VOLUME-III, ISSUE-XII ISSN (Online): 2350-0476 ISSN (Print): 2394-207X
IMPACT FACTOR: 4.205
INTERNATIONAL JOURNAL OF MULTIFACETED AND MULTILINGUAL STUDIES

Challenges of Big Data Processing and Scheduling of Processes Using Various Hadoop Schedulers: A Survey

[1]Mrs. Bareen Shaikh
Assistant Professor
MIT Arts Commerce Science College, Alandi (D), Pune-412105

[2]Mrs. Kavita Shinde
Assistant Professor
MIT Arts Commerce Science College, Alandi (D), Pune-412105

[3]Mrs. Sangeeta Borde
Assistant Professor
MIT Arts Commerce Science College, Alandi (D), Pune-412105

Abstract:

Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. In order to process large amounts of data in an inexpensive and efficient way, open source software called Hadoop is used. Aim of Hadoop is to provide parallelize job execution across multiple nodes. Hence many scheduling algorithms have been proposed in the past. Hadoop comes with three types of schedulers namely FIFO, Fair and Capacity Scheduler. The common aim of scheduling algorithms is to reduce the execution time of parallel applications and also to solve issues related to data processing. The primary purpose of this paper is to survey on Big data management and to provide an overview on various algorithms related to job scheduling and its challenges in Hadoop and the latest advancements.

Key Words: Big Data, Hadoop, Map, Reduce, Scheduler, Job Tracker, Task Tracker.

Introduction: Big data management (BDM) is the process of collecting, storing, analyzing and visualization of large volumes of data, which can be in the form of structured, unstructured and semi-structured formats The Hadoop framework enables distributed 'big data' processing across servers that can improve application performance and offer up redundancy. Hadoop uses HDFS (Hadoop Distributed File System) for storing data and to process these data it uses MapReduce Programming model introduced by Google. The Map Reduce concept is fairly simple to understand for those who are familiar with clustered scale-out data processing solutions. It is a programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.