Big data analytics and its Security issues

Khan Mohammad Wafa[1*] and Feroz Ahmad Baloch[2]

[1]Lecturer and Dean of Faculty, Department of Information Technology, Computer Science Faculty, Bost University.

[2]Lecturer, Department of Information Technology, Computer Science Faculty, Bost University.

Email: khanmohammad_wafa@yahoo.com

Abstract

Data is currently one of the most important assets for companies in every field. The continuous growth in the importance and volume of data has created a new problem: it cannot be handled by traditional analysis techniques. This problem was, therefore, solved through the creation of a new paradigm: Big Data. However, Big Data originated new issues related not only to the volume or the variety of the data, but also to data security and privacy. The question that arises now is, how to develop a high performance platform to efficiently analyze big data and how to design an appropriate mining algorithm to find the useful things from big data. To deeply discuss this issue, this paper begins with a brief introduction to data analytics, followed by the discussions of big data analytics. Some important open issues and further research directions will also be presented for the next step of big data analytics. Over the last few years, data has become one of the most important assets for companies in almost every field. Not only are they important for companies related to the computer science industry, but also for organizations, such as countries' governments, healthcare, education, or the engineering sector. Data are essential with respect to carrying out their daily activities, and also helping the businesses' management to achieve their goals and make the best decisions on the basis of the information extracted from them.

Keywords: Big data, data analytics and data mining

Introduction

As the information technology spreads fast, most of the data were born digital as well as exchanged on internet today. According to the estimation of (Lyman and Varian, 2004) the new data stored in digital media devices have already been more than 92 % in 2002, while the size of these new data was also more than five Exabyte's. In fact, the problems of analyzing the large-scale data were not suddenly occurred but have been there for several years because the creation of data is usually much easier than finding useful things from the data. Even though computer systems today are much faster than those in the 1930s, the large-scale data is a strain to analyze by the computers we have today. In response to the problems of analyzing large-scale data, quite a few efficient methods (Xu, 2009), such as sampling, data condensation, density-based approaches, grid-based approaches, divide and conquer, incremental learning, and distributed computing, have been presented. Of course, these methods are constantly used to improve the performance of the operators of data analytics process. The results of these methods illustrate that with the efficient methods at hand, we may be able to analyze the large-scale data in a reasonable time. The dimensional reduction method (e.g., principal components analysis; PCA (Ding c, 2004) is a typical example that is aimed at reducing the input data volume to accelerate the process of data analytics. Another reduction method that reduces the data computations of data clustering is sampling (Kollios, 2003), which can also be used to speed up the computation time of data analytics.

Big data analytics:

Nowadays, the data that need to be analyzed are not just large, but they are composed of various data types, and even including streaming data (russom, 2011). Since big data has the unique features of "massive, high dimensional, heterogeneous, complex, unstructured, incomplete, noisy, and erroneous," which may change the statistical and data analysis approaches. Although it seems that big data makes it possible for us to collect more data to find more useful information, the truth is that more data do not necessarily mean more useful information. It may contain more ambiguous or abnormal data. For instance, a user may have multiple accounts, or an account may be used by multiple users, which may degrade the accuracy of the mining results (Boyd, 2012). Therefore, several new issues for data analytics come up, such as privacy, security, storage, fault tolerance, and quality of data.

The big data may be created by handheld device, social network, internet of things, multimedia, and many other new applications that all have the characteristics of volume, velocity, and variety. As a result, the whole data analytics has to be re-examined from the following perspectives:

From the volume perspective, the deluge of input data is the very first thing that we need to face because it may paralyze the data analytics. Different from traditional data analytics, for the wireless sensor network data analysis, (Baraniuk, 2011) pointed out that the bottleneck of big data analytics will be shifted from sensor to processing, communications, storage of sensing data, this is because sensors can gather much more data, but when uploading such large data to upper layer system, it may create bottlenecks everywhere.

In addition, from the velocity perspective, real-time or streaming data bring up the problem of large quantity of data coming into the data analytics within a short duration but the device and system may not be able to handle these input data. This situation is similar to that of the network flow analysis for which we typically cannot mirror and analyze everything we can gather.

From the variety perspective, because the incoming data may use different types or have incomplete data, how to handle them also bring up another issue for the input operators of data analytics.

