REPORT REPRINT

Big data, machine learning shape performance-monitoring developments

NANCY GOHRING

28 FEB 2017

Vendors that are taking a centralized approach to IT data collection, analytics or both, have been shaping the conversation around performance monitoring. Scale, openness and advanced analytics are key elements in IT operations analytics tools.

451 Research®

Vendors that are taking a centralized approach to IT data collection, analytics or both have been shaping the conversation around performance monitoring. We continue to see new entrants – including legacy vendors and startups – emerge with tools for collecting and analyzing the full stack of application and infrastructure data. The ability to scale, a willingness to embrace openness across several fronts, and development of machine-learning techniques are key elements of the IT operations analytics (ITOA) tools that will define the performance-monitoring space.

THE 451 TAKE

To handle the growing volume of operations data that businesses collect, vendors are building big-data back ends or advanced analytics tools that form, in essence, ITOA 2.0. The ability to scale, a willingness to embrace openness across several fronts, and development of machine-learning techniques are key elements of the new breed of tools that will define the performance-monitoring space. We expect competition in this space to intensify, with legacy vendors such as CA and BMC pursuing the opportunity, in addition to startups. Vendors with point products, such as those that do application monitoring or server monitoring, will continue to have a role as collectors of specialized data that feed the big-data and analytics platforms. However, pressure on these vendors to prove their value and differentiate will increase.

CONTEXT

Where once applications were monolithic and comprised just a few services, modern applications are much more complex. They might include hundreds or thousands of microservices built on containers that spin up and down as needed, spanning on-premises servers, cloud workloads and even serverless deployments. Today's apps might include code pushes that happen weekly, daily, hourly or even more frequently.

Traditional performance-monitoring tools may not cut it when applied to these complex environments. One reason is that those tools tend to be siloed, looking after networking, servers, application code, databases and cloud performance separately, without understanding the interconnections between those resources. Another is that they weren't designed to handle the volume and variety of data issued from a modern application environment.

We've seen a number of responses to this problem, including the platform approach, where individual vendors attempt to deliver a comprehensive set of data about the full stack. Another is the DIY approach, which typically stitches together some combination of open source and homegrown monitoring tools. IT operations analytics also responds to this demand, although early versions fell short because they lacked sophisticated analytics and big-data capabilities. More recent tools for ITOA aim to allow users to collect data from almost any source – including directly from apps and infrastructure, as well as from third-party monitoring tools – in order to run advanced analytics across this dataset that represents much of the full stack.

These are the key elements that we think are important to look for in these ITOA 2.0 offerings that define this space:

- *Openness:* Openness comes in several 'flavors,' such as the ability to ingest data from virtually any source, including potentially competitive third-party tools. This is key and overcomes one of the shortcomings of previous generations of ITOA products that sometimes limit data ingestion and analysis to data collected by a single vendor. The reverse is true, too – the vendors must make it easy for customers to ship data to another tool so that, for example, they can use the analytics tool of their choice.

  Many, although not all, of the new tools rely on open source technologies, including Hadoop or pieces of the Elasticstack, on the back end. We think that strategic use of open source technologies will be key to enabling individual vendors to focus on differentiators and technology advances.

- *Data agnostic:* Some of these solutions are able to consume structured, semi-structured and unstructured data. Offerings that are able to combine and correlate logs, metrics, business data such as revenue, and social sentiment from sites like Twitter will deliver valuable insights to customers. Combining such disparate data sources isn't easy. End users should consider the techniques vendors employ to normalize and analyze the data. Do they convert logs to metrics and only retain the metric? Was the back end designed for metrics and, thus, only samples logs? The downsides to each approach may or may not impact utility for individual use cases.
- *Scale and speed:* Not every business will be collecting huge volumes of data, but large enterprises and web-scale businesses will, and some of the offerings were designed with these customers in mind. Vendors serving these segments discuss volume in terms of terabytes per day (and some, terabytes per hour) and search speeds in terms of seconds.
- *Advanced analytics:* Monitoring tools have long included predefined dashboards that visualize common analyses, such as error and traffic rates, over time. However, advanced analytics tools do more: they use machine-learning techniques to predict problems in advance; automatically build and adjust thresholds, issue alerts when key performance indicators fall out of normal range; and correlate disparate data streams to develop meaningful alerts. They also typically allow users to perform sophisticated queries.

  We think that analytics capabilities will prove to be important differentiators. The vendors we talk to aren't just using tried-and-true machine-learning algorithms; instead, many are attempting to add techniques that inject their experience in operations monitoring to influence analytics outcomes. While we think that analytics will be important, some of the vendors we spoke with are concentrating on the back end, leaving open the possibility that customers will use their products primarily for data collection and normalization, relying on other tools for analytics.

VENDORS

This list isn't comprehensive but includes vendors that we've recently heard position themselves as offering a central IT operations data store or central IT operations analytics tool, or that we've seen being used by customers as such. We've seen some new entrants to this space, including from legacy vendors.

BMC

BMC's TrueSight Intelligence collects and analyzes business and operations data from a variety of sources. Using a REST API, Intelligence can ingest data from third-party products such as Splunk and AppDynamics, BMC products like TrueSight IT Data Analytics, and sources of business data such as a social sentiment app. BMC envisions a wide array of metrics and events that could be pulled into Intelligence beyond infrastructure performance metrics, including the number of tweets with a certain hashtag and alerts from IoT sensors. It built the system for scale, using open source tools, including Storm and Spark for real-time processing and streaming, as well as Cassandra.

