Chapter 16
Privacy and Big Data Analytics

16.1 Introduction

While the personal information in the realm of Big Data [1] is considered the most valuable resource of the twenty-first century, the 'new oil', [2] its analytics, i.e., Big Data Analytics can be described as the 'new engine' for economic and social value creation. At the same time, the threat of loss of privacy of the personal data looms large. Enterprises or other organizations keen to harness the power of Big Data with its vast potential are also cautious and hence recognizing their responsibility to protect the privacy of personal data collected and analyzed in the Big Data framework [3].

Therefore, in any Big Data initiative, the mechanisms to govern and protect the privacy [4] in the context of risk, and thereby maintaining the adequate mechanisms to mitigate the same assume importance. The delicate balance between realizing the benefits on one side and optimizing the risk levels and resource levels on the other side is to be carefully carved out by using business frameworks such as COBIT [1].

16.2 Privacy Protection [4]

The velocity, volume and variety features of Big Data demand that the enterprises seek new ways to adequately effectively address legal, regulatory, business and operational requirements and needs. To obtain adequate measurable ROI (Return on Investment) on Big Data initiatives on the data stored in online or offline repositories, enterprises are performing analytics tasks for correlating, aggregating and statistically analyzing huge chunks of terabytes and petabytes of data in real time. Also, when the enterprises are deciding to move data into the cloud and to use cloud-based analytics services using Massively Parallel Processing (MPP) or Symmetric Multiprocessing (SMP) analytical databases, the issues of cloud security become reinforced with the issues and laws of data privacy protection. Currently, each region

C. S. R. Prabhu et al., *Big Data Analytics: Systems, Algorithms, Applications*,
https://doi.org/10.1007/978-981-15-0094-7_16

(EU or USA) is handling privacy in a different way, enterprises are forced to reconsider the methods they adopt to handle and protect the privacy of individuals and the information collected about them and how they implement cloud-based Big Data solutions. This, in turn, has impact on the software project execution and delivery. The growth of Big Data has led to distributed and disparate storages of personally identifiable health records and also online transaction such as credit card transactions. The processing of such data is liable to regulatory privacy mechanisms as Payment Card Industry Data Security Standard (PCI DSS), 1998 UK Data Protection act, US Health Insurance Portability and Accountability Act (HIPPA) COBIT offers a systematic comprehensive framework for compliance.

16.3 Enterprise Big Data Privacy Policy and COBIT 5 [1]

In any enterprise, Big Data is a given asset, and the enterprise has a responsibility to maintain privacy of data and individual as much as its commitment to derive and deliver value from its own Big Data assets. Big Data is an asset of the enterprise that will be required to fit within the domain of COBIT 5 principle, i.e., meeting the stakeholder needs COBIT 5 distinguishes clearly between corporate governance and corporate management. The corporate governance is the responsibility of the board. The needs of stakeholders of Big Data are ensured and maintained at a high level in the enterprise. The 'RACI' chart of COBIT 5 framework represents the respective responsibilities of the Board and Management with the respective roles as R (Responsible), A (Accountable), C (Consulted) and I (Informed). 'R' indicates 'taking responsibility' and takes the main operational stakes in fulfilling the activity list and creating the intended outcomes. 'A' indicates the role that is 'accountable' for the success of the task given; 'C' indicates 'Consulted' or provided impact; 'I' or 'Informed' receiving information about achievements and/or deliverable of the task. Each role (one of the R, A, C, I,) is assigned to each stakeholder. For example, Board has all others accountable ('A') to it while CEO has the responsibility ('R') to execute under consultation ('C') with CFO while giving information ('I') to the Business Process Owner and Chief Information Security Officer.

The following questions are important to be answered in the context of privacy of Big Data in the enterprises.

1. Are the sources of Big Data trust worthy?
2. What structure and skills are existing in the enterprises to govern Big Data privacy?
3. Are there the right tools to maintain Big Data privacy requirements?
4. How to ensure and maintain authenticity of data?
5. How and in what way the information will be used by whom?
6. How to improve the processes of acquiring data?
7. How to protect the sources of the data?
8. What are the insights we need to derive from Big Data in the enterprise?

9. Which are the legal and regulatory requirements containing or affecting the enterprise for which the particular data is collected?
10. Which trends are we creating which can be exploited by the rivals of the enterprise?
11. How to assure the secrecy and privacy of data and protect it from employees?

16.4 Assurance and Governance

The main steps required to drive assurance in an enterprise are:

1. Ensuring that interested parties are provided with positive substantiated opinions in governance and management of the enterprise-IT in accordance with the objectives of Assurance.
2. Defining clearly the objectives of Assurance in consistence with the Enterprise objectives so as to maximize the value of Assurance initiatives.
3. Ensuing regulatory and contractual requirement for Enterprise to provide Assurance over their own IT management.