Big data input:

The problem of handling a vast quantity of data that the system is unable to process is not a brand-new research issue; in fact, it appeared in several early approaches (agrawal, 1993) e.g., marketing analysis, network flow monitor, gene expression analysis, weather forecast, and even astronomy analysis. This problem still exists in big data analytics today; thus, preprocessing is an important

task to make the computer, platform, and analysis algorithm be able to handle the input data. The traditional data preprocessing methods (family, 1997),(e.g., compression, sampling, feature selection, and so on) are expected to be able to operate effectively in the big data age. However, a portion of the studies still focus on how to reduce the complexity of the input data because even the most advanced computer technology cannot efficiently process the whole input data by using a single machine in most cases. By using domain knowledge to design the preprocessing operator is a possible solution for the big data. the domain knowledge, B-tree, divide-and-conquer to filter the unrelated log information for the mobile web log analysis. A later study considered that the computation cost of preprocessing will be quite high for massive logs, sensor, or marketing data analysis. Sampling and compression are two representative data reduction methods for big data analytics because reducing the size of data makes the data analytics computationally less expensive, thus faster, especially for the data coming to the system rapidly. To avoid the application-level slow-down caused by the compression process, in (Jun et al., 2012) attempted to use the FPGA to accelerate the compression process. The I/O performance optimization is another issue for the compression method. For this reason, (Zou et al., 2014) employed the tentative selection and predictive dynamic selection and switched the appropriate compression method from two different strategies to improve the performance of the compression process. To make it possible for the compression method to efficiently compress the data, a promising solution is to apply the clustering method to the input data to divide them into several different groups and then compress these input data according to the clustering information.

In summary, in addition to handling the large and fast data input, the research issues of heterogeneous data sources, incomplete data, and noisy data may also affect the performance of the data analysis. The input operators will have a stronger impact on the data analytics at the big data age than it has in the past. As a result, the design of big data analytics needs to consider how to make these tasks (e.g., data clean, data sampling, data compression) work well.

Big data analysis frameworks and platforms

most of the studies on the traditional data analysis are focused on the design and development of efficient and/or effective “ways” to find the useful things from the data. But when we enter the age of big data, most of the current computer systems will not be able to handle the whole dataset all at once; thus, how to design a good data analytics framework or platform and how to design analysis methods are both important things for the data analysis process. In this section, we will start with a brief introduction to data analysis frameworks and platforms, followed by a comparison of them.

Comparison between the frameworks/platforms of big data:

(Talia, 2013) pointed out that cloud-based data analytics services can be divided into data analytics software as a service, data analytics platform as a service, and data analytics infrastructure as a service. A later study (Lu et al., 2014) presented a general architecture of big data analytics which contains multi-source big data collecting, distributed big data storing, and intra/inter big data processing. Since many kinds of data analytics frameworks and platforms have been presented, some of the studies attempted to compare them to give a guidance to choose the applicable frameworks or platforms for relevant works. To give a brief introduction to big data analytics, especially the platforms and frameworks,

In (Hu et al., 2014), in addition to defining that a big data system should include data generation, data acquisition, data storage, and data analytics modules, Hu et al. also mentioned that a big data system can be decomposed into infrastructure, computing, and application layers. Moreover, a promising research for NoSQL storage systems was also discussed in this study which can be divided into key-value, column, document, and row databases. Since big

data analysis is generally regarded as a high computation cost work, the high performance computing cluster system (HPCC) is also a possible solution in early stage of big data analytics.

Big data analysis algorithms:

Clustering algorithms In the big data age, traditional clustering algorithms will become even more limited than before because they typically require that all the data be in the same format and be loaded into the same machine so as to find some useful things from the whole data. Although the problem (chiang et al., 2011) of analyzing large-scale and high-dimensional dataset has attracted many researchers from various disciplines in the last century, the characteristics of big data still brought up several new challenges for the data clustering issues. Among them, how to reduce the data complexity is one of the important issues for big data clustering. In,(Shirkhorshidi et al., 2014) divided the big data clustering into two categories: single-machine clustering (i.e., sampling and dimension reduction solutions), and multiple-machine clustering (parallel and Map Reduce solutions). This means that traditional reduction solutions can also be used in the big data age because the complexity and memory space needed for the process of data analysis will be decreased by using sampling and dimension reduction methods. More precisely, sampling can be regarded as reducing the "amount of data" entered into a data analyzing process while dimension reduction can be regarded as "downsizing the whole dataset" because irrelevant dimensions will be discarded before the data analyzing process is carried out.

