Journal of Advances in Science and Technology
Vol. 13, Issue No. 1, (Special Issue) March-2017, ISSN 2230-9659

Big Data Analysis Solutions A Review

Priyanka P. Shinde*

Department of Computer Application, Government College of Engineering Karad

Abstract – Big Data analytics plays a key role in reducing the data size and complexity in Big Data applications. Visualization is an important approach to helping Big Data get a complete view of data and discover data values. Big Data analytics and visualization should be integrated seamlessly so that they work best in Big Data applications. Conventional data visualization methods, as well as the extension of some conventional methods to Big Data applications, are introduced in this paper. The challenges of Big Data visualization are discussed. New methods, applications, and technology progress of Big Data visualization are presented.

Keywords— Big Data, Visualization, Interactive Visualization, Virtual Reality, Networks, Cloud Computing, Information Technology, Telecommunication Systems

---------------------------♦-----------------------------

INTRODUCTION

As the information technology spreads fast, most of the data were born digital as well as exchanged on the internet today. According to the estimation of Lyman and Varian[1], 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 exabytes. 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[2], 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.[1]

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 [3] 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 [5], which can also be used to speed up the computation time of data analytics.

Although the advances of computer systems and internet technologies have witnessed the development of computing hardware following Moore's law for several decades, the problems of handling the large-scale data still exist when we are entering the age of big data. That is why Fisher et al. [5] pointed out that big data means that the data is unable to be handled and processed by most current information systems or methods because data in the big data era will not only become too big to be loaded into a single machine, it also implies that most traditional data mining methods or data analytics developed for a centralized data analysis process may not be able to be applied directly to big data. In addition to the issues of data size, Laney [6] presented a well-known definition (also called 3Vs) to explain what is the “big” data: volume, velocity, and variety. The definition of 3Vs implies that the data size is large, the data will be created rapidly, and the data will exist in multiple types and captured from different sources, respectively. Later studies [7,8] pointed out that the definition of 3Vs is insufficient to explain the big data we face now. Thus, veracity, validity, value, variability, venue, vocabulary and vagueness were added to make some complement explanation of big data [8].

LITERATURE REVIEW

Land set, Sara, et al (2015) [9] proposed big data-driven solutions to predict the 30-day risk of readmission for congestive heart failure (CHF) incidents. In this technique, extract useful factors from National Inpatient Dataset (NIS) and augment it with patient dataset from Multicare Health System (MHS). Then, develop scalable data mining models to predict a risk of readmission using the integrated dataset. Effectiveness and efficiency of the open-source predictive modeling framework describe the results from various modeling algorithms tested and compare the performance against baseline non-distributed, non-parallel, non-integrated small data results previously published to demonstrate comparable accuracy over millions of records[9].

Siriweera, T. H. A. S., et al (2015) [10] proposed a solution reference that gives guidance to organizations that want to innovate using big data technology and predictive analytics techniques for improving their performance. The reference architecture is the result of an iteration of Hevner's framework for designing information [10].

Ambade, Suhas V., and Priya R. Deshpande [11] (2015) proposed solution presents six software solutions for big data of different types of applications including statistical analysis, data acquisition social networks, power grid, healthcare, and telecom applications. It demonstrates for each solution the reasons behind choosing Map Reduce, the structure of datasets, description of the evaluation environment, software and hardware, the method of the application implementation on Map Reduce, and the evaluation results [11].

Glover, Fred, and Manuel Laguna [12] (2013) Proposed optimal route discovery (TORD), which is an NP-hard problem. Since the rise of GPS-equipped mobile devices has led to the emergence of big trajectory data, we propose a hybrid evaluation indicator of routes called preference combined by the distance from the road network and popularity from trajectories. To achieve this goal, they first develop an improved compact minimum bound rectangle algorithm to get the association between trajectory data and road network, which can update the association conveniently for new trajectories. Then, construct a two-layer preference network that is used for mining the preference and travel time. Finally, devise three approximation algorithms to answer TORD queries. The results of empirical studies show that all the proposed algorithms are capable of answering TORD queries efficiently [12].

Ordonez, Carlos, et al. [13] (2014) proposed solution presents a new data model aimed at solving this issue. Starting from the well-known E-R model, technique introduce some additional components to identify data and ”big data” in the system, in order to drive the implementation of SQL-like solutions to manage data, and No-SQL solutions to manage big data. The paper also discusses a Hospital Information System use case, to clearly show how the proposed enriched E-R model can be successfully adopted [13].

Yan, Zheng, et al. (2016) [14] proposed analytic agriculture framework that identify disease based on symptoms similarity and recommend a solution based on high similarity. To achieve this objective Hadoop and Hive tools has been used. The data is collected, cleansed and normalized. Data is collected from laboratory reports, websites, etc. then cleansing of data is done that is important information is extracted from unstructured redundant data. In the next step, normalization is done and features are extracted from cleaned data. Normalized data is uploaded on HDFS and saved in a file supported by the hive. HiveQL is an SQL-like query language and used to analyze the agricultural data. It finds out disease name based on crop disease symptoms and purposes a solution based on evidence from historical data. The result is represented in the form of graphs that will be useful for recommending a solution that is highly used for high symptoms similarity [14].

