International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 06, June 2017
www.ijates.com
ijates
ISSN 2348 - 7550

Systematic Mapping Study of Big Data Analytics Tools and Techniques

Sheena Rewari[1], Dr. Williamjeet Singh[2]

[1]*Research scholar, Punjabi University, Patiala, Punjab, (India)*

[2]*Assistant Professor, UCOE Punjabi University, Patiala, Punjab, (India)*

ABSTRACT

Big data is a huge dataset demonstrating the aspects of volume, velocity, variety, veracity, validity, value, variability and vagueness in an OR relationship. More than sufficient new insights are discovered while dealing with big data. There are many software and hardware solutions available in the technology prospect that facilitate grabbing, saving and consecutive analysis of big data. Hadoop and its correlated high-tech solutions are one of them.

Keywords: Big data analytics, tools, Hadoop, challenges, application areas.

I. INTRODUCTION

With all the devices available today to collect data such as RFID readers, microphones, cameras, sensors and so on, we are seeing an explosion in data being collected worldwide. Big data is a term that is used to describe these large number of datasets that may be unstructured and grow so quickly that it is impossible to manage with a regular database or stastical tools. [1]

1.1. Big Data Adoption

The big data adoption process goes through a number of phases are given in figure1.

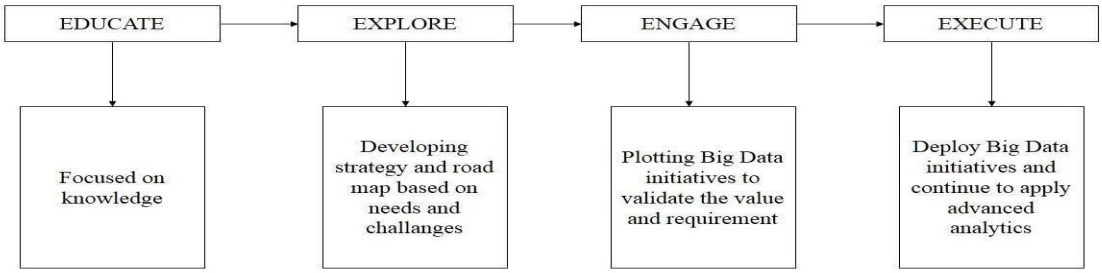

Figure 1big data adoption

1.2. Big Data Analytics

Big data analytics is the process of analyzing and acquiring intelligence from data to gain meaningful patterns in big data. Big data analytics helps a business acknowledge the needs of a customer so that businesses can expand their customer base and maintain the existing ones with admissible offerings of their product. [2]

The three supreme types of analytics are shown in figure 2.