Cloud Vista (Xu et al., 2012) is a representative solution for clustering big data which used cloud computing to perform the clustering process in parallel. Classification algorithms Similar to the clustering algorithm for big data mining, several studies also attempted to modify the traditional classification algorithms to make them work on a parallel computing environment or to develop new classification algorithms which work naturally on a parallel computing environment. In (Tekin and van, 2013) the design of classification algorithm took into account the input data that are gathered by distributed data sources and they will be processed by a heterogeneous set of learners. In this study, presented a novel classification algorithm called "classify or send for classification" (CoS). They assumed that each learner can be used to process the input data in two different ways in a distributed data classification system. One is to perform classification function by itself while the other is to forward the input data to another learner to have them labeled. The information will be exchanged between different learners. In brief, this kind of solutions can be regarded as a cooperative learning to improve the accuracy in solving the big data classification problem. An interesting solution uses the quantum computing to reduce the memory space and computing cost of a classification algorithm

Frequent pattern mining algorithms Most of the researches on frequent pattern mining (i.e., association rules and sequential pattern mining) were focused on handling large-scale dataset at the very beginning because some early approaches of them were attempted to analyze the data from the transaction data of large shopping mall. Because the number of transactions usually is more than "tens of thousands", the issues about how to handle the large scale data were studied for several years, such as FP-tree (han et al., 2000) using the tree structure to include the frequent patterns to further reduce the computation time of association rule mining. In addition to the traditional frequent pattern mining algorithms, of course, parallel computing and cloud computing technologies have also attracted researchers in this research domain.

Summary of process of big data analytics:

This discussion of big data analytics in this section was divided into input, analysis, and output for mapping the data analysis process of KDD. For the input (see also in "Big data input") and output (see also "Output the result of big data analysis") of big data, several methods and solutions proposed before the big data age (see also "Data input") can also be employed for big data analytics in most cases.

However, there still exist some new issues of the input and output that the data scientists need to confront. A representative example we mentioned in "Big data input" is that the bottleneck will not only on the sensor or input devices, it may also appear in other places of data analytics (baraniuk, 2011). Although we can employ traditional compression and sampling technologies to deal with this problem, they can only mitigate the problems instead of solving the problems

completely. Similar situations also exist in the output part.

Although several measurements can be used to evaluate the performance of the frameworks, platforms, and even data mining algorithms, there still exist several new issues in the big data age, such as information fusion from different information sources or

information accumulation from different times.

Several studies attempted to present an efficient or effective solution from the perspective of system (e.g., framework and platform) or algorithm level. A simple comparison of these big data analysis technologies from different perspectives is described in Table 1, to give a brief introduction to the current studies and trends of data analysis technologies for the big data. The "Perspective" column of this table explains that the study is focused on the framework or algorithm level; the "Description" column gives the further goal of the study; and the "Name" column is an abbreviated name of the methods or platform/framework. From the analysis framework perspective, this table shows that big data framework, platform, and machine learning are the current research trends in big data analytics system. For the mining algorithm perspective, the clustering, classification, and frequent pattern mining issues play the vital role of these researches because several data analysis problems can be mapped to these essential issues.

Table 1: The big data analysis frameworks and methods

P	Name	References	Year	Description	T
Analysis framework	DOT	[88]	2011	Add more computation resources via scale out solution	Framework
	GLADE	[89]	2011	Multi-level tree-based system architecture	
	Starfish	[92]	2012	Self-tuning analytics system	
	ODT-MDC	[96]	2012	Privacy issues	
	MRAM	[91]	2013	Mobile agent technologies	
	CBDMASP	[94]	2013	Statistical computation and data mining approaches	
	SODSS	[97]	2013	Decision support system issues	
	BDAF	[93]	2014	Data centric architecture	
	HACE	[95]	2014	Data mining approaches	
	Hadoop	[83]	2011	Parallel computing platform	Platform
	CUDA	[84]	2007	Parallel computing platform	
	Storm	[85]	2014	Parallel computing platform	
	Pregel	[125]	2010	Large-scale graph data analysis	
	MLPACK	[86]	2013	Scalable machine learning library	ML
	Mahout	[87]	2011	Machine-learning algorithms	
	MLAS	[124]	2012	Machine-learning algorithms	