CA

Using a set of primarily open source technologies, CA has developed a modular big-data platform that it hopes will power many of its products and potentially be productized for customers to use as they like. Known as Project Jarvis, the engine follows a lambda data-processing architecture. It uses Elasticsearch as the service layer, HDFS, Spark and Spark streaming for data processing, and Kafka as the data bus.

CA's plan is to allow customers to use RESTful APIs to ingest data, such as logs, directly from the technology stack, as well as from many of CA's products, such as App Experience Analytics, API Management and APM. Additionally, it expects to allow customers to ingest data from sources such as proprietary systems, and external sources such as Twitter. The focus is on positioning Jarvis as a centralized IT data repository, with additional CA products offering an analytics layer.

CISCO/APPDYNAMICS

As part of its planned acquisition of AppDynamics, Cisco revealed that it's been building an analytics software product that will ingest data from a variety Cisco and third-party sources, including Tetration, network monitoring, Cisco Umbrella and AppDynamics. It has not revealed much about the product yet but plans to unveil more within the next couple of quarters. We think that Cisco is in a strong position to deliver on this idea given its position in networking, security and monitoring.

DATADOG

With its long list of integrations for collecting metrics from infrastructure, and now applications, Datadog is being used as a central IT operations analytics tool. It focuses on aggregating and correlating metrics from a wide variety of resources and uses machine learning for anomaly detection.

LOOM

Loom is a new market entrant with a clear focus on analytics, running machine-learning technologies on data it collects primarily from log tools (it ingests metrics as text) without retaining much of that data at all. The bulk of Loom's back end is proprietary, although it uses Elasticsearch for a search function and saves graphs in Graphite. While users can execute searches, Loom largely defers to popular log management vendors for the capability since most of its customers already have a log management tool. In fact, Loom doesn't store all the logs it ingests. Instead, it only saves the graphs for later reference, including a set of logs generated before and after an anomaly.

ROCANA

Rocana can ingest a large volume and variety of data – its largest customer collects roughly 1TB per hour – and deliver advanced analytics that can be particularly useful when applied to a large and potentially historical set of data. Rocana's technology is built on open source projects including Kafka, Impala, Avro and Parquet, with data stored in Hadoop Distributed File System.

Rocana is trying to position itself as a centralized data repository that different teams – including site reliability, security and compliance, app developers, and business units – can access using its front end or the tools of their choice. The goal is to become the data warehouse for all machine data within an organization, from which different products can be deployed to access and analyze the data depending on the use case.

Although it can't name it publicly, Rocana has a large and recognizable retailer as a customer, which is using the product in this centralized manner, indicating that Rocana has successfully sold the concept in a large, complex environment. That type of case study is key for a young company such as Rocana that is tackling the challenging sale of a centralized data repository.

SCALYR

While it's not there yet, Scalyr's vision is to serve as a centralized repository for all operational data. Users could potentially tap into the data via third-party tools, or Scalyr might end up developing its own analytics and visualization front ends for monitoring application performance, for example.

Even though Scalyr has work to do before it can achieve its vision, we think its technology shows promise that might be particularly useful to large deployments. Rather than use open source tools such as Elasticstack or Hadoop, Scalyr built a custom data store for its log management service to try to set itself apart from the pack. It claims that 95% of queries on the system return in under a second and that it searches at a lightning-fast 750GB per second. This homegrowm back end also supports large volumes of data, with one e-commerce customer regularly collecting 2TB per day and spiking to 10TB on the busiest day of the year.

SUMO LOGIC

Sumo Logic argues that it has advantages over other vendors that combine metrics and logs because it was designed from the start to ingest time-series data, and it retains that data in its native format, where others – namely log vendors that correlate metrics and logs – may convert time-series data to logs and retain just the logs. Compared to application performance management (APM) vendors that are now doing some log ingest, Sumo Logic is pushing the advantages of designing its systems to ingest varying types and volumes of data as it is produced, compared to APM tools, which tend to ingest at set increments. The result is that Sumo Logic can do real-time data streaming, which it argues is useful in scenarios such as preventing a distributed denial-of-service attack, where a customer was able to quickly learn about and shut down a suspicious server.

Sumo Logic is also evolving its message to position itself not necessarily as a single tool that IT operations, DevOps, security and those in other roles use, but as a central data platform for all types of IT operations data that can be accessed by end users in various ways. An example of how that might work is its recent partnership with New Relic, where customers can feed data from Sumo into New Relic Insights if they prefer New Relic's analytics and dashboarding capabilities. We think Sumo Logic is well positioned to establish itself as a central repository for IT operations and that it's making the right partnerships to enable it.

WAVEFRONT

Wavefront is a high-volume, low-latency platform for collecting, visualizing and analyzing metrics. It can ingest one million points per second and query at the same rate, putting it at the high end in terms of volume compared with other vendors. While about 70% of Wavefront customers instrument their code to generate custom metrics, customers also commonly pull in metrics from sources such as APM and NPM tools. A key use case that Wavefront targets is tying in metrics from a variety of sources, including all layers of an infrastructure, to create a centralized place to do visualizations and analysis.

Wavefront is working on adding machine-learning technology to boost its analytics capabilities, and it's also developing better ways to incorporate logs. Both are key capabilities for its ability to compete as a centralized data and analytics tool for IT operations.