Domain Specific Big Data Analytics: A Review

Ch. Sridharababu[1], K. Srinivasa Rao[2]
[1]*Asst.Professor, Pydah College of Engineering and Technology, Visakhapatnam*
[2]*Research Scholar, Dept of CS&SE, AU College of Engineering, Andhra University, Visakhapatnam*
(E-mail: ch.sridhar38@gmail.com, sri.kurapati@gmail.com)

***Abstract*—**Big Data is an evolving topic and has attracted the attention of many researchers and practitioners of every domain. Big Data are now quickly mounting in all science and engineering domains, including physical, biological and biomedical sciences. The era of Big Data is underway. Big data analytics plays a key role through reducing the data size and complexity in big data applications. Big data analytics would definitely lead to valuable knowledge for many organizations. Various technologies can be used to handle the big data. In this paper, we give the comprehensive characteristics of big data. We also investigate the current definitions of these concepts available in the literature. Finally, this paper provides a survey of domain specific big data analytics for innovative web applications.

Keywords*—Big Data, Domain, Big Data Analytics, Web Applications*

I. INTRODUCTION

Domain is an area of interest. Domains of knowledge are interdependent, and training in one domain may well affect performance in another domain. There are several domains such as financial, banking, marketing, social media, medical, weather and e-commerce etc. Nowadays every domain generates petabytes of data. For example, social media (Facebook, You Tube, Twitter, and WhatsApp) generated texts, images, photos, videos contain an unexpected and rapidly growing quantity of data. Data alone is worthless without analysis. Analysis of the information contained in these data sets already led to major development in any domain. The distributed and heterogeneous nature of the data provides the significant challenges for in the physical and life sciences and also in commerce, medicine, defense, finance, telecommunication and other industries. These challenges are overcome by using data mining techniques.

Data Mining [1] is defined as extracting information from huge sets of data. In brief, data mining is the procedure of mining knowledge from data. Data mining is a method of pattern discovery against a pool of data using specialized data mining tools. These tools use a sophisticated blend of classical and advanced components like artificial intelligence, pattern recognition, databases, traditional statistics, and graphics to present hidden relationships and patterns they find in any given data pool. With the ever growing of datasets, data mining tasks has significantly increased. Additionally data reduction, data selection, feature selection is an essential task especially when dealing with large datasets [2]. This presents an unprecedented challenge for researchers. It is because, existing algorithms may not always respond in an adequate time when dealing with these high dimensional data.

The major challenge is automation of this process and developing new machine learning algorithms to ensure consistency. In addition to all these clustering of large datasets that help in analyzing the big data is of major concern. However exact definition for big data is not defined and there is a belief that it is problem specific [3].

Big data represents the general form of problems and techniques used for application domains that collect and maintain massive volumes of raw data for domain specific data analysis. Traditional architectures and infrastructures process large amounts of data within an acceptable time and resources there by generating huge data. Organizations need to find new tools and methods specialized for big data processing for extracting useful information. So, big data analytics has become a key factor for companies to reveal hidden information and achieve competitive advantages in the market scenario. Modern data- intensive technologies, increased computational and data storage resources have contributed heavily to the development of big data analytics. In this paper, we considered on domain specific big data analytics and its available techniques. The remainder of this paper is organized as follows. The characteristics of big data are described in section 2.Section 3outlines a literature review on domain specific big data analytics carried out by various researchers. Finally, Section 4 gives the conclusion.

II. CHARACTERISTICS OF BIG DATA

Big data is different from other any data in terms of ten parameters-volume, variety, velocity, veracity, value, variability, validity, venue, vocabulary and vagueness (10V's), are the major challenges of big data management:

i. Volume:

Most organizations were already struggling with the increasing size of their databases as the Big Data tsunami hit the data stores. Volume of data is quantity of generated and stored data. The size and value of the data determines whether it can be considered big data or not.

ii. Variety:

Data can be stored in multiple formats. For example database, excel, csv, access or for the matter of the fact, it can be stored in a simple text file. The type and nature of the data helps people to analyze data effectively and use the resulting insight. Variety ranges from neatly-structured tabular data, to

unstructured data containing items such as images, emails, spreadsheets, social media conversations and streaming media.

