Big Data: The New Challenges in Data Mining

Mrs. Deepali Kishor Jadhav

Abstract—**Big Data is a new term used to identify the datasets but due to their large size and complexity, we cannot manage them with our current methodologies or data mining software tools. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This paper represents a broad overview of the topic, Big Data challenges, Data Mining Challenges with Big Data, Big Data processing framework and forecast to the future.**

Index Terms— **Big Data, New Challenges, Data Mining, Future Challenges, Big Data Problems.**

I. INTRODUCTION

Big Data is *too big, too fast,* or *too hard* for existing tools to process [1]. Here, "too big" means that organizations increasingly must deal with petabyte-scale collections of data that come from click streams, transaction histories, sensors, and elsewhere. "Too fast" means that not only is data big, but it must be processed. "Too hard" is a catchall for data that doesn't fit neatly into an existing processing tool or that needs some kind of analysis that existing tools can't readily provide.

Big Data is currently defined using three data characteristics: volume, variety and velocity [2]. At some point when the volume, variety and velocity of the data are increased, the current techniques and technologies may not be able to handle storage and processing of the data. At that point the data is defined as Big Data. The term Big Data Analytics is nothing but the process of analyzing and understanding the characteristics of massive size datasets by extracting useful geometric and statistical patterns. These three characteristics of a dataset increase the complexity of the data.

Many applications involve the Big Data problem, including network traffic risk analysis, geospatial classification and business forecasting. Network intrusion detection and prediction are time sensitive applications and they require highly efficient Big Data techniques and technologies to tackle the problem. In this paper some of the problems and challenges associated with the Big Data technologies and techniques are discussed.

The current definition of Big Data defined on a 3D space, V^3, formed by three parameters, volume, variety and velocity cannot provide a suitable platform for the early detection of Big Data characteristics for Big Data classification. Figure 1 shows the 3D space defined for Big V^3, formed by three parameters, volume, variety and velocity cannot provide a suitable platform for the early detection of Big Data characteristics for Big Data classification.

Figure 1 shows the 3D space defined for Big Data, where the axis of volume represents the growth of data size, the axis of velocity represents the increase in speed in which the data must be processed, and the axis of variety represents the increase in various types of data.

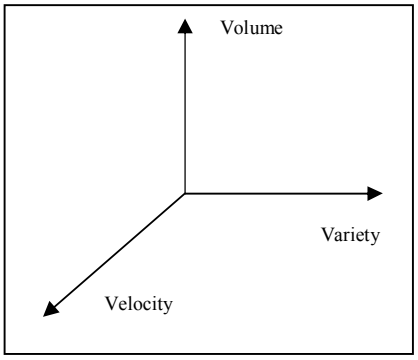

Figure 1: Current definition (V^3) of Big Data characteristics

A new definition for Big Data as a 3D space, C^3, as shown in Figure 2, which is defined based on three new parameters: cardinality, continuity, and complexity. In C^3 space the cardinality defines the number of records in the dynamically growing dataset at a particular instance. The continuity defines two characteristics and they are: (i) representation of data by continuous functions, and (ii) continuously growth of data size with respect to time. The complexity defines three characteristics and they are: (i) large varieties of data types, (ii) high dimensional dataset; and (iii) the speed of data processing are very high[3].

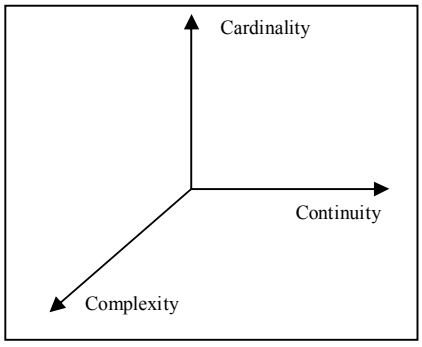

Figure 2: Proposed definition (C^3) of Big Data characteristics

Manuscript received November 19, 2013.
Mrs. Deepali Kishor Jadhav, Assistant professor, Department of Computer Science and Engineering, K.I.T.'s College of Engineering, Kolhapur. Maharahtra, India.(e-mail:Deepkjadhav80@gmail.com)

II. NEW CHALLENGES

While big data can yield extremely useful information, it also presents new challenges with respect to how much data to store, how much this will cost, whether the data will be secure, and how long it must be maintained.

A. *Storage Challenges:*

To cope with growing information volumes, organizations will increasingly use techniques such as data compression, duplication, object storage, and storage virtualization. Nonetheless, this process will continue to be a challenge. Data managers will thus have to become more discriminating about what to save and for how long, said the Taneja Group's Matchett.

It's also clear that current technologies won't be able to provide the necessary capacity or performance to handle the growing amount of data, said Micron's Kilbuck. This will require new storage approaches, and economics will affect the development of these technologies. Given the current volume of data, the cost per byte of various types of storage will be a primary driver in companies' decision making, noted Don Brown, senior architect and lead field engineer for big data applications vendor WibiData [4].

