Chapter 15
Security in Big Data

15.1 Introduction

In the previous chapters, we have seen how techniques of the Big Data Analytics can be applied to various application domains such as Social Semantic Web, IOT, Financial Services and Banking, Capital Market and Insurance. In all these cases, the success of such application of the techniques of Big Data Analytics will be critically dependent on security. In this chapter, we shall examine how and to what extent it is possible to insure security in Big Data.

The World Economic Forum recently dubbed data as 'the new oil.' There is a new age gold rush in which companies such as IBM, Oracle, SAS, Microsoft, SAP, EMC, HP and Dell are aggressively organizing to maximize profits from the Big Data phenomenon [1]. Since the new oil, the most valuable resource is data now and those who are in possession of the greatest amounts of data will have enormous power and influence. Thus, companies such as Face book, Google and Acxion are creating the largest datasets about human behavior, ever created in history and they can leverage this information for their own purposes, whatsoever they may be, whether for profit, surveillance or medical research.

Similar to other valuable resources, this new most valuable resource, 'data' should be adequately protected and safeguarded with due security provisions, as called for similar resources. Unfortunately, this is missing; we do not have adequate security mechanisms presently to adequately safeguard this most valuable resource. The millions or trillions of records stored say in a supermarket database or data warehouse of a supermarket are not having any serious security for protection. The database storing such data is vulnerable and can be accessed and hacked by illegal or criminal elements. This is also true for the Big Data stored in companies like Face book and Google, with possibly a better situation, but still vulnerable for access, hacking and misuse or abuse. When the 'big data' is being stored in a vulnerable manner, our ability to capture and store information is greatly outfacing our ability to understand it or its implications. Even though the cost of storing the Big Data is coming down

C. S. R. Prabhu et al., *Big Data Analytics: Systems, Algorithms, Applications*,
https://doi.org/10.1007/978-981-15-0094-7_15

drastically, the social costs are much higher, posing huge future liabilities for Society and our World.

The more data we produce and store, the more organized crime is happy to consume.

This situation is very similar to a bank which stores too much money in one place and this will be of very much greater interest to robbers and thieves who will have much easier and better opportunity to rob. Eventually, our personal details will fall (if not already fallen) into the hands of criminal cartels, the competition or even the foreign governments. Examples of this scale of leakages are: 2010 wiki-leaks debacle, (which leaked millions of classified diplomatic information), Snowden leaks of National Security Agency (NSA) (of classified security files).

15.2 Ills of Social Networking—Identity Theft

Social media provide ready-made provisions for identity theft since all the information that the criminals are looking for is readily available online: date of birth, mother's maiden name, etc. in every Face book account. Contrary to the trust of the subscribers, the criminals will have free access to all this information in the Face book account. While resetting the TOS, Face book can override privacy options given by the subscriber and make available all the information to anyone, such as advertisers and therefore to the data brokers, including criminals. With more than 600,000 Face book accounts compromised daily, anyone's account is as vulnerable as anyone else's. In fact, criminals have created specialized tools such as targeted viruses and Trojans to take over the personal data in Face book and other social media accounts, without any personal permission. In reality, about 40% of all the social networking sites have been compromised and at least 20% of the email accounts have been compromised and taken out by criminals, without anybody's permissions. A technique called 'Social Engineering' is used by criminals posing as friends or colleagues and thereby illegally exploits the trust we repose on our trusted friends and colleagues. In a single click on a masquerading, friend 'request' or 'message' will lead virus spreading across. Sensational news stories with their links being clicked can also be misleading into viruses. For example, Koob face is a targeted Face book worldwide virus [2–8].

15.3 Organizational Big Data Security

Individual organizations also may maintain their own Big Data repositories and yet they may not have made adequate arrangements from the security perspective. If the security of Big Data is breached, it would be resulting in substantial loss of credibility and its consequent effects from the provisions of law; more than whatever is the immediate damage.

In today's new era of Big Data, various companies are using the latest technology to store and analyze petabytes of data about their own company business and their own customers. As a result, the classification of information becomes even more critical. For insuring that Big Data becomes secure, techniques such as encryption, logging, honey pot detection must be necessarily implemented. Big Data can play a crucial role in detecting fraud, in banking and financial services sectors. We can also deploy techniques for analyzing patterns of data originating from multiple independent sources to identify or anomalies and possible fraud [9–12].

