batch.ly

Log Analysis for Audit and Fraud Detection using Spark on EMR is 73% cheaper with Batchly

HIGHLIGHTS

Use Case: Periodically parse Apache Log for audit purposes and fraud detection and prevention.

Technology & Dataset Used

Amazon EMR

Apache Spark

Scala Log Processor

320GB Apache Server logs

Result per Run

- 350 instances
- 92% spot instance usage
- 73% savings over on-demand AWS costs
- No IT overhead

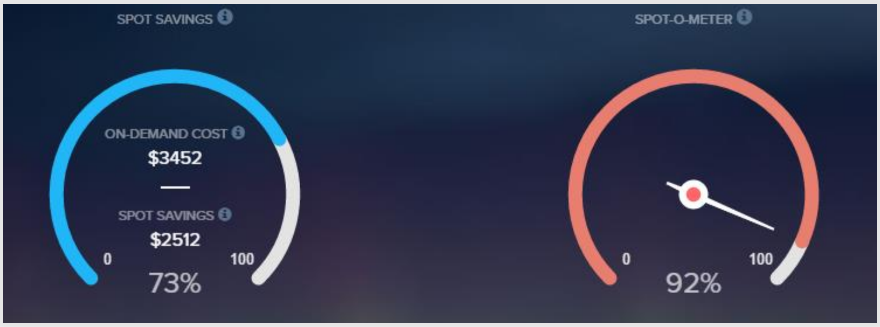
LOG ANALYSIS

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.

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.

Amazon EMR 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 EMR cluster was initially configured in Amazon EMR management console. This cluster information was then used in Batch.ly, a solution that fully automates AWS infrastructure (including spot instances) provisioning and management, to configure a Spark job.

The Apache Log dataset was periodically loaded from S3 onto EMR and Spark log processor was scheduled to run everyday through Batchly to compute parameters for audit and fraud detection. Batchly also provided autoscaling option for quick turnaround time on ad-hoc reporting. Overall, with 92% of the instances running on Spot, Batchly helped save 73% over on-demand AWS costs with no IT overhead.

Batchly - Cost effectively manage your Hadoop, Stream, Batch, Load Testing, Log Processing and Transcoding workloads on AWS

www.batchly.net