Big Data Challenges and Security Issues

Priyanka.R & Bhavani.V.R
TRP Engineering College
Tamilnadu, India

Shanmugasundaram Hariharan
TRP Engineering College
Tamilnadu, India

***Abstract*— Big Data has drawn huge attention among researchers in information sciences, policy decision making in governments and enterprises. It is used to describe the exponential growth and availability of data, both structured and unstructured. Big Data produces evolutionary breakthroughs in scientific disciplines, which give us a lot of opportunities to make great progresses in many fields. The future competitions in business productivity and technologies would surely converge into the big data explorations. On the other hand, Big Data also arises with many challenges, such as difficulties in data capturing, storage, analytics and visualization. The most challenging task in big data is ensuring the privacy and security of big data repositories. This paper brings in the view of big data-its opportunities and applications, its challenges. It also focuses on the techniques to assure the privacy and security of big data.**

***Keywords*— Big data, Security, Hadoop, Holistic approach**

I. INTRODUCTION

Big data is generally defined in the form of three Vs [1] : Volume, Velocity and Variety as shown in Fig 1. The term volume is the size of the data set, velocity indicates the speed of data in and out, and variety describes the range of data types and sources. The fourth 'V' can be value, variability, or virtual [2]. Several dimensions can be added with prior definitions of 'big data' in order to specify the importance of quality data and the various level of trust in different data sources [3]. If data is of insufficient quality, then it has been integrated with other data and information, a false correlation could result in the organization making an incorrect analysis of a business opportunities. Big data is a collection of huge data sets which becomes difficult to process using traditional processing system or using state of art approaches. In 2012, Gartner gave a more detailed definition as: ''Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enhance the decision making, insight discovery and process optimization". A data set can be called Big Data if it is formidable to capture, perform curation, analysis and visualization on it at the current technologies.

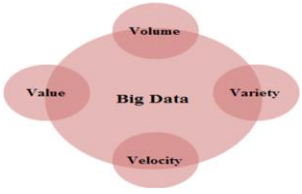

Fig. 1. Three Vs of Bigdata

II. BIG DATA APPLICATIONS

Big Data makes a prominent growth of the world economy by enhancing the productivity and competitiveness of enterprises, industries and also the public administrations. Big Data has a deep relationship with e-Science **[4]** which is computationally intensive science which usually is implemented in distributed computing systems. Many issues on Big Data applications can be resolved by e-Science which requires Grid computing **[5]**. Other Big Data applications lies in many scientific disciplines like astronomy, atmospheric science, social computing **[6],** biological and other complex scientific researches [7].

Web-based applications prompt to use Big Data frequently, such as recent hot spots social computing, social network analysis, and online communities. There are countless sensors around us, they generate enormous sensor data that need to be utilized, and for e.g. intelligent transportation systems (ITS) are based on the analysis of large volumes of complex sensor data. The recent advances in remote sensing (RS) and computer techniques give birth to the explosive growth of remote sensing (RS) data. RS data are regarded as RS 'Big Data'. The RS data gathered by a single satellite data centre are dramatically increasing by several terabytes per day [9].But large amount of proliferations of data has posted many challenges in managing, processing and interpreting these RS Big data.

Big data connects and integrates the physical world, the human society and cyberspace. Here the physical world has a reflection in cyberspace in the form of big data, through Internet and other information technologies, while human society does its big data-based mapping in cyberspace by using some mechanisms like human-computer interfaces.

III. RELATED WORK

Big Data bring many attractive opportunities but we are also facing a lot of challenges [13] in handling Big Data. The challenges lies in data capture, storage, searching, sharing, analysis, and visualization. If these challenges are not overcome we do not have the capabilities to explore big data.

Work related to the Challenges in Big Data analysis like data inconsistence and incompleteness, scalability, timeliness and data security, privacy issues [14, 15] are going on. To obtain the value from Big Data, we need to develop new techniques and technologies to analyse it. Until now, scientists have developed variety of techniques and technologies to capture, analyze and visualize Big Data, but still they have some limitations over meeting variety of needs.

These techniques and technologies include number of disciplines, like computer science, economics, mathematics, statistics and other expertises. There are some tools to make sense of Big Data. Current tools focus on three classes, namely, batch processing tools, stream processing tools, and interactive analysis tools. Most batch processing tools are based on the Apache Hadoop infrastructure, such as Mahout and Dryad [16,17].

