ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Survey on Improving Privacy Preserving in Big Data Mining

Swapnil Shinde, Sandip.A.Kahate

Master of Engineering Student, Department of Computer Engineering, Sharadchandra Pawar College of Engineering, Dumbarwadi, Otur, Maharastra,India

Assistant Professor, Dept. of C.E., Sharadchandra Pawar College Of Engineering, Dumbarwadi, Otur, Maharastra, India

ABSTRACT: Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. I analyze the challenging issues in the data-driven model and also in the Big Data revolution.

KEYWORDS: Big Data, data mining, heterogeneity, autonomous sources, complex and evolving associations

I. INTRODUCTION

This Better Practice Guide aims to improve government agencies' competence in big data analytics by informing government agencies about the adoption of big data including:

- Identifying the business requirement for big data capability including advice to assist agencies identify where big data analytics might support improved service delivery and the development of better policy;
- Developing the capability including infrastructure requirements and the role of cloud computing, skills, business processes and governance;
- Considerations of information management in the big data context including assisting agencies in identifying high value datasets,advising on the government use of third party datasets, and the use of government data by third parties,promoting privacy by design,promoting Privacy Impact Assessments (PIA) and articulating peer review and quality assurance processes; and
- Big data project management including necessary governance arrangements for big data analytics initiatives.

Government agencies have extensive experience in the application of information management principles that currently guide data management and data analytics practices, much of that experience will continue to apply in a big data context.

This better practice guide is intended initially as an introductory and educative resource for agencies looking to introduce a big data capability and the specific challenges and opportunities that accompany such an implementation. Often there will be elements of experience with implementing and using big data to a greater or lesser degree across government agencies. In this guide we aim to highlight some of the changes that are required to bring big data into the mainstream of agencies operations. More practical guidance on the management of big data initiatives will be developed subsequent to this better practice guide as part of a guide to responsible data analytics.

As outlined greater volumes and a wider variety of data enabled by new technologies presents some significant departures from conventional data management practice. To understand these further we outline the meaning of big data and big data analytics contained and explore how this is different from current practice.

II. LITERATURE REVIEW

Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of Big Data is to carry out computing on the petabyte (PB), even theexa-byte (EB)-level data with a complex computing process. Therefore, utilizing a parallel computing

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

infrastructure, its corresponding programming language support, and software models to efficiently analyse and mine the distributed data are the critical goals for Big Data processing to change from "quantity" to "quality."Currently, Big Data processing mainly depends on parallel programming models like MapReduce, as well as providing a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of MapReduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, withMapReduce parallel programming being applied to many machine learning and data mining algorithms.

Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently. To improve the efficiency of algorithms,proposed a general-purpose parallel programming method, which is applicable to a large number of machine learning algorithms based on the simple MapReduce programming model on multicore processors. Ten classical data mining algorithms are realized in the framework, including locally weighted linear regression, k-Means, logistic regression, naive Bayes, linear support vectormachines, the independent variable analysis, Gaussian discriminant analysis, expectation maximization, and back-propagation neural networks [5].With the analysis of these classical machine learning algorithms, we argue that the computational operations in the algorithm learning process could be transformed into a summation operation on a number of training data sets. Summation operations could be performed on different subsets independently and achieve penalization executed easily on the MapReduce programming platform. Therefore, a large-scale data set could be divided into several subsets and assigned to multiple Mapper nodes. Then, various summation operations could be performed on the Mapper nodes to collect intermediate results. Finally, learning algorithms are executed in parallel through merging summation on Reduce nodes. Ranger et al. [1] proposed a MapReduce-based application programming interface Phoenix, which supports parallel programming in the environment of multicore and multiprocessor systems, and realized three data mining algorithms including k-Means, principal component analysis, and linear regression. In paper [3] improved the MapReduce's implementation mechanism in Hadoop, evaluated the algorithms' performance of single-pass learning, iterative learning, and query-based learning in the MapReduce framework, studied data sharing between computing nodes involved in parallel learning algorithms, distributed data storage, and then showed that the MapReduce mechanisms suitable for large-scale datamining by testing series of standard data mining tasks on medium-size clusters. Papadimitriou and Sun [2] proposed a distributed collaborative aggregation (DisCo)framework using practical distributed data pre-processing and collaborative aggregation techniques. The implementation on Hadoop in an open source MapReduce project showed that DisCo has perfect scalability and can process and analyse massive data sets (with hundreds of GB). To improve the weak scalability of traditional analysis software and poor analysis capabilities of Hadoop systems,

Das et al. conducted a study of the integration of R (open source statistical analysis software) and Hadoop. The in-depth integration pushes data computation to parallel processing, which enables powerful deep analysis capabilities for Hadoop.

The rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible.

The unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

III. PROPOSED ALGORITHM

The proposed a HACE theorem to model Big Data characteristics. The characteristics of HACH make it an extreme challenge for discovering useful knowledge from the Big Data.

The HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations.

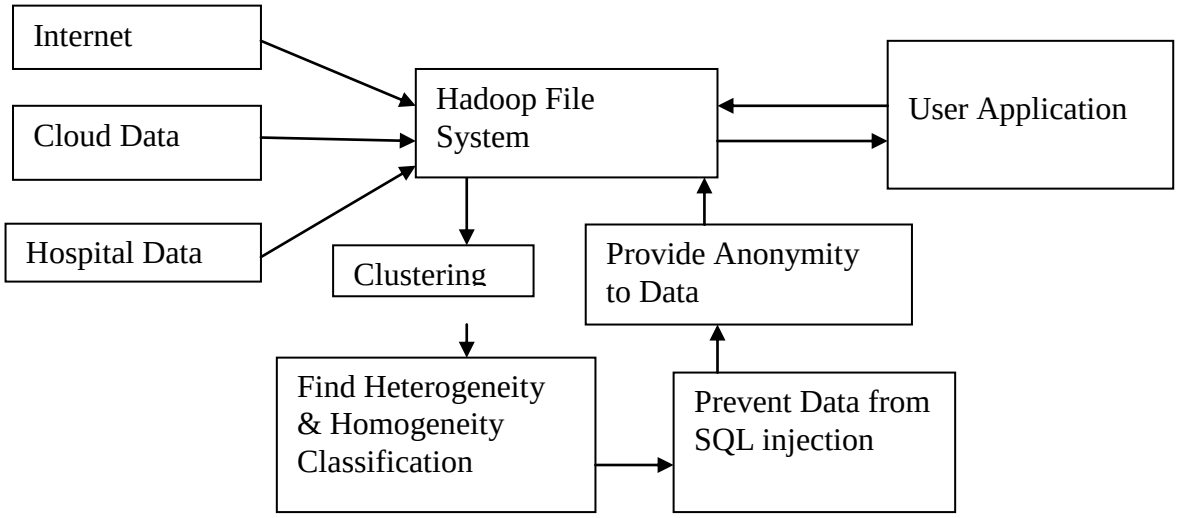
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data.



IV. PSEUDO CODE

HACE Theorem-

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.

These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data.

- **Huge Data with Heterogeneous and Diverse Dimensionality**

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors prefer their own schemata or protocols for data recording, and the nature of different applications also results in diverse data representations. For example, eachsingle human being in a biomedical world can be represented by using simple demographic information such as gender, age, family disease history, and so on. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. For a DNA or genomic-related test, microarray expression images and sequences are used to represent the genetic code information because this is the way that our current techniques acquire the data. Under such circumstances, the heterogeneous features refer to the different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each single observation. Imagine that different organizations (or health practitioners) may have their own schemata to represent each patient, the data heterogeneity and diverse dimensionality issues become major challenges if we are trying to enable data aggregation by combining data from all sources.

- **Autonomous Sources with Distributed and Decentralized Control**

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data-related applications, such as Google, Flicker, Facebook, and Walmart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technical designs, but also the results of the legislation and the regulation rules in different countries/ regions.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

- **Complex and Evolving Relationships**

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. This is similar to using a number of data fields, such as age, gender, income, education background, and so on, 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. Our friend circles may be formed based on the common hobbies or people are connected by biological relationships. Such social connections commonly exist not only in our daily activities, but also are very popular in cyber-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 on the data. 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 (nonlinear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

V. ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere gratitude to my Project GuideProf.S.A.Kahate (Assistant Professor, Computer Engineering Department) for his encouragement, guidance, and insight throughout the research and in the preparation of this dissertation. He truly exemplifies the merit of technical excellence and academic wisdom.

VI. CONCLUSION AND FUTURE WORK

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. Such combined characteristics suggest that Big Data require a “big mind” to consolidate data for maximum values.

To explore Big Data, I have analysed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyse model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future.

I regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, I will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real-time.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

REFERENCES

[1]Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE, "Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.
[2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.
[3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.
[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.
[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.
[6] T. Mitchell, "Mining our Reality," Science, vol. 326, pp. 1644-1645,2009.
[7] J. Mervis, "U.S. Science Policy: Agencies Rally to Tackle Big Data,"Science, vol. 336, no. 6077, p. 22, 2012
[8] D. Luo, C. Ding, and H. Huang, "Parallelization with Multiplicative Algorithms for Big Data Mining," Proc. IEEE 12th Int'lConf. Data Mining, pp. 489-498, 2012.