NETWORK PERFORMANCE ANALYSIS USING SPARK

Abstract

In a huge computer cluster, it becomes very important to monitor the network traffic and modify the network components to best use the resources in the network. One way to monitor the network in real time is by producing network logs which are huge in size. This RFP proposes using cluster computing algorithms on the network logs coupled with distributed storage model such as Spark and Hadoop can be useful. Preprocessing the network logs which are mostly text files and extracting required data can be done in Spark itself as Spark API supports streaming data from network in real time. “Required Data” in this context can be the request/response time, bandwidth used by each device, or heartbeat pings of all the devices in the network. Since Spark runs on distributed storage systems, various algorithms can be used simultaneously to detect which devices are not performing well, which devices have failed in the network etc. Normally this is done in a server, but managing such huge data and various algorithms can be too much for a single server. Extending this concept to predict the needed resources by an algorithm to run on a machine, various machine learning algorithms using the MLib library can be used. A combination of these strategies to monitor the network traffic can be used to reach standards of a network in an organization.