COMPARABLE SURVEY OF BIG DATA TECHNOLOGIES

Ashif Ali
Research Scholar (Ph.D), College of Engineering
Department of Computer Science & Engineering
Dr. A.P.J Abdul Kalam University, Indore (M.P), India-452014

Abstract - The term Big data is a great buzzword in IT industries these days. It refers to large volume having both structured and unstructured format. This huge volume of data is generated from multidirections through various channels and usually of great concern to us. So we can analyze the insights of this data and can use it for the betterment of decision making and for profitable business strategies. Volume of data coming through direction is so huge that it cannot be processed using traditional procedures and technologies. Managing such volume requires the standard framework like Hadoop. It is also an open source which is attracting the mass audience for its management and popularity. Along with Hadoop big technologies like Pig, Hive and lot of other products also came into picture. Upcoming technologies like Spark, NoSQL databases and Google's Map reduce are also hitting the tech giants to solve complex problems. There are lot of proprietary and open source technologies in market which could be used to manage the data handling problems in big data environment. In this paper we will discuss the few technologies and the level application in small scale, mid scale and big scale industries.
Keywords: Big data, Hive, Pig, Spark, Framework, Technologies.

I. INTRODUCTION

Big Data technologies are transforming the way data is used to be analyzed. One reason is the massive amount of data that is being generated from different sources such as social networks, sensors; search engines, banks, telecommunication and web, handling this massive amount of data take us in the era of Big Data.

According to the YouTube statistics 100 hours of video are being uploaded to YouTube [1] servers in every minute. Facebook is dealing with more than 500 terabytes of data daily [2], companies such as a Google and Yahoo are recording search engine results for analyzing the searching trends; crawling different web sources to analyze for any important events; gathering marketing data for analyzing the current and future trends which all results the generation of large data sets also referred to as Big Data.

Data is everywhere, from social sciences to physical science, business and commercial world, for example, digitizing the past fifty year's newspapers will results the massive amount of data, in astronomy storing billions of astronomical objects, in biology storing genes, proteins and small molecules results in massive amounts of data. In business world such as handling millions of call data records in telecommunication, handling millions of transactions in banking and handling millions of transactions for multinational grocery store results in large data sets. Analyzing these large datasets and getting out meaningful information from it is a challenging in itself.

1.1 Big Data

Big Data can be described in 5 V's such as variety, volume and velocity [3].

1.1.1 Variety

Data has different variations, for example semi-structured or unstructured, such as data, generated from web sites, social networks, emails, sensors and web logs is unstructured. Structured data refers to as data generated in result of conversion from call data record to tabular format in order to calculate the monetary value out of it or banks transactions data or data generated from the airline ticketing system are different varieties in the data.

1.1.2 Volume

Volume refers to the amount of data or size of the data set. Nowadays figures are in Tera and Peta bytes. For instance

Airbus can generate half of terabytes of data in one flight [4].

1.1.3 Velocity

Velocity refers to the speed of data generation which is very fast nowadays. For example weather sensors are kept on generating data as new updates comes, Twitter is generating data at 9100 tweets per second and on Facebook users is sending 3 million messages to each other every 20 minutes [5].

There are different technologies to analyze the Big Data. Hadoop [6] is one of the most popular among them. There are many others, such as Cloudera and Cassandra and technologies for warehousing such as Hive and HBase [6] which can be used in conjunction with Hadoop to ease the analysis and provides the abstraction over Hadoop platform.

1.1.4 Veracity

Veracity is not just about data quality, it is about data understand ability. Data governance initiatives have little sex appeal and most data stewards already have a primary job role. Too many users are waiting for data nirvana—perfectly clean data. Veracity is all about making sure the data is accurate, which requires processes to keep the bad data from accumulating in your systems. [6]

1.1.5 Value

Value starts and ends with the business use case. The business must define the analytic application of the data and its potential associated value to the business. Use cases are important both to define initial —"Big Data" pilot justification and to build a road map for transformation.

The most important element of the big data we call the Sage Blue Book is value. Value that includes a large volume and variety of data that is easy to access and delivers quality analytics that enables informed decisions. Providing a fair market valuation on used technology - one piece or an entire portfolio at a time. This validates the investment return on investment (ROI) and promotes future funding [7].

Big data can deliver value in almost any area of business or society:

a) It helps companies to better understand and serve customers: Examples include the recommendations made by Amazon or Netflix.
b) It allows companies to optimize their processes: Uber is able to predict demand, dynamically price journeys and send the closest driver to the customers.
c) It improves our health care: Government agencies can now predict flu outbreaks and track them in real time and pharmaceutical companies.
d) It helps us to improve security: Government and law enforcement agencies use big data to foil terrorist attacks and detect cyber crime.

It allows sport stars to boost their performance: Sensors in balls, cameras on the pitch and GPS trackers on their clothes allow athletes to analyze and improve upon what they do.