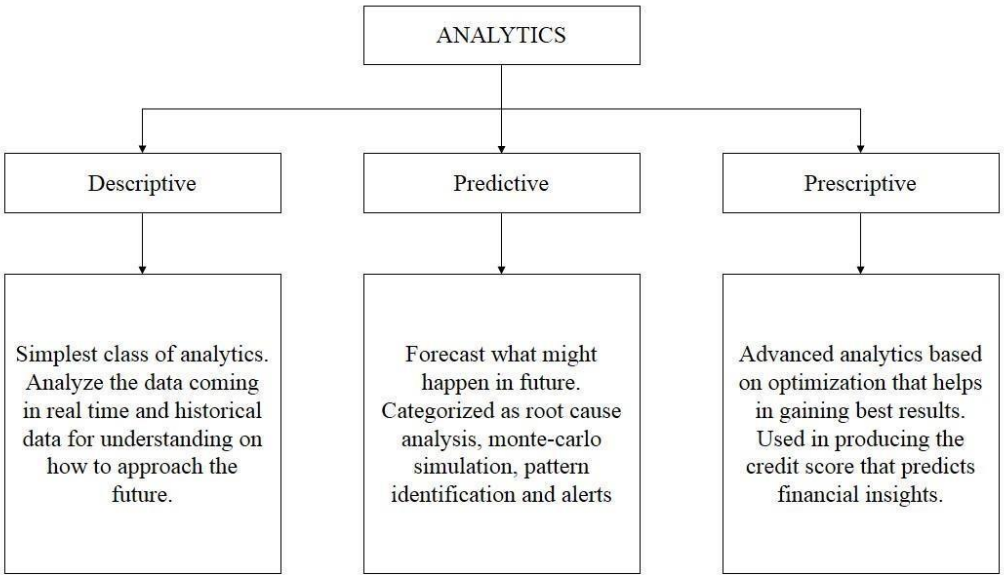

Figure 2 big data analytics types

1.3. Motivations and Objectives

This study is carried through in order to evaluate the existing big data analytics tools and techniques which are more applicable. This analysis is necessary to make it possible to know which class of big data analytics tools, techniques and application areas have been shielded in past research and helps to identify gaps.

This study plans at periodically reviewing the big data analytics tools, techniques and application areas used in existing studies. The results may help the researchers to get an outline of the state of big data analytics and the feature the research gaps.

This paper is structured as follows. Section 2 describes the research methodology used in this study.

Section 3 gives the classification of big data analytics research papers considering the following criteria: (1) application domains; (2) techniques used; and (3) platform/framework used. Section 4 presents the mapping of studies. Section 5 discusses the paper. Section 6 presents the conclusions and future directions.

II. RESEARCH METHODOLOGY

The research methodology is composed of two stages. The first stage involves the research of works related to big data analytics. The second stage is concerned with establishing a classification scheme described in Section 3.

1) RQ.1: What are the different categories of tools and techniques in the area of big data analytics? Section 3.2 of this paper answers this question.

2) RQ.2: What are the application areas in big data analytics research?

The section 3.1 describes the hot big data analytics application areas with their future directions.

3) RQ.3: What are the various platforms/framework used in big data analytics?

To answer this question table in section 3.3 describes the platform involved. The research is initialized with these queries and then follows the steps described.

2.1. Search Strategy and Screening

Sources of information

To increase the probability of relevant articles, a set of appropriate databases must be chosen. For this review, the major databases of electronic journals are searched. The digital libraries considered are:

- IEEE Explore (http://ieeexplore.ieee.org)
- ACM Digital Library (www.acm.org/dl)
- Science Direct (www.sciencedirect.com)
- Springer (www.springerlink.com) Additional sources

IBM's Big Data University website: bigdatauniversity.com

Technical Reports.

2.2. Study Selection

Research papers published by journals, conference proceedings and workshops are thought to be worthy and reliable. Keyword based search is employed to select the most relevant works. The keywords used are big data analytics, big data analytics tools and techniques and big data analytics applications.

The Table 1 shows the defined search strategy and number of results obtained. From the returned studies, firstly irrelevant studies are excluded on the basis of title. Certain studies could not be estimated from the title, then their abstract is considered. If even abstract is not evident then after reading the full text of papers, irrelevant studies are excluded.

Table 1 Search Selection

S. no.	E-resource	Studies returned	Excluded			Search string
			based on title	based on abstract	based on full text	
1.	ieeexplore.ieee.org	115	60	30	12	big data analytics + tools + techniques
2.	www.acm.org	136	70	31	25	big data analytics + tools + techniques
3.	www.sciencedirect.com	133	79	35	16	big data analytics + tools + techniques + applications
4.	www.springerlink.com	146	126	12	8	big data analytics + tools + techniques + applications

2.3. Establishing a Classification Scheme

The selection process resulted in 359 papers selected from five different digital libraries as on 27th January 2017,

05.10 pm (IST). Each paper is carefully assessed and classified. The selected research papers are classified according to the criteria established in section 3 of this article.

Distribution of papers and Mapping of studies

The results of the classification offer important guidelines for future research on big data analytics. Literature related to big data analytics has increased enormously in the last 4 years, papers from 2013 to 2017 are reviewed. The distribution of reviewed papers over the years is depicted in Table 2.

