Big Data Analysis of Cloud Storage Logs using Spark

Shelly Garion, Hillel Kolodner, Allon Adir, Ehud Aharoni, Lev Greenberg
IBM Research - Haifa, Israel
{shelly, kolodner, adir, aehud, levg}@il.ibm.com

Motivation
- Investigate the logs of an operational cloud object store service to understand how it is used
- Requires going over very large amounts of historical data (e.g., PBs of records) collected over long periods
- Existing tools, such as Elasticsearch, Logstash, or Kibana are good for presenting short-term metrics, but cannot perform advanced analytics
- Apache Spark is good for log analysis and advanced analytics, but we still need to use it smartly
- Our techniques include sampling, smart grouping and aggregation, and the use of machine learning methods targeted at log data

Use-case 1: Latency analysis
- Problem: Identify time frames in which the performance decreased
- Challenge: Impractical to collect all the latencies, sort them, and calculate the exact percentiles
- Methods:
  - Focus on HEAD operations
  - Divide latencies into a histogram using the “Map/Reduce” method

Use-case 2: Archiving potential
- Problem: Estimate the potential for archiving, e.g., estimate the number of candidate objects and the expected archive size
- Challenge: Impractical to compile information for all objects that have ever been created, used, rewritten, or erased
- Methods:
  - Take a random sample of the objects
  - Two passes over the data – daily reduction and a final analysis on the daily summaries

Use-case 3: Anomaly detection
- Problem: Detect security threats and anomalies in object accesses
- Challenge: Large volume of operations on an object store and very large number of objects
- Methods:
  - Train a model of “normal” customer behavior over long time spans
  - Detect activities with significant deviations from the trained models and report alerts

Algorithm for latency analysis

Figure 1: Distribution of latency of HEAD object

Figure 2: The probability that an object will be touched again if it has not been touched for T days (for T=7, 14, 28, 56, 84) as a function of the day number

Figure 3: Abnormalities in access to objects for two accounts with high Z-scores