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.

INTRODUCTION

MAPREDUCE is a widely used programming model for processing and generating vast data sets on large-scale compute clusters. Hadoop is the most popular open-source MapReduce framework, using which a broad set of applications have been developed, including web indexing, machine learning, log file analysis, financial analysis [4] and bioinformatics processing. A typical characteristic of these applications is that they run repeatedly with different input data sets.

EXISTING SYSTEM

The Hadoop framework has up to 190 configuration parameters, and overall performance is highly sensitive to the settings of these parameters. Because the Hadoop configuration for optimum performance is application-specific, applying the default or a single set of configuration settings optimized for a certain application to a wide range of
Do Your Projects With Domain Experts...

applications leads to suboptimal Performance

DisADVANTAGE OF Existing SYSTEM

✓ Application is extremely tedious and time-consuming, and may even cause serious performance degradation.

PROPOSED SYSTEM

In Proposed System RFHOC, a novel methodology to optimize Hadoop performance by leveraging the notion of a random forest to build accurate and robust performance prediction models for the phases of the map and reduce stage of a Hadoop program of interest.

ADVANTAGE OF PROPOSED SYSTEM

✓ Hadoop configuration setting that leads to optimized application performance. We evaluate RFHOC using 5 Hadoop benchmarks, each with 5 input data sets ranging from 50 GB to 1 TB. The results show that RFHOC speeds up Hadoop programs

✓ RFHOC’s performance benefits to increase with increasing input data set sizes.

ARCHITECTURE:
HARDWARE REQUIREMENTS:

- System: Pentium IV 2.4 GHz.
- Hard Disk: 40 GB.
- Floppy Drive: 44 Mb.
- Monitor: 15 VGA Colour.

SOFTWARE REQUIREMENTS:

- Coding Language: Java 1.7, Hadoop 0.8.1
- Database: MySql 5
- IDE: Eclipse