Table 2 Distribution of Papers

e E - Resourc					
IEEE	1	4	9	9	
ACM	2			1	
Elsevier			2	1	
Springer			1		1

III. CLASSIFICATION METHOD

The research papers are classified by giving consideration to following criteria: (1) application areas of big data analytics, (2) techniques used, and (3) platform used.

3.1. Classification based on application areas

On the basis of research papers studied, the operative big data analytics research fields are explained in the table below.*Table 3 Big Data Analytics Application Areas*

APPLICATION AREA	ROLE OF BIG DATA ANALYTICS
HEALTHCARE **[3] [4] [5] [6]**	Provides a comprehensive view of treatment delivery for meeting future needs. Access to historical lab results and medications to monitor patient's condition on daily basis.

EDUCATION **[7]** **[8]**	Allow the information to be shared in cloud based data warehouse with other medical institutions. For grade analysis of students. Personalized learning.
GEOGRAPHICAL SYSTEMS **[9] [10] [11] [12]**	To introduce weather forecasting for improving yield productivity and usage of pesticides. For remote sensing observatory earth spacecraft to detect the volcanic plumes and initiate an
CRIME ANALYTICS/ GOVERNANCE **[13] [14] [15]**	To collect real time crime data to identify crime trends. For better decision making and e- governance policies. For the safety and wellbeing among the citizens of the country.
TRAFFIC ANALYSIS **[14] [1]**	Automatic decision making systems in order to catch various kind of traffic violators. For connected and autonomous vehicles.
FINANCE AND MARKETING **[16] [1]**	Satisfying customer expectations and optimize customer engagement. Using digital sensing on social media with social media monitoring tools to analyze data about competitors and their marketing strategies.

3.2. Classification based on techniques used

Table 4Techniques Used for Big Data Analytics

TECHNIQUE	DESCRIPTION
ASSOCIATION RULE MINING [2]	It looks for relationship between variables or objects. It is a popular and well researched method for discovering interesting relationships between variables in large databases.
CLASSIFICATION [2]	Classification is a method of recognizing categories that an observation belong to, based on its attributes.
GENETIC ALGORITHMS [2]	Genetic algorithm is a method for solving optimization problems that are based on natural selection.
MACHINE LEARNING [17] [4]	It provides the computers the ability to learn without being clearly programmed.
REGRESSION ANALYSIS [2]	Regression analysis involves examining independent variables to see their impact on dependent variables.

CLUSTERING [2] [18]	It is the process of making a group of objects into classes of similar objects. A cluster of data objects can be treated as one group
OUTLIER ANALYSIS [2]	It is also pronounced as anomaly detection. It is the process of identifying and excluding the items or observations that do not match to an expected pattern.

3.3. Classification based on platform used

Platforms engaged are used as the third standard for classification of research papers.

Table 5 Platforms used for Big Data Analytics

PLATFORM USED	DESCRIPTION	ADVANTAGES
HADOOP [7] [19] [20] [13] [21] [22] [11] [23] [24] [25] [2]	Hadoop is an open source project of apache foundation. It uses map reduce technology as its foundation. Hadoop replicates its data across different computers so that if one goes down, the data is processed on one of the replicated computers.	Hadoop handles massive quantities of data using commodity hardware that is relatively inexpensive computers
MONGODB [19] [3]	Mongo DB is an open source document and nosql database. Mongo DB uses sharding technique for S splitting the data evenly across the cluster for	It handles large amounts of structured, nstructured and semi structured data with ease. cale out architecture.
RAPID MINER [25]	Rapid miner is a platform for data mining and machine learning. It uses operator tree modeling knowledge discovery process. Mostly database and excel files are smoothly	Data transformation, data integration, data modeling as well as visualization methods are incorporated by rapid miner. Rapid miner provides fully integrated platform for data analysis
HIVE [9]	HIVE provides data warehousing tools to extract, transform and load data. This query data is stored in Hadoop files.	Using HIVE the data is portioned into tables to improve performance.
POSTGRE SQL [18]	Postgre sql is an open source object- relational database system.	Postgresql runs stored procedures in more than 12 languages. It uses multiple row data storage (MVCC) for

