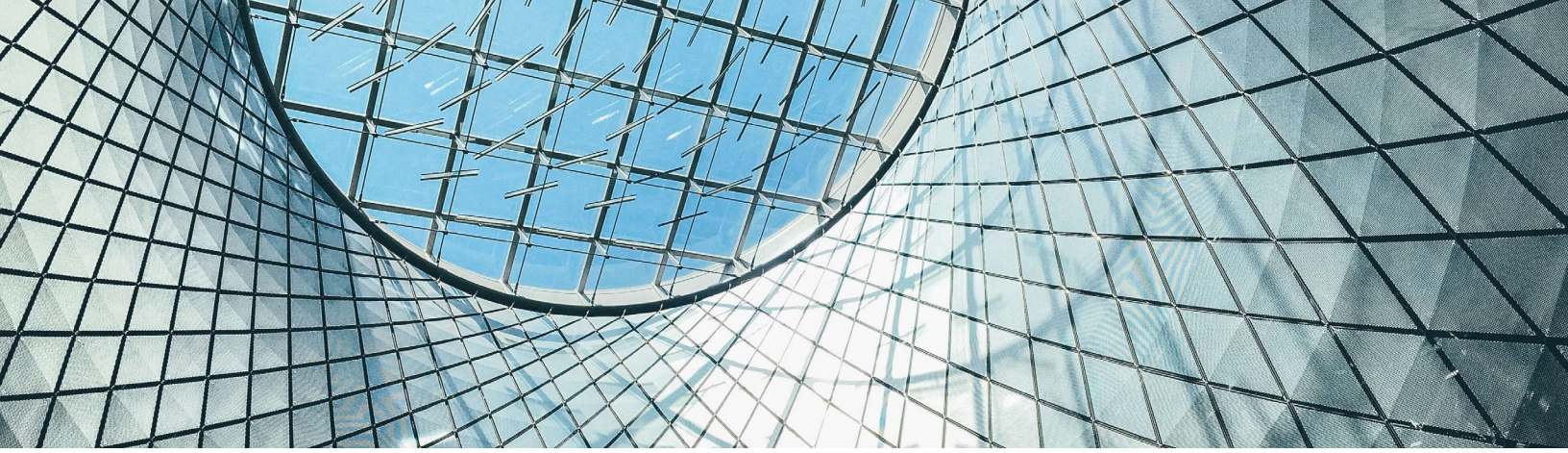


451 Research Market Insight Report Reprint

Observability for AI platforms helps IT understand AI software performance

September 9, 2024

by Mike Fratto

AI and ML is being used to mine the deluge of data that is collected for insight, for predictions of outages, and to build comprehensive visualizations of complex applications. Observability of AI platforms and their foundational models are currently managed by data scientists tasked with selecting and developing the models used for the enterprise applications.

S&P Global
Market Intelligence

Introduction

Artificial intelligence and machine learning are significantly impacting all facets of IT, including observability. AI and ML are being used to mine the deluge of data that is collected for insight and predictions of outages, and to build comprehensive visualizations of complex applications. AI also helps IT operate more efficiently with generative AI, allowing users to create queries, dashboards and other reports using natural language rather than the obscure query languages.

THE TAKE

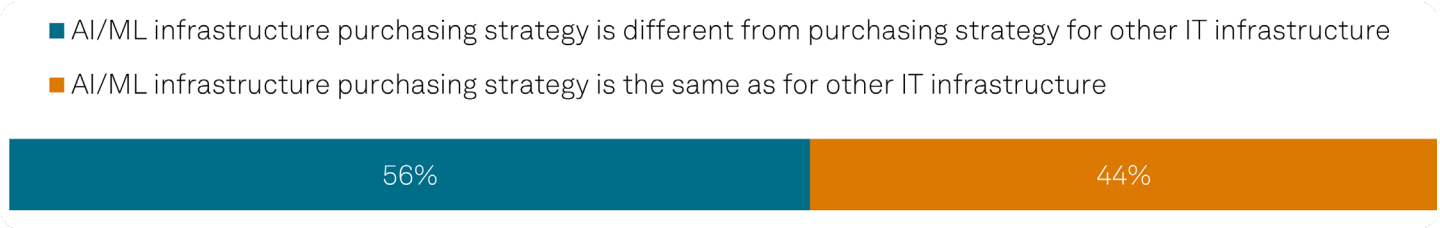
The observability of AI platforms and their foundational models is currently managed by data scientists tasked with selecting and developing the models used for enterprise applications. Eventually, IT operations will take over the life cycle management. It will need the tools to observe model performance and adapt the AI platform as it grows and changes. That will start with the ability to collect data from all layers of the AI application, then analyze it for anomalies, report on and manage expenses for the AI platform, and optimize performance and workload placement. These goals are very different from those of the data scientists developing AI applications, but we expect feature convergence to bring all teams together to better collaborate on ongoing operations.

