Design and evaluation of scalable log search engine and the anomaly detection method on large scale network

Hiroshi Abe
School of Information Science,
Japan Advanced Institute of Science and Technology

Abstract

Network administrators usually collect and store logs generated by servers, networks, and security appliances to stably manage the network systems they operate. When they find network troubles and security incidents, they investigate the contents of log information to identify the source of problems. The size of the system to store and search the log messages tends to be larger when the size of the target managed network becomes more substantial. The large-scale networks usually integrate the multi vendor's network equipment, such as routers/switches and security devices. And recent virtualization technology such as Virtual Machine, NFV, Containers cause of output the large log data. Network administrators need to search the keyword as soon as possible to solve the network troubles and incidents. But the high-speed log search system usually make cluster systems up and its hard to manage, they need to have a massive management cost to maintain the search system.

And the next problems is to detect anomaly event on a vast scale network. Network administrators monitor their networks using log data manually or using monitoring systems pay attention to the values of target devices threshold. But it's hard to integrate the automatic anomaly detection system if the network is so vast, tune thresholds of too many target devices are also difficult.

In this research, we proposed a fast log storing and searching system "Hayabusa" which is optimized for the time-dimensional search operation. We usually use a general-purpose database to store log data such as RDBMS to store the parsed data or Hadoop ecosystems to store original data. Our method treats log data like time series data, but the data is not parsed and store original data at high speed by the time structured directory design. And search speed is also fast because of the design of directory, particular database structure for full-text search and core scale processing mechanism. We also propose a simple distributed system which adds scalability to the existing Hayabusa system. The time required to perform a full-text search over 14.4 billions of records data is just 6 seconds, which is fast enough for daily operations of administrators managing a vast scale network. The distributed Hayabusa had some problems about request management and storage capacity, and we improved these problems using request management system and network storage. And we also released the distributed Hayabusa system as the Open Source Software.

About the network anomaly, we propose a method to detect abnormalities by analyzing the total amount of syslog data which collected in the event network using the Bollinger Bands algorithm used in stock trading. We performed anomaly detection by lightweight calculation in the statistical method using real data of syslog that collected by ShowNet constructed by Interop Tokyo 2016. And we evaluated the effectiveness of the Bollinger Bands algorithm.

This research result such as Hayabusa and detection of anomalies using the Bollinger band will be a proposal to contribute to reducing the operational burden for network administrators. And since Hayabusa is released as open source software, many people can benefit from using a fast log retrieval system.

Key Words: time serise data, log search engine, distributed system, anomaly detection, network management