A PERUSAL ON HADOOP SMALL FILE PROBLEM

BINCY P ANDREWS[1] & BINU A[2]

[1]Post Graduate Student, Department of IT, Rajagiri School of Engineering and Technology, Kochi, Kerala, India

[2]Assistant Professor, Department of IT, Rajagiri School of Engineering and Technology, Kochi, Kerala, India

ABSTRACT

The Hadoop Distributed File System (HDFS) is developed to store and process large data sets over the range of tera bytes & peta bytes. However, storing a large number of small files in HDFS is inefficient. Also too many small files increase the number of mappers, job coordination effort (task scheduling), less work for each map task and overall processing time. This paper lists some of the existing solutions for handling small file problem. Thus, the main objective of this survey is to help tool designers/developers and experienced end users understand the details of particular file problem & to select a possible solution, enabling them to make the best choices for their particular research interests.

KEYWORDS: Hadoop, Clusters, Small Files