ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SURVEY ON BIG DATA ANALYSIS & PROCESSING WITH DATA MINING METHODOLOGY

Bharati Punjabi*, Prof. Sonal S. Honale
* Scholar - MTech. CSE,Abha Giakwad Patil College of Engineering,Nagpur.
Professor - MTech. CSE,Abha Giakwad Patil College of Engineering,Nagpur.

ABSTRACT

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.

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. 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. The Big Data challenge is becoming one of the most exciting opportunities for the next years.

KEYWORDS: Big Data, data mining, heterogeneity, autonomous sources.

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 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. 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