Building Blocks of Big Data Analytics: A Review

Mrs.Kavithamani[1], Mr.Jayakrishnan Chathu[2], Mr.Harshanth K Prakash[3], Srinevasan Krishnamurthy[4]
[1,]Bharathiar University, Coimbatore.
[2]Sunnyvale, CA, USA
[3]Vellore Institute of Technology, Chennai.
[4]Santa Clara, California, USA.

Abstract- *Big data analytics is the process of examining large data sets containing a variety of data. The data can be textual or image data. This paper specifies the key building blocks of big data analytics. Moreover, it provides various aspects of information security.*

Keywords- Big Data, Information security, Bigdata analytics, Hadoop HDFS, Data visualization.

I. INTRODUCTION

With each passing day, we are witnessing a rapid increase of sheer amount of data produced and being communicated in every walk of life- whether it is technology driven or not. This data is being produced from multiple sources and applications- web, knowledge bases, online transactions, user interactions, system logs etc, and that too in various formats- unstructured, semi-structured, and structured.

Traditional structured data stores, like the RDBMS technologies, are designed to handle transactional and structured data. When the volume of data reaches petabytes and is also unstructured, it is simply not feasible to handle it with the available structured database management systems. Though structured large volume data access was addressed by Column oriented DBMS with very effi--cient performance, the problem of large and unstructured data remained unaddressed until the invention of Big Data. This is primarily because semi-structured or unstructured data defeats the fundamental RDBMS concepts like querying, indexing , and referential integrity etc. Also it does not fit in the relational theory and rigid schema concepts. Invention of Big Data technologies, mainly Hadoop, a distributed data storage system, and MapReduce, a programming paradigm for massive clusters of distributed storage, paved a new way for addressing these problems.

Fig 1.

As illustrated in Fig 1, three V's that characterize Big Data are velocity, volume, and variety. It is essentially designed to handle data that is produced at a much faster rate, not structured to store in a table, and well beyond the tera byte ranges.[16] In the traditional approach, in real time, data is either copied from storage to memory and operated upon, or committed from memory to storage. This renders itself impractical in the case of very large amounts of data (volume) . What Big Data does is to store the data in different systems and the code is copied to those systems, while processing is being done in parallel. [15].

II. BUILDING BLOCKS OF BIG DATA

Let's delve into the building blocks of Big Data. The key building blocks of Big Data Analytics could be defined as Storage, Parallel Programming Paradigm and Visual Presentation.

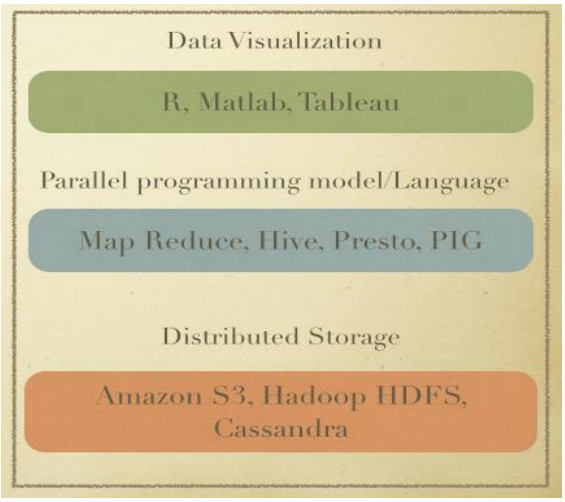

Fig 2

2.1 Distributed Storage

The core of a Big Data stack is the distributed scalable storage file system, which has the ability to distribute and store data across a cluster of machines/servers. 2006 was a significant year, when Google researchers in OSDI'2006 introduced, "Bigtable: A Distributed Storage System for Structured Data", [2] while Doug Cutting and Mike Cafarella's went a step ahead and created Hadoop HDFS. These were the first ones to appear on the distributed storage horizon.[21,22]

Architecture of HDFS cluster contains a NameNode, a DataNode, and Client Machines. A NameNode is a master node that tracks and directs the storage of the cluster. A DataNode is what makes up the majority of the machines within a cluster. A Client Machine is responsible for loading data into the cluster. NameNode performs the key job of splitting the data files into blocks and gets them replicated and stored across the machines in the cluster. Thus providing fault tolerance, and scalability.