PIG [13]	Pig is a high level language that generates Map Reduce code to analyze large data sets.	Pig is one of the best tool to make unstructured data into structured one.
CASSANDRA [8]	Provides high scalability and availability without compromising performance.	Cassandra is decentralized which means no network bottlenecks.

IV. DISCUSSION

The systematic mapping study is derived from 30 publications. 5 papers are published in journals and rest of the papers is published in conferences and workshops. Any technique such as classification, clustering, association rule mining, regression analysis, genetic algorithms, and outlier analysis can be used for big data analytics. A large number of studies confirmed that application of big data analytics tool depends on the situation and objective. The result of same research may vary with the use of different tool, result depends on tool used.

V. CONCLUSION

Big data analytics is an active area of research. The result of this study may help new potential users in understanding the range of available big data analytics tools and techniques. This study presents big data analytics active application areas. One of its most well liked application area is crime analysis and governance for better law and order. Crime data analytics is major example that shows how big data analytics can be used by law enforcement communities to take full advantage to retain public support. Crime data analytics is used to provide whole crime statistics of the region that provides benefit to the society by striking the government that why the crime is increasing The government can take better decisions for better living of the citizens that would naturally add up to lot of lives.

REFERENCES

[1] "https://courses.bigdatauniversity.com/courses/course-v1:BigDataUniversity+BD0101EN+2016/info," [Online]. Available: https://bigdatauniversity.com/. [Accessed 7 Feburary 2017].

[2] F. Armour, S. Kaisler and A. Espinosa, "Introduction to Big Data Analytics: Concepts, Methods, Techniques and Applications," in IEEE 48th Hawaii International Conference on System Sciences, Hawaii, 2015.

[3] P. Dhaka and R. Johri, "Big Data Application: Study and Archival of Mental Health Data, using MongoDB," in International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 3-5 March,2016.

[4] R. Mennour and M. Batouche , "Drug discovery for breast cancer based on big data analytics techniques," in 5th internationalconference on Information and Communication Technology and Accessibility, Marrakech, Morocco, 21-23 December,2015.

[5] P. J. A. Patel and D. P. Sharma, "Big data/or Better Health Planning," in IEEE International Conference on Advances in Engineering & Technology Research (ICAETR) , Unnao, India , August 01-02, 2014.

[6] Y. Wang and N. Hajli, "Exploring the path to big data analytics success in healthcare," Journal of Business Research Elsevier, pp. 287-299, 15 August 2016.

[7] L. Cen, D. Ruta and J. Ng, "Big Education: Opportunities for Big Data Analytics," in IEEE InternationalConference on Digital Signal Processing, Singapore, 21-24 July 2015.

[8] S. A. Hossain, "Big Data Analytics in Education: Prospects and Challenges," in 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), Noida, India, 2-4 Sept. 2015.

[9] A. U. Abdullahi, R. Ahmad and N. M. Zakaria, "Big data:Performance profiling of metrological and ocenographic data on HIVE.," in IEEE 3rd International Conference On Computer And Information Sciences (ICCOINS), India, 2016.

[10] S. Early, "Really, Really Big Data NASA at the Forefront of Analytics," IT Professional, vol. 18, no. 1, pp. 58-61, 2016.

[11] M. M. Rathore, A. Ahmad, A. Paul and A. Daniel, "Hadoop based real time big data architecture for remote sensing earth observatory system," in 6th ICCCNT, Denton, USA, july 13-15 2015.

[12] M. Bendre, R. Thool and V. Thool, "Big Data in Precision Agriculture : Weather Forecasting for Future Farming," in 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 4-5 September, 2015.

