ISSN:0975-9646

A Survey on Big Data Analytics Using Hadoop Ecosystem Tools

Monika Yadav,

All Saints' of technology, Bhopal

Sonal Chaudhary

All Saints' of technology, Bhopal

Abstract-Big data is the term for any assortment of data sets thus massive and complex that it becomes tough to process using traditional processing applications. The challenges include analysis, capture, curation, search, sharing, storage, transfer, image, and privacy violations. The trend to larger data sets is because of the extra info derived from analysis of one massive set of connected data, as compared to separate smaller sets with identical total quantity of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime so on." huge data is difficult to figure with exploitation most electronic database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or perhaps thousands of servers". Big data sometimes includes data sets with sizes on the far side the power of commonly used software tools to capture, curate, manage, and process data inside a tolerable time period. huge data "size" is a perpetually moving target, as of its starting from many dozen terabytes to several petabytes of data. huge data may be a set of techniques and technologies that need new varieties of integration to uncover massive hidden values from massive datasets that are various, complex, and of an enormous scale. Big data environment is employed to amass, organize and analyze the various varieties of data. There is an observation concerning Map Reduce framework that framework generates great deal of intermediate data. Therefore, in addition because the tasks finishes there is would like of throwing that rich data, because MapReduce is unable to utilize them.

INTRODUCTION

We produce a pair of 2.5 quintillion bytes of data — such a lot that 90 percent of the information within the world nowadays has been created within the last two years alone. This abundant quantity of data comes from everywhere: sensors accustomed gather climate data, posts to social media sites, digital photos and videos, purchase dealing records, and cellular phone GPS signals to name a few. This immense quantity of the information is understood as "Big data"[14]. massive information could be a meaninglessness, or catch-phrase, utilizes to describe an enormous volume of each structured and unstructured data that's thus immense that it's difficult to process exploitation traditional database and software techniques. In most enterprise situations the data is just too giant or it moves too quick or it exceeds current process capability. Big data has the potential to assist organizations to boost operations and create quicker, a lot of intelligent decisions[15]. Big Data, currently a days this term becomes common in IT industries. As there's an enormous quantity of data lies within the industry however there's nothing before massive information comes into picture [3]. massive data is really an evolving term that describes any voluminous quantity of structured, semistructured and unstructured data that has the potential to be mined for data. though massive data does not visit any specific amount, thus this term is commonly used once speaking concerning petabytes and exabytes of data[16]. Big data is an broad term for big assortment of the data sets thus this immense and complicated that it becomes troublesome to operate them using traditional processing applications. once handling larger datasets, organizations face difficulties in having the ability to make, manipulate, and manage massive data. massive data is especially a problem in business analytics as a result of normal tools and procedures don't seem to be designed to look and analyze large datasets.

The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional info derived from analysis of one large set of connected information, as compared to separate smaller sets with an equivalent total quantity of information, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on"[10].Scientists often encounter limitations because of giant data sets in several areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. the constraints additionally have an effect on web search, finance and business information processing. data sets grow in size partly because they're more and more being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software package logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensing element networks. Big data outlined as way back as 2001, analyst Doug Lucy Craft Laney (currently with Gartner) articulated the currently mainstream definition of huge data because the 3 Vs of big data: volume, velocity and variety [18]. massive data is characterized by welknown 3Vs: the acute volume of data, the big variety of kinds of data and therefore the rate at which the data should be must processed. though massive data doesn't refer to any specific amount, the term is commonly used when speaking regarding petabytes and exabytes of data, much of that cannot be integrated simply. [16]

PREVIOUS WORK IN THE FIELD OF MAP/REDUCE

DumitrelLoghinet. al. 2015 presents a time–energy performance analysis of MapReduce on heterogeneous systems with GPUs. To execute MapReduce on heterogeneous systems with GPUs, we introduce a novel lazy processing technique which simplifies application development and requires no modifications to the underlying Hadoop framework. Based on this experiment, the wimpy (performance improvements in low-power) nodes achieve similar execution times compared to a single brawny node and also exhibit energy savings of up to two-thirds.

RazvanNituet. al. 2014 proposes An Improved GPU MapReduce Framework for Data Intensive Applications. This framework improves the MapReduce performance by adding GPU capabilities by implementing a hybrid CPU-GPU framework for heterogeneous environments. All the functionalities regarding GPU programming are already implemented. The users just have to define the functions specific to the MapReduce paradigm, without having advanced knowledge about GPU programming. The GPU tasks are implemented using the OpenCL library. Since Hadoop is written in Java, we used the JOCL (OpenCL Java language binding) solution to integrate these two languages.

