

Abstract

Hadoop is an open source platform based on Java that is under the Apache license and is used to support applications running on big data. HDFS is a component of Hadoop that can store very large files with streaming data access and run on hardware commodity groups. HDFS is designed for large files in the number of files up to petabytes, exabytes. In today's era, Big Data has become an interesting topic to discuss with various forms of data, speed and volume data. In Big Data, there are large and small files that will be processed in HDFS. However, there are problems when HDFS files are small in large numbers, so that there are several solutions offered for large numbers of small files in HDFS using HAR (Hadoop Archive) and Combine File Input Format methods. With these two methods the small files in HDFS are resolved. then do a comparison which method is used to produce blocks and the right time is used a little.

Keywords: Hadoop, HDFS, HAR, combine file format input, small file, block.