

A Research on Big Data Analysis & Processing With Data Mining Techniques

Bharati Punjabi, Prof. Sonal Honale

Scholar - MTech. CSE, Abha Giakwad Patil College of Engineering, Nagpur.

bharati.ki.p@gmail.com

Prof. Sonal S. Honale – Deptt. Of Mtech CSE Abha Giakwad Patil College of Engineering, Nagpur.

sonalhonale@gmail.com

ABSTRACT

Today increasing number of organizations are facing the problem of explosion of data and the size of the databases used in today's enterprises has been growing at exponential rates. Data is generated through many sources like business processes, transactions, social networking sites, web servers, etc. and remains in structured as well as unstructured form. Today's business applications are having enterprise features like large scale, data-intensive, web-oriented and accessed from diverse devices including mobile devices. 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. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many peta bytes of data in a single data set. Difficulties include capture, storage, search, sharing, analytics and visualizing. Typical examples of big data found in current scenario includes web logs, sensor networks, satellite and geo-spatial data, 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 cluster, Hadoop Distributed File System, Parallel Processing, MapReduce

INTRODUCTION

Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).

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