On the other hand Google's Bigtable is described as "a sparse, distributed, persistent multidimensional sorted map." Data is assembled in order by row key, and indexing of the map is arranged according to row, column keys and timestamps. High capacity is achieved by compression algorithms.[18]

Other new entrants are Hbase ,mongoDB [7] and Cassandra. Hbase is an abstraction over HDFS, to provide Bigtable like capabilities to Hadoop.[7,17] Cassandra[4] derived its data model from Bigtable but follows Amazon's Dynamo for its storage model.[3]

2.2 Parallel Programming Model

The parallel programming paradigm of Big Data is what taps into the power of the core distributed storage system, and allows for massive scalability across hundreds or thousands of servers in the cluster. When Big Data became popular and widely available with Hadoop, MapReduce was the first parallel programming model of the Hadoop stack. [19].MapReduce is basically the software framework for writing applications for processing vast amounts of data in-parallel on large clusters in a reliable, fault-tolerant manner.[5] A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system.[13] The framework manages the task scheduling and re-executes the failed tasks. But the MapReduce framework does not offer simpler query interface like SQL that is more popular with the RDBMS technology. Hive addressed this issue by providing a SQL like layer over MapReduce. Hive basically translates SQL queries into multiple stages of MapReduce, and it is optimized for query throughput. [12,14].Both Hive and MapReduce are file based and fault tolerant. Presto, which was open sourced by Facebook, differentiated it by offering memory based model, thus providing fast low latency but non fault tolerant parallel programming model more suited for interactive queries.[21]

2.3 Data Visualization

Conveying the insight obtained from processing Big Data is an interesting functionality that is left to tools like Matlab, R, and GNU Octave to name a few. Various commercial products like, Tableau, Silk, DataWrapper are becoming more popular because of the intuitive graphing and other visual charts that they provide.

III. BIG DATA ANALYTICS

This new paradigm of Big Data provides a new opportunity, which did not exist earlier, that is the ability to analyze and correlate large volumes of data for analytic purposes. Traditional analytics, which is the data crunching, question-answering phase leading up to the decision-making phase in the overall Business Intelligence process[11], was mostly done with samples of real time data, that mainly addressed problems of statistical nature. That is predictions based on a sample. The success and failure of this methodology depended on how precise and accurate the statistical model is.

For example, in the field of internet security Intrusion Detection Systems is one of the key component, that sits right behind the first wall of defense, the firewall. Usually the deep packet inspection role is offloaded on to the IDS system. But soon we found that the IDS raised lot of false positives because they only had limited view of the network data, and IDS systems are not in a position to monitor all the data. This is one such an area where a Big Data Analytic model could play a significant role. Consider a Big Data powered security infrastructure that gets all the real time data fed into it, at the same time various attack signatures could be played in parallel on big clusters. That could uncover all the previously undetected security threats, because of lack of complete data and parallel processing [8,9].

What Big Data analytics brings to the table is a way of examining large data sets containing a variety of data types to uncover hidden patterns, unknown correlations, market trends, data security, fraud detection, customer preferences and other useful business information. [23,24].The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits, that did not exist before with the conventional model.[10]

IV. CONCLUSION

Currently, Big Data field is ripe with various choices of tools and technologies, each suitable for a different set of applications as dictated by Cap Theorem. In this paper, we explained the Building Blocks of Big Data and the Big Data itself. We also looked at one of the key defining use of Big Data, Data Analytics. This novel field of Big Data Analytics is in its nascent stage, with plenty of research coming up and its various applications. One powerful application of Big Data Analytics would be to achieve comprehensive information security based on spectrum of information from various sources.

REFERENCES

[1] Michael Cox and David Ellsworth, "Application-controlled demand paging for out-of-core visual-izetion",October 1997.

[2] Fay Chang, Jeffrey Dean, et al, "Bigtable: A Distributed Storage System for Structured Data",OSDI'06.

[3] Konstantin Shvachko, HairongKuang, Sanjay Radia, Robert Chansler , "The Hadoop Distributed FileSystem",Yahoo!Sunnyvale, California USA.

[4] AvinashLakshman, Prashant Ma- lik, "Cassandra - A Decentralized Structured Storage System", Facebook.

[5] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simpli fied Data Processing on Large Clusters",Google, Inc.

[6] Rico Suter, "MongoDB An introduction and performance analysis"HSR HochschulefürTechnikRapperswil.

[7] Tyler Harter, DhrubaBorthakur† ,Siying Dong ,AmitanandAiyer, LiyinTang , Andrea C. Arpaci-Dusseau, Remzi H. Arpaci, "Analysis of HDFS UnderHBase: A Facebook Messages Case Study", Dusseau University of Wisconsin, Madison Facebook Inc.