iii. Velocity:

The data movement is now almost real time and the retrieval time has reduced to fractions of seconds. This high velocity data represent Big Data**.** Velocity of data refers to the speed at which the data is generated and processed, to meet the demands and challenges that lie in the path of growth and development.

iv. Veracity:

Veracity represents credibility of the data source and suitability of the data for the target audience. The quality of captured data can vary greatly, affecting accurate analysis.

v. Value:

Value deals with what value should come out which data.

vi. Variability:

Dataset inconsistencies can impede processes to handle and manage it. Dataset inconsistencies may exist in dynamic, evolving, spatiotemporal data, time series, seasonal, and any other type of non-static behavior in data sources, customers, objects of study, etc.

vii. Validity:

Validity is a result of the logical inferences drawn from matching data having a sound basis in logic or fact

Volume -Validity = Insignificance?

viii. Venue:

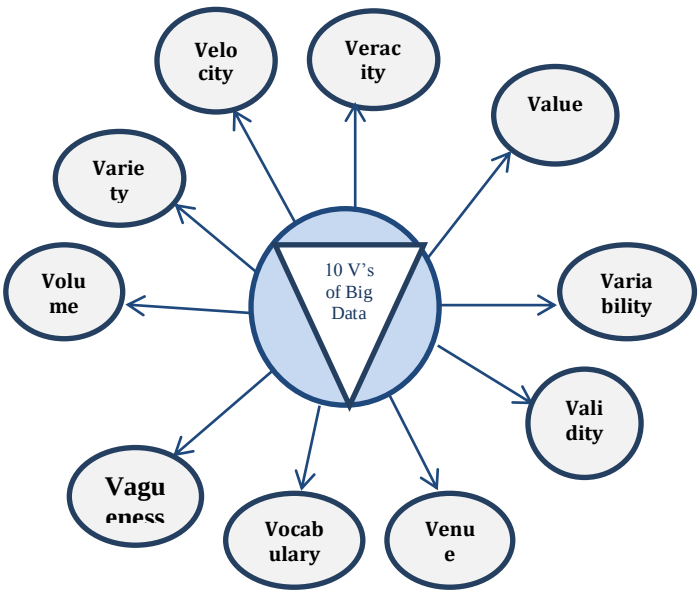
It is a heterogeneous data from multiple platforms and different owners' systems.

ix. Vocabulary:

It refers the data's structure, content, semantics, syntax, and schema.

x. Vagueness:

Confusion over the meaning of big data (What's new about it? What are the tools? etc.) .



Big Data could be analyzed by these above mentioned characteristics are depicted in figure 1. Although it might sound simple, but it is observed that none of the data scientists have specified a complete example of big data. There are several characteristics on which data scientist agrees.

III. LITERATURE REVIEW

Big data oriented techniques have emerged as an important research topic for smart cities and urban surveillance systems examined by Ling Hu et al. (2018). They proposed a novel automatic background subtraction algorithm for urban surveillance systems [4].

According to Choi et al. (2017) big data research is still in infancy. They presented the challenges and opportunities of big data analytics in this unique application domain (domain specific big data analytics). And also technological development and advances for industrial based business systems, reliability and security of industrial systems, and their operational risk management are examined [5].

Lakshmi.C et al. (2016) outlined definition and challenges of big data. They explained 5 V's of big data. They defined a standardized framework to decompose big data system into four subsequent modules, which are data generation, data acquisition, data storage and data analytics. These four modules form a big data value chain. In addition, they presented a detailed survey of materials and methods used in research and industry communities [6].

In another study, Achariya D.P. et al. (2016) sketched big data as 4 V's. They deliberated the potential impact of big data challenges, open research issues (i.e., IoT for Big Data Analytics, Cloud Computing for Big Data Analytics, Bio-inspired Computing for Big Data Analytics, and Quantum Computing for Big Data Analytics) and various tools associated with it [3].

Security challenges and privacy issues of big data analytics are addressed by Aditya Dev Mishra et al. (2016). They discussed various comparisons of security challenges [7].

Rajeswari D. (2015) emphases on the various states of art studies towards big data analytic techniques and gives a comprehensive analysis of various applications. The author also discussed about research gap of existing studies of big data analytics [8].

