

An Efficient Key Partitioning Scheme for Heterogeneous MapReduce Clusters

Muhammad Hanif*, Choonhwa Lee*

* Division of Computer Science and Engineering, Hanyang University, Seoul, Republic of Korea

honeykhan@hanyang.ac.kr, lee@hanyang.ac.kr

Abstract— Hadoop is a standard implementation of MapReduce framework for running data-intensive applications on the clusters of commodity servers. By thoroughly studying the framework we find out that the shuffle phase, all-to-all input data fetching phase in reduce task significantly affect the application performance. There is a problem of variance in both the intermediate key's frequencies and their distribution among data nodes throughout the cluster in Hadoop's MapReduce system. This variance in system causes network overhead which leads to unfairness on the reduce input among different data nodes in the cluster. Because of the above problem, applications experience performance degradation due to shuffle phase of MapReduce applications. We develop a new novel algorithm; unlike previous systems our algorithm considers a node's capabilities as heuristics to decide a better available trade-off for the locality and fairness in the system. By comparing with the default Hadoop's partitioning algorithm and leen algorithm, on the average our approach achieve performance gain of 29% and 17%, respectively.

Keyword— Hadoop; MapReduce; Cloud Computing; Context-awareness; Heterogeneous system



Muhammad Hanif was born in 1989. He is a combined Master and PhD candidate in Computer Software Engineering at Hanyang University. His research interest includes Cloud & Distributed Computing, Big Data and Distributed Scheduling.



Choonhwa Lee is an associate professor in the Division of Computer Science and Engineering at Hanyang University, Seoul, South Korea. He received his B.S. and M.S. degrees in computer engineering from Seoul National University, South Korea, in 1990 and 1992, respectively, and his Ph.D. degree in computer engineering from the University of Florida, Gainesville, in 2003. His research interests include cloud computing, peer-to-peer and mobile networking and computing, and services computing technology.