NOVEL APPROACH FOR AUTO TUNING HADOOP CONFIGURATION

A project report submitted in partial fulfilment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by;

B.MANASWINI
Regd.No.14811A0506

G.SUKANYA
Regd.No.14811A0519

S.NANDINI
Regd.No.14811A0561

P.CHAITANYA
Regd.No.14811A0512

Under the Esteemed Guidance of

Mr. K. VARA PRASAD, M.Tech

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

AVANTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Approved by AICTE, New Delhi & Permanently affiliated to JNTU Kakinada)
(Accredited by NAAC, UGC & NBA, AICTE)
MAKAVARAPALEM, NARSIPATNAM,
VISAKHAPATNAM DIST

(2014-2018)

AVANTHI INSTITUTE OF ENGINEERING & TECHNOLOGY

(Approved by AICTE, Permanently affiliated to JNTU Kakinada)

(Accredited by NAAC, UGC & NBA, AICTE)

MAKAVARAPALEM, NARSIPATNAM,

VISAKHAPATNAM-531113

CERTIFICATE

This is to certify that the project entitled "NOVEL APPROACH FOR AUTO TUNING HADOOP CONFIGURATION" in partial fulfilment for the of degree of Bachelor of Technology in COMPUTER SCIENCE AND ENGINEERING, at AVANTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY, MAKAVARAPALEM, VISAKHAPATNAM is an bonafied work carried out by B.MANASWINI (14811A0506), G.SUKANYA (14811A0519), P.CHAITANYA (14811A0512) S.NANDINI (14811A0561), under the guidance and supervision during 2017-2018.

(K.VARAPRASAD)

PROJECT GUIDE

(Dr. G. SATYANARAYANA)

HEAD OF THE DEPARTMENT

Head of the Department
Computer Science and Engineering
Avanthi Institute of Engg. & Technology
Tamaram (V), Makavarapalem (MD),
Narsipatnam-531113

6/4/18

EXTERNAL EXAMINER

Scanned with CamScanner

ABSTRACT

Hadoop is a widely-used implementation framework of the Map Reduce 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.

In this project, 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.

Map Reduce is a widely used programming model for processing and generation vast data sets on large scale compute clusters. The Hadoop framework has up to 190 configuration parameters, and overall performance is highly sensitive to the settings of these parameters.