

**CHENNAI – PONDICHERRY**

**HADOOP PERFORMANCE MODELING FOR JOB ESTIMATION AND RESOURCE PROVISIONING**

**ABSTRACT**

MapReduce has become a major computing model for data intensive applications. Hadoop, an open source implementation of MapReduce, has been adopted by an increasingly growing user community. Cloud computing service providers such as Amazon EC2 Cloud offer the opportunities for Hadoop users to lease a certain amount of resources and pay for their use. However, a key challenge is that cloud service providers do not have a resource provisioning mechanism to satisfy user jobs with deadline requirements. Currently, it is solely the user's responsibility to estimate the required amount of resources for running a job in the cloud. This paper presents a Hadoop job performance model that accurately estimates job completion time and further provisions the required amount of resources for a job to be completed within a deadline. The proposed model builds on historical job execution records and employs Locally Weighted Linear Regression (LWLR) technique to estimate the execution time of a job. Furthermore, it employs Lagrange Multipliers technique for resource provisioning to satisfy jobs with deadline requirements. The proposed model is initially evaluated on an in-house Hadoop cluster and subsequently evaluated in the Amazon EC2 Cloud. Experimental results show that the accuracy of the proposed model in job execution estimation is in the range of 94.97 and 95.51 pe