International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 06
May 2017

Technical and Security Issues ofBig Data

[1]Sowmya Koneru
[1]Assistant professor, Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh, India
konerusowmya@gmail.com

***Abstract*—Now a days, organizations have started to depend on visions of their customers, internal processes and business operations to uncover new opportunities for growth of organizations. Big data refers to data volumes in the range of exabytes and beyond. The sheer size of the data is a major challenge and other attributes being variety, velocity, value, complexity.privacy and usability.Data security is a vital issue of the bigdata analytics. This paperanalyzes various technical and security issues of Big Data.**

***Keywords*—Big data, Security, Issues,**

I. INTRODUCTION

Since 1990sBig data is in formal literature of Database Systems. It refers not only to specific, large datasets, but also to data collections that consolidate many datasets from multiple sources.The capability of Big Data isto revolutionize much more beyond just research.

Google File System, MapReduce and Hadoop led to possibly the most extensive development of Big Data technologies, and lead the companies focused on the Web, such as Facebook, LinkedIn, Microsoftand Twitter. They have become the indispensable foundation for applications ranging from Web search to content recommendation and computational advertising. There have been persuasive cases made for the value of Big Data for healthcare, urban planning, intelligent transportation, environmental modeling, energy saving, smart materials, machine translation between natural languages, education, computational social sciences, systemic risk analysis in finance, homeland security, computer security, and so on.

In the scientific domain, by revealing the genetic origin of illnesses, such as mutations related to cancer, the Human Genome Project, completed in 2003, is one project that's a testament to the promises of big data. Consequently, researchers are embarking on two major efforts, the Human Brain Project and the US BRAIN Initiative, in a quest to construct a supercomputer simulation of the brain's inner workings, in addition to mapping the activity of about 100 billion neurons in the hope of unlocking answers to Alzheimer's and Parkinson's. In August 2010, the White House, OMB, and OSTP proclaimed that Big Data is a National challenge and priority along with healthcare and national security [1]. The National Science Foundation, the National Institutes of Health, the U.S. Geological Survey, the Departments of Defense and Energy, and the Defense Advanced Research Projects Agency announced a joint R&D initiative in March 2012 that will invest more than $200 million to develop new big data tools and techniques[14].

II. TECHNICAL ISSUES

There are three fundamental issue areas that need to be addressed in dealing with big data: storage issues, management issues, and processing issues. Each of these represents a large set of technical research problems in its own right.

A. Storage Issues

The quantity of data has increased each time as invented a new storage medium. Due to the largely improve in the utilization of the social media the explosion of the storage issue has raised– is that there has been no new storage medium. Moreover, data is being created by everyone and everything (e.g., devices, etc) – not just, as heretofore, by professionals such as scientist, journalists, writers, etc.

Current disk technology limits are about 4 terabytes per disk. So, 1 exabyte would require 25,000 disks. Even if an exabyte of data could be processed on a single computer system, it would be unable to directly attach the requisite number of disks. Access to that data would overwhelm current communication networks. Assuming that a 1 gigabyte per second network has an effective sustainable transfer rate of 80%, the sustainable bandwidth is about 100 megabytes. Thus, transferring an exabyte would take about 2800 hours, if we assume that a sustained transfer could be maintained. It would take

longer to transmit the data from a collection or storage point to a processing point than it would to actually process it!

Two solutions manifest themselves. First, process the data "in place" and transmit only the resulting information. In other words, "bring the code to the data", versus the traditional method of "bring the data to the code." Second, perform triage on the data and transmit only that data which is critical to downstream analysis. In either case, integrity and provenance metadata should be transmitted along with the actual data.

B. Management Issues

Management will, perhaps, be the most difficult problem to address with big data. Unlike the collection of data by manual methods, where rigorous protocols are often followed in order to ensure accuracy and validity, digital data collection is much more relaxed. The richness of digital data representation prohibits a bespoke methodology for data collection. Data qualification often focuses more on missing data or outliers than trying to validate every item. Data is often very fine-grained such as clickstream or metering data. Given the volume, it is impractical to validate every data item: new approaches to data qualification and validation are needed.

