LOGENTRIES INSIGHTS:

The State of Log Management & Analytics for AWS

Trevor Parsons Ph.D
Co-founder & Chief Scientist

logentries™

Introduction

The Log Management industry was traditionally driven by regulatory compliance and security concerns resulting in a multi-billion dollar market focused on security and information event management (SIEM) solutions. However, log management has evolved into a market that is focused on both the management and analytics of log data. Log management technologies are becoming more powerful and dynamic, allowing for data to be easily extracted and analyzed from logs for a much wider range of use cases. For example, unstructured events can be parsed in real-time for important field values, which can be subsequently analyzed and rolled up into metrics dashboards.

As a result, today's log management technologies can take millions of unstructured events per second, analyze them in real-time and extract key insights for:

- *Debugging during development*
- *System monitoring for IT operations*
- *Answering questions from support queries*
- *Product Usage Analytics*
- *Web and Mobile Analytics*
- *Business Analytics*

> **"Amazon Web Services offer a complete set of infrastructure and application services that enable you to run virtually everything in the cloud.**

Historically, one of the challenges of Log Management and Analytics solutions has been the requirement for end users to have deep technical skills in order to be able to extract such insights. Most solutions have focused on providing users with a powerful, yet complex, query language that can be applied to extract insights from log data. Thus, these solutions have been limited to usage by large enterprise organizations with specialist data analysts and the budget and resources required to up-skill on these technologies.

But the Log Management and Analytics industry is changing and customers today are requiring a better approach to using log management technology; one that is focused on ease of use and quick time to value. Removing the requirement for experts to operate Log Management and Analytics solutions is imperative, and will allow for the extraction of insights from log data to be accessible by a much wider range of organizations of any size. Furthermore, this will be particularly important for users of the cloud i.e. those running systems on Infrastructure as a Service (IaaS), Platform as a Service (PaaS) or Software as a Service (SaaS) components, since log data is a key resource for better understanding of these systems.

This paper will outline why Log Management and Analytics is an important technology for cloud computing. It will also do a deep dive on logging on Amazon Web Services (AWS) in particular, outlining the different sources of log and machine generated data from the available AWS services and components, as well as detailing how this data can be applied by AWS users for a range of different use cases. Finally, it will review common use cases across AWS end users.

> **"Logs give you the data you need to understand the true behavior of the system in a production environment.**

Why log management is important for the cloud

Traditionally, system operators ran their infrastructure in-house or had complete control of servers in the data center. Thus, it was possible to apply different point solutions to fully instrument the system including network devices, hardware, hypervisors, operating systems as well as the middleware and application tiers. However, today it is becoming more and more common for systems to be deployed entirely on the cloud or to at least make use of numerous cloud components. For cloud-based systems full instrumentation is often not an option since many parts of the stack may no longer be under your full control; the access required to apply instrumentation may not be available. For example, with IaaS you only have access from the operating system and up, i.e. the operating system, the middleware and application tier. The provider will control everything below the operating system such as the hypervisor layer, the hardware and the network. For those using PaaS, the situation is even more constrained since PaaS vendors tend to manage the OS and middleware components on behalf of their users. You, therefore, only have access to the application tier from an instrumentation perspective. Finally, with SaaS components, you generally do not have any ability to instrument and are required to rely on any instrumentation APIs or endpoints provided by the SaaS vendor.

As a result, it is common for traditional Application Performance Monitoring (APM) solutions, which gave 100% end-to-end visibility for on-premise systems, to only give a fraction of that visibility for cloud-based systems. It is difficult to instrument the cloud and, thus, alternative approaches are required to give visibility into cloud-based components which otherwise can become black boxes from a performance or system monitoring perspective.