15.4 Security in Hadoop

In the highly popular Hadoop framework (a Java-based distributed parallel processing system), significant security vulnerabilities can be identified. Specific techniques for handling such security issues in Hadoop environment are suggested below:

1. The W3C had identified SPARQL for protecting data originating from divergent sources. A 'secured query' concept was proposed for privacy protection.
2. Jolene proposed that processing of queries may be performed in accordance with the service provider's security policy. This will insure that only those queries which are acceptable according to the security policy will be processed, while the others will be not processed, for security reasons.
3. Access control for XML documents [13] was proposed by Bertino by adopting techniques from cryptography and digital signatures [14]. Another approach proposed by IBM researchers is that query processing of queries may be performed in a secured environment, using the mechanism 'Kerberos' (of MIT). 'Kerberos' uses an encryption technology along with a trusted third party, an arbitrator, to be able to perform a secure authentication on an open network. 'Kerberos' uses cryptographic tickets to prevent transmission of plain text passwords over the network (and 'Kerberos' is based on Needhan Shouder Protocol).
4. Airavat [15] is an access control mechanism (by Roy et al.) along with privacy, which aims at preventing leakage of information beyond the security policy of the data provider [16–21].

15.5 Issues and Challenges in Big Data Security

Data security involves not only encryption of data as a primary requirement but it shall also depend upon the enforcement of security policies for access and sharing. Also, it is required to provide security for the algorithms deployed in memory management and allocation of resources.

In industry sectors as telecom, marketing, advertising, retail and financial services, Big Data security becomes crucial.

In e-governance sector also the issues of security in Big Data scenarios assume great importance. Data explosion in the Big Data scenario will make life difficult for many industries if they do not take adequate measures of security.

15.6 Encryption for Security

Since the data is present in the clusters of Hadoop environment, it is possible for the critical information stored in it to be stolen by a data thief or a hacker. Encryption of all the data stored will be insuring security. Keys used for encryption should be different for different servers and the key information may be stored centrally, under the protection of firewall.

15.7 Secure MapReduce and Log Management

Both mappers and data are required to be accessed in the presence of an entrusted mapper.

For all MapReduce jobs which may manipulate the data, we may maintain logs along with individual user ID's of those users who executed those jobs. Auditing the logs regularly helps protecting the data.

15.8 Access Control, Differential Privacy and Third-Party Authentication

It is effective to integrate differential privacy along with access control to achieve better security. The owners or providers of Big Data sources will define the security policy and control privacy violations if they take place. Thus, the users able to perform the execution of their jobs without any data leakage and S.E Linux (Security-Enhanced Linux) [22] can be deployed for prevention of data leakages.

Security policy can be specified and supported using the Linux Security Module (LSM). By modifying the Java Virtual Machine (JVM) and MapReduce framework, it is possible to enforce differential privacy. In a cloud service, the user identity pool can be stored, so that individual identities for each application will not be required to be stored.

In addition to the above, third-party authentication is also supported by cloud service provider. The third party will be trusted by both cloud service provider and the user who is accessing the data offered in the cloud service. This third-party authentication will add an additional layer of security to the cloud service. Third-party publication of data required for outsourcing of data also is for external publication

purposes. The machine itself serves and plays the role of a third-party publisher when the data is stored in the cloud.

15.9 Real-Time Access Control

Operational control within a database in the cloud can be used to prevent configuration drift and/or unauthorized changes to the application. For this purpose, the parameters such as IP address, time of the day, authentication methods—all can utilize. It will also be better to keep the security administrator different from a database administrator. For protecting sensitive data, label security method can be implemented by affixing data labels or by classifying data as public, confidential or sensitive. The user will also have labels affixed to them similarly. When the user attempts to access, the user's label can be matched with data classification label and only then the access can be permitted to the user. The prediction, detection and prevention of possible attacks can be achieved by log tracking and auditing. Fine-grain auditing (such as column auditing) also is possible by deploying appropriate tools (such as those offered in DBMSs such as Oracle).

