (2009a) "Tutorial: Business Intelligence – Past, Present, and Future," *Communications of the Association for Information Systems* (25)39. Available at: http://aisel.aisnet.org/cais/vol25/iss1/39