The sources of this data are varied - both temporally and spatially, by format, and by method of collection. Individuals contribute digital data in mediums comfortable to them: documents, drawings, pictures, sound and video recordings, models, software behaviors, user interface designs, etc., – with or without adequate metadata describing what, when, where, who, why and how it was collected and its provenance. Yet, all this data is readily available for inspection and analysis.

Going forward, data and information provenance will become a critical issue. JASON has noted [10] that "there is no universally accepted way to store raw data, reduced data, code and parameter choices that produced the data." Further, they note:

"We are unaware of any robust, open source, platform-independent solution to this problem." As far as we know, this remains true today. To summarize, there is no perfect big data management solution yet. This represents an important gap in the research literature on big data that needs to be filled.

C. Processing Issues

Assume that an exabyte of data needs to be processed in its entirety. For simplicity, assume the data is chunked into blocks of 8 words, so 1 exabyte = 1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for end-to-end processing would be 20 nanoseconds. To process 1K petabytes would require a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable information.

III. SECURITY ISSUES OF BIG DATA

Major big data security issues and challenges are discussed by some of the researchers are given below:

A. Fake data generation

To undermine the quality of your big data analysis, attackers can change data and send it into available data streams. Attackers may penetrates into the companies system and may generate wrong reports regarding the employee queries[26]. This may leads to miss understand the system and failed to identify the trends and also may miss the opportunity to solve problems before senior damage. Such challenges can be solved through applying fraud detection approach.

B. Presence of untrusted mappers:

If an attacker got access to your mappers' code, can change the settings of the existing mappers or add vulnerable ones[24]. This way, a data processing can be damaged: attackers can make mappers prepared lists of key/value pairs. Which is why the results brought up by the Reduce process will be faulty.

The problem here is that getting such access may not be too difficult since generally big data technologies don't provide an additional security layer to protect data. They usually tend to rely on perimeter security systems. But if those are faulty, your big data becomes a low hanging fruit.

C. Troubles of cryptographic protection:

Even encryption is a way of protecting user'sinformation; it is also a big data security issue[27]. Even though availability of existing encryption and decryption mechanism available data

stored in clouds is possible, data is not properly maintained with encryption because encryption and decryption if huge amounts of data chunks Despite the possibility to encrypt big data and the essentiality of doing so, this security measure is often ignored. Sensitive data is generally stored in the cloud without any encrypted protection. And the reason for acting so recklessly is simple: *constant encryptions and decryptions of huge data chunks slow things down*, which entails the loss of big data's initial advantage – speed.

D. Sensitive information mining:

Perimeter-based security is typically used for big data protection. It means that all 'points of entry and exit' are secured. But what IT specialists do inside your system remains a mystery[26]..Such a lack of control within your big data solution may let your corrupt IT specialists or evil business rivals mine unprotected data and sell it for their own benefit. Your company, in its turn, can incur huge losses, if such information is connected with new product/service launch, company's financial operations or users' personal information.

E. Struggles of granular access control:

Sometimes, data items fall under restrictions and practically no users can see the secret info in them, like, personal information in medical records (name, email, blood sugar, etc.). But some parts of such items (free of 'harsh' restrictions) could theoretically be helpful for users with no access to the secret parts, say, for medical researchers. Nevertheless, all the useful contents are hidden from them. And this is where talk of granular access starts. Using that, people can access needed data sets but can view only the info they are allowed to see.

F. Lack of security audits:

Big data security audits help companies gain awareness of their security gaps. And although it is advised to perform them on a regular basis, this recommendation is rarely met in reality. Working with big data has enough challenges and concerns as it is, and an audit would only add to the list. Besides, the lack of time, resources, qualified personnel or clarity in business-side security requirements makes such audits even more unrealistic.