15.10 Security Best Practices for Non-relational or NoSQL Databases

Non-relational databases or NoSQL databases are not yet evolved fully with adequate security mechanisms. Robust solutions to NoSQL injunction are still not matured, as each NoSQL database is aimed at a different modeling, objective, where security was not exactly a consideration. Developers using NoSQL databases are usually dependent on security embedded in the middleware only, as NoSQL databases do not explicitly provide for support for enforcing security.

15.11 Challenges, Issues and New Approaches Endpoint Input, Validation and Filtering

Many Big Data systems acquire data from endpoint devices such as sensors and other IOT devices. How to validate the input data to create trust that the data received is not malicious and how to filter the incoming data?

Real-Time Security Compliance Monitoring
Given the large number of alerts that may be generated by security devices, real-time security monitoring is a challenge. Such alerts correlated or not may lead to

many false positives which may be ignored or 'clicked away' by humans who cannot cope up with the large numbers. This problem is going to be serious in Big Data scenario where the input data streams are large and are incoming with high velocity. Appropriate security mechanisms for data stream processing are to be evolved.

Privacy-Preserving Analytics
Big Data can be viewed as big brother, invading privacy with invasive marketing, decreased civil freedom and increased state control. Appropriate solutions are required to be developed.

15.12 Research Overview and New Approaches for Security Issues in Big Data

The security research in the context of the Big Data environment can be classified into four categories according to NIST group on Big Data security: Infrastructure security, data privacy, data management and integrity/reactive security. In the context of infrastructure security for Big Data, the Hadoop environment becomes the focus. There is a proposal for G-Hadoop, an extension of the MapReduce framework to run multiple clusters that simplifies user authentication and offer mechanisms to protect the system from traditional attacks [23]. There are also new proposals for a new scheme of [24], a secure access system [25] and encryption scheme [26]. High availability is proposed for Hadoop environment [27] wherein multiple active node names are provided at the same time. New infrastructures of storage system for improving high availability and fault tolerance are also provided [27, 28]. Alternative architectures for Hadoop file system which when combined with network coding and multimode reading enable better security [29]. By changing the infrastructure of the nodes and by the deploying certain specific new protocols, better secure group communication in large scale networks is achieved by Big Data systems.

Authentication
An identity-based sign cryption scheme for Big Data is proposed in [30].

In the context of the Big Data, the access control problem is addressed and techniques are proposed for enforcing security policies at key, value level [31] and also a mechanism of integrating all access control problem features is proposed [32].

In the context of data management, security provision can be made at collection or storage. One solution proposed [33] suggests that we can divide the data stored in Big Data system into sequenced parts and storing them in different cloud storage providers.

In the context of integrity or reactive security, the Big Data environment is characterized by its capacity to receive streams of data from different origins and with distinctive formats whether structural or unstructured. The integrity of data needs to be checked that it can be used properly. On the other hand, Big Data itself can be

applied for monitoring security so as to detect whether a system is newly attacked or not.

Traditionally, integrity is defined as the maintenance of consistency, accuracy and trustworthiness of data. It protects the data from unauthorized alteration during its life cycle.

Security comprises of integrity, confidentiality and availability. While insuring integrity is critical, the management of integrity in Big Data scenario is very difficult. Proposals have been made for external integrity verification of the data [34] or a framework to insure it during a MapReduce process [35].

In the context of the possible attacks on Big Data systems by malicious users, where detection [36] can be made by provenance data related to the MapReduce process [37].

Recovery from disaster in a Big Data system also is an important problem to solve by providing adequate mechanisms for recovery.

15.13 Conclusion

In this chapter, we have identified the security vulnerabilities and threats in Big Data and also summarized the possible techniques as remedial measures.

15.14 Review Questions

1. How the Big Data scenario in the context of social networking is vulnerable and what are the security risks?
2. Is there adequate protection insured for data in Big Data?
3. Explain the problems of Identity theft in social networks.
4. Explain organizational Big Data security threads and protection mechanisms.
5. Explain social engineering thread.
6. Explain security provisions in Hadoop.
7. Explain ‘Kerboros.’
8. Explain the role of encryption in Big Data security.
9. How can we deploy secure MapReduce and log management?
10. Explain access control, deferential privacy and third-party indication.

