International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869, Volume-3, Issue-6, June 2015

Understanding Big Data: Framework and Tools for Massive Data Storage and Mining

Amit Bhagat

ABSTRACT:
The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. This useful information for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

Keywords: big data, data mining, analytics, decision making.

I. INTRODUCTION

With the increase in storage capabilities and methods of data collection, huge amounts of data have become easily available. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data has become cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data.The size, variety, and rapid change of such data require a new type of big data analytics,as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.

During the last 35 years, data management principles such as physical and logical independence, declarative querying and cost-based optimization have led, during the last 35 years, to a multi-billion dollar industry. More

importantly, these technical advances have enabled the first round of business intelligence applications and laid the foundation for managing and analyzing Big Data today. The many novel challenges and opportunities associated with Big

Data necessitate rethinking many aspects of these data management platforms, while retaining other desirable aspects. We believe that appropriate investment in Big Data will lead to a new wave of fundamental technological advances that will be embodied in the next generations of Big Data management and analysis platforms, products, and systems.

Big Data has as a term appeared literally first times towards the end of the 1990's [1]. In the year 2012, [2] crystallizes the term as follows: "Big Data symbolizes the aspiration to build platforms and tools to ingest, store and analyze data that can be voluminous, diverse, and possibly fast changing". It is not clear who are professionals and who are laypeople in Big Data era. For example, computer scientists, statisticians, mathematicians, and informatics have Big Data capabilities [3]. Usually, laypeople are not interested in platforms and tools – they are interested in results of analyzed data. Furthermore, laypeople might be worry about growing data masses ("90% of the data in the world today has been created in the last two years alone" [4]) and their analysis ("3% of the potentially useful data is tagged, and even less analyzed" [5]).

Under the explosive increase of global data, the term of big data is mainly used to describe enormous datasets. Compared with traditional datasets, big data typically includes masses of unstructured data that need more real-time analysis. In addition, big data also brings about new opportunities for discovering new values, helps us to gain an in-depth understanding of the hidden values, and also incurs new challenges, e.g., how to effectively organize and manage such datasets. Recently, industries become interested in the high potential of big data, and many government agencies announced major plans to accelerate big data research and applications [2]. In addition, issues on big data are often covered in public media, such as The Economist [3, 4], New York Times [5], and

National Public Radio [6, 7]. Two premier scientific journals, Nature and Science, also opened special columns to discuss the challenges and impacts of big data [8, 9]. The era of big data has come beyond all doubt [10].

II. BIG DATA

Three Vs of Big Data Volume of data: Volume refers to amount of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes. Variety of data: Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc. Velocity of data:Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.There is not only more data, but it also comes from a wider variety of sources and formats. As described in the report by the President's Council of Advisors of Science & Technology, some data is "born digital," meaning that it is created specifically for digital use by a computer or data processing system. Examples include email, web browsing, or GPS location. Other data is "born analog," meaning that it emanates from the physical world, but increasingly can be converted into digital format. Examples of analog data include voice or visual information captured by phones, cameras or video recorders, or physical activity data, such as heart rate or perspiration monitored by wearable devices.11 With the rising capabilities of "data fusion," which brings together disparate sources of data, big data can lead to some remarkable insights.

III. BIG DATA MINING

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress" [9]. Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [34] . However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold [8]. The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad [11] in his invited talk at the KDD BigMine'12 Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 milion tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. Anew large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, Twitter are starting to look carefully to this data to find useful patterns to improve user experience. Alex 'Sandy' Pentland in his 'Human Dynamics Laboratory' at MIT, is doing research in finding patterns in mobile data about what users do, and not in what people says they do [28]

IV BIG DATA SOLUTIONS

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and MapReduce are explained in following points.

HDFS Architecture Hadoop includes a fault-tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers.

