
Automated Finger Pointing in Data-Intensive Distributed Systems

Priya Narasimhan
(Carnegie Mellon U.)

Abstract

Problem diagnosis (or fingerprinting) involves instrumenting systems to yield meaningful data, automatically detecting errors, failures and anomalous behavior within these systems, and ascertaining their root-cause, i.e., the underlying fault. Fingerprinting is difficult because the distributed interactions, protocols and inter-component dependencies in computer systems can cause a problem to change ``shape" or manifestation, leading to potential red herrings in problem determination. There can be many root causes of an outward manifestation of a problem and there might be insufficient information to distinguish between the various root causes. On the other hand, too much monitoring and too many error messages might overwhelm the system, obscure the root cause, and lead to increased latencies and additional resource costs. This talk describes the data and visual analysis techniques that we have developed for automated fingerprinting in data-intensive distributed systems such as MapReduce (Hadoop) -- the aim was to perform online and offline root-cause analyses in order to identify a problem node/process, to diagnose the source of the problem, and report it to the user or administrator in a meaningful/useful manner. This talk will cover both the black-box and white-box problem-diagnosis techniques that we have developed so far, and will highlight our lessons learned and experiences with these systems.