References

1. M. Goodman, *Future Crimes* (Bantam Press, 2015)
2. H.S. Rekha, C. Prakash, G. Kavitha, Understanding trust and privacy of Big Data in social networks—a brief review. In *Proceedings of the 2014 3rd International Conference on Eco-Friendly Computing and Communication Systems (ICECCS 2014)*, Bangalore, India, 18–21 December 2014, pp. 138–143
3. A. Mantelero, G. Vaciago, Social media and Big Data, in *Cyber Crime and Cyber Terrorism Investigator's Handbook* (Syngress: Boston, MA, USA, 2014), pp. 175–195
4. V. Estivill-Castro, P. Hough, M.Z. Islam, Empowering users of social networks to assess their privacy risks, in *Proceedings of the 2014 IEEE International Conference on Big Data*, Washington, DC, USA, 27–30 Oct 2014, pp. 644–649
5. H. Ren, S. Wang, H. Li, Differential privacy data aggregation optimizing method and application to data visualization, in *Proceedings of the 2014 IEEE Workshop on Electronics, Computer and Applications (IWECA 2014)*, Ottawa, ON, Canada, 8–9 May 2014, pp. 54–58
6. L. Xu, C. Jiang, Y. Chen, Y. Ren, K.J.R. Liu, Privacy or utility in data collection? A contract theoretic approach. IEEE J. Sel. Top. Signal Proc. **9**, 1256–1269 (2015)
7. A.S. Weber, Suggested legal framework for student data privacy in the age of big data and smart devices, in *Smart Digital Futures*, vol. 262 (IOS Press: Washington, DC, USA, 2014)
8. D. Thilakanathan, R. Calvo, S. Chen, S. Nepal, Secure and controlled sharing of data in distributed computing, in *Proceedings of the 16th IEEE International Conference on Computational Science and Engineering (CSE 2013)*, Sydney, Australia, 3–5 Dec 2013, pp. 825–832
9. J.B. Frank, A. Feltus, The widening Gulf between genomics data generation and consumption: a practical guide to Big Data transfer technology. Bioinf. Biol. Insights **9**(Suppl. 1), 9–19 (2015)
10. J.J. Stephen, S. Savvides, R. Seidel, P. Eugster, Program analysis for secure big data processing, in *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering*, Vasteras, Sweden, 15–19 Sept 2014, pp. 277–288
11. J. Chen, Q. Liang, J. Wang, Secure transmission for big data based on nested sampling and coprime sampling with spectrum efficiency. Secur. Commun. Netw. **8**, 2447–2456 (2015). [CrossRef]
12. V. Chang, Towards a Big Data system disaster recovery in a private cloud. Ad Hoc Netw. **35**, 65–82 (2015). [CrossRef]
13. E. Bertino, S. Castano, E. Ferari, M. Mesiti, Specifying and enforcying access control policies for XML documents sources 139–151 (2004)
14. E. Bertino et al., Specifying and enforcing security polices in XML document sources, 139–151. Open Circus Summit (OCS), 2012 seventh, Beijing, 19–29 June 2012). For imposing one additional trusted security layer, authentic third party distribution of XML documents was also propos [3]
15. A. Kilzer, E. Witchel, I. Roy, V. Shmatikov S.T.V. Setty, Airavat security and privacy for map reduce
16. C.-T. Yang, W.-C. Shih, L.-T. Chen, C.-T. Kuo, F.-C. Jiang, F.-Y. Leu, Accessing medical image file with co-allocation HDFS in cloud. Future Gener. Comput. Syst.**43**–**33**, 61–73 (2015). [CrossRef]
17. Z. Wang, D. Wang, NCluster: using multiple active name nodes to achieve high availability for HDFS, in *Proceedings of the 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC)*, Zhangjiajie, China, 13–15 Nov 2013, pp. 2291–2297
18. J. Kepner, V. Gadepally, P. Michaleas, N. Schear, M. Varia, A. Yerukhimovich, R.K. Cunningham, Computing on masked data: a high performance method for improving big data veracity, in *Proceedings of the 2014 IEEE High Performance Extreme Computing Conference (HPEC)*, Waltham, MA, USA, 9–11 Sept 2014, pp. 1–6
19. Z. Quan, D. Xiao, D. Wu, C. Tang, C. Rong, TSHC: trusted scheme for Hadoop cluster, in *Proceedings of the 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)*, Xi'an, China, 9–11 Sept 2013, pp. 344–349

