Apr 1st, 12:20 PM - 12:40 PM

Workload Prediction with Cost-aware Data Analytics System

Anshuman Das Mohapatra
adasmohapatra@unomaha.edu

Workload Prediction with Cost-aware Data Analytics System

Anshuman Das Mohapatra, Graduate Student, Computer Science
Faculty Mentor: Kwangsung Oh

Serverless compute resources have proved vital for managing dynamic workloads in Cloud applications due to its agility and pure pay-as-you-go pricing model. A recent work, Cocoa: Compute Cost-aware Data Analytics System, focussed on the cost-performance trade-off space by determining optimal compute resource configurations encompassing virtual machines (VMs) and serverless, i.e., how many VM and serverless instances are needed for data analytics applications to meet cost-performance goals. While this work demonstrated that exploiting heterogeneous compute resources together is beneficial for data analytics workloads, it relied on the observed (static) input values such as number of tasks, mean execution time per task, and maximum number instances while determining configurations with a simple assumption that they are valid for diverse workloads, which may be false. That is, any incorrect inputs may significantly inflate cost and degrade performance. Thus, **the first goal of this research is to offer a novel way of predicting data analytics workloads** to determine the correct aforementioned inputs by using Machine Learning techniques on top of Spark, one of the de-facto data processing engines. Events from various job (query) runs are parsed into meaningful features with which a Random Forest model is trained for predicting the query execution time and other inputs in real-time. Cocoa then consumes these predictions to determine the compute resource configurations based on the cost-performance goals. To this end, we modified Spark to incorporate real-time statistics into model tuning through the enhanced metrics. Furthermore, historical information from experiments to validate our approaches is used to design and build a new prediction model that establishes black-box optimization for determining optimal compute resources configurations that exploits both VMs and serverless without Cocoa's decision framework and any static inputs. Preliminary experimental results with 100 GB data on AWS shows that a richer and well-grounded cost-performance trade-off space can be traversed by assimilating crucial task information into data analytics systems.