

**CHENNAI – PONDICHERRY**

**RFHOC: A RANDOM-FOREST APPROACH TO AUTO-TUNING HADOOP'S CONFIGURATION**

**ABSTRACT**

Hadoop is a widely-used implementation framework of the MapReduce programming model for large-scale data processing. Hadoop performance however is significantly affected by the settings of the Hadoop configuration parameters. Unfortunately, manually tuning these parameters is very time-consuming, if at all practical. This paper proposes an approach, called RFHOC, to automatically tune the Hadoop configuration parameters for optimized performance for a given application running on a given cluster. RFHOC constructs two ensembles of performance models using a random-forest approach for the map and reduce stage respectively. Leveraging these models, RFHOC employs a genetic algorithm to automatically search the Hadoop configuration space. The evaluation of RFHOC using five typical Hadoop programs, each with five different input data sets, shows that it achieves a performance speedup by a factor of 2.11 × on average and up to 7.4 × over the recently proposed cost-based optimization (CBO) approach. In addition, RFHOC's performance benefit increases with input data set size.