While it can be difficult to instrument cloud-based components, in general they tend to produce log data streams or provide access to APIs that can be polled to generate data streams. These data streams can be analyzed to give visibility into your systems:

- **SaaS:** hosted services such as database as a service [1], email as a service [2], as well as modern day CDNs [3] provide log forwarding and event metrics via APIs that can be consumed by Log Management and Analytics solutions to provide visibility from a KPI tracking perspective.
- **PaaS:** PaaS vendors (e.g. Heroku[4], Cloud Foundry[5], and Engineyard[6]) provide log data that can contain both application and system level log events [4], as well as run time metrics [7] from the PaaS middleware that give detailed performance and error tracking information into your system. Heroku, one of the leading PaaS vendors, showcases how log data can be utilized in their Log2Viz open source project which provides a performance dashboard based on real-time analysis of performance metrics recorded in Heroku logs [8].
- **IaaS:** IaaS providers tend to provide monitoring APIs (e.g. Rackspace[9] and Amazon CloudWatch[10]). Such APIs can be polled to provide a stream of performance information on cloud server instances as well as on the different services provided by the IaaS vendor.

While instrumentation may not be possible, the existing log data and API data streams provided by cloud vendors can be analyzed by Log Management and Analytics solutions to provide real- time KPI dashboards giving deep visibility into what are often otherwise perceived as black box components.

Logging on AWS

AWS provides a range of cloud services including compute and network services, storage and CDN, database services, analytics and many others. This section outlines log data and monitoring APIs from a number of AWS services that can be polled and analyzed to give further visibility into your AWS environments.

EC2

By far the most popular service used by AWS customers is Elastic Compute Cloud (EC2). EC2 provides the ability to scale server instances on demand. Log data can be monitored on the instance operating system to capture operating system, middleware and application level data. Log data can be collected via syslog or using an agent [11], and can then be forwarded to a centralized logging server or service.

Instance level resource usage information can be captured by the CloudWatch API, which can be polled periodically to create a stream of data which can be forwarded on to a log analytics service [12].

S3

Amazon's Simple Storage Service (S3) is a reliable, fast, and inexpensive data storage infrastructure. The S3 service records access to data via access log records which can be useful for security or access audits, as well as for understanding your system data access patterns or error conditions. Access logging can be enabled [13] such that access logs are written to S3. Read only access can be subsequently granted to Log Management and Analytics services such that this data can be collected periodically.

CloudFront

Amazon CloudFront is a content delivery service which integrates with other AWS services to provide an easy way to distribute content to end users with low latency and high data transfer speeds. Log data can be enabled for CloudFront [14] such that it is written to S3. CloudFront logs record details about each request. Read only access can be subsequently granted to Log Management and Analytics services such that this data can be collected periodically.

ELB

Amazon Elastic Load Balancing (ELB) automatically distributes incoming application traffic across multiple instances to achieve greater levels of fault tolerance. Logging can be enabled for ELB such that log events are written to S3 [15]. ELB access logs capture detailed information for all requests made to your load balancer. Each log contains details such as the time a request was received, client's IP address, latencies, request path, and server responses. You can use ELB access logs to analyze traffic patterns and to troubleshoot your back-end applications. Read only access can be granted to Log Management and Analytics services such that this data can be collected periodically.

"Logentries provides an automatic setup script to get all of your AWS logs into your Logentries account in a simplified process.

RDS

Amazon Relational Database Service (RDS) provides a relational database as a service. RDS supports MySQL, Oracle and SQL Server Database engines and captures the following log data for each DB:

- *MySQL: Error, Slow Query and General Log*
- *Oracle: Alert Log and Trace Files*
- *SQL Server: Error Log, Alert Log and Trace Files*

The RDS logs allow you to identify long running queries and any database errors and can be accessed via the RDS API [16].

Elastic Beanstalk

AWS Elastic Beanstalk is Amazon's PaaS and provides application servers and app containers pre-deployed and ready for use, allowing users to reduce management of their environment. Log data is accessible from your Elastic Beanstalk server instance i.e. OS, middleware and application generated log data. AWS Elastic Beanstalk can be configured to forward log data to a centralized logging server or service [17].

CloudWatch

Amazon CloudWatch provides monitoring for AWS cloud resources and the applications customers run on AWS. CloudWatch provides resource usage information for AWS EC2 instances as well as for other AWS services such as EBS, RDS, Dynamo DB, SQS, SNS etc. [18] and also allows for the collection of custom application level metrics.