Apache Hadoop is one of the most known software platforms that support data-intensive distributed applications as shown in Figure 2. It implements the computational technique named Map/Reduce[18]. Apache Hadoop platform consists of the Hadoop kernel, Map/Reduce and Hadoop distributed file system (HDFS), Storm and S4 are examples for data analytic tools. The interactive analysis processes the data in an interactive environment allowing users to directly connect to the computer and hence can interact with it in real time e.g., Google's Dremel are Big Data platforms based on interactive analysis.

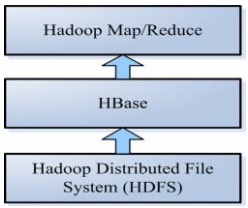

Fig. 2. Hadoop Architecture

Security issues are a great deal with big data. One of the key security issues with big data is that organisations collect and process a great deal of sensitive information regarding customers and employees, intellectual property, trade secrets and financial information. As organisations look to gain value from such information, they are increasingly seeking to combine data from a wider range of stores and applications to provide more contexts to increase the value of the data. By centralising data in one place, it becomes a target for attackers, which leads to the information exposed easily and could mitigate trust in the organisation and damage its reputation. It is due to this reason that makes it vital that big data stores are properly controlled and protected. The solution would be to create a new approach to assure the big data security and privacy policies.

IV. PROPOSED WORK

There are also security advantages to big data. When data is stored and centralised, organisations should first classify the information and apply proper controls to it, such as imposing retention periods as specified by the regulations that they face. This will allow organisations to remove data that has little value or that no longer needs to be kept so that it can be disposed of and is no longer available for theft or some illegal actions demanding presentation of records. Another security advantage is that large swathes of data can be mined for security events, such as malware, spear phishing attempts or frauds, such as account take over. Apart from this a new approach should be proposed to enhance the security of big data. A Holistic Approach is being developed; it denotes that in most organisations, the volume of big data generated and stored can be a major challenge, with searching such vast amounts of data – most of which is unstructured – often taking weeks or more using traditional tools. MeriTalk, an online IT community for the US Government, recently surveyed 151 professionals regarding big data and found that nine out of ten see challenges on the path to harness big data [19]. When asked what they have in place today compared to what they believe will be needed for successful big data management, respondents stated that they had, on average, 49% of the data storage and access technology that they need, 46% of the computational power. Figure 3 presents the most significant challenges in large volume of data in analytics.

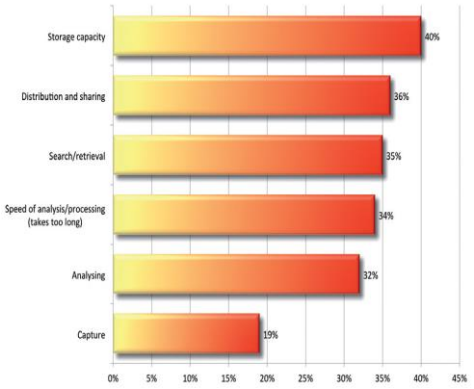

Fig. 3. Most significant challenges in managing large volume of information

Prior to the start of any big data projects, organisations need to locate and identify all of the data sources in their network, from where they originate, who created them and who can access them. This should be an enterprise effort, with input from security and risk managers, legal & policy teams that locates and index data. This also needs to be a continuous process so that both the existing data and also new data that is created throughout the network becomes uncovered [20]. The next step is to classify the data that has been discovered according to its sensitivity and business complexity. However, data classification can be a complex and long process which remains as a factor that has been a significant problem for many to implement technologies that rely on data classification, such as data leakage prevention systems. Organisations need to take into considerations about the industry standards and government regulations to which they must depend and follow for ensuring that records are retained and archived for the time periods specified and that data is protected according to the guidelines contained in some standards such as PCI DSS, which specifies that payment cardholder data is held in a secure manner. To make ease the classification process, organisations should look for automated database and network discovery tools, which can be used to scan networks to identify all data sets and analyse it. . Data warehouses are popular technologies for managing large volumes of data. But most of them depend on a relational format for storing data, which works fine for structured data, but less for unstructured data. An alternative for organisations looking to get a handle on big data is to use an open source software framework that supports data

intensive distributed applications and can work with thousands of systems in a network and peta bytes of data Hadoop [21] is one such software which is the most popular choices among organisations. Hadoop is suited for storing the large amounts of unstructured data in data stores.

V. CONCLUSION AND FUTURE WORK

As data volumes continue to expand, much of which is in unstructured form, organisations are looking forward o extract value from that data to open the opportunities for the business that it contains. However, traditional data storage and analysis tools are not up to the task of processing and analysing the information the data contains, focusing not just to the volume of data, but also to the unstructured, adhoc nature of much of the content. In addition, the centralised nature of big data stores creates new security challenges to which organisations must respond, which requires that controls are placed around the data itself, rather than the applications and systems that store the data.

