Chapter 3
Privacy Preserving Big Data Publishing: Challenges, Techniques, and Architectures

Nancy Victor
VIT University, India

Daphne Lopez
VIT University, India

ABSTRACT

Data privacy plays a noteworthy part in today's digital world where information is gathered at exceptional rates from different sources. Privacy preserving data publishing refers to the process of publishing personal data without questioning the privacy of individuals in any manner. A variety of approaches have been devised to forfend consumer privacy by applying traditional anonymization mechanisms. But these mechanisms are not well suited for Big Data, as the data which is generated nowadays is not just structured in manner. The data which is generated at very high velocities from various sources includes unstructured and semi-structured information, and thus becomes very difficult to process using traditional mechanisms. This chapter focuses on the various challenges with Big Data, PPDM and PPDP techniques for Big Data and how well it can be scaled for processing both historical and real-time data together using Lambda architecture. A distributed framework for privacy preservation in Big Data by combining Natural language processing techniques is also proposed in this chapter.

DOI: 10.4018/978-1-5225-2863-0.ch003

INTRODUCTION

"Data is the new oil", declared Clive Humby, a Sheffield mathematician ('Tech giants may be huge, but nothing matches big data', 2013). Michael Palmer expanded the quote as: "Data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals etc. to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value". This is true in the case of Big Data. Data is the natural resource growing bigger and bigger each and every second. Big Data is so large amount of data that cannot be processed using traditional systems. Big Data analytics is the process of generalizing values from large data sets through which hidden patterns, unknown correlations and other useful information can be uncovered ('The state of the enterprise cloud and prepping for AWS re:Invent 2013').

The main characteristics of Big Data (4 V's) are: Volume, Velocity, Variety and Veracity (The Four V's of Big Data, 2015).

- **Volume:** The word "big" in Big Data defines the volume. The various sources of Big data include sensors, social media, activity generated data, data warehouse appliances, archives, business apps etc. (The Big 9 big data sources, 2014; Top 10 categories for big data sources and mining technologies, 2012).
- **Velocity:** This refers to the speed at which the data flows in and out of the system. Some of the examples for data generation points include mobile devices, microphones, sensors, social media etc.
- **Variety:** Big Data includes structured, semi-structured and unstructured data, which is being produced from various sources.
- **Veracity:** It refers to the inconsistencies and incompleteness in data which is collected from various sources.

In order to derive value out of this massive data, it should be collected and processed efficiently. This itself brings in a lot of challenges which includes preserving the privacy of data that is collected from various data sources at very high rates, in a variety of data formats. For processing and managing Big data, various technologies are used in the Hadoop ecosystem. This includes HDFS for storage and replication, MapReduce for distributed processing, Mahout for machine learning, Pig for scripting and so on (Khan, N et al., 2014). Data publishing plays a major role in the case of Big data as the data which is collected can be publicized for use or reuse by researchers in order to obtain valuable research output. The data can then be used for performing various data mining tasks, which helps to gain better insights about the data which is collected.