Chapter 5
Challenges for Big Data Security and Privacy

M. Govindarajan
Annamalai University, India

ABSTRACT

Security and privacy issues are magnified by the volume, variety, and velocity of big data, such as large-scale cloud infrastructures, diversity of data sources and formats, the streaming nature of data acquisition and high volume inter-cloud migration. In the past, big data was limited to very large organizations such as governments and large enterprises that could afford to create and own the infrastructure necessary for hosting and mining large amounts of data. These infrastructures were typically proprietary and were isolated from general networks. Today, big data is cheaply and easily accessible to organizations large and small through public cloud infrastructure. The purpose of this chapter is to highlight the big data security and privacy challenges and also presents some solutions for these challenges, but it does not provide a definitive solution for the problem. It rather points to some directions and technologies that might contribute to solve some of the most relevant and challenging big data security and privacy issues.

INTRODUCTION

Big data refers to collections of data sets with sizes outside the ability of commonly used software tools such as database management tools or traditional data processing applications to capture, manage, and analyze within an acceptable elapsed time. Big data sizes are constantly increasing, ranging from a few dozen terabytes in 2012 to today many petabytes of data in a single data set. Big data creates tremendous opportunity for the world economy both in the field of national security and also in areas ranging from marketing and credit risk analysis to medical research and urban planning. The extraordinary benefits of big data are lessened by concerns over privacy and data protection.

As big data expands the sources of data it can use, the trust worthiness of each data source needs to be verified and techniques should be explored in order to identify maliciously inserted data. Information security is becoming a big data analytics problem where massive amount of data will be correlated, analyzed and mined for meaningful patterns.

DOI: 10.4018/978-1-5225-7598-6.ch005

Security of big data can be enhanced by using the techniques of authentication, authorization, encryption and audit trails. There is always a possibility of occurrence of security violations by unintended, unauthorized access or inappropriate access by privileged users.

To protect privacy, two common approaches used are the following. One is to restrict access to the data by adding certification or access control to the data entries so sensitive information is accessible to a limited group of users only. The other approach is to anonymize data fields such that sensitive information cannot be pinpointed to an individual record. For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconduct by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals (Xindong Wu et al., 2014).

BACKGROUND

Today we are living in an era of digital world. With the rapid increase in digitization the amount of structured, semi structured and unstructured data being generated and stored is exploding. Usama Fayyad (2012) has presented amazing data numbers about internet usage like "every day 1 billion queries are there in Google, more than 250 million tweets are there in Twitter, more than 800 million updates are there in Face book, and more than 4 billion views are there in You tube". Each day, 2.5 quintillion bytes of data are generated and 90 percent of the data in the world today were created within the past two years. The data produced nowadays is estimated in the order of zeta bytes, and it is growing around 40% every year. International Data Corporation (IDC) terms this as the "Digital Universe" and predicts that this digital universe is set to explode to an unimaginable 8 Zetabytes by the year 2015. The above examples demonstrate the rise of big data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to manage, capture, and process.

From a privacy and security perspective, the challenge is to ensure that data subjects (i.e., individuals) have sustainable control over their data, to prevent misuse and abuse by data controllers (i.e., big data holders and other third parties), while preserving data utility, i.e., the value of big data for knowledge/patterns discovery, innovation and economic growth.

Cloud protection alliance big data working group identify top protection and seclusion problems that need to confine for making the big data computing and infrastructure more secure. Most of these issues are linked to the big data storage and computation. There having some challenges which are related to secure data storage (Cloud Security Alliance White paper, 2012). Different security challenges related to data security and privacy are discussed in (A. A. Soofi et al., 2014) which include data breaches, data reliability, data accessibility and data support. Privacy is major concern in outsourced data. Recently, some controversies have revealed how some security agencies are using data generated by individuals for their own benefits without permission. Therefore, policies that cover all user privacy concerns should be developed. Furthermore, rule violators should be identified and user data should not be misused or leaked. The following sections describe some relevant challenges to security and privacy in the context of big data.