Cloudwatch information is available via the CloudWatch API that can be polled periodically and streamed into your Log Management and Analytics Service [12].

CloudTrail

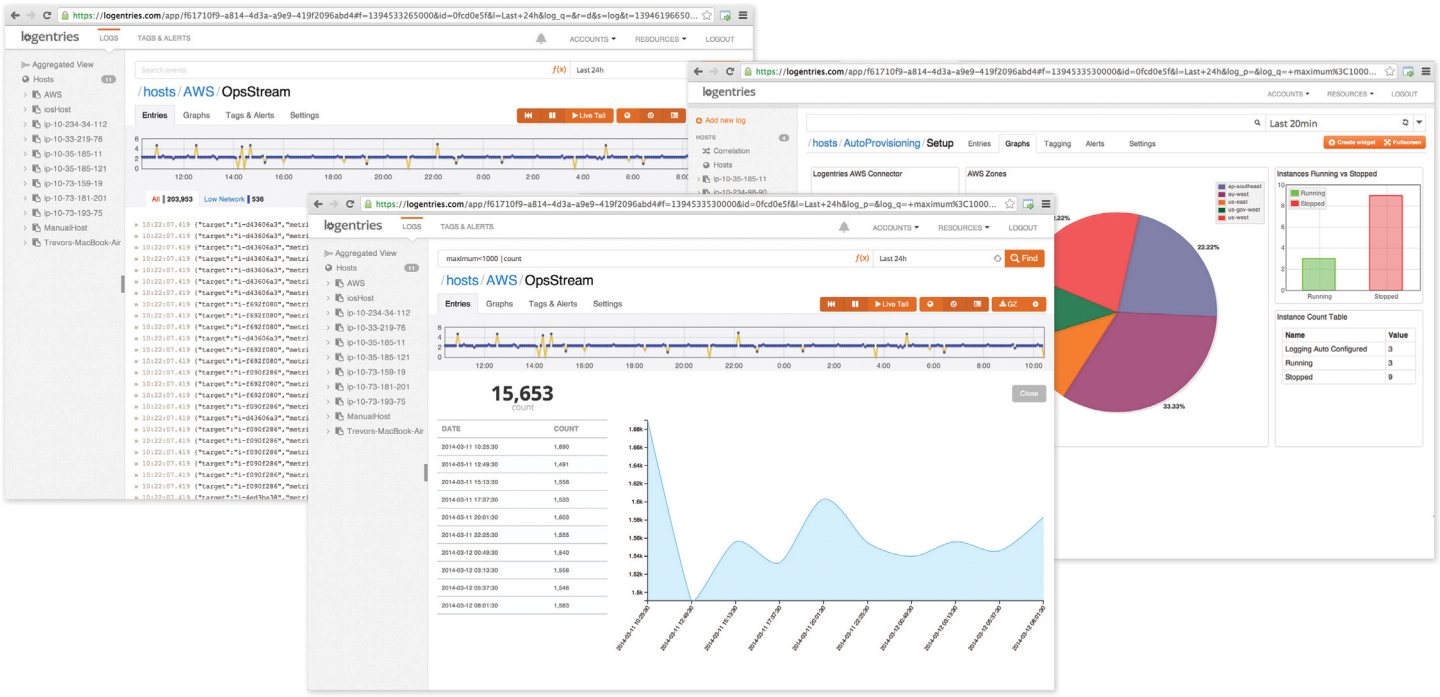
Amazon's CloudTrail is a web service that records AWS API calls within your AWS environment and stores the records in a log file. Log files are stored on S3 and read-only access can be provided to Log Management and Analytics providers such that this data can be collected periodically [19].

CloudTrail logs can be utilized for security analysis, audit trails for compliance purposes, or to simply track changes made to AWS services for IT Operation troubleshooting purposes.

Trusted Advisor

Amazon's Trusted Advisor is a service that inspects your AWS environment and makes recommendations in relation to cost efficiencies, system performance and reliability, or security.