20. M. Kuzu, M.S. Islam, M. Kantarcioglu, Distributed search over encrypted Big Data, in *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, San Antonia, TX, USA, 2–4 March 2015 pp. 271–278
21. A. Irudayasamy, L. Arockiam, Scalable multidimensional anonymization algorithm over big data using map reduce on public cloud. J. Theor. Appl. Inf. Technol. **74**, 221–231 (2015)
22. *Security Enhenced Linux*. Security Enhanced Linux N.P. Web 13 Dec 2013
23. J. Zhao, L. Wang, J. Tao, J. Chen, W. Sun, R. Ranjan, J. Kolodziej, A. Streit, D. Georgakopoulos, A security framework in G-Hadoop for big data computing across distributed Cloud data centres. J. Comput. Syst. Sci. **80**, 994–1007 (2014). [CrossRef]
24. Y.-S. Jeong, Y.-T. Kim, A token-based authentication security scheme for Hadoop distributed file system using elliptic curve cryptography. J. Comput. Virol. Hacking Tech. **11**, 137–142 (2015). (Cross Ref)
25. B.A. Kitchenham, D. Budgen, O. Pearl Brereton, Using mapping studies as the basis for further research—A participant-observer case study. Inf. Softw. Technol. **53**, 638–651 (2011). [CrossRef]
26. J.C. Cohen, S. Acharya, Towards a trusted HDFS storage platform: mitigating threats to Hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection. J. Inf. Secur. Appl. 19, 224–244 (2014). [CrossRef]
27. M.A. Azeem, M. Sharfuddin, T. Ragunathan, Support-based replication algorithm for cloud storage systems, in *Proceedings of the 7th ACM India Computing Conference*, Nagpur, India, 9–11 Oct 2014, pp. 1–9
28. P. Meye, P. Raipin, F. Tronel, E. Anceaume, Mistore: a distributed storage system leveraging the DSL infrastructure of an ISP, in *Proceedings of the 2014 International Conference on High Performance Computing & Simulation (HPCS)*, Bologna, Italy, 21–25 July 2014, pp. 260–267
29. Y. Ma, Y. Zhou, Y. Yu, C. Peng, Z. Wang, S. Du, A novel approach for improving security and storage efficiency on HDFS. Procedia Comput. Sci. **52**, 631–635 (2015). [CrossRef]
30. G. Wei, J. Shao, Y. Xiang, P. Zhu, R. Lu, Obtain confidentiality or/and authenticity in Big Data by ID-based generalized signcryption. Inf. Sci. **318**, 111–122 (2015). [CrossRef]
31. H. Ulusoy, P. Colombo, E. Ferrari, M. Kantarcioglu, E. Pattuk, GuardMR: fine-grained security policy enforcement for MapReduce systems, in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, Singapore, 14–17 April 2015, pp. 285–296
32. P. Colombo, E. Ferrari, Privacy aware access control for Big Data: a research roadmap. Big Data Res. **2**, 145–154 (2015). [CrossRef]
33. H. Cheng, C. Rong, K. Hwang, W. Wang, Y. Li, Secure big data storage and sharing scheme for cloud tenants. China Commun. **12**, 106–115 (2015). [CrossRef]
34. C. Liu, C. Yang, X. Zhang, J. Chen, External integrity verification for outsourced big data in cloud and IoT. Future Gener. Comput. Syst. **49**, 58–67 (2015). [CrossRef]
35. Y. Wang, J. Wei, M. Srivatsa, Y. Duan, W. Du, Integrity MR: integrity assurance framework for big data analytics and management applications, in *Proceedings of the 2013 IEEE International Conference on Big Data*, Silicon Valley, CA, USA, 6–9 Oct 2013, pp. 33–40
36. Z. Tan, U.T. Nagar, X. He, P. Nanda, R.P. Liu, S. Wang, J. Hu, Enhancing big data security with collaborative intrusion detection. IEEE Cloud Comput. **1**, 27–33 (2014). [CrossRef]
37. C. Liao, A. Squicciarini, Towards provenance-based anomaly detection in MapReduce, in *Proceedings of the 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, Shenzhen, China, 4–7 May 2015, pp. 647–656