BIG DATA AND HADOOP

ABSTRACT

"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days." – Eric Schmidt, of Google, said in 2010.

The number of people who have access to the internet today equals to the world population in 1960 i.e. 3 billion. We live in the age of BIG DATA where almost 50,000GB is created per second. To deal with this problem of Big Data, the companies like Google, yahoo where are all trying to find a solution and Google published a paper called "Google File System" which is a distributed storage system. We can say that Apache Hadoop is inspired from "Google File system". I call it this way Big data is a problem and Hadoop is one of the solution to it. Hadoop has two core components HDFS (distributed storage system), MapReduce (processing). The motive of ths tutorial is to give an overview of the Apache Hadoop framework, discussing the architectures, applications of different technologies such as PIG, HIVE, MAPREDUCE, HBASE which are built on the Hadoop file system (HDFS). The tutorial also introduces you to the Cloudera which provides the services of Apache Hadoop.