MapReduce Architecture The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur on multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. In a traditional data warehousing scenario, thismight entail applying an ETL operation on the data to produce something usable by the analyst. In Hadoop, these kinds of operations are written as MapReduce jobs in Java. There are a number of higher level languages like Hive and Pig that make writing these programs easier. The outputs of these jobs can be written back to either HDFS or placed in a traditional data warehouse. There are two functions in MapReduce as follows: map – the function takes key/value pairs as input and generates an intermediate set of key/value pairs reduce – the function which merges all the intermediate values associated with the same intermediate key

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop was inspired by Google's MapReduce Programming paradigm [16]. Hadoop is a highly scalable compute and storage platform. But on

the other hand, Hadoop is also time consuming and storage-consuming. The storage requirement of Hadoop is extraordinarily high because it can generate a large amount of intermediate data. To reduce the requirement on the storage capacity,Hadoop often compresses data before storing it. Hadoop takes a primary approach to a single big workload, mapping it into smaller workloads. These smaller workloads are then merged to obtain the end result. Hadoop handles this workload by assigning a large cluster of inexpensive nodes built with commodity hardware. Hadoop also has a distributed, cluster file system that scales to store massive amounts of data, which is typically required in these workloads. Hadoop has a variety of node types within each Hadoop cluster; these include DataNodes, NameNodes, and EdgeNodes. The explanations are as follows:
NameNode: The NameNode is the central location for information about the file system deployed in a Hadoop environment. An environment can have one or two NameNodes, configured to provide minimal redundancy between the NameNodes. The NameNode is contacted by clients of the Hadoop Distributed File System (HDFS) to locate information within the file system and provide updates for data they have added, moved, manipulated, or deleted. b. DataNode: DataNodes make up the majority of the servers contained in a Hadoop environment. Common Hadoop environments will have more than one DataNode, and oftentimes they will number in the hundreds based on capacity and performance needs. The DataNode serves two functions: It contains a portion of the data in the HDFS and it acts as a compute platform for running jobs, some of which will utilize the local data within the HDFS. c. EdgeNode: The EdgeNode is the access point for the external applications, tools, and users that need to utilize the Hadoop environment. The EdgeNode sits between the Hadoop cluster and the corporate network to provide access control, policy enforcement, logging, and gateway services to the Hadoop environment. A typical Hadoop environment will have a minimum of one EdgeNode and more based on performance needs iii. IBM InfoSphere BigInsights It is an Apache Hadoop based solution to manage and analyze massive volumes of the structured and unstructured data. It is built on an open source Apache Hadoop with IBM big Sheet and has a variety of performance, reliability, security and administrative features

V.OPEN SOURCE REVOLUTION

The Big Data phenomenon is intrinsically related to the open source software revolution. Large companies as Facebook, Yahoo!, Twitter, LinkedIn benefit and contribute working on open source projects. Big Data infrastructure deals with Hadoop, and other related software as: • Apache Hadoop: software for data-intensive distributed applications, based in the MapReduce programming model and a distributed file system called Hadoop Distributed Filesystem (HDFS). Hadoop allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job. • Apache Hadoop related projects: Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra, Cascading, Scribe and many others. • Apache S4: platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time. • Storm: software for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter. In Big Data Mining, there are many open source initiatives. The most popular are the following: • Apache Mahout: Scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining. • R: open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets. • MOA: Stream data mining open source software to perform data mining in real time. It has implementations of classification, regression, clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software. The streams framework provides an environment for defining and running stream processes using simple XML based definitions and is able to use MOA, Android and Storm. SAMOA is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA. • Vowpal Wabbit: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface when doing linear learning, via parallel learning More specific to Big Graph mining we found the following open source tools: • Pegasus : big graph mining system built on top of MapReduce. It allows to find patterns and anomalies in massive real-world graphs. See the paper by U. Kang and Christos Faloutsos in this issue. GraphLa: high-level graph-parallel system built without using MapReduce. GraphLab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex-programs which are executed

in parallel on each vertex and can interact with neighbouring vertices[9-26].

