ISSN:0975-9646

Data Mining Application for Big Data Analysis

[1]Bharati Punjabi [2]Prof. Sonal Honale

[1]Scholar - M.Tech. CSE,Abha Giakwad Patil College of Engineering,Nagpur.
[2]Professor – Deptt. Of Mtech CSE Abha Giakwad Patil College of Engineering,Nagpur.

Abstract : Data mining is the application of specific algorithms for extracting patterns from data. Big Data is a new term used to identify the datasets that due to their large size and complexity, we cannot manage them with our current methodologies or data mining software tools. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. Processing or analyzing the huge amount of data or extracting meaningful information is a challenging task. The term "Big data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Difficulties include capture, storage, search, sharing, analytics and visualizing. Typical examples of big data includes web logs, RFID generated data, sensor networks, , social data from social networks, Internet text and documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and/or interdisciplinary scientific research, military. Surveillance, medical records, photography archives, video archives, and large-scale ecommerce.

Keywords— Big Data Problem, Hadoop Distributed File System, Parallel Processing, Hadoop cluster, MapReduce.

I. INTRODUCTION

Applications where data collection has grown tremendously and is beyond the capability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions . In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible.

Data is being produced at an ever increasing rate. This growth in data production is being driven by individuals and their increased use of media; organisations; the switch from analogue to digital technologies; and the proliferation of internet connected devices and systems.

There has also been an acceleration in the proportion of machine-generated and unstructured data (photos , videos, social media feeds and so on) compared to structured data such that 80% or more of all data holdings are now unstructured and new approaches and technologies are required to access, link, manage and gain insight from these data sets.

The commonly accepted definition of big data comes from Gartner who define it as high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, decision making, and process optimization. These are known as the "three Vs". Some analysts also discuss big data in terms of value (the economic or political worth of data) and veracity (uncertainty introduced through data quality issues). Government agencies hold or have access to an ever increasing wealth of data including spatial and location data, as well as data collected from and by citizens. Experience suggests that such data can be utilised in ways that have the potential to transform service design and delivery so that personalised and streamlined services, that accurately and specifically meet individual's needs, can be delivered to them in a timely manner. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it.

Apache Hadoop – It is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are commonplace and thus should be automatically handled in software by the framework.

The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop MapReduce transfers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality—nodes manipulating the data that they have on hand—to allow the data to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking.

The base Apache Hadoop framework is composed of the following modules:

Hadoop Common – contains libraries and utilities needed by other Hadoop modules;

Hadoop Distributed File System (HDFS) – is a distributed file system providing fault tolerance and designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. Hadoop provides a distributed file system (HDFS) that can store data across thousands of servers, and a means of running work (Map/Reduce jobs) across those machines, running the work near the data. HDFS has master/slave architecture. Large data is automatically split into chunks which are managed by different nodes in the hadoop cluster.