Can Basaran et.al 2013 present a new MapReduce framework, called Grex (a new GPU-based MapReduce framework), designed to leverage general purpose graphics processing units (GPUs) for parallel data processing. The experimental results show that our system is up to 12.4× and 4.1× faster than two state-of-the-art GPU-based MapReduce frameworks for the tested applications.

Miao Xinet. al. 2012 presents an approach of MapReduce improvement with GPU acceleration, which is implemented by Hadoop and OpenCL. As a heterogeneous multi-machine and multicore architecture, it aims at both data- and compute-intensive applications. Java language is the best practice in Hadoop programing, for achieving a better seamless-integration; we select an OpenCL Java language binding (JOCL) to integrate these two frameworks together. JOCL use Java Native Interface (JNI) to call the kernel program that drives the GPUs. An almost 2 times performance improvement has been validated, without any farther optimization.

Wenbin Fang et. al. 2011 proposes Accelerating MapReduce with Graphics Processors: MARS. Mars is a MapReduce runtime system accelerated with graphics processing units (GPUs).It runs on NVIDIA GPUs, AMD GPUs as well as multicore CPUs.It is implemented MarsCUDA using NVIDIA CUDA. The experimental results show that, the GPU-CPU co-processing of Mars on an NVIDIA GTX280 GPU and an Intel quad-core CPU outperformed Phoenix, the state-of-the-art MapReduce on the multicore CPU with a speedup of up to 72 times and 24 times on average, depending on the applications. Additionally, integrating Mars into Hadoop enabled GPU acceleration for a network of PCs.

CONCLUSION

In this paper a survey is done on the previous work done in the field of map/reduce. Since data is growing so fast these days and Hadoop is widely used everywhere. So in this paper several work done in the field of map/reduce acceleration has been discussed.

REFERENCES

[1] Riedl, J., Konstan, J. and Terveen, L. (1992) 'GroupLens Research', available from Internet <http://www.grouplens.org/> (12 March 2006).

[2] Montaner, M., Lopez, B. and De la Rosa J.L. (2003) 'A Taxonomy of Recommender Agents on the Internet', *Artificial Intelligence Review*, Kluwer Academic Publisher, 19, 285 – 330.

[3] Shardanand, U. (1994) 'Social Information Filtering for Music Recommendation', Massachusetts Institute of Technology.

[4] Shardanand, U. and Maes, P. (1995) 'Social Information Filtering: Algorithms for Automating "Word of Mouth"', Proc. Conf. Human Factors in Computing Systems.

[5] Kangas, S. (2002) 'Collaborative Filtering and Recommendation Systems', Research report TTE4 – 2001 – 35.

[6] S. Meng, W. Dou, X. Zhang and J. Chen, "KASR: A keyword-aware service recommendation method on MapReduce for big data application", IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 12, pp. 3221-3231, 2014.

[7] Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F., Reitererand, S., Stettinger, M.: Basic approaches in recommendation systems. In: Robillard, M., Maalej, W., Walker, R.J., Zimmermann, T. (eds.) Recommendation Systems in Software Engineering, Chap. 2. Springer, Heidelberg (2014).

[8] Cheikh Kacfah Emani, Nadine Cullot and Christophe Nicolle "Understandable Big Data: A survey" in Computer Science Review Volume 17, August 2015, Pages 70–81.

[9] H. Hu, Y. Wen, T.-S. Chua, and X. Li, ''Towards scalable systems for big data analytics: A technology tutorial,'' IEEE Access, vol. 2, pp. 652–687, 2014.

[10] Rodrigo Agerri,Xabier Artola, Zuhaitz Beloki, German Rigau, Aitor Soroa "Big data for Natural Language Processing: A streaming approach" in Knowledge-Based SystemsVolume 79, May 2015, Pages 36–42.

[11] A. Rabkin and R. H. Katz, "How Hadoop Clusters Break," IEEE Software, vol. 30, pp. 88-94, 2013.

[12] T. Jiang, Q. Zhang, R. Hou, L. Chai, S. A. McKee, Z. Jia, and N. Sun, "Understanding the behavior of in-memory computing workloads," in Workload Characterization (IISWC), IEEE International Symposium on, 2014, pp. 22–30.

[13] Martin Zwilling. What Can Big Data Ever Tell Us About Human Behavior [online]. Available: http://www.forbes.com/sites/martinzwilling/2015/03/24/what-can-big-data-ever-tell-us-about-human-behavior/#1580346f1bed. Date accessed: (March 24, 2015).

[14] Saeed Shahrivari, "Beyond Batch Processing: Towards Real-Time and Streaming Big Data", Computers, Vol. 3, pp. 117.129, 2014.