	PIMRU	[124]	2012	Machine Learning algorithms	
	Radoop	[129]	2011	Data analytics, machine learning algorithms, and R statistical tool	
Mining algorithm	DBDC	[144]	2004	Parallel clustering	CLU
	PKM	[145]	2009	Map-reduce-based *k*-means clustering	
	CloudVista	[111]	2012	Cloud computing for clustering	
	MSFCUDA	[113]	2013	GPU for clustering	
	BDCAC	[127]	2013	Ant on grid computing environment for clustering	
	Corest	[114]	2013	Use a tree construction for generating the coresets in parallel for clustering	
	SOM-MBP	[126]	2013	Neural network with CGP for classification	CLA
	CoS	[115]	2013	Parallel computing for classification	
	SVMGA	[72]	2014	Using GA for reduce the number of dimensions	
	Quantum SVM	[116]	2014	Quantum computing for classification	
	DPSP	[121]	2010	Applied frequent pattern algorithm to cloud platform	FP
	DHTRIE	[120]	2011	Applied frequent pattern algorithm to cloud platform	
	SPC, FPC, and DPC	[117]	2012	Map-reduce model for frequent pattern mining	
	MFPSAM	[119]	2014	Concerned the specific interest constraints and applied map-reduce Model	

Security issues:

Since much more environment data and human behavior will be gathered to the big data analytics, how to protect them will also be an open issue because without a security way to handle the collected data, the big data analytics cannot be a reliable system. In spite of the security that we have to tighten for big data analytics before it can gather more data from everywhere, the fact is that until now, there are still not many studies focusing on the security issues of the big data analytics. According to our observation, the security issues of big data analytics can be divided into fourfold: input, data analysis, output, and communication with other systems. For the input, it can be regarded as the data gathering which is relevant to the sensor, the handheld devices, and even the devices of internet of things. One of the important security issues on the input part of big data analytics is to make sure that the sensors will not be compromised

by the attacks. For the analysis and input, it can be regarded as the security problem of such a system. For communication with other system, the security problem is on the communications between big data analytics and other external systems. Because of these latent problems, security has become one of the open issues of big data analytics (Raghav et al., 2015).

Privacy issues:

The privacy concern typically will make most people uncomfortable, especially if systems cannot guarantee that their personal information will not be accessed by the other people and organizations. Different from the concern of the security, the privacy issue is about if it is possible for the system to restore or infer personal information from the results of big data analytics, even though the input data are anonymous. The privacy issue has become a very important issue because the

data mining and other analysis technologies will be widely used in big data analytics; the private information may be exposed to the other people after the analysis process. For example, although all the gathered data for shop behavior are anonymous (e.g., buying a pistol), because the data can be easily collected by different devices and systems (e.g., location of the shop and age of the buyer), a data mining algorithm can easily infer who bought this pistol. More precisely, the data analytics is able to reduce the scope of the database because location of the shop and age of the buyer provide the information to help the system find out possible persons. For this reason, any sensitive information needs to be carefully protected and used. The anonymous, temporary identification, and encryption are the representative technologies for privacy of data analytics, but the critical factor is how to use, what to use, and why to use the collected data on big data analytics (Raghav et al., 2015).

Conclusion

In this paper, we reviewed studies on the data analytics from the traditional data analysis to the recent big data analysis. From the system perspective, the KDD process is used as the framework for these studies and is summarized into three parts: input, analysis, and output. From the perspective of big data analytics framework and platform, the discussions are focused on the performance-oriented and results-oriented issues. From the perspective of data mining problem, this paper gives a brief introduction to the data and big data mining algorithms which consist of clustering, classification, and frequent patterns mining technologies. To better understand the changes brought about by the big data, this paper is focused on the data analysis of KDD from the platform/framework to data mining. The open issues on computation, quality of end result, security, and privacy are then discussed to explain which open issues we may face. Last but not least, to help the audience of the paper find solutions to welcome the new age of big data, the possible high impact research trends are given below:

For the computation time, there is no doubt at all that parallel computing is one of the important future trends to make the data analytics work for big data, and consequently the technologies of cloud computing, Hadoop, and map-reduce will play the important roles for the big data analytics. To handle the computation resources of the cloudbased platform and to finish the task of data analysis as fast as possible, the scheduling method is another future trend.

Using efficient methods to reduce the computation time of input, comparison, sampling, and a variety of reduction methods will play an important role in big data analytics. Because these methods typically do not consider parallel computing environment, how to make them work on parallel computing environment will be a future research trend. Similar to the input, the data mining algorithms also face the same situation that we mentioned in the previous section, how to make them work on parallel computing environment will be a very important research trend because there are abundant research results on traditional data mining algorithms.