[13] A. Jain and V. Bhatnagar, "Crime Data Analysis Using Pig with Hadoop," in International Conference on Information Security & Privacy (ICISP2015), Nagpur, INDIA, 11-12 December 2015.

[14] J. W. Lee, J. S. jeong, M. Kim and K. H. Yoo, "Safe-Return-Home Service based on Big Data Analytics," in ACM Proceedings of the 2015 International Conference on Big Data Applications and Services, Jeju Island, Republic of Korea, October 20 - 23, 2015.

[15] D. T. Mahmood and U. Afzal, "Security Analytics: Big Data Analytics for Cybersecurity," in IEEE 2nd National Conference on Information Assurance (NCIA), Rawalpindi, Pakistan, 2013.

[16] D. Schmidt, W. C. Chen and G. Ostrouchov, "Introducing a New Client/Server Framework for Big Data Analytics with the R Language," in XSEDE16, Miami, USA, 12-21 July, 2016.

[17] J. Lluis and B. Garcia, "A Quick View on Current Techniques and Machine Learning Algorithms for Big Data Analytics," in IEEE ICTON 2016, Europe, 206.

[18] G. Bordogna, L. Fringerio, A. Cuzzocrea and G. Psaila, "Clustering Geo-Tagged Tweets for Advanced Big Data Analytics," in IEEE International Congress on Big Data, Italy, 2016.

[19] E. Dede, M. Govindaraju, R. S. Canon and L. Ramakrishnan, "Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis," in Proceedings of the 4th ACM workshop on Scientific cloud computing, New York, USA, June 17 - 17, 2013.

[20] I. "Hadoop as a service," IBM Corporation, Somers, New York,USA, 2015.

[21] A. Mukherjee, J. Datta, R. Jorapur, R. Singhvi, S. Haloi and W. Akram, "Shared disk big data analytics with apache hadoop," IEEE, p. 26, 2012.

[22] A. Pal, P. Agrawal, K. Jain and S. Aggarwal, "A performance analysis of map reduce task with large number of files dataset in big data using hadoop," in Fourth International Conference on Communication Systems and Network Technologies, 2014.

[23] E. Sivaraman and D. Manichachezian, "High performance and fault tolerant distributed file system for big data storage andprocessing using hadoop," in International Conference on Intelligent Computing Applications, Coimbatore, India, 2014.

[24] K. Singh and R. Kaur, "Hadoop: Addressing challanges of Big Data," in International Advance computing conference IAAC, India, 2014.

[25] M. Utmal and R. K. Pandey, "Taxanomy on the integration of hadoop and rapid miner for big data analytics," in International Conference on Computational Intelligence and Communication Networks, India, 2015.

[26] J. Zhu, E. Zhuang, J. Fu, J. Baranowski, A. Ford and J. Shen, "A Framework-Based Approach to Utility BigData Analytics," IEEE TRANSACTIONS ON POWER SYSTEMS, vol. 31, no. 3, pp. 2455-2462, 2015.

[27] C. Zhang, X. Shen, X. Pei and Y. Yao, "Applying Big Data Analytics Into Network Security: Challenges, Techniques and Outlooks," in IEEE International Conference on Smart Cloud, New York, NY, USA, 2016.

[28] C. Verma and D. R. Pandey, "Big data representation forgrade analysis through hadoop framework," in IEEE6th InternationalConference - Cloud System and Big Data Engineering, India, 2016.

[29] P. Vashisht and V. Gupta, "Big Data Analytics Techniques: A Survey," in IEEE InternationalConference on Green Computing and Internet of Things (ICGCIoT), Noida, India, 8-10 October,2015.

[30] C. W. Tsai, C. F. Lai, H. C. Chao and A. V. Vasilakos, "Big Data analytics: a survey," Journalof big data, a springer open journal, pp. 1-32, 2015.