VI CONCLUSION

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. This data is going to be more diverse, larger, and faster. We discussed in this paper some insights about the topic, and what we consider are the main concerns, and the main challenges for the future. Big Data is becoming the new Final Frontier for scientific data research and for business applications. We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before. Everybody is warmly invited to participate in this intrepid journey.

REFERENCES

[1] A. Vailaya, "What's All the Buzz Around "Big Data?"", IEEE Women in Engineering Magazine, December 2012, pp. 24-31,

[2] B. Brown, M. Chui and J. Manyika, "Are you Ready for the era of 'Big Data'? " McKinsey Quarterly, McKinsey Global Institute, October 2011

[3] B.Gerhardt, K. Griffin and R. Klemann, "Unlocking Value in the Fragmented World of Big Data Analytics", Cisco Internet Business Solutions Group, June 2012,

[4] C. Eaton, D. Deroos, T. Deutsch, G. Lapis and P.C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, Mc Graw-Hill Companies, 978-0-07-179053-6, 2012

[5] C. Tankard, "Big Data Security", Network Security Newsletter, Elsevier, ISSN 1353-4858, July 2012

[6] E. Begoli and J. Horey, "Design Principles for Effective Knowledge Discovery from Big Data", Software Architecture (WICSA) and European Conference on Software Architecture (ECSA) Joint Working IEEE/IFIP Conference on, Helsinki, August 2012

[7] Intel IT Center, "Peer Research: Big Data Analytics", Intel's IT Manager Survey on How Organizations Are Using Big Data, August 2012,

[8] Intel IT Center, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics, June 2012

[9] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011,

[10] K. Bakshi, "Considerations for Big Data: Architecture and Approach", Aerospace Conference IEEE, Big Sky Montana, March 2012

[11] M. Smith, C. Szongott, B. Henne and G. Voigt , "Big Data Privacy Issues in Public Social Media", Digital Ecosystems Technologies (DEST), 6th IEEE International Conference on, Campione d'Italia, June 2012

[12] P. Russom, "Big Data Analytics ", TDWI Best Practices Report, TDWI Research, Fourth Quarter 2011, http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics/asset.aspx

[13] R.D. Schneider, Hadoop for Dummies Special Edition, John Wiley&Sons Canada, 978-1-118-25051-8, 2012

[14] R. Weiss and L.J. Zgorski, "Obama Administration Unveils "Big Data" Initiative:Announces $200 Million in new R&D Investments", Office of Science and Technology Policy Executive Office of the President, March 2012

[15] S. Curry, E. Kirda, E. Schwartz, W.H. Stewart and A. Yoran, "Big Data Fuels Intelligence Driven Security", RSA Security Brief, January 2013

[16] S. Madden, "From Databases to Big Data", IEEE Internet Computing, June 2012, v.16, pp.4-6

[17] S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011

[18] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, AI Magazine, Fall 1996, pp. 37- 54

[19] V. Borkar, M.J. Carey and C. Li, "Inside "Big Data Management": Ogres, Onions, or Parfaits?", EDBT/ICDT 2012 Joint Conference Berlin Germany, 2012

[20] D. boyd and K. Crawford. Critical Questions for Big Data. Information, Communication and Society, 15(5):662–679, 2012.

[21] F. Diebold. "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. Discussion Read to the Eighth World Congress of the Econometric Society, 2000.

[22] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.

[23]. Han J, Lee J-G, Gonzalez H, Li X (2008) Mining massive rfid, trajectory, and traffic data sets. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, p 2

[24]. Garg MK, Kim D-J, Turaga DS, Prabhakaran B Multimodal analysis of body sensor network data streams for real-time healthcare. In: Proceedings of the international conference on multimedia information retrieval(2010). ACM, pp 469– 478.

[25]Park Y, Ghosh J A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium(2012). ACM, pp 445–454

[26]. Tasevski P Password attacks and generation strategies. Tartu University: Faculty of Mathematics and Computer Sciences(2011)