Jiang, Fan (2016) [15] proposed distributed dictionary learning framework based on rank-1 matrix decomposition with sparseness constraint (D-r1DL framework). The framework was implemented using the spark distributed computing engine and deployed on three different processing units: an in-house server, in-house high-performance clusters, and the Amazon Elastic Compute Cloud (EC2) service. The whole analysis pipeline was integrated with our neuroinformatics system for data management, user input /output, and real-time visualization. Performance and accuracy of D-r1DL on both individual and group-wise fMRI Human Connectome Project (HCP) dataset shows that the proposed framework is highly scalable. The Resulting group-wise functional network decompositions are highly accurate, and the fast processing time confirm this claim. In addition, D-r1DL can provide real-time user feedback and results from visualization which is vital for large-scale data analysis [15].

Cuzzocrea, Alfredo, Fan Jiang, and Carson Kai-Sang Leung [16] (2015) proposed a data science solution that uses Map Reduce to mine uncertain big data for frequent patterns satisfying user-specified anti-monotonic constraints. Experimental results show the effectiveness of our data science solution for mining interesting patterns from uncertain big data[16].

Yang, Zhi, et al [17] (2014) proposed solution, according to business logic and hardware configuration of cluster nodes, the data deployment strategy can be established. Then, the deployment scheme can be implemented with interface operation. Lastly, cluster nodes load data according to the deployment scheme. The solution has been applied to the Objectification Parallel Computing (OPC). The application result shows that OPC can achieve the best performance which can meet the demand of system efficiency and the operation of data deployment is simple [17].

Ahmed, Ejaz, et al. [18] (2015) proposed a hybrid entropy method combined with Analytical Hierarchical Process (AHP) to select appropriate cloud solution to manage big data projects in group decision-making environment. In order to collate individual opinions of decision makers for rating the importance of various criteria and alternatives, we employed usability analysis using the proposed hybrid AHP-Entropy method [18].

DISCUSSION

Semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive, and Pig as well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems.

CONCLUSIONS

This big data provides the enterprise with more choices because of its lots of related technologies and tools, which will continue to be developed and become innovative hotspots in the future, such as Hadoop distribution, the next generation of data warehouse, advanced data visualization, etc. In recent years, academia pays more attention to cloud computing. Big data focuses on "data", like data service, data acquisition, analysis, and data mining, which pays more attention to the ability of data storage. Cloud computing focuses on computing architecture and practices. Big data and cloud computing are two sides of the same issue. It is more accurate to analyze and forecast big data by using cloud computing and release more hidden value of data; in order to meet the service demand of big data, we can find even better practical applications to the cloud computing. Nowadays, more and more enterprises hope that they can transfer their own applications and infrastructures to the cloud platform. Cloud computing brings great changes to the big data. First, cloud computing provides a quite cheap storage place for the big data and makes medium-sized and small enterprises complete big data analysis. Second, cloud heterogeneous system to process data accurately. Although this paper clearly has not resolved the entire subject about this substantial topic, hopefully, it has provided some useful discussion and a framework for researchers. Computing has huge IT resources, distributes widely and becomes an effective way for enterprises.

REFERENCES

[1] 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.

[2] Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009. Google Scholar

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

[4] 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. View ArticleGoogle Scholar

[5] Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9. View ArticleGoogle Scholar

[6] Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001.

[7] [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[8] van Rijmenam M. Why the 3v's are not sufficient to describe big data, BigData-Startups, Tech. Rep. 2013. [Online]. Available: http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/.

[9] big data v's, Tech. Rep. 2014. [Online]. Available: https://www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.

[10] Landset, Sara, et al. "A survey of open source tools for machine learning with big data in the Hadoop ecosystem." Journal of Big Data 2.1 (2015): 24.

[11] Ambade, Suhas V., and Priya R. Deshpande. "Heterogeneity-Based Files Placement in

Hadoop Cluster." Computational Intelligence and Communication Networks (CICN), 2015 International Conference on. IEEE, 2015.

[12] Siriweera, T. H. A. S., et al. "Intelligent big data analysis architecture based on automatic service composition." Big Data (BigData Congress), 2015 IEEE InternationalConference on. IEEE, 2015. Glover, Fred, and Manuel Laguna. Tabu Search∗. Springer New

[13] Ordonez, Carlos, et al. "Extending ER models to capture database transformations to build data sets for data mining." Data & Knowledge Engineering 89 (2014): 38-54.

[14] Yan, Zheng, et al. "Deduplication on encrypted big data in a cloud." IEEE Transactions on Big Data 2.2 (2016): 138-150.

[15] Jiang, Fan. "Efficient frequent pattern mining from big data and its applications." (2014).

[16] Cuzzocrea, Alfredo, Fan Jiang, and Carson Kai-Sang Leung. "Frequent Subgraph Mining from Streams of Linked Graph Structured Data." EDBT/ICDT Workshops. 2015.

[17] Yang, Zhi, et al. "A Real-Time Distributed Query Solution Based on the OPC." Cloud Computing and Big Data (CCBD), 2014 International Conference on. IEEE, 2014.

[18] Ahmed, Ejaz, et al. "Network-centric performance analysis of runtime application migration in mobile cloud computing." Simulation Modelling Practice and Theory 50 (2015): 42-56.

[19] Kabir G. Kharade, Asawari A. Patwardhan, “Enhancing the Effectiveness of Public Transport by Implementing IVRS Technology”, at 3rd International Conference on “Dynamics of Business in Emerging markets(INCONRIT-2014)”, 21st& 22nd February,2014

[20] Kabir G. Kharade, S.K.Kharade, “Comparative Analysis of RSA and DES Algorithms” in “International Journal of Advanced Computer Technology & Management (IJACTM)”, ISSN: 2343-662X, Volume: I, Issue: I May.2016 Page 20-22

Corresponding Author

Priyanka P. Shinde*

Department of Computer Application, Government College of Engineering Karad

E-Mail – priyanka.shinde@gcekarad.ac.in