Abstract

Cloud computing has emerged as a model that harnesses massive capacities of data centers to host services in a cost-effective manner. MapReduce has been widely used as a Big Data processing platform, proposed by Google in 2004 and has become a popular parallel computing framework for large-scale data processing since then. It is best suited for embarrassingly parallel and data-intensive tasks. It is designed to read large amount of data stored in a distributed file system such as Google File System (GFS), process the data in parallel, aggregate and store the results back to the distributed file system. Scheduling is one of the most critical aspects of MapReduce. Also three important scheduling issues in MapReduce such as locality, synchronization and fairness exist. This paper tries to illustrate and analyze the overview of thirteen different aware scheduling algorithms with different techniques and approaches for MapReduce in Hadoop and their scheduling issues and problems. At the end, Advantages and disadvantages of these algorithms are identified.


Index Terms

Computer Science Distributed Systems

Keywords

Cloud Computing, MapReduce, Scheduling algorithms