Trusted Advisor information is available via the AWS Support API [20] which can be polled periodically and streamed into your Log Management and Analytics Service [12].



How to use log management to optimize your AWS environment

As outlined above, Amazon provides a wide range of log data and APIs to better understand your systems running within the AWS environment. Next we outline some examples of combining the above data for real world use cases.

Troubleshooting & Performance Monitoring

To effectively troubleshoot issues in your environment, or to fully understand system performance, it is always better to have the complete picture. When running your systems within AWS you can use a Log Management and Analytics solution to collect the following information to give an end-to-end view of your entire system:

- *Operating system, middleware and application level log data from your server instances*
- *Log data produced by any system components run as a service (e.g. RDS, Cloudfront, S3, ELB)*
- *Client side logs produced when code executes in the client side browser or on mobile applications [21]*
- *Resource usage information (via CloudWatch) on your server instance, AWS services, or any application specific metrics.*

Collecting this information using a Log Management and Analytics service provides a single dashboard where data can be investigated and correlated from different sources providing a number of different views into your application stack. This allows for the identification of the source of exceptions or errors within your system and the ability to pinpoint performance bottlenecks or resource usage issues. System transactions can be tracked from the client tier through the backend system components and services to identify exceptions or performance issues. Real User Monitoring (RUM) can be achieved by analyzing client side log events to better understand the perceived user experience.

Compliance

For compliance purposes it can help if you have the ability to:

- *Collect, analyze and archive all log data produced by the various system components*
- *Track any changes or access to the system*

This can be achieved within your Amazon environment by:

- *Collecting log data from your EC2 server instances and from any AWS services being used*
- *Collecting CloudTrail logs*
- *Archiving the above data for long term storage*

Log Management and Analytics solutions provide for the collection and analysis of data from your EC2 instances, AWS services, and CloudTrail logs. Furthermore they allow for archiving of this data to S3 for long-term data storage [22].

Web and App Usage Analytics

Analysis of how clients are accessing your system, or how certain parts of you system are being utilized, can be achieved by collecting and analyzing the following information:

- *Logs generated by the client's browser or mobile application*
- *Load balancer, content delivery service or web server request logs*
- *Log data from your back end system components and services*
- *Log data recording access to backend system component (e.g.via Cloudtrail)*

Combining the above information gives the ability to track and understand:

- *Client side application usage*
- *The source of web requests and related latencies or errors*
- *Activity in backend components related to app usage or internal system processes*
- *Inter-component communication as a result of application usage or internal system processes*
- *The flow of user generated transactions across client-side, front end and backend components*

What's important for AWS users

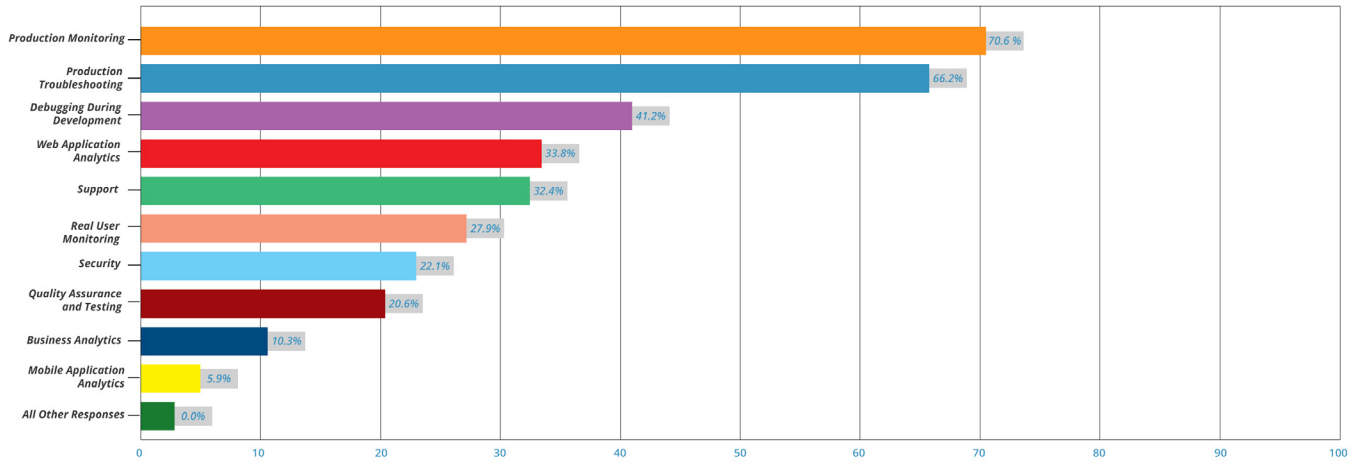
In the context of the log and API data available from AWS, and the many applicable use cases it can be applied to, Logentries conducted a survey across a sample of AWS users to understand their log management needs within the AWS environment. The survey asked more than 100 AWS users:

