Chapter 4
Security and Privacy Challenges in Big Data

Dharmpal Singh
JISCE, India

Ira Nath
JISCE, India

Pawan Kumar Singh
Honeywell Labs, India

ABSTRACT

Big data refers to enormous amount of information which may be in planned and unplanned form. The huge capacity of data creates impracticable situation to handle with conventional database and traditional software skills. Thousands of servers are needed for its processing purpose. Big data gathers and examines huge capacity of data from various resources to determine exceptional novel awareness and recognizing the technical and commercial circumstances. However, big data discloses the endeavor to several data safety threats. Various challenges are there to maintain the privacy and security in big data. Protection of confidential and susceptible data from attackers is a vital issue. Therefore, the goal of this chapter is to discuss how to maintain security in big data to keep your organization robust, operational, flexible, and high performance, preserving its digital transformation and obtaining the complete benefit of big data, which is safe and secure.

INTRODUCTION

BIG DATA is a term used for a collection of data sets so large and complex that it is difficult to process using traditional applications/tools. It is the data exceeding Terabytes in size. Because of the variety of data that it encompasses, big data always brings a number of challenges relating to its volume and complexity. A recent survey says that 80% of the data created in the world are unstructured. One chal-

DOI: 10.4018/978-1-5225-9742-1.ch004

lenge is how these unstructured data can be structured, before we attempt to understand and capture the most important data. Another challenge is how we can store it. Here are the top tools used to store and analyze Big Data. We can categorize them into two (storage and Querying/Analysis).

Big data is often characterized by the 3Vs: the extreme *volume* of data, the wide *variety* of data types and the *velocity* at which the data must be processed. Those characteristics were first identified by Gartner analyst Doug Laney in a report published in 2001. More recently, several other Vs have been added to descriptions of big data, including *veracity*, *value* and *variability*. Although big data doesn't equate to any specific volume of data, the term is often used to describe terabytes, petabytes and even exabytes of data captured over time.

Such voluminous data can come from myriad different sources, such as business transaction systems, customer databases, medical records, internet clickstream logs, mobile applications, social networks, the collected results of scientific experiments, machine-generated data and real-time data sensors used in internet of things (IoT) environments. Data may be left in its raw form or preprocessed using data mining tools or data preparation software before it's analyzed.

Big data is a collection of data from various sources ranging from well-defined to loosely defined, derived from human or machine sources.

Big data also encompasses a wide variety of data types, including structured data in SQL databases and data warehouses, unstructured data, such as text and document files held in Hadoop clusters, or NoSQL systems, and semi-structured data, such as web server logs or streaming data from sensors. Further, big data includes multiple, simultaneous data sources, which may not otherwise be integrated. For example, a big data analytics project may attempt to gauge a product's success and future sales by correlating past sales data, return data and online buyer review data for that product.

Velocity refers to the speed at which big data is generated and must be processed and analyzed. In many cases, sets of big data are updated on a real- or near-real-time basis, compared with daily, weekly or monthly updates in many traditional data warehouses. Big data analytics projects ingest, correlate and analyze the incoming data, and then render an answer or result based on an overarching query. This means data scientists and other data analysts must have a detailed understanding of the available data and possess some sense of what answers they're looking for to make sure the information they get is valid and up to date. Velocity is also important as big data analysis expands into fields like machine learning and artificial intelligence (AI), where analytical processes automatically find patterns in the collected data and use them to generate insights.

A.P.Plageras et al (2018) describes how Internet of Things (IoT) supplies to everybody with latest types of services with the aim of development in our day by day living. With this innovative skill, other currently constructed technologies such as Big Data, Cloud Computing, and careful observing could be accomplished. In this work, we study the aforesaid technologies for searching their common functions, and merging their operations, in order to have advantageous situations of their usage. Instead of the boarder perception of a smart city, we will attempt to explore new systems for gathering and controlling sensors' information in a smart building which functions in IoT domain. For the proposed work, a cloud server could provide the service for gathering the information that generated from each sensor in the smart building. This information is not very hard to be controlled from distance, by a distant (mobile) device working on a network arrangement in IoT technology. As an outcome, the proposed results for gathering and controlling sensors 'information in a smart building could move us to an energy efficient green smart building.