IV. CONCLUSION

The world has entered into an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. Many security issues and technical challenges described in this paper. These security issues are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. As a future enhancement of the

REFERENCES

[1] American Institute of Physics (AIP). 2010. College Park, MD, (http://www.aip.org/fyi/2010/)
[2] Ayres, I. 2007. *Supercrunchers*, Bantam Books, New York, NY
[3] Boyd, D. and K. Craford. 2011. "Six Provocations for Big Data", Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society"
[4] The Economist. 2010. "Data, Data Everywhere", (online edition, Feb28, http://www.economist.com/node/15557443
[5] Felten, E. 2010. "Needle in a Haystack Problems", https://freedom-to-tinker.com/blog/felten/needle-haystack
[6] Fox, B. 2011. "Leveraging Big Data for Big Impact", Health Management Technology, http://www.healthmgttech.com/
[7]Freeman, K. 2011. http://en.wikipedia.org/wiki/File:Kencf0618Facebook Network.jpg
[8] Gantz, J. and E. Reinsel. 2011. "Extracting Value from Chaos", IDC's Digital Universe Study, sponsored by EMC
[9] Jacobs, A. 2009. "Pathologies of Big Data", *Communications of the ACM*, 52(8):36-44
[10] JASON. 2008. "Data Analysis Challenges", The Mitre Corporation, McLean, VA, JSR-08-142
[11] Kaisler, S. 2012. "Advanced Analytics", CATALYST Technical Report, i_SW Corporation, Arlington, VA
[12] Kaisler, S., W. Money, and S. J. Cohen. 2012. "A Decision Framework for Cloud Computing", *45th Hawaii International Conference on System Sciences*,Grand Wailea, Maui, HI, Jan 4-7, 2012
[13] Kang, U. 2012. "Mining Tera-scale Graphs with MapReduce: Theory, Engineering, and Discoveries", PhD. Thesis, Computer Science, Carnegie-Mellon University, Pittsburgh, PA
[15] Mervis, J. 2012. "Agencies Rally to Tackle Big Data", Science, 336(4):22, June 6, 2012 Popp, R., S. Kaisler, et al. 2006. "Assessing Nation-State Fragility and Instability", *IEEEAerospace Conference*, 2006, Big Sky, MT
[16] Ritchey, T. 2005. "Wicked Problems: Structuring Social Messes with Morphological Analysis", Swedish Morphological Society, http://www.swemorph.com/wp.html
[17] Rittel, H. and M. Webber. *1973.* "Dilemmas in a General theory of Planning*", in Policy Sciences*, Vol. 4, Elsevier Scientific, Amsterdam, the Netherlands, pp. 155-169
[18] Stonebraker, M. and J. Hong. 2012. "Researchers' Big Data Crisis; Understanding Design and Functionality", Communications of the ACM, 55(2):10-11
[19] Taleb, N. 2010. *The Black Swan: The Impact ofthe Highly Improbable*, Random House, New York,NY
[20] Big. James Manyika,Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and AngelaHung Byers. Big Data: The next frontier for innovation, competition, and productivity McKinsey Global Institute. May 2011.

[21] Daniel Fasel and Andreas Meier, editors.Big Data.Springer Publisher, Heidelberg, Germany, 2013.
[22]Lisa Kart. Market Trends: Big Data Opportunities in Vertical Industries. Technical report, Gartner, 2012.
[23]IBM.IBM Research Dublin. http://www.research.ibm.com/labs/ireland/, March 2014.
[24]Wikibon Report, 2011.
[25] Liang, Qilian, et al. "Security in big data." Security and Communication Networks 8.14 (2015): 2383-2385.
[26] Xu, Lei, et al. "Information security in big data: privacy and data mining." IEEE Access 2 (2014): 1149-1176.
[27] Mahmood, Tariq, and UzmaAfzal. "Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools." Information assurance (ncia), 2013 2nd national conference on.IEEE, 2013.