Cloud Audit Experimentation: An efficient facility for storing and mining of the cloud event logs

Sejun Song, Praveen Rao, Baek-Young Choi, and Deep Medhi
{sjsong, raopr, choiby, dmedhi}@umkc.edu
University of Missouri-Kansas City

Cloud computing has been appealing its highly scalable, convenient, efficient, and cost effective services to become an alternative to the traditional infrastructure-oriented services. Despite its promising advantages, the wide adoption of the cloud service, however, has been hesitated by users mainly due to the fear of ***security*** including the leakage of the confidential data and loss of privacy in the cloud. Furthermore, it has been challenged by the regulatory compliance auditors to ensure ***accountability*** and ***auditability*** of the applications and data in the cloud.

For ensuring the practical cloud deployment, in addition to the traditional security components such as building firewalls to close up ports and enhancing the authentication to regulate the access, ***a detailed event tracking and recording component*** is one of the most fundamental and inevitable building blocks. Cloud event logs could be designed as free-form texts with a minimal structure, and their formats vary among different cloud providers and device vendors. Furthermore, since cloud event logs are tracking and debugging the fundamental access action events, they are often too low-level from cloud service perspectives. Due to the sheer volume (e.g., National Industrial Security Program Operating Manual (NISPOM) requirement is to record every time someone touches anything in the system, which may generate thousands of event logs for a simple access on a system.), diversity, and complexity, the typical recording process may require significant storage and network resources as well as the examination required by an auditor or on-going troubleshooting investigation could cause heavy system performance overheads.

We propose to experiment a new paradigm of cloud event record and storage system architecture that serves an essential component for the secure and accountable cloud computing. The following two issues are to be investigated from a system wide perspective:

1) Automatically transform such low-level minimally-structured event logs into meaningful and prioritized high-level cloud service events, using powerful data mining techniques tailored to the problem domain.
2) For ad hoc analytics over cloud event logs, we will leverage popular NoSQL (Not Only SQL) systems such as Apache Hive, HBase, and Pig, designed for loosely structured data. These NoSQL systems have become popular due to the limitations of relational database systems for storing and processing massive, heterogeneous datasets. For mining cloud event logs, we will employ tools such as Apache Mahout and Spark, which provide a suite of scalable machine learning algorithms.

Using Chameleon and CloudLab, we would like to evaluate the performance and scalability of the aforementioned systems and tools to store, retrieve, and analyze cloud event logs. We are also hopeful that other users of Chameleon and CloudLab will allow us to use their event logs that we want to process and understand. Through this research, we will identify potential performance bottlenecks and scalability issues that may arise in large-scale deployments. Our goal is to provide service providers with actionable insights from the event logs—in near real-time—to ensure secure and accountable usage of cloud resources. The long term vision of the research is not limited to the data description technology development and the enhancement, but 1) to replace the traditional system-wide protection concepts with the service-oriented concepts, 2) to make seamless event processing with true cloud virtualization, and eventually, 3) to equip the secure cloud computing infrastructure.

Acknowledgments
Praveen Rao would like to thank the National Science Foundation for the support under Grant No. 1115871.