2 RELATED WORKS

Big Data is present in almost every domain today. Lots of research has been done in this vast field. In [1] authors have proposed that sensing technologies, cloud computing, internet of things and big data analytics systems as emerging technologies which has made it possible to achieve impressive progress in the field of computational field and storage which play a major role in efficiency and effectiveness of the healthcare systems. Cloud computing [2] is nowhere left behind in big data as well. System consisting of monitoring agents, cloud infrastructure, and operation centre by using Hadoop Map Reduce and Spark to enhance the processing by splitting and processing data streams concurrently. But with every positive thing comes the challenges and issues related as well. There is a model [3] proposed to tackle challenges like data complexity, computational complexity, and system complexity and also presented suggestions for carrying out Big Data projects. The emergence of big data, considered both the opportunities and the

ethical challenges for the market research [7] as proposed in this research.

There are various applications that are developed in this domain as well. Big data is suitable for every trend these days. Therefore, applications like [4] an electricity generation forecasting system which can predict the required power generation close to 99% of the actual usage by using Big Data Analytics was proposed. Also a Location-Aware Analytics.

System [5] using effective spatio-textual indexes and incremental algorithms that has been deployed and widely accepted. Another application is where authors proposed a model [17] for monitoring and analyzing Internet Traffic which is in the form of Big Data. They provided analysis and forecasts, including traffic management and network upgrade to enhance the quality that can be used to promote investments as well.

Data privacy is another big issue these days. For this author have proposed [6] six data management research challenges for Big Data and Cloud such as Data Privacy, Approximate Results, Data exploration to enable Deep Analytics, Enterprise data enrichment with web and social media. Another application is where authors have reported [9] a bibliometric study of critical BI&A publications, researchers, and research topics based on more than a decade of related academic and industry publications.

One of the bigger research was proposed in one the publication where authors described the HACE theorem [12] that characterizes the features of the Big Data Revolution and also proposes a big data processing model, from the data mining perspective.

Harnessing the full potential of any technology is necessary as well. In a paper [13] authors proposed a comprehensive overview of the applications of data processing platform designed to harness the potential of big data in the field of road transport policies in Europe. Also, to implement big data algorithms efficiently, authors proposed [14] FB+-tree that builds fast indexing structure using multi-level Key ranges, which is explained on the basis of B+ trees. Point searches and range searches are helped by early termination of searches for non-existent data.

For any technology to perform up to its full potential it should be as robust it can be. In one of the paper, authors proposed a model [15] based on robust data analytics, high performance computing, efficient data network management and cloud computing techniques that are critical towards optimized performance of Smart Grid's so as to reduce the cost for customer.

Big data is surely very important and big terms these days. Organizations have understood the importance of big data for managing their data efficiently and to obtain day to day analysis of the data generated. In one the paper this functionality was proposed and a deep understanding of this whole concept was given to the users.

In another research [18] authors proposed an overview of big data, significance of big data, how hadoop works and systems, which is based on analysis of published implementation architectures of big data use cases. Different flavours of Hadoop.

There are many researches that have been proposed in this domain and also so many are being going on to make the people aware of any new and unseen facts of this technology.

3 PROBLEM DESCRIPTIONS

There are different technologies to deal with Big Data analysis, but most of them are complex and requires expertise to deal with them. Especially for non-computer scientists such as social scientists, they require good programming skills and knowledge of configuring and maintaining the infrastructure which almost makes it impossible for them to explore the large data sets or to perform ad hoc analytics on it. For example, how a social scientist can explore the data to find an event in 1975 by having the previous fifty years of newspaper data? Or how a social scientist can predict the human behavior by analyzing its previous 5 years of data gathered from different sources such as cell phone records with GPS tracking, search engine queries, internet

transaction data, consumer behavior or its social network activity? [8]

There are plenty of Big Data analysis platforms or frameworks are available nowadays in the market, but the problem for non-computer scientists is to master them because of the complexity involves with them and where necessary to take training in order to use them for exploring Big Data in adhoc manners and doing analytics on large data sets. These systems inherit the problem of maintaining them as well, which might include at application or infrastructure level.

4 GRAND CHALLENGES IN BIG DATA

There are many challenges in harnessing the potential of big data today, ranging from the design of processing systems at the lower layer to analysis means at the higher layer, as well as a series of open problems in scientific research. Among these challenges, some are caused by the characteristics of big data, some, by its current analysis models and methods, and some, by the limitations of current data processing systems. In this section, we briefly describe the major issues and challenges.

A. Data complexity

The study of data complexity metrics is an emergent area in the field of data mining and is focus on the analysis of several data set characteristics to extract knowledge from them. This information used to support the election of the proper classification algorithm

B. Computational complexity

Three of the key features of big data, namely, multi-sources, huge volume, and fast-changing, make it difficult for traditional computing methods (such as machine learning, information retrieval, and data mining) to effectively support the processing, analysis and computation of big data. Such computations cannot simply rely on past statistics, analysis tools, and iterative algorithms used in traditional approaches for handling small amounts of data. New approaches will need to break away from assumptions made in traditional computations based on independent and identical distribution of data and adequate sampling for generating reliable statistics. When solving problems involving big data, we will need to re-examine and investigate its computability, computational complexity, and algorithms.