[15] Sachin Agarwal. Monitoring and Troubleshooting Apache Storm [online]. Available: https://dzone.com/articles/monitoring-and-troubleshooting-apache-storm-with-o. Date accessed: (April 7, 2016).

[16] M. R. Evans, D. Oliver, K. Yang, X. Zhou, S. Shekhar, "Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities," Ed. S. Wang, M. F. Goodchild, CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery. Springer, 2014.

[17] Wikipedia. "Spatial Database" [online]. Available: https://en.wikipedia.org/wiki/Spatial_database.

[18] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy and Dhruba Kumar Bhattacharyya "Big Data Analytics in Bio-informatics: A Machine Learning Perspective "in Journal of LATEX CLASS FILES, vol. 13, no. 9, September 2014.

[19] H. Nordberg, K. Bhatia, K. Wang and Z. Wang, "BioPig: a Hadoop-based analytic toolkit for large-scale sequence data," Bioinformatics, vol. 29, no. 23, pp. 3014–3019, 2013.

[20] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing," Genome Biol, vol. 10, no. 11, p. R134, 2009.

[21] Cynthia Burghard. Big Data and Analytics Key to Accountable Care Success, IDC Health Insights [online]. Available: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IML14338USEN&appname=skmwww. Date accessed: (October 2012).
[22] Raghupathi W, Raghupathi V: Big data analytics in healthcare: promise and potential. Health Inform Sci Syst. 2014, 2 (1): 3-10.1186/2047-2501-2-3.
[23] Anjali P P and Binu A "A Comparative Survey Based on Processing Network Traffic Data Using Hadoop Pig and Typical MapReduce" in International Journal of Computer Science & Engineering Survey (IJCSES) Vol.5, No.1, February 2014.
[24] Y. Lee and Y. Lee, Toward scalable internet traffic measurement and analysis with Hadoop, ACM SIGCOMM Computer Communication Review, vol. 43, no. 1, pp. 5-13,
2012.
[25] Hossein Hassani and Emmanuel Sirimal Silva "Forecasting with Big Data: A Review "in Annals of Data Science March 2015, Volume 2, Issue 1, pp 5-19.
[26] Veershetty Dagade, Mahesh Lagali, Supriya Avadhani and Priya Kalekar, Big Data Weather Analytics Using Hadoop, International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume 14 Issue 2 –APRIL 2015.
[27] Intel IT center. Big Data in the Cloud: Converging Technologies [online]. Available: http://www.intel.com/content/www/us/en/big-data/big-data-cloud-technologies-brief.html. Date accessed: (April 2015).
[28] D. Zhang, F. Sun, X. Cheng, and C. Liu,(2011) "Research on hadoop-based enterprise file cloud storage system," Proc. 3rd International Conference on Awareness Science and Technology (iCAST 11), IEEE Press, pp. 434-437.
[29] H. Hu, Y. Wen, T.-S. Chua, and X. Li, ''Towards scalable systems for big data analytics: A technology tutorial,'' IEEE Access, vol. 2, pp. 652–687, 2014.
[30] Carmen C. Y. Poon, Senior Member, IEEE, Benny P. L. Lo, Mehmet Rasit Yuce, Senior Member, IEEE, Akram Alomainy, Senior Member, IEEE, and Yang Hao, Fellow, IEEE "Body Sensor Networks: In the Era of Big Data and Beyond" in IEEE REVIEWS IN BIOMEDICAL ENGINEERING, VOL. 8, 2015.
[31] A. Rabkin and R. H. Katz, "How Hadoop Clusters Break," IEEE Software, vol. 30, pp. 88-94, 2013.
[32] J. Dean and S. Ghemawat, MapReduce: Simplified data processing on large clusters, in Operating Systems Design and Implementation (OSDI) 04', 2004, pp. 137-150.
[33] M. Grossman, M. Breternitz and V. Sarkar, "HadoopCL: MapReduce on Distributed Heterogeneous Platforms Through Seamless Integration of Hadoop and OpenCL", Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing Workshops and PhD Forum, pp. 1918-1927.
[34] P. Jaaskelainen, C. Lama, P. Huerta, and J. Takala, "OpenCL-based design methodology for application-specific processors," Embedded Computer Systems (SAMOS), 2010 International Conference, pp. 223- 230, 2010.
[35] Gupta, K.G., Agrawal, N. and Maity, S.K., "Performance analysis between aparapi (a parallel API) and JAVA by implementing sobel edge detection Algorithm," in PARCOMPTECH, Bangalore, Feb. 2013, pp. 1-5.
