Chapter 19
Statistical Visualization of Big Data Through Hadoop Streaming in RStudio

Chitresh Verma
Amity University, India

Rajiv Pandey
Amity University, India

ABSTRACT

Data Visualization enables visual representation of the data set for interpretation of data in a meaningful manner from human perspective. The Statistical visualization calls for various tools, algorithms and techniques that can support and render graphical modeling. This chapter shall explore on the detailed features R and RStudio. The combination of Hadoop and R for the Big Data Analytics and its data visualization shall be demonstrated through appropriate code snippets. The integration perspective of R and Hadoop is explained in detail with the help of a utility called Hadoop streaming jar. The various R packages and their integration with Hadoop operations in the R environment are explained through suitable examples. The process of data streaming is provided using different readers of Hadoop streaming package. A case based statistical project is considered in which the data set is visualized after dual execution using the Hadoop MapReduce and R script.

INTRODUCTION

This chapter highlights the points related to data visualization and its steps involved in the process. Use of R programming language and RStudio as an integrated development environment will be highlighted. R language statistical feature is commonly used for data analytics and visualization with active support of RStudio or Rattle for user friendly graphical environment.

DOI: 10.4018/978-1-5225-3142-5.ch019

Traditional system of data analytics is not able to meet the demands of the "Big Data Analytics". The term "Big Data" can trace its origin to data mining which is referred by statisticians. It attempts to extract information from the data set which is not support by the traditional systems. It involves construction of statistical data model which can be visualized with underlying data pattern that lays down the idea.

The reader of this chapter is assumed to have all the basic knowledge of Hadoop and its components. It is expected that he knows the Hadoop framework setup and its operations in profundity. Therefore only R and its data visualization part of Big Data Analytics will be explored in depth.

The amalgamation of "R" programming language and Hadoop framework had developed as a solution for Big Data Analytics. The recording of unstructured data collection by various industries and institutes has led to the employment of Hadoop framework. This framework is used for storing and data computation of the records. The conjugation of R and Hadoop system appears as the ideal solution for Big Data Analytics. R and Hadoop are open source solutions available on the web and both are data driven technologies. Usage of R and Hadoop in tandem has some fundamental problems. These problems and their solutions are discussed in the upcoming sections of this chapter.

DATA VISUALIZATION

Data visualization is not only done by standard charts and graphs but also by technologically more advanced ways such as info-graphics, real-time dials and gauges, heat maps (Spakov&Miniotas, 2015). The visualization results like charts and bars are also interactive and they can be changed with a click of button. The data visualization is a well-developed domain where accomplished designers and data scientists have worked to build combination of the excellent visualization for data interpretation. It can be said that data visualization is not only creative but also decoding the data to the viewer is meaningful. In other words, connecting the gap between the actual data and logical inference is possible only by data visualization. A data designer uses his imagination to build the representation of the data which can easily be comprehended by the audience. All the combinations of data and its illustrations have the above mentioned sole purpose.

What Is Data Visualization?

Data visualization is the process of extracting the meaningful information from vast amount of data and then showing them in pictorial representation form for better understanding of the end users (Chen et al., 2007). Data visualization is science of filtering and isolating the data and then visualizing in different representation techniques.

The product of data visualization to the viewer may look as information moving from point A to point B. The data visualization process does not only involve designing the reports and charts but presenting it in a way that spectator can interpret the with least amount of effort.

Applications of Data Visualization

Data visualization is useful in areas of science and technologies. The data visualization can be broadly found in five areas: