Hewlett Packard
Enterprise

Business white paper

Analyzing machine data—the best way forward

Get to root cause faster by applying machine learning and automation to log data analysis.

When something goes wrong with key IT infrastructure, applications, or services, speed is of the essence. IT needs to identify the root cause and rectify the situation in the shortest possible time frame to avoid the diverse and very real costs of downtime—from poor service levels and frustrated users to a tarnished reputation for IT and the business.

Table of contents

Executive summary

There is no shortage of data in today's IT environments to provide clues about the underlying cause of a problem or possible remedies. With the profusion of performance metrics, alerts, SNMP traps, and even non-IT data, operations analysis teams have many data sources to draw from. However, one source of valuable data has not yet been fully and effectively utilized by most companies in troubleshooting performance problems: log data.

While logs have played a role in troubleshooting and raw data analysis, the overwhelming growth in log data in recent years—and the unstructured nature of that data—has made it more difficult to search through log messages manually. It can be nearly impossible to find that crucial insight, "the needle in the haystack," that can accelerate the remediation of the problem and prevent future problems.

This paper describes a more effective approach to harvesting the insights that hide within log data—an approach based on automating log analysis and applying sophisticated machine learning to the analysis process, so that IT operations teams can pinpoint the root cause of performance issues in minutes rather than hours or days.

The promise—and the challenge—of analyzing log data

How do you solve a problem where the cause is unknown and you don't know where to start? Then there are those intermittent problems that come and go and can impact your performance and availability greatly, yet it's nearly impossible for you to fix the problem because you can't catch it. And how do you catch human error?

The answer lies in your data. Specifically, machine log data contains an enormous volume of information about system usage, performance, events, configuration changes, customer data, and so on. If you have a way to analyze that information, your IT operations analysis team can glean value to provide your business more predictable operational performance.

While the sheer volume of log data keeps growing, the nature of log data is that it is unstructured and can take on different formats—all of which makes it extremely difficult to make use of that data. For example, IT operations analysis now typically includes public, private, and hybrid clouds, on-premises and SaaS applications, BYOD and Internet of Things (IoT)-related projects, and all of their associated infrastructure. This volume and complexity mean your IT operations can break in more places, and identifying the root cause is more difficult than ever.

Recent improvements in technology have helped address this problem. The increased efficiencies and capacities of modern storage systems have reduced the cost per record for data storage and made it financially practical to capture more log data and store it for longer periods of time. Tools have also been developed to quickly search through logs—if you know what you're looking for.

However, analyzing massive volumes of logs can still be a very manual process, and traditional log analysis techniques are simply not keeping pace.

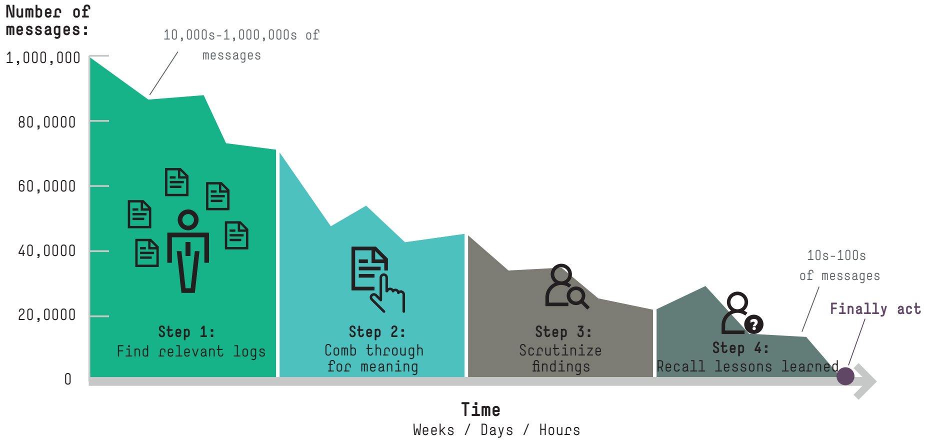


Figure 1. Human-driven, manual troubleshooting is time consuming and inefficient.

In today's typical "search-based" methodology, the steps are:

1. Decide which log source(s) are relevant for the problem at hand. Logs can be from the network, servers, databases, Web servers, applications, and so on.
2. Decide what text strings or messages to search for in the logs, to try to find which messages are important.
3. Scrutinize findings and narrow down probable causes.
4. Remember what you learned from last time to help you pinpoint the cause.

The IT administrator or investigator types in search terms, adjusting and updating them based on the results each previous search leads. This process tends to be investigative and iterative.

In addition, a search for the obvious terms, such as "error," "exception," or "critical" will usually return a large number of results, many of which occur every day during normal activity. In many cases a simple word isn't adequate to describe the target of the search, so the user needs to build regular expressions to gradually pinpoint the source of the problem. Frequently, especially if this is a new problem, the administrator would not know exactly which terms to search. As a result, meaningful messages may be missed.

The net result of the manual search process is that valuable IT resources are spent troubleshooting for days or even weeks—combing through millions of log messages with iterative searches, scrutinizing them to see if they are of value—and sometimes never finding what they're looking for.

A more intelligent approach to log analysis

1. Group similar logs together to expedite processing.
The volume, variety, and velocity of messages have made it difficult for investigators to know whether they're looking at the same message multiple times or lots of different messages. A better approach is to let the system analyze all the log messages and determine the appropriate groupings, which will significantly reduce the "noise" of millions of messages. For example, if every login generates a log-in message, and there are three million logins per day, these three million messages should be "clustered" or grouped together so that it's not necessary to go through each of them during a problem investigation.

2. Use machine learning to determine relevance of logs.
Once messages have been effectively clustered, the next step is to determine which ones are of interest or relevant to the issue at hand. There are multiple ways to determine relevance of logs: key words, frequency of logs encountered, historical relevance, patterns and anomalies, and correlation of log messages.

Interesting messages could include messages that have not been encountered, or those that are only seen rarely. Such messages could indicate that something special or different is happening and may be worthy of investigation.

For example, if the log-in messages mentioned above are common and probably of no interest, they should be disregarded. However, understanding patterns and anomalies is important. If on normal days there are three million successful logins and 1,000 failed logins, a dramatic change in these numbers could indicate a problem. A machine-learning solution would be able to highlight such patterns and the anomalies.

Correlations can also be helpful in determining relevance. Administrators should be able to see what else happened or changed in the system—such as an uptick in the event count, slower response times, reduced availability levels, or a spike in memory usage—and then find messages that occurred when this change happened in the system.

Using the log-in example again, if the number of failed log-in messages increased at roughly the same time as response time became slow, there may be a highly relevant correlation. There may be additional messages during this time that may be of interest too. The solution should help identify these relevant correlations.

3. Tune log analytics with SME expertise to optimize accuracy.
It is important that the log analysis system be able to leverage the knowledge of your subject matter experts (SMEs), including interesting messages encountered in the past, irrelevant messages ignored in previous investigations, or key words that were relevant for troubleshooting other problems. Ideally, the solution should continuously improve its effectiveness with input from SMEs.

For example, if a specific database log message was encountered the last time a service outage occurred, it should enable the database administrator to flag it so the machine remembers to flag it if it happens again. Or if a component of the service logs critical messages every time it cannot connect to a remote server to check for updates, it should be possible to flag this message as irrelevant for troubleshooting purposes.

Automated log analytics from Hewlett Packard Enterprise Operations Analytics

Embodying the best-practice log analysis methodologies described above and using patented intellectual property from HPE Labs, HPE Operations Analytics is uniquely capable of addressing the core challenges of log analysis and deriving value from operational data.

In addition to log data, HPE Operations Analytics can analyze other data types such as performance metrics, events, topology, and non-IT data.

HPE Operations Analytics applies machine learning to automatically sort through the massive volumes of log messages. It quickly and efficiently finds and identifies messages that are truly relevant, applies powerful analysis algorithms that self-learn over time, and leverages the knowledge of experts, enabling it to provide fresh insights to find the root cause of the problem every time. These insights can be applied to accelerate problem resolution and help prevent future issues.

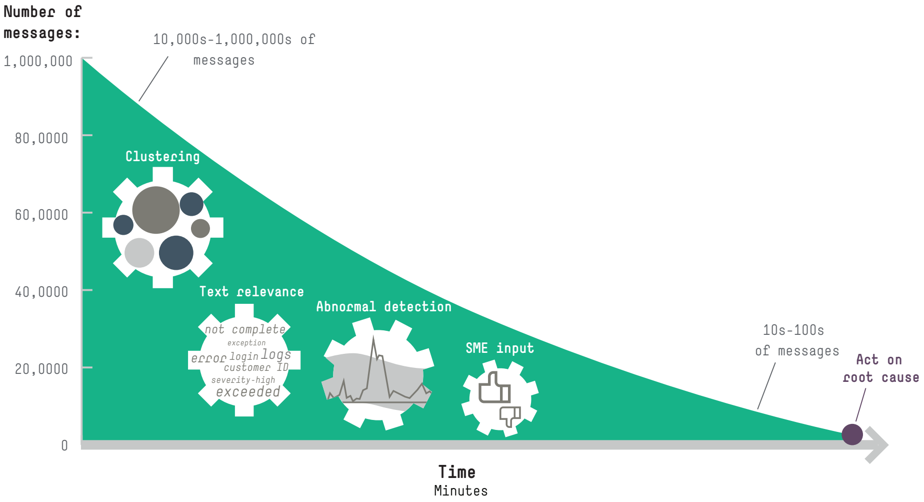


Figure 2. Automated log analytics quickly finds patterns and root cause.

Intelligent log clustering

HPE Operations Analytics uses machine learning algorithms to automatically identify messages that are similar and groups them based on pattern matching, with variations in one or more parameters. The grouping parameters are the variables within the messages; they can be singular, such as user ID, device type, or error type; or they can contain two or more compounded parameters.

For example, if there are three million daily logins to the system, and every login generates the message "User <userid> has logged into the system," HPE Operations Analytics automatically groups the three million log-in messages into a single cluster, so it knows they are essentially one type of message. HPE Operations Analytics also shows how many times this message has appeared and what values the parameter (userid) had.

HPE Operations Analytics log clustering capabilities enable users to troubleshoot more effectively with visual analysis that can uncover developing issues.

The clustering capabilities of HPE Operations Analytics provide additional value beyond troubleshooting. By grouping similar messages together, the product can isolate parameters within the clustered messages and graph them. For example, by analyzing log messages that write the connection time of a user or show the customer ID in an error log, HPE Operations Analytics gives users the ability to see behaviors over time and potentially uncover developing issues. Similarly, by allowing the user to see customer IDs in an error message, the cluster group can provide an indication whether this error is widespread or limited to a certain number of customers.

Equally important, the parameters uncovered by HPE Operations Analytics can correlate business-related data with IT data. For example, log messages may have information regarding shipments, user functionality, and other application factors that are impacted by IT performance and reliability. The line-of-business owner can use this information to guide better business decisions.

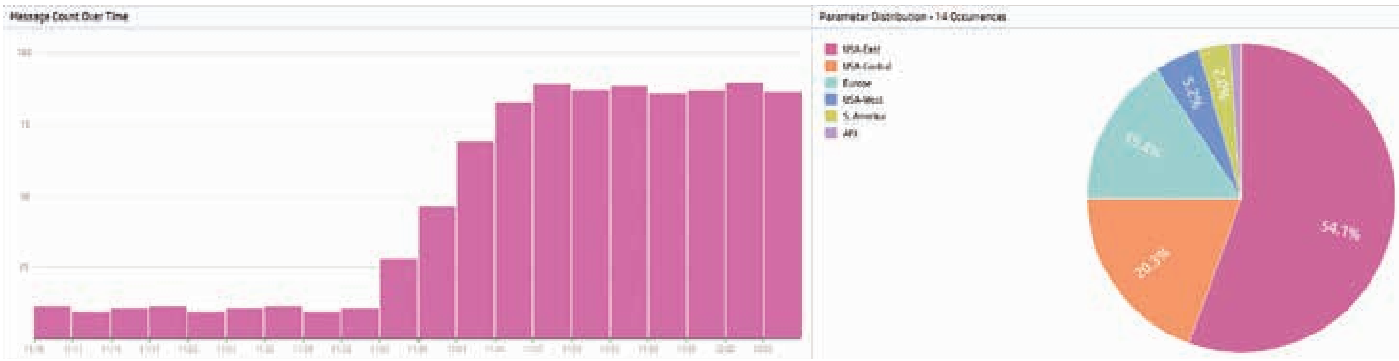


Figure 3. These sample log analytics charts show error messages over time (left) and error messages by region (right); and can help you look for patterns in your log data.

Self-learning log relevance calculation

Once clustering has narrowed down the number of unique messages to investigate, HPE Operations Analytics calculates a relevance score of the clusters of log messages based on key words, time frames, prior relevance, correlations of parameters, pattern and anomaly detection. Over time, this algorithm self-learns the patterns and parameters of what is relevant in solving the problem.

HPE Operations Analytics shows the results of its log analysis with a visual chart, automatically identifying a list of the most relevant logs.

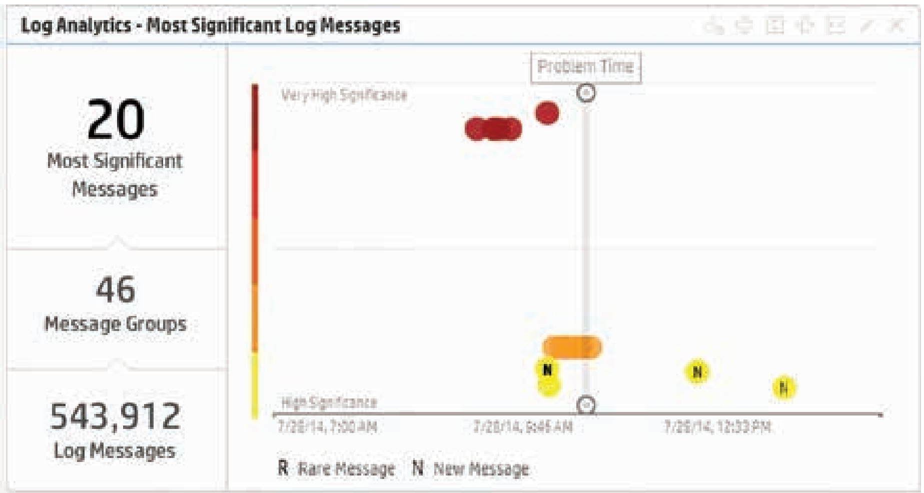


Figure 4. The user is provided with visual representation of message relevance based on sophisticated clustering, pattern matching, and ranking mechanisms.

HPE Operations Analytics identifies messages the user has never seen before (New Message), as well as messages that appear more frequently or less often than usual (Rare). Abnormal message behavior may be indicative of a message that's more significant in analyzing a problem. From here, the user can click through and drill down directly to the log to get additional details about each message to remediate the problem.

Harnessing SME expertise

All machine learning algorithms need to be tuned to work optimally to meet specific operational needs. However, HPE Operations Analytics provides machine learning root cause analysis out of the box without any need for initial setup and configuration. Instead, HPE Operations Analytics utilizes an expert-sourcing feature that allows SMEs to provide input to fine tune the algorithms as they go about their normal course of troubleshooting.

The SME can instruct the system to "ignore" a message because it is irrelevant to problem resolution; he or she also has the option to "like" a given message, an expert-sourcing feature that leverages the collective knowledge of SMEs to give that message type additional weight in relevance calculations. In addition, the SME can pre-specify the relevance of key words and text patterns. HPE Operations Analytics harnesses the collective knowledge of the organization, using its input to refine the relevance of messages.

Figure 5. Log analytics uses human input to rank the importance of messages.

Conclusion

The increasing complexity of IT operations brings with it large volumes of log data that can be extremely valuable in improving IT performance and lowering MTTR. However, processing the large volume, variety, and velocity of log data brings IT operations analysis teams to a critical choice. They can continue to search for the root causes of problems manually, in an ad hoc way, detective-style, hoping to find the needle in the haystack. Or they can make the move to an automated, systematic analysis of log data and arrive at the true root cause in minutes.

HPE Operations Analytics helps you extract highly relevant, actionable insights from your log data right out of the box, without the need for a data scientist. It brings powerful new capabilities to log analysis, and in turn brings unique advantages to the business, including:

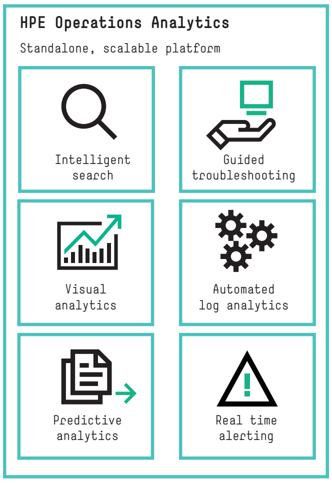
- **A better customer experience,** because problems can be identified, resolved, and prevented in far less time than with traditional techniques, so end users can complete their transactions with a faster system response time and higher availability
- **More efficient and proactive IT operations teams,** thanks to the ability of the machine learning algorithm to quickly and accurately pinpoint the root cause of problems in minutes and prevent problems from ever happening in the first place through predictive analytics
- **More value to the business with IT data,** by using the patterns automatically detected in log analytics, and by correlating non-IT and IT data to provide valuable insights to improve business revenue and efficiency

About HPE Operations Analytics

HPE Operations Analytics helps IT organizations proactively manage their operational performance by realizing the value of all their data, structured and unstructured. HPE Operations Analytics has unique log analytics and predictive analytics that utilize patented technology and “expert sourcing” intelligence. HPE Operations Analytics finds the root causeof an IT issue in minutes rather than hours or days. With HPE Operations Analytics, IT organizations can reduce MTTR by gaining actionable insights and identifying operational issues before end users are impacted.

Learn more at
hp.com/go/opsanalytics

Watch our product videos to learn more about HPE Operations Analytics and its log analytics capabilities.



Sign up for updates

★ Rate this document

Hewlett Packard Enterprise

4AA5-7132ENN, November 2015, Rev. 1