To achieve the above objectives, the Assurance Professionals should be made the part of Big Data initiatives in the enterprise right from their inception. Such Assurance Professionals should have adequate knowledge about the business and also have expertise in using analytics tools such as R, Hadoop, Greenplum, Teradata, etc., along with the skills to interpret the data correctly to the stakeholders concerned. They shall keep abreast with the new development in tools and techniques of Big Data and also keep the management and audit teams updated on the tools to be used. The Assurance Professionals shall also ensure that Big Data privacy, and security solutions are implemented and also that adequate Big Data privacy governance exists by taking action in the following directions:

1. That data annoymization/sanitization or de-identification is ensured.
2. Adequate and relevant privacy policies and processes/procedures and supporting structures are implemented for Big Data scenarios.
3. Involve senior management and ensure their commitment to implementing the above.
4. Define and implement clear cut data destruction policies including comprehensive data management policy, data disposal, ownership and accountability.
5. Ensuring legal and regulatory compliance on Big Data and privacy assurance requirement.
6. Continuous education and training on Big Data policies, process and procedures.

Data governance is the exercise and enforcement of authority over management of data assets and performance of data functions. It comprises of pragmatic and practical steps to formalize accountability for the management of data assets: deciding who should do what and ensure that they do so; identifying assigning data stewards; a 3-d approach: De Facto, Discipline and Database.

Data stewardship is formalizing accountability for management of data resources. Data stewards need to be provided with knowledge, tools, forums and processes to become more effective and efficient in data management.

Metadata plays an important role–metadata is for management of data recorded in IT tools that improve the business and technical understanding of data and data-related processes.

Data governance requires commitment to enforcing behavioral improvements focused on managing the effectiveness and efficiency with which data will be managed. Data governance formalizes the processes for providing proactive and reactive data issue escalation and resolution. All activities of data governance should become part of the daily work process. Best practices of data governance are to be defined, and the standard processes are to be adopted for implementation. The assessment of gap between the best practices on one hand and the existing practices on the other has to be identified along with risks associated with the identified gap. Based on the gap assessment, the nature of the implementation plan for the best practices can be identified. Accordingly, an action plan can be drawn for delivering data governance program.

The best practices for implementing data governance and data stewardship include:

1. Commitment and support of the management of the organization after duel assessment and understanding by sponsoring the activities of data governance program and endorsing the efforts of the data governance team.
2. Management should ensure that the data governance team shall focus on the issues that have been identified and planned for resolution.
3. The responsibilities of data stewards should be identified and recognized.
4. The goals, scope, expectations and measurement of success of the data governance program should be well defined and communicated to all business units, functional teams, project teams and IT areas/departments of the organization for compliance.
5. It should be realized that data governance is not a single process or discipline or change of behavior.
6. Accountability towards Management about the data definition, production and usage will be the responsibility every individual identified with one or more responsibilities.
7. The data governance team will be given the responsibility to implement the planned program; it will be applicable to business data, technical data and metadata, consistently across the organization.

16.5 Conclusion

Big Data's depth represents its value and its challenge. By collecting data from different sources into a single source, Big Data promises new answers to new questions [5] hitherto never asked. But it is also fundamentally challenging regulations based on collection, identification and aggregation. At the same time, we also need to focus on transparency, expert review, efficacy, security, audit and accountability. In this chapter, we surveyed the scenario of enterprises Big Data privacy policy definition and management implementation. We have also surveyed the processes involved in enterprise data governance.

16.6 Review Questions

1. What is the threat of privacy in Big Data? How it can be instigated.
2. How privacy protection is possible in Big Data?
3. Explain enterprises Big Data privacy and COBIT.
4. Explain RACI profile.
5. Explain Assurance and governance in enterprises.
6. Explain what are the challenges and issues in privacy protects.
7. What is Trusted Scheme for Hadoop Cluster (TSHC)?
8. What is anonymization?
9. Is privacy possible in social networking scenario?

References

1. *Privacy and Big Data*, An ISACA White Paper, Aug 2013
2. D. Hirsch, The glass house effect: why Big Data is the new oil and what to do about it? *Big Data and Privacy: Making Ends Meet Digest*, pp. 44–46. ISACA White Paper, Aug 2013
3. *Big Data Impacts and Benefits*, ISACA log, March 2013
4. M. Birnhack, S-M-L-XLDATA: Big Data as a new international privacy paradigm. *Big Data and Privacy: Making Ends Meet Digest*, ISACA White Paper, Aug 2013
5. S. Freiwald, Managing the muddled mass of Big Data. *Big Data and Privacy Making Ends Meet Digest*, pp. 31–33. ISACA White Paper Aug 2013