Abstract

Modern internet applications, scientific applications have created a need to manage immense amounts of data quickly. According to a Study, the amount of information created and replicated is forecasted to reach 35 zettabytes (trillion gigabytes) by the end of this decade. The exponentially growing dataset is known as Big Data. Big Data is generated by number of sources like Social Networking and Media, Mobile Devices, Internet Transactions, Networked Devices and Sensors. Data mining is the process of extracting interesting, non-trivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amount of data [9]. Traditional mining algorithms are not applicable to Big data as the algorithms are not scalable. In many of these types of applications, the data is extremely large and hence there is an ample opportunity to exploit parallelism, in the management & analysis of this type of data. Earlier methods of dealing with massive data were by using the concepts of parallel processing /
An Approach to mining massive Data

computing, with a setup of multiple nodes / processors. With the advent of Internet, distributed processing using the powers of multiple servers located on the internet became popular. This led to the development of S/w frameworks, to deal with analysis & management of massive datasets. These s/w frameworks use the concept of a distributed file system, where data & computations on it can be distributed across a large collection of processors. In this paper, we propose a method for dealing with large data sets, using the concept of distributed file systems and related distributed processing, The Apache HADOOP (HDFS). HDFS is a software framework that supports data-intensive distributed applications and enables applications to work with thousands of nodes and petabytes of data. Hadoop MapReduce [1] is a software framework for distributed processing of large data sets on compute clusters, which enable most of the common calculations on large scale data to be performed on large collections of computers, efficiently & tolerant to h/w failures during computations. We include in this paper, a case study of a mining application, for mining a large data set (a Email Log) that uses the Apache Hadoop framework for preprocessing the data & converting it into a form, acceptable as input to traditional mining algorithms.

References

- Jitesh Shetty, Jafar Adibi, The Enron Email Dataset Database schema and Brief Statistical Report.
- Druba Borthakur, 2009 Microsoft Research Seattle, Hadoop Architecture and its usage at Facebook
- Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden, Michael Stonebraker, Copyright 2009 ACM, A Comparison of Approaches to Large-Scale Data Analysis
- Jimmy Lin, Chris Dyer, 2010, Data-Intensive Text Processing with MapReduce
- S. Jiawei Han and Micheline Kamber. 2006. Data mining Concepts and techniques: Morgan/Kauffman publishers.

Index Terms

Computer Science

Emerging Trends in Technology

Keywords

Hadoop Hdfs  Mapreduce  Dendograms  Clusters  Web Services  Cloud