B. *Security Challenges:*

The security mechanism in cloud technology is generally weak. Hence tampering of data at the public cloud is predictable and it is a big concern. Finding a robust security mechanism for the purpose of using the public cloud like Cloud Computing Storage System (CCSS) is a challenging problem. In cloud technology, an attacker can easily tamper the data that is being exchanged between the CCSS server and the Hadoop Distributed File System (HDFS) and Network Traffic Recording System (NTRS) units; the attacker can spoof the reply between them and shut down the server (CCSS) using DOS attack [5]. These problems can lead to challenges in implementing Big Data analytics tool with the suggested network topology.

C. *Communication Challenges:*

In computer networking research and applications, the communication cost is the major concern compared to the processing cost of the data. The challenge here is to minimize that communication cost while satisfying the additional storage and data requirement from public cloud for processing Big Data. The bandwidth and latency are the two major network features that will affect the communication between the clients and the cloud server [6] and [7]. These problems and associated challenges to find solutions will adversely affect the timing requirements of the Big Data processing at HDFS and User Interaction and Learning System (UILS).

D. *Ethical Challenges:*

Big data also presents new ethical challenges. Corporations are using big data to learn more about their workforce, increase productivity, and introduce revolutionary business processes. However, these improvements come at a cost: tracking employees' every move and continuously measuring their performance against industry benchmarks introduces a level of oversight that can quash the human spirit. Such monitoring might be in the best interest of a corporation but is not always in the best interest of the people who make up that corporation.

In addition, as big multimedia datasets become common place, the boundaries between public and private space will blur. Emerging online applications will not only enable users to upload video via mobile social networking but will soon incorporate wearable devices in the form of a digital watch or glasses to allow for continuous audiovisual capture. People will essentially become a camera [18].

III. DATA MINING CHALLENGES WITH BIG DATA

For an intelligent learning database system to handle Big Data, the essential key is to scale up to the exceptionally large volume of data Figure 3 shows a conceptual view of the Big data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III).

The challenges at Tier I focus on data accessing and actual computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing. For example, while typical data mining algorithms require all data to be loaded into the main memory, this is becoming a clear technical barrier for Big Data because moving data across different locations is expensive (e.g., subject to intensive network communication and other IO costs), even if we do have a super large main memory to hold all data for computing.

The challenges at Tier II center on semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III). For example, depending on different domain applications, the data privacy and information sharing mechanisms between data producers and data consumers can be significantly different. In addition to the above privacy issues, the application domains can also provide additional information to benefit or guide Big Data mining algorithm designs. In a social network, on the other hand, users are linked and share dependency structures. The knowledge is then represented by user communities, leaders in each group, and social influence modeling etc. Therefore, understanding semantics and application knowledge is important for both low-level data access and for high level mining algorithm designs.

At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by complex and dynamic data characteristics. The circle at Tier III contains three stages. Firstly, sparse, heterogeneous, uncertain, incomplete, and multi-source data are preprocessed by data fusion techniques. Secondly, complex and dynamic data are mined after pre-processing. Thirdly, the global knowledge that is obtained by local learning and model fusion is tested and

relevant information is fed back to the pre-processing stage. Then the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing.

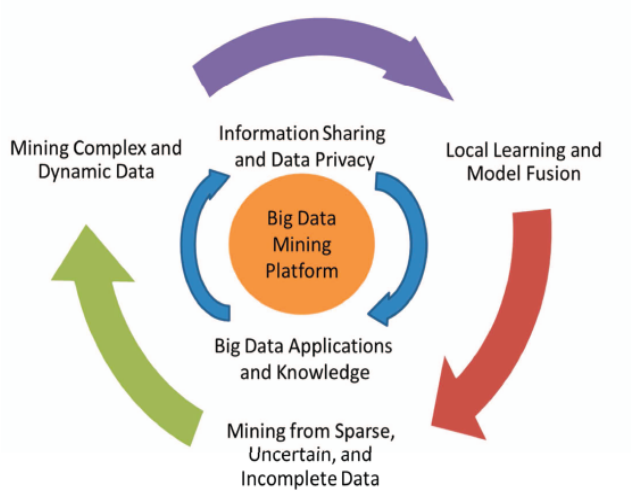

Figure 3: A Big Data processing framework

The research challenges form a three tier structure center around the "Big Data mining platform" (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high level semantics, application domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms [8].

IV. APPLICATIONS

While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation[8]. This is similar to using a number of data fields, such as age, gender, income, education background etc., to characterize each individual. This type of sample-feature representation inherently treats each individual as an independent entity without considering their social connections which is one of the most important factors of the human society. People form friend circles based on their common hobbies or connections by biological relationships. Such social connections commonly exist in not only our daily activities, but also are very popular in virtual worlds. For example, major social network sites, such as Facebook or Twitter, are mainly characterized by social functions such as friend connections and followers (in Twitter). The correlations between individuals inherently complicate the whole data representation and any reasoning process. In the sample-feature representation, individuals are regarded similar if they share similar feature values, whereas in the sample-feature-relationship representation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature domains at all. In a dynamic world, the features used to represent the individuals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. Such a complication is becoming part of the reality for Big Data applications, where the key is to take the complex (non-linear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

V. FUTURE CHALLENGES

There are many future important challenges in Big Data management and analytics, that arise from the nature of data: large, diverse, and evolving [9; 10; 11]. These are some of the challenges that researchers and practitioners will have to deal during the next years:

1) Analytics Architecture: It is not clear yet how an optimal architecture of an analytics system should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz [12]. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, extensible, allows ad hoc queries, minimal maintenance, and debuggable.
2) Statistical significance: It is important to achieve significant statistical results, and not be fooled by randomness. As Efron explains in his book about Large Scale Inference [14], it is easy to go wrong with huge data sets and thousands of questions to answer at once.
3) Distributed mining: Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.
4) Time evolving data: Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first. For example, the data stream mining fieldhas very powerful techniques for this task [15].
5) Compression: Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two main approaches: compression where we don't loose anything or sampling where we choose what is the data that is more representative. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling, we are losing information, but the gains in space may be in orders of magnitude. For example Feldman et al. [14] use coresets to reduce the complexity of Big Data problems. Coresets are small sets that provably approximate the

original data for a given problem. Using merge-reduce the small sets can then be used for solving hard machine learning problems in parallel.

6) Visualization: A main task of Big Data analysis is how to visualize the results. As the data is so big, it is very difficult to find user-friendly visualizations. New techniques, and frameworks to tell and show stories will be needed, as for example the photographs, infographics and essays in the beautiful book "The Human Face of Big Data" [17].
7) Hidden Big Data: Large quantities of useful data are getting lost since new data is largely untagged file based and unstructured data. The 2012 IDC study on Big Data [16] explains that in 2012, 23% (643 exa bytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

VI. CONCLUSION

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. This data is going to be more diverse, larger, and faster. We discussed in this paper a change to the basic definition V^3 of Big Data to C^3 so that the Big Data analytics maybe better explained and understood with mathematical and statistical techniques. Research on Big Data techniques and technologies evolving and at the same time new problems and challenges are emerging, hence the hope is to develop better and better techniques and technologies towards finding solutions for Big Data classification problem. We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before.

ACKNOWLEDGMENT

I would like to thank all those who contributed to this paper: Dr. Preeti Patil, Shivani Kale, Ranjita Pandhare and Yogita Narule for their valuable comments and sharing their knowledge. This work has been funded by K.I.T.'s College of Engineering, Kolhapur.

REFERENCES

[1] Sam Madden • Massachusetts Institute of Technology From Database to Big Data,IEEE,2012

[2] P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis, Understanding big data – Analytics for enterprise class Hadoop and streaming data, McGraw-Hill, 2012.

[3] Shan Suthaharan , Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning.

[4] Neal Leavitt,"Storage Challenges: Where Will Att That Big Data Go?", 0018-9162/13/$31.00 © , 2013 IEEE.

[5] I. Muttik and C. Barton. "Cloud security technologies," information security technical report 14.1: 1-6, 2009.

[6] S. Carlin and K. Curran. "Cloud Computing Technologies." International Journal of Cloud Computing and Services Science (IJ-CLOSER) 1.2: 59-65, 2012.

[7] P. C. Wong, H. W. Shen, C. R. Johnson, C. Chen, and R. B. Ross, "The Top 10 Challenges in Extreme-Scale Visual Analytics," Computer Graphics and Applications, IEEE, 32(4), 63-67, 2012.

[8] Xindong Wu,, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", 1041-4347/13, 2013 IEEE

[9] Wei Fan, Albert Bifet, " Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2

[10] V. Gopalkrishnan, D. Steier, H. Lewis, and J. Guszcza.Big data, big business: bridging the gap. In Proceed-ings of the 1st International Workshop on Big Data,Streams and Heterogeneous Source Mining: Algorithms,Systems, Programming Models and Applications, Big-Mine '12, pages 7{11, New York, NY, USA, 2012. ACM.

[11] C. Parker. Unexpected challenges in large scale machine learning. In Proceedings of the 1st International Work- shop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine '12, pages 1{6, New York,NY, USA, 2012. ACM.

[12] N. Marz and J. Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications, 2013.

[13] B. Efron. Large-Scale Inference: Empirical Bayes Meth- ods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010.

[14] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In SODA, 2013.

[15] J. Gama. Knowledge Discovery from Data Streams. Chapman & all/Crc data Mining and Knowledge Discovery. Taylor & Francis Group, 2010.

[16] J. Gantz and D. Reinsel. IDC: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. December 2012.

[17] R. Smolan and J. Erwitt. The Human Face of Big Data. Sterling Publishing Company Incorporated, 2012.

[18] Katina Michael, Keith W. Miller, " Big Data: New Opportunities and New Challenges", IEEE Computer Society 0018-9162/13/$31.00 © 2013 IEEE