ON TRAFFIC-AWARE PARTITION AND AGGREGATION IN MAPREDUCE FOR BIG DATA APPLICATIONS

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

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

Submitted by:

B.NAVEEN KUMAR
Regd. No.14811A0508

P.V.PRATHYUSHA
Regd.No.14811A0549

N.S.JYOTHI
Regd.No.14811A0546

B.TARUN RAJ
Regd.No.14811A0505

E.HARISH
Regd.No.15815A0503

Under the guidance of

Miss. D. SANTHOSHI
Assistant Professor
Department of Computer Science and Engineering

AVANTHI INSTITUTE OF ENGINEERING &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, New Delhi & Permanently affiliated to JNTU Kakinada)
(Accredited by NAAC, UGC & NBA, AICTE)
MAKAVARAPALEM, NARSIPATNAM,
VISAKHAPATNAM-531113

CERTIFICATE

This is to certify that the project entitled "ON TRAFFIC-AWARE PARTITION AND AGGREGATION IN MAPREDUCE FOR BIG DATA APPLICATIONS" in partial fulfillment for the of degree Bachelor in COMPUTER S CIENCE AND ENGINEERING, at AVANTHI INSTITUTEOF ENGINEERING AND TECHNOLOGY, MAKAVARAPALEM, VISAKHAPATNAM is an bonafied work carried out by B.NAVEEN KUMAR(14811A0508), P.V.PRATHYUSHA (14811 A0549),N.S.JYOTHI(14811A0546),B.TARUNRAJ(14811A0505),E.HARISH(148 11A0509) under the guidance and supervision during 2017-2018.

Project Guide

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

6/4/18
External Examiner

ABSTRACT

MapReduce is a simple programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a commodity cluster by exploiting parallel map tasks and reduce tasks. Although many efforts have been made to improve the performance of MapReduce jobs, they ignore the network traffic generated in the shuffle phase, which plays a critical role in performance enhancement. Traditionally, a hash function is used to partition intermediate data among reduce tasks, which, however, is not traffic-efficient because network topology and data size associated with each key are not considered.

In this project, a study on reducing network traffic cost for a MapReduce job by designing a novel intermediate data partition scheme is performed. Furthermore, this project jointly considers the aggregator placement problem, where each aggregator can reduce merged traffic from multiple map tasks. To accomplish this, a decomposition-based distributed algorithm is proposed to deal with the large-scale optimization problem for big data application and an online algorithm is also designed to adjust data partition and aggregation in a dynamic manner. The efficiency and working of the proposed algorithm has been clearly presented in the experimental results that demonstrate the extensive simulation of algorithms by reducing network traffic cost under both offline and online cases.