Future work should mainly focus on aspects like as they go through the data classification process, organisations should also try to develop or update policies regarding data handling, such as defining the types of data that must be stored and for how long, where they should be stored and how data will be accessed when they are in need. Implementing such policies will prevent users from creating their own data stores that are outside the control of the IT department.

ACKNOWLEDGMENT

The authors would like to express our special thanks to insightful suggestions for the anonymous reviewers. Also the authors would extend the support rendered by the management for carrying out the research work successfully.

REFERENCES

1. Doug Laney, 3d Data managment: controlling data volume, velocity and variety, Appl. Delivery Strategies Meta Group (949), 2001.
2. Paul Zikopoulos, Chris Eaton, Paul. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill Professional, 2011.
3. Soares, L., "The rise of big data. Educ. Rev", 47 (3), 60, 2012.
4. Tony Hey, Anne E. Trefethen, "The uk e-science core programme and the grid", Future Gener. Comput. Syst. 18 (8), 1017–1031, 2002.
5. Bart Jacob, Michael Brown, Kentaro Fukui, Nihar. Trivedi, "Introduction to Grid Computing", IBM Redbooks Publication, 2005.
6. Fei-Yue Wang, Daniel Zeng, Kathleen M. Carley, Wenji Mao, "Social computing: from social informatics to social intelligence", IEEE Intell. Syst. 22 (2), 79–83, 2007.
7. Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White, Seung Yon Y. Rhee, Doug Howe, Maria Costanzo, "Big data: the future of biocuration", Nature 455 (7209), 47–50, 2008.
8. Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, Cheng Chen, "Data-driven intelligent transportation systems: a survey", IEEE Trans. Intell. Trans. Syst. 12 (4), 1624–1639, 2011.
9. P. Gamba, Peijun Du, C. Juergens, D. Maktav, "Foreword to the special issue on human settlements: A global remote sensing challenge", IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 4 (1), 5–7, 2011.
10. G. Li, X. Cheng, "Research status and scientific thinking of big data", Bulletin of the Chinese Academy of Sciences, 27 (6), 647–657, 2012.
11. Y. Wang, X. Jin, Xueqi, Network big data: Present and future, Chinese Journal of Computers 36 (6), 1125–1138, 2013.
12. X.-Q. Cheng, X. Jin, Y. Wang, J. Guo, T. Zhang, G. Li, Survey on big data system and analytic technology, Journal of Software 25 (9), 1889–1908, 2014.
13. James P. Ahrens, Bruce Hendrickson, Gabrielle Long, Steve Miller, Robert Ross, Dean Williams, Data-intensive science in the us doe: case studies and future challenges, Comput. Sci. Eng. 13 (6) ,14–24, 2011.
14. Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwas Dayal, Michael Franklin, Johannes Gehrke, Laura Haas, Jiawei Han AlonHalevy, H.V. Jagadish, Alexandros Labrinidis, Sam Madden, Yannis Papakon stantinou, Jignesh Patel, Raghu Ramakrishnan, Kenneth Ross, Shahabi Cyrus, Dan Suciu, Shiv Vaithyanathan, Jennifer Widom, "Challenges and Opportunities with Big Data", CYBER CENTER TECHNICAL REPORTS, Purdue University, 2011.
15. Richard T. Kouzes, Gordon A. Anderson, Stephen T. Elbert, Ian Gorton, Deborah K. Gracio, "The changing paradigm of data-intensive computing", Computer 42 (1), 26–34, 2009.
16. Grant Ingersoll, Introducing apache mahout: scalable, commercial-friendly machine learning for building intelligent applications, IBM Corporation(2009)
17. Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, Dennis Fetterly, Dryad: distributed data-parallel programs from sequential building blocks, in: EuroSys '07 Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems, vol. 41(3), 2007, pp. 59–72.
18. Jeffrey Deam, Sanjay Ghemawat, Mapreduce: simplified data processing on large clusters, Commun. ACM 51 (1) (2008) 107–113.
19. MeriTalk, "Big data gap", 2012. www.meritalk.com/bigdatagap. (last accessed 25.02.2015)
20. 'Big data and infosecurity'. Varonis, 2012. Accessed June 2012. http://blog.varonis.com/big-datasecurity. (last accessed 27.02.2015).
21. Kajeepeta, Sreedhar. 'Strategy:Hadoop and big data'.InformationWeek, 2012. .http://reports.informationweek.com/abstract/81/8670/Business-Intelligence-and Information-Management/strategy-hadoop-andbig-data.html. (last accessed 27.02.2015).