[8] RasimAlguliyev,YadigarImamverdiyev, "Big Data: Big Promises for Information Security", Institute ofInformation Technology, Azerbaijan National Academy of Sciences .

[9] Samuel Marchal, Xiuyan Jiang, RaduState,Thomas Engel, "A Big Data Architecture for LargeScale Security MonitoringFaculty of Sciences", Technology and Communication– University of Luxembourg, Luxembourg .

[10] Alvaro A. Cárdenas,Sreeranga P. Rajan,Pratyusa K. Manadhata, "Big Data Analytics for Security", University of Texas at Dallas,HPLabs, Fujitsu Laboratories of America.

[11] Andrew McAfee and Erik Brynjolfsson, "Big Data: The Management Revolution",Harvard Business Review.

[12] Chen, Yanpei and Alspaugh, Sara and Katz,Randy. Interac- tive Analytical Processing in Big Data Systems: A Cross-industry Study of MapReduce Workloads.Proc. VLDB Endow., August 2012.

[13] A.Thusoo,J.S.Sarma,N.Jain,Z.Shao,P.Chakka,S.Anthony, H.Liu,P. Wyckoff, and R. Murthy, "Hive:a warehousing solution over a mapreduce framework," Proc. VLDB Endow., vol. 2, no. w2, pp. 1626–1629, Aug. 2009.

[14] C. Olston, B. Reed, U. Srivasta- wva, R. Kumar,and A. Tomkins, "Pig latin: a not-so-foreign language for data processing," in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ser. SIGMOD'08. New York, NY, USA: ACM, 2008, pp.1099– 1110.

[15] A. Labrinidis, and H. V. Jagadish, "Challengesand Opportunities with Big Data," Journal Pro- 2ceedings of the VLDB Endowment, vol. 5, no.12, pp. 2032-2033, August 2012.

[16] D. Agrawal, S. Das, and A. El Abbadi, "Bigdata and cloud computing: current state and future opportunities," Proc. of the 14th International Conference on Extending DatabaseTechnology, pp. 530-533, 2011.

[17] Mittelstadt S., Behrisch M., Weber S., SchreckT. et al, "Visual analytics for the big data era A comparative review of state-of- the-art commercial systems," Proc. of the IEEE Conference on Visual Analytics Science and Technology(VAST), pp. 173-182, 2012.18. T. White,

Hadoop: The Definitive Guide. O'Reilly Media, Yahoo! Press, June 5, 2009 2009, pp. 922-933.

[18] T. White, Hadoop: The Definitive Guide. O'Reilly Media, Yahoo! 22Press, June 5, 2009.

[19] Abouzeid, A., Bajda-Pawlikowski, K., Abadi,D. et al., "HadoopDB:An Architectural Hybrid of MapReduce and DBMS Tech-nologies for Analytical Workloads"Proceedings of the VLDB Endowment 2 ,2009, pp. 922 933.

[20] "Hadoop Eco System for Big Data Security andPrivacy" Pradeep Adluru, SrikariSindhooriDatla, Xiaowen Zhang*Computer Sci2enceDepartmentCollege of 2Staten Island, CUNY2800 Victory Blvd, Staten Island, NY 10314 978-1-4577-1343-9/12/$26.00 2©2015 IEEE

[21] M.H.Padgavankar et al, / (IJCSIT) International Journal of Computer WScience and Information Technologies, Vol. 5 (2) , 2014, 2218-2223

[22] Big Data Storage and Challenges 2M.H.Padgavankar, Dr.S.R.GuptaME, CSE, PRMIT&R, Amravati, Maharashtra, IndiaAssistant Professor, CSE, PRMIT&R, Amravati, S`WAMaharashtra, India .

[23] Big Data: Big Promises for Infor- mation Security IEEERasimAlguliyev Institute of Information TechnologyAzerbaijan National SAcademy of Sciences Baku,Azerbaijandirector@iit.ab.azYadigarImamverdiyevI nstitute of Information TechnologyAzerbaijan National Academy of Scienc- es Baku, Azerbaijan.

[24] Big Data Analytics for SecurityAlvaro A. Cárdenas | University of Texas at DallasPratyusa K. Manadhata | HP LabsSreeranga P. Rajan | Fujitsu Laboratories of America IEEE 2013.