How to model the mining problem to find something from big data and how to display the knowledge we got from big data analytics will also be another two vital future trends because the results of these two researches will decide if the

data analytics can practically work for real world approaches, not just a theoretical stuff.

References

Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Proc ACM SIGMOD Int Conf Manag Data. 1993;22(2):207–16.

Baraniuk RG. More is less: signal processing and the data deluge. Science. 2011;331(6018):717–9.

Boyd D, Crawford K. Critical questions for big data. Inform Commun Soc. 2012;15(5):662–79.

Chiang M‑C, Tsai C‑W, Yang C‑S. A time‑efficient pattern reduction algorithm for k‑means clustering. Inform Sci. 2011;181(4):716–31.

Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the Twenty-first Interna- tional Conference on Machine Learning, 2004, pp 1–9.

Famili A, Shen W‑M, Weber R, Simoudis E. Data preprocessing and intelligent data analysis. Intel Data Anal. 1997;1(1–4):3–23.

Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In : Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000. pp. 1–12.

Hu H, Wen Y, Chua T‑S, Li X. Toward scalable systems for big data analytics: a technology tutorial. IEEE Access. 2014;2:652–87.

Jun SW, Fleming K, Adler M, Emer JS. Zip‑io: architecture for application‑specific compression of big data. In: Proceedings of the International Conference on Field‑Programmable Technology, 2012, pp 343–351.

Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15(5):1170–87.

Lu R, Zhu H, Liu X, Liu JK, Shao J. Toward efficient and privacy‑preserving computing in big data era. IEEE Netw. 2014;28(4):46–50.

Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. Available: http://www2.sims.berkeley. edu/research/projects/how-much-info-2003/printable_report.pdf.

Russom P. Big data analytics. TDWI: Tech. Rep ; 2011.

Shirkhorshidi AS, Aghabozorgi SR, Teh YW, Herawan T. Big data clustering: a review. In: Proceedings of the International Conference on Computational Science and Its Applications, 2014. pp 707–720.

Talia D. Clouds for scalable big data analytics. Computer. 2013;46(5):98–101.

Tekin C, van der Schaar M. Distributed online big data classification using context information. In: Proceedings of the Allerton Conference on Communication, Control, and Computing, 2013. pp 1435–1442.

Xu H, Li Z, Guo S, Chen K. Cloudvista: interactive and economical visual cluster analysis for big data in the cloud. Proc VLDB Endowment. 2012;5(12):1886–9.

Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.

Zou H, Yu Y, Tang W, Chen HM. Improving I/O performance with adaptive data compression for big data applications. In: Proceedings of the International Parallel and Distributed Processing Symposium Workshops, 2014. pp 1228–1237.

Raghav Toshniwal, Kanishka Ghosh Dastidar, Asoke Nath. Big Data Security Issues and Challenges, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2 (February 2015

د غټو معلوماتو تحلیلونه او د هغوی امنیتي مسئلې

خان محمد وفا[1*] او فیروز احمد بلوچ[2]

۱،۲ معلوماتي ټکنالوژۍ څانګه، کمپیوټرساینس پوهنځی، بُست پوهنتون

مسؤل ایمیل ادرس: khanmohammad_wafa@yahoo.com

لنډیز

د غټو معلوماتو زمانه اوس راتلونکي ده. مګر د پخواني معلوماتو تحلیل نه سي کولای چې د معلوماتو دومره لویه اندازه اداره کړي. هغه پوښتنه چې اوس راپیدا کیږي دا ده، چې څنګه د ښه اجراتو پلیټ فارم یا تګلاره جوړه کړو ترڅو په ګټوره توګه غټه ډاټا تحلیل کړو او څنګه د کان جوړولو یو مناسب الګوریتم ډیزاین کړو تر څو د غټو ډاټا څخه ګټور شیان پیدا کړو. ددې موضوع د ژور بحث لپاره، دا پاڼه د معلوماتو د تحلیلونو په لنډې پېژندګلوی سره پیل کیږی او بیا د غټو ډیټا تحلیلونو خبرې کیږي. ځینې مهمې خلاصې مسئلې او د څېړنې نور جهتونه به هم د غټو معلوماتو د تحلیلونو د بل ګام لپاره وړاندې سي.

کلیدي کلیمې: لوی معلومات، د معلوماتو تحلیل او د معلوماتو کان کیندنه