## Log Analysis for Fraud Detection using Spark

### Log Analysis

Acts of fraud frequently involve a series of legitimate activities that individually do not warrant notice. However when they are observed in the right sequence over time, pattern recognition can detect suspicious activity. Patterns of internal or external fraud often lie in the massive amounts of unstructured machine data and log files generated by business applications and systems.

## The Challenge

To accomplish the log analysis goals, an Apache Spark system was chosen for its data resilience and faster execution time for iterative computations. In our monitoring workflow, we needed to periodically query and compute several statistics including average, min and max content size of responses, count of response codes, inbound IP addresses that breached certain thresholds.  
  
Cloudera\_Vm provides a managed service to run Spark clusters. However, the idea was to make this processing extremely cost efficient and repeatable with little to no IT overhead.

## Solution

The cloudera cluster was initially configured in Cloudera management console. A solution that fully automates Cloudera infrastructure (including spot instances) provisioning and management, to configure a Spark job.  
  
The Apache Log dataset was periodically loaded from HDFS onto Cloudera and Spark log processor was scheduled to run every day through HDFS to compute parameters for audit and fraud detection. Spark also provided auto scaling option for quick turnaround time on ad-hoc reporting. Overall, with 92% of the instances running on Spot, Spark helped save 73% over on-demand Cloudera costs with no IT overhead.