"What are your top 3 Log Management & Analytics needs?"

Respondents were asked to select 3 from the following options:

- *Web Application Analytics*
- *Debugging During Development*
- *Mobile Application Analytics*
- *Support*
- *Quality Assurance and Testing*
- *Security*
- *Compliance - HIPPA, PCI, SOX*
- *Production Troubleshooting*
- *Real User Monitoring*
- *Production Monitoring*
- *Business Analytics*
- *Other*

Top 10 Log Management & Analytics Needs for AWS Users



Production Monitoring (70.6%), Production Troubleshooting, (66.2%) and Debugging During Development (41.2%) were the most commonly selected logging needs for AWS users. While troubleshooting and debugging are traditional use cases for log data, using log data for production monitoring shows how AWS users expect Log Management and Analytics solutions to allow for system monitoring via more advanced capabilities (e.g. real-time alerting and performance metrics dashboards). Furthermore, the need to use log data for Web Analytics (33.8%) and Support (33.4%) suggests that there is a recognition that logs can now be applied to a wider set of use cases – i.e. beyond applying logs to the traditional troubleshooting, security and compliance tasks. Interestingly, security was actually ranked 7th (22.1%) with Compliance not selected by any respondents – further highlighting that log data is being recognized as valuable beyond these traditional use cases. Finally, other non-traditional uses also featured with 10.3% and 5.9% of respondents selecting using log data for Business Analytics and Mobile Application Analytics respectively.

The above results highlights the requirement for:

1. Log Management and Analytics technologies with capabilities beyond just "search." There is now a requirement to not only identify events in your logs, but to also be able to build metrics dashboards and reports from field values/important events and to be able to generate real-time notifications upon threshold breaches or important system events.

2. Log Management and Analytics solutions that are easy-to-use and do not require deep technical skills to operate. For example, end users looking at support issues, web analytics or business metrics will often not have deep programming skills or the ability to learn complex query languages.

Conclusions

This paper outlines why Log Management and Analytics is an important technology for cloud computing and, in particular, how log data can be applied to a new set of use cases beyond traditional use cases such as security and compliance. Due to the fact that cloud-based systems are more difficult to instrument, traditional point solutions (e.g. APM and network monitoring tools) are in many cases no longer applicable and alternative approaches to system understanding are required. Log data and data streams generated by cloud service APIs are a valuable asset that can be consumed, analyzed and correlated by Log Management and Analytics solutions for better understanding of cloud-based components and services and provide visibility into cloud systems which might otherwise be considered as black boxes.

