Data Stream Processing Management in the Cloud

Eva Kalyvianaki

Department of Computer Science, City University London

Abstract

As users of "big data" applications expect fresh results, we witness a new breed of stream processing systems (SPS) that are designed to scale to large numbers of cloud-hosted machines. Such systems face new challenges: (i) to benefit from the "pay-as-you-go" model of cloud computing, they must scale out on demand, acquiring additional virtual machines (VMs) and parallelising operators when the workload increases; (ii) failures are common with deployments on hundreds of VMs—systems must be fault-tolerant with fast recovery times, yet low per-machine overheads. An open question is how to achieve these two goals when stream queries include stateful operators, which must be scaled out and recovered without affecting query results.

In this talk I will describe a novel approach to externalise operator state explicitly to the SPS though a set of state management primitives. State externalisation enable us to handle both scale out and recovery from operators' failures using the same primitives. Our system periodically checkpoints operator state and saves it to upstream VMs. Failed operators are recovered by restoring checkpointed state on a new VM.