Chapter 2
An Overview of Big Data Security With Hadoop Framework

Jaya Singh
Indian Institute of Information Technology Allahabad, India

Ashish Maruti Gimekar
Indian Institute of Information Technology Allahabad, India

S. Venkatesan
Indian Institute of Information Technology Allahabad, India

ABSTRACT

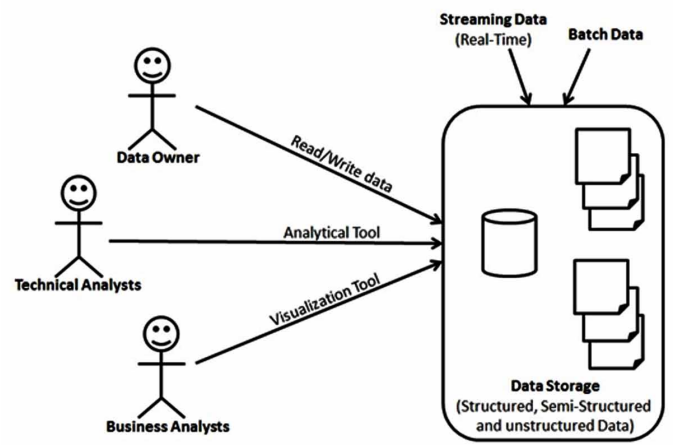
Big Data is a very huge volume of data which is beyond the storage capacity and processing capability of traditional system. The volume of data is increasing at exponential rate. Therefore, there is the need of such mechanism to store and process such high volume of data. The impressiveness of the Big data lies with its major applicability to almost all industries. Therefore, it represents both, the tremendous opportunities and complex challenges. Such omnipotent eminence leads to the privacy and security related challenges to the big data. Nowadays, security of big data is mainly focused by every organization because it contains a lot of sensitive data and useful information for taking decisions. The hostile nature of digital data itself has certain inherited security challenges. The aim of Big data security is to identify security issues and to find the better solution for handling security challenges. The observation and analysis of different security mechanism related to the issues of big data and their solutions are focused in this chapter.

1. INTRODUCTION

Big data is a phenomenon that is defined by very rapid expansion of raw data. It refers to the large volume of data which is more than the storage capacity and requires more processing power than the traditional systems. Currently we are living in the world where data is the most valuable thing. So, the important

DOI: 10.4018/978-1-5225-7501-6.ch002

Figure 1. Big Data architecture



thing is how to store, process and analyse the data, to get more knowledge from it. This large volume of data comes from many applications like sensors, social networks, online shopping portals and Government agencies. Storing and processing such data is a challenging task.

Big data is distributed everywhere across the multiple machines. It is a massive or vast collection of not only great quantity of data but also various kinds of complex data which previously never would have been considered together and it exceeds the processing capacity of conventional database system to capture, store, manage and analyse. Figure 1 shows the framework of Big Data through two data sources (real-time streaming data & batch data) and three data analysts (Data owner, technical analysts & business analysts) along with data storage infrastructure.

There are mainly three categories of data: structured data, semi-structured data and unstructured data (Bill Vohries, 2013). Structured data are highly organized data which have a pre-defined schema like relational database management system. Semi-structured data are those data which cannot be stored in rows and tables in a typical database. They have inconsistent structure like logs, tweets, sensor feeds. Unstructured data lack structure or are not structured like free form text, reports, and customer feedback forms. Big data is the combination of all the three types of data. It has to face three important challenges (B. Gerhardt et al., 2012):

- **Volume:** The volume of data is very large and cannot be processed on a single system. Its size may be in Terabytes, Petabytes and so on.
- **Velocity:** We need to fetch and process that data again and again. So we need to access it several times. So velocity is the speed to fetch data stored on particular node and the speed of the data coming in from various sources.
- **Variety:** It consists of structured, unstructured and semi-structured data. Hence managing different types of data is the main challenge.

In addition to the 3 V's there are some other challenges of big data that are presented below:

- **Veracity:** It is the quality of captured data, which can change dynamically. Veracity of data affects the accuracy of data analysis results.