New approaches for big data computing will need to address big data-oriented, novel and highly efficient computing paradigms, provide innovative methods for processing and analyzing big data, and support value-driven applications in specified domains. New features in big data processing, such as insufficient samples, open and uncertain data relationships, and unbalanced distribution of value density, not only provide great opportunities, but also pose grand challenges, to studying the computability of big data and the development of new computing paradigms.

C. System complexity

Big data processing systems suitable for handling a diversity of data types and applications are the key to supporting scientific research of big data. For data of huge volume, complex structure, and sparse value, its processing is confronted by high computational complexity, long duty cycle, and real-time requirements. These requirements not only pose new challenges to the design of system architectures, computing frameworks, and processing systems, but also impose stringent constraints on their operational efficiency and energy consumption.

The design of system architectures, computing frameworks, processing modes, and benchmarks for highly energy-efficient big data processing platforms is the key issue to be address in system complexity. Solving these problems can lay the principles for designing, implementing, testing, and optimizing big data processing systems. Their solutions will form an important foundation for developing hardware and software system architectures with energy-optimized and efficient distributed storage and processing.

5 CONCLUSIONS

Big data has made a strong impact in almost every sector and industry today. In this paper, we have briefly reviewed the grand challenges that big data brings us. We close by a few suggestions on how to make a big data project successful. It is no secret that in big data research and applications, industry is ahead of academia. The successful applications of big data in industry point to the following necessary conditions for a big data project to be successful. Firstly, there must be very clear requirements, regardless of whether they are technical, social, or economic. Secondly, to efficiently work with big data, we will need to explore and find the kernel structure or kernel data to be processed. Finding kernel data and structures, which are small enough and yet can characterize the behavior and properties of the underlying big data, is non-trivial because it is very domain-specific. Thirdly, a top-down management model should be adopted. Although a bottom-up approach may allow us to solve some niche problems, the isolated solutions often cannot be put together into a complete solution. Finally, the goal should be to solve the entire problem by an integrated solution, rather than striving for isolated successes in a few aspects.

REFERENCES

1. M. A. Beyer and D. Laney, “The importance of □big data□: A definition,” Gartner, Tech. Rep., 2012.
2. X. Wu, X. Zhu, G. Q. Wu, et al., “Data mining with big data,” IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.Rajaraman and J. D. Ullman, “Mining of massive datasets,” Cambridge University Press, 2012.
3. Z. Zheng, J. Zhu, M. R. Lyu. “Service-generated Big Data and Big Data-as-a-Service: An Overview,” in Proc. IEEE BigData, pp. 403-410, October 2013. A . Bellogín,I. Cantador, F. Díez, et al., “An empirical comparison of social, collaborative filtering, and hybrid recommenders,” ACM Trans. on Intelligent Systems and Technology, vol. 4, no. 1, pp. 1-37, January 2013.
4. W. Zeng, M. S. Shang, Q. M. Zhang, et al., “Can Dissimilar Users Contribute to Accuracy and Diversity of Personalized Recommendation?,” International Journal of Modern Physics C, vol. 21, no. 10, pp. 1217- 1227, June 2010.
5. X. Liu, G. Huang, and H. Mei, “Discovering homogeneous web service community in the user-centric web environment,” IEEE Trans. on Services Computing, vol. 2, no. 2, pp. 167-181, April-June 2009.
6. Zielinnski, T. Szydlo, R. Szymacha, et al., “Adaptive soa solution stack,” IEEE Trans. on Services Computing, vol. 5, no. 2, pp. 149-163, April-June 2012.
7. F. Chang, J. Dean, S. mawat, et al., “Bigtable: A distributed storage system for structured data,” ACM Trans. on Computer Systems, vol. 26, no. 2, pp. 1-39, June 2008.
8. V. Gupta, G. S. Lehal, “A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages,” Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 2, pp. 157-161, May 2013.
9. T. Niknam, E. Taherian Fard, N. Pourjafarian, et al., “An efficient algorithm based on modified imperialist competitive algorithm and K-means for data clustering,” Engineering Applications of Artificial Intelligence, vol. 24, no. 2, pp. 306-317, March 2011.
10. M. J. Li, M. K. Ng, Y. M. Cheung, et al. “Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters,” IEEE Trans. on Knowledge and Data Engineering, vol. 20, no. 11, pp. 1519-1534, November 2008.
11. AbdelrahmanElsayed, Osama Ismail, and Mohamed E. El-Sharkawi, MapReduce: State-of-the-Art and Research Directions.
12. Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, Data Mining with Big Data (IEEE), Vol. 26, NO. 1, JANUARY 2014
13. Michele De Gennaro, Elena Paffumi, Giorgio Martini, Big Data for Supporting Low-Carbon Road Transport Policies in Europe: Applications, Challenges and Opportunities, Intl Journal of Big Data Research (Elsevier), 2 June 2016
14. Cui Yu, Josef Boyd, FB+- tree for Big Data Management, Intl Big Data Research (Elsevier), Pg: 25-36, Vol. 4, June 2016