[36] P. WillettThe Porter stemming algorithm: then and now Program: Electr Libr Inform Syst, 40 (3) (2006), pp. 219–223.
[37] Niwattanakul S, Singthongchai J, Naenudorn E, Wanapu S. Using of Jaccard coefficient for keywords similarity. In: Proc. of the international multi conference of engineers and computer scientists, vol I; 2013. p. 380–4.
[38] A. Huang, Similarity measures for text document clustering, in: Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 2008, pp. 49–56.
[39] A. Chu, R. Kalaba, and K. Spingarn, "A comparison of two methods for determining the weights of belonging to fuzzy sets", Journal of Optimization Theory and Applications, Vol. 27, No.4, pp.531-538, 1979.
[40] G. Salton, "Automatic Text Processing," Addison-Wesley, 1989.
[41] Wang Jun, Li Lei and Ren Fuji, "An Improved method of Keywords Extraction Based on Short Technology Text", International Conference on Natural Language processing and Knoledge Engineering (NLP-KE), pp. 1-6.
[42] Adomavicius G., Kwon Y.: New recommendation techniques for multicriteria rating systems. IEEE Intel. Syst. **22**(3), 48–55 (2007).
[43] X. Zhang, J.-J. Lu, X. Qin and X.-N. Zhao, "A high-level energy consumption model for heterogeneous data centers", Simul. Model. Pract. Theory, vol. 39, pp. 41-55, 2013.
[44] X. Peng and Z. Sai, ``A low-cost power measuring technique for virtual machine in cloud environments," Int. J. Grid Distrib. Comput., vol. 6, no. 3, p. 69, 2013.
[45] Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Recommender Systems Handbook, pp. 217–253 (2011).
[46] B. He, W. Fang, Q. Luo, N. K. Govindaraju, and T. Wang, "Mars: A MapReduce framework on graphics processors, " in PACT '08: Proceedings of the 17th international conference on Parallel architectures and compilation techniques, 2008, pp. 260-269.
[47] Koichi Shirahata, Hitoshi Sato, and Satoshi Matsuoka. Hybrid Map Task Scheduling for GPU-Based Heterogeneous Clusters. In Proceedings of CloudCom, pages 733-740, 2010.
[48] M. Xin and H. Li, "An implementation of gpu accelerated mapreduce: Using hadoop with opencl for data- and compute-intensive jobs", International Joint Conference on Service Sciences, pp. 6-11.
[49] Malakar, R.; Vydyanathan, N., "A CUDAenabled Hadoop cluster for fast distributed image processing, " Parallel Computing Technologies (PARCOMPTECH), 2013 National Conference on , vol., no., pp.1, S, 21-23 Feb. 2013.
[50] Zhai Yanlong, Guo Ying, Chen Qiurui, Yang Kai and E. Mbarushimana, "Design and Optimization of a Big Data Computing Framework Based on CPU/GPU Cluster", pp. 1039-1046, 2013.
[51] J. Zhu, Li Juanjuan, E. Hardesty, H. Jiang and L. Kuan-Ching, "GPU-in-Hadoop: Enabling MapReduce across distributed heterogeneous platforms", IEEE/ACIS 13th International Conference of Computer and Information Science (ICIS), 2014, pp. 321-326.
[52] Sufeng Niu, Guangyu Yang, Nilim Sarma, Pengfei Xuan " Combining Hadoop and GPU to preprocess large Affymetrix microarray data", in IEEE International Conference on Big Data (Big Data), 2014, pp. 692-700.
[53] Yuhong Feng ; Junpeng Wang ; Zhiqiang Zhang ; Haoming Zhong "The Edge Weight Computation with MapReduce for Extracting Weighted Graphs" IEEE Transactions on Parallel and Distributed Systems Volume:PP , Issue: 99 , 2016 pp: 1-1.
[54] SungYe Kim, Jeremy Bottleson, Jingyi Jin, Preeti Bindu " Power Efficient MapReduce Workload Acceleration Using Integrated-GPU", in IEEE First International Conference onBig Data Computing Service and Applications (BigDataService),2015 pp. 162 - 169.
[55] Reza, M.; Sinha, A.; Nag, R. & Mohanty, P. (2015), CUDA-enabled Hadoop cluster for Sparse Matrix Vector Multiplication., in 'ReTIS' , IEEE, , pp. 169-172 .
[56] Abhaya Kumar Sahoo , Sundar Sourav Sarangi , Rachita Misra "A comparison study among GPU and map reduce approach for searching operation on index file in database query processing", in IEEE International Conference on Man and Machine Interfacing (MAMI), 2015, pp. 1 - 5.