Chapter 15
Big Data Analytics Using Apache Hive to Analyze Health Data

Pavani Konagala
Vaagdevi College of Engineering, India

ABSTRACT

A large volume of data is stored electronically. It is very difficult to measure the total volume of that data. This large amount of data is coming from various sources such as stock exchange, which may generate terabytes of data every day, Facebook, which may take about one petabyte of storage, and internet archives, which may store up to two petabytes of data, etc. So, it is very difficult to manage that data using relational database management systems. With the massive data, reading and writing from and into the drive takes more time. So, the storage and analysis of this massive data has become a big problem. Big data gives the solution for these problems. It specifies the methods to store and analyze the large data sets. This chapter specifies a brief study of big data techniques to analyze these types of data. It includes a wide study of Hadoop characteristics, Hadoop architecture, advantages of big data and big data eco system. Further, this chapter includes a comprehensive study of Apache Hive for executing health-related data and deaths data of U.S. government.

INTRODUCTION

In today's life, web is playing an important role. A large amount of data is available online. These data are getting generated from various sources such as twitter, face book, cell phone GPS data, healthcare etc. Big data analytics (Chen et al, 2014) is the process of collecting and analysing large complex data sets containing a variety of data types to find customer preferences and other useful information. The processing of such data is difficult using traditional data processing applications. Therefore, to manage and process these types of data requires a new set of frameworks. Hadoop is an open software project for structuring Big Data and for making this data useful for analytics purposes. The creator of this software is Doug Cutting. He is an employee at Yahoo for the Nutch search engine project. He named it after seeing his son's toy elephant. The symbol for Hadoop is a yellow elephant. Hadoop serves as a core platform to enable the processing of large data sets over cluster of servers. These servers are designed to be scalable with high degree of fault tolerance.

DOI: 10.4018/978-1-5225-5852-1.ch015

- **Seven V's of Big Data Analytics:** The Big Data (Sagiroglu et al, 2013) is broken into seven dimensions: Volume, Variety, Velocity, Veracity, Visualisation, Variability and Value.
  - **Volume:** Volume is the amount of data. The volume of data stored in an organisation has grown from megabytes to petabytes. The big volume represents Big Data.
  - **Variety:** Variety refers to the many sources and types of data such as structural, semi structural and un structural.
  - **Velocity:** It deals with the speed at which data flows from different sources such as social media sites, mobile device, business process, networks and human interaction etc. This velocity of data should be handled to make valuable business decisions.
  - **Veracity:** It is virtually worthless, if the data set being analysed is incomplete and inaccurate. This may happen due to the collection of data set from various sources with different formats, with noise and errors. Large amount of time may be involved to clean up this noisy data rather than analysing it.
  - **Visualisation:** Once the data set is processed it should be presented in readable format. Visualisation may contain many parameters and variables which cannot be represented using normal graphical formats or spread sheets. Even three-dimensional visualisations also may not help. So, the visualisation has become a new challenge of Big Data Analytics. AT & T has announced a new package called Nanocubes for visualisation.
  - **Variability:** Variability refers to the data set whose meaning and interpretations changes constantly. These changes occur depending on the context. Particularly this is true with Natural Language Processing. A single word may have different meanings. Over time new meanings may be created in place of old one. Interpreting them is essential in the applications like social media analytics. Therefore, the boundless variability of Big Data presents a unique challenge for Data scientists.
  - **Value:** There is a high potential value for Big Data Analytics. In the applications such as US health care system, Big Data Analytics have reduced the spending to 12-17 percent. The Big Data offers not only new and effective methods of selling but also new products to meet previously undetected market demands. Many industries use Big Data for reducing the cost of their organisations and their customers.

Although the popular 3 V's (Volume, Velocity, and Variety) of Big Data Analytics are intrinsic but the other V's (Variability, Veracity, Value and Visualisation) are also important attributes. All of them are useful to analyse and benefit from Big Data Analytics.

- **Hadoop Advantages:** The characteristics or advantages of Hadoop which makes it best solution to handle the data is as listed below.
- **Scalability:** Depending on amount of client data more systems are added to store any amount of data. i.e. Hadoop can scale up incrementally.
- **Flexibility:** Hadoop can store any variety of data i.e. structured and un structured data or semi structured data.
- **Cost Effective:** Hadoop is an open source and can be downloaded freely.
- **Fault Tolerance:** By using the facility of Replication factor the data can replicated or duplicated on two, three or more systems. If one system crashes also the data is available on other system.
- **High Performance:** Hadoop provides high performance in presence of failures also.