In particular, the paper looks in detail at Amazon Web Services, the log data and cloud service APIs available and how this data can be consumed and applied by Log Management and Analytics solutions. It provides details on different use cases that log data can be applied for - such as Performance Monitoring, Troubleshooting, Web Analytics, Usage Tracking and Compliance. The paper also gives details on a survey carried out across a sample of AWS users that outlines their top 10 needs in relation to Log Management and Analytics for AWS. Non- traditional logging such as Performance Monitoring, Web analytics, Support, and Business Analytics feature in the top 10 which suggests that capabilities beyond powerful search are required by Log Management and Analytics solutions going forward. Furthermore the results also suggest that, since support teams or analysts interested in web or business analytics may not have deep technical skills, log management solutions must provide a set of features that are easy-to-use and accessible to non-technical users alongside more their advanced capabilities suitable for the more technical power users

About Logentries

Logentries is a leading SaaS-delivered log management and analytics service, developed to make machine-generated log data easily accessible to developers, operations, and business analytics customers. While traditional log management and analytics solutions require advanced technical skills to use, and are costly to set-up, Logentries provides an accessible alternative for managing huge amounts of data, visualizing insights that matter, and sharing that information across the business. With more than 35,000 users in over 100 countries, Logentries is processing billions of events everyday. To get started for free at, www.logentries.com

Trevor Parsons

Co- Founder & Chief Scientist:

Trevor Parsons is Chief Scientist and Co-founder of Logentries. Trevor spent more than ten years researching enterprise software and, in particular, specializing in developing enterprise monitoring and performance tools for distributed systems. He is also a research fellow at the Performance Engineering Lab Research Group and was formerly a Scientist at the IBM Center for Advanced Studies and holds a PhD from University College Dublin, Ireland. Today, Trevor leads the Logentries team in making our customers' jobs easier by delivering valuable business insight from log data management and analysis.

References

[1] Heroku PostGres, Metric Logs: https://devcenter.heroku.com/articles/heroku-postgres-metrics-logs
[2] SendGrid, Event webhooks: http://sendgrid.com/docs/API_Reference/Webhooks/event.html
[3] Faslty, Logging Integration: https://www.fastly.com/blog/new-fastly-logging-features
[4] Logging on Heroku: https://devcenter.heroku.com/articles/logging
[5] Cloud Foundry Logging: http://docs.cloudfoundry.com/docs/running/managing-cf/logging.html
[6] Cooking up EngineYard logs with Chef: https://blog.logentries.com/2012/10/cooking-up-engine-yard-logs-with-chef/
[7] Heroku Labs log-runtime-metrics: https://devcenter.heroku.com/articles/log-runtime-metrics
[8] Log2Viz, Logs as Data for Performance Visibility: https://blog.heroku.com/archives/2013/3/19/log2viz
[9] Monitoring on Rackspace, http://www.rackspace.com/cloud/monitoring/howitworks/,
[10] Amazon CloudWatch: http://aws.amazon.com/cloudwatch/
[11] The Logentries Agent: https://logentries.com/doc/agent/
[12] Announcing OpStream for AWS: https://blog.logentries.com/2013/11/announcing-opsstream-for-aws- combining-cloudwatch-cloudtrail-trusted-advisor/
[13] Amazon S3 Access Server Logging: http://docs.aws.amazon.com/AmazonS3/latest/dev/ServerLogs.html
[14] Amazon CloudFront Request Logging: http://aws.typepad.com/aws/2009/05/amazon-cloudfront- request-logging.html
[15] Amazon Elastic Load Balancing, Access Logging: http://docs.aws.amazon.com/ElasticLoadBalancing/latest/DeveloperGuide/access-log-collection.html
[16] Announcing point-and-click access to Amazon RDS database logs: http://aws.amazon.com/about- aws/whats-new/2013/03/04/amazon-rds-db-log-access/
[17] Configuring Elastic Beanstalk with Logentries: https://logentries.com/doc/amazon/
[18] Amazon CloudWatch Supported Services: http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/supported_services
[19] AWS Launches Cloudtrail: http://techcrunch.com/2013/11/13/aws-launches-cloudtrail-a-new-logging-api-for-governance-and-compliance/
[20] AWS Support API: http://docs.aws.amazon.com/awssupport/latest/APIReference/Welcome.html
[21] Real User Monitoring: https://blog.logentries.com/2014/02/real-user-monitoring-i-need-insight-and-i-need-it-now/
[22] Archiving Data Are you Vulnerable: https://blog.logentries.com/2014/03/archiving-data-are-you- vulnerable/