Benjelloun F. Z. et al. (2015) presented big data projects, opportunities, examples and models in many sectors such as healthcare, commerce, tourism and politics. They also discussed examples of technologies and solutions developed to face big data challenges [9].

Al Nuatami et al. (2015) compared different definitions of big data, smart city and benefits of incorporating big data applications. They identified the requirements that support the implementation of big data applications for smart city services [10].

O.Y.Al-Jarrah et al. (2015) presented a comprehensive review of state-of-the-art sustainable/energy-efficient machine-learning literatures, including theoretical, empirical and experimental studies pertaining to the various needs and recommendations. They addressed domain specific big data

challenges (i.e. Geo, climate and environment; Bio, medicine, and health; Stars, galaxies, and the universe) [11].

Vibhavari Chavan et al. (2014) defined big data as 3 V's. They considered the big data challenges such as analysis, search, capture, visualization, sharing, storage, transfer and privacy violations [12].

Stephen Kaisler et al. (2014) suggested a fifth V: value, which is the contribution big data has to decision making [13].

In [14], Huang and Liu (2014) claimed that next-generation computing systems for big data analytics need innovative designs in both hardware and software that would provide a good match between big data algorithms and the underlying computing and storage resources. They discussed overview of the Lambda Architecture.

Dominic Breuker (2014) argued the process of analyzing big data both easier and more efficient, thereby helping to close the gap between supply and demand of data scientists [15].

Wu et al (2014) presented a HACE Theorem that characterizes the features of the big data. From the data mining perspective, they explained the characteristics of a big data processing model, [16].

T.K.Das et al (2013) developed a frame work for analyzing unstructured data. They completed the first phase where unstructured data is pulled from public tweets of twitter and the XML data parsed to store in a NOSQL database like HBASE [17].

The overview of big data and it's applications in health care systems are outlined by Raghunath Nambiar et al. (2013). They discussed about cisco architecture for connected health [18].

It is crucial to think of big data as a phenomenon rather than a singular technology proposed by David Corrigan (2013). He recommended six best practices for successful big data governance initiatives [19].

Boyd D. et al. (2012) perspective big data is cultural, technological, and scholarly phenomenon. They asked difficult questions of big data's models of intelligibility before they crystallize into new orthodoxies. They argued bigger data are not always better data [20].

J. Manyika et al. (2011) discussed big data techniques and technologies. They are also explained big data domains such as Health care (United States), Public sector administration (European Union), Retail (United States), Manufacturing (global), Personal location data (global) [21].

Bill Hamilton (2010) concentrated on the definition of big data. He investigates the opportunities and challenges of big data analytics in health care industry [22].

A.Jacobs (2009) given a meta- definition for big data: big data should be defined at any point in time as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time" [23].

Bryant R. E., et al.(2008) discussed the big data technology and application challenges. Big data computing is conceivably the biggest innovation in computing in the last decade. They have only begun to see its potential to collect, organize, and process data in all walks of life [24].

The current research on the big data is still in an early stage so significant research efforts are needed for analysis and storage management associated with the domain specific big data.

IV. CONCLUSION

In the context of discussing key works in the literature and providing our understandings on those specific topics, the study focuses on domain specific big data analytics. Big Data Analytics is a wide area that integrates techniques from various fields including Data mining,statistics,machine learning, artificial intelligence,pattern recognition and deep learning, for the analysis of large volumes of data. The main goal of our paper was to make a survey of various big data handling techniques those handle a domain specific massive amount of data from different sources.

REFERENCES

[1] Jiawei Han., Micheline Kamber.,Jian Pei., "Data Mining: Concepts and Techniques", 3rd Edition, The Morgan Kaufmann Series in Data Management Systems, 2011.

[2] A. Rajaraman and J. Ullman, "Mining of Massive Data Sets", Cambridge Univ. Press, 2011.

[3] Achariya.D.P.,Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", International Journal of Advanced Computer Scienceand Applications, vol. 7(2), pp.511–518, 2016.

