Capacity-aware Key Partitioning Scheme for Heterogeneous Big Data Analytic Engines

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*— Big data and cloud computing became the centre of interest for the past decade. With the increase of data size and different cloud application, the idea of big data analytics become very popular both in industry and academia. The research communities in industry and academia never stopped trying to come up with the fast, robust, and fault tolerant analytic engines. MapReduce becomes one of the popular big data analytic engine over the past few years. 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 problems, applications experience performance degradation due to shuffle phase of MapReduce applications. We develop a new novel algorithm; unlike previous systems our algorithm considers each 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 partitioning algorithm: a). In case of 2 million key-value pairs to process, on the average our approach achieve better resource utilization by about 19%, and 9%, in that order; b). In case of 3 million key-value pairs to process, our approach achieve near optimal resource utilization by about 15%, and 7%, respectively.**

***Keyword*— Cloud and Distributed Computing, Context-aware Partitioning, Hadoop MapReduce, Heterogeneous Systems**

Muhammad Hanif was born in Pakistan. He received his B.S. degrees in computer and software engineering from University of Engineering and Technology (UET), Peshawar, Pakistan in 2012. He is currently pursuing his MS leading to PhD degree in Computer Software Engineering at Hanyang University, Seoul, South Korea. His current research interest includes Cloud & Distributed Computing, Big Data Analytic Engines, Stream Processing Frameworks, and Distributed Scheduling.

Choonhwa Lee was born in South Korea. He has been with the Division of Computer Science and Engineering at Hanyang University, Seoul, South Korea since 2004, and currently as Professor. He received his B.S. and M.S. degrees in computer engineering from Seoul National University (SNU), South Korea, in 1990 and 1992, respectively, and his Ph.D. degree in computer engineering from the University of Florida, Gainesville, in 2003. He worked as senior research engineer at LGIC Ltd from 1992 to 1998. He is a member of IEEE since 2004. His research interests include cloud computing, peer-to-peer and mobile networking and computing, and services computing technology.