[4] Ling Hu, Qiang Ni, and Feng Yuan, "Big Data oriented Novel background Subtraction Algorithm for Urban Surveillance Systems", Big Data Mining and Analytics, vol 1, pp.137-145, June 2018.

[5] Tsan-Ming Choi, Hing Kai Chan, XiaohangYue, "Recent Development in Big Data Analytics for Business Operations and Risk Management", IEEE Transactions on Cybernetics, vol. 47(1), pp.81-92, 2017.

[6] C Lakshmi,V.V Nagendra Kumar, "Survey Paper on Big Data", International Journal of Advanced Research in Computer Science and Software Engineering, vol6(8), pp.368-381, August 2016.

[7] AdityaDev Mishra, Youddha Beer Singh, "Big Data Analytics for Security and Privacy Challenges", International Conference on Computing, Communication and Automation, pp.50-53, 2016.

[8] Rajeshwari.D, "State of the Art of Big Data Analytics: A Survey", International Journal of Computer Applications, vol 120(22), pp.39-46, June 2015.

[9] Benjelloun, F. Z., Lahcen, A. A., &Belfkih, S., "An Overview of Big Data Opportunities, Applications and Tools", In Intelligent Systems and Computer Vision (ISCV), pp. 1-6, IEEE, 2015.

[10] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed &Jameela Al-Jaroodi, "Applications of big data to smart cities", Journal of Internet Services and Applications, vol 6(25), pp-1-15, 2015.

[11] O. Y. Al-Jarrah,P. D. Yoo,S. Muhaidat,G. K. Karagiannidis & K.Taha, "Efficientmachinelearningforbigdata: Areview", Big Data Research, vol2(3), pp.87-93,2015.

[12] VibhavariChavan, Rajesh. N. Phursule, "Survey Paper on Big Data", International Journal of Computer Science and Information Technologies, Vol. 5 (6), pp. 7932-7939, 2014.

[13] Stephen Kaisler, Frank Armour& J. Albert Espinosa, "Introduction to Big Data: Challenges, Opportunities and Realities", 47th Hawaii International Conference on System Science, IEEE, pp. 728, 2014.

[14] H. Howie Huang, Hang Liu, "Big Data Machine Learning and Graph Analytics: Current State and Future Challenges", IEEE International Conference on Big Data, pp.16-17, 2014.

[15] Dominic Breuker, "Towards Model-Driven Engineering for Big Data Analytics – An Exploratory Analysis of Domain-Specific Languages for Machine Learning", 47th Hawaii International Conference on System Science, IEEE, pp. 758-767, 2014.

[16] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, vol. 26(1),pp.97-107, January 2014.

[17] T.K.Das, P.Mohan Kumar, "BIG Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology, vol. 5 (1), PP. 153-156, 2013.

[18] R. Nambiar, R. Bhardwaj, A. Sethi, and R. Vargheese," A look at challenges and opportunities of big data analytics in healthcare ", IEEE International Conference on Big Data, pp. 17–22 , 2013.

[19] David Corrigan,"Integrating and governing big data", IBM Whitepaper, pp.1-13, 2013.

[20] Danahboyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon", Information, Communication, & Society, vol.15 (5), pp. 662-679, 2012.

[21] J. Manyika., "Big data: The Next Frontier for Innovation, Competition, and Productivity",McKinsey Global Institute, pp. 1–137, 2011.

[22] Bill Hamilton, "Big Data is the Future of Healthcare", Cognizant white paper, 2010.

[23] A. Jacobs, "The Pathologies of Big Data", ACM Queue, vol. 7(6), pp. 1-12, 2009.

[24] Bryant R. E., Katz R. H., &Lazowska E. D., "Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society", A white paper prepared for the Computing Community Consortium committee of the Computing Research Association,2008.

Ch. Sridharababu completed M.Tech (CST) from Andhra University, Visakhapatnam, India and is currently working as Assistant Professor in Pydah College of Engineering and Technology, Visakhapatnam. His area of research interest includes Big Data, Machine learning, Deep learning.

K. Srinivasa Rao completed M.Tech (CST) from Andhra University, Visakhapatnam, India and is currently pursuing his PhD from Andhra University, Visakhapatnam. His area of research includes Big Data, Machine learning, Deep learning.