Context

According to 451 Research's Voice of the Enterprise: AI & Machine Learning, Infrastructure 2024 survey, 14% of organizations are developing applications on public cloud infrastructure, 40% of organizations that are developing AI/ML applications are doing so on private cloud infrastructure, 30% are doing so in private noncloud infrastructure, and 17% are deploying on hybrid cloud.

Enterprises building AI and ML applications will want to have the same operational visibility they have with the rest of their application estate. Observability vendors are just starting to tackle the difficult issue of measuring and reporting on AI infrastructure performance. Enterprises are approaching their purchasing of AI infrastructure independently from the rest of their IT infrastructure, but IT will likely want to consolidate performance measurement and management where it can.

The bulk of enterprises are purchasing dedicated infrastructure for AI workloads



Q. Is your organization's purchasing strategy for AI/ML infrastructure the same or different from its strategy for purchasing other IT infrastructure?
Base: All respondents, abbreviated fielding (n=681).
Source: 451 Research's Voice of the Enterprise: AI & Machine Learning, Infrastructure 2024.

This report focuses primarily on observability vendors in the IT operations segment, not those targeting data scientists and AI experts. Observability vendors primarily address roles in IT operations, cloud operations and developer and DevOps teams. The data collection, analysis and presentation are well suited for operational roles, and observability insights kick off other workstreams like troubleshooting, application optimization and IT automation.

With more enterprises deploying applications that rely on AI — and even exploring the idea of hosting AI platforms on their own infrastructure — observability vendors traditionally in the operations space are turning to providing observability of AI applications for operations roles. Some, like Dynatrace Inc., aspire to attract data scientist and AI developer roles that are using AI and ML operations software and services to assist data scientists with developing and test-monitoring AI applications and stacks (such as Arize, Fiddler.ai, LangChain and Traceloop).

Approaches to observability for AI

Observability vendors are approaching observability for AI platforms from different angles, in large part dependent on where they start from. The approaches are differentiating because they determine what aspect of AI observability the vendor is prioritizing at any point in the development. Bear in mind, observability for AI product strategies is fluid, and based on what product teams perceive as customer demand. Observability vendors are also at different stages, with some having almost no observability for AI capabilities by choice, and others already shipping generally available AI observability capabilities. What is important is understanding a vendor's road map.

Observability for AI is primarily used for monitoring the AI hardware and software infrastructure, and reporting on performance issues. AI environments themselves are in charge of scheduling jobs, distributing workloads and managing the infrastructure in response to program demands. AI platforms can respond quickly to issues within the AI environment to ensure job completion times are met, and react when they are not. Observability provides a view into this opaque world.

Full stack monitoring starts with hardware

The importance of hardware monitoring to enterprise IT will depend on who owns the hardware. If an AI cloud service is used for training and inferencing, access to hardware counters is unlikely due to the opacity of the cloud service. However, for enterprises that are running their own AI infrastructure either on-premises or acquiring cloud services with dedicated GPUs and other AI processors, monitoring of the hardware becomes more important.

This is the starting point for observability like Hewlett Packard Enterprise Co.'s OpsRamp, which provides detailed monitoring of server performance and status, including the network and GPUs, up through the OS and Kubernetes environment. Infrastructure monitoring is the basis of full stack observability when using your own hardware and software infrastructure. Performance metrics for services start at a higher level, with the assumption that the cloud provider will monitor and maintain the underlying infrastructure. The critical new metrics will be in monitoring the GPU and TPU cores reporting on processor hardware utilization, heat generation and power consumption.

Observability for AI is about software

Infrastructure monitoring is fine for operations teams managing hardware, but does not inform IT and data scientist teams about the performance of the AI platform in terms of model behavior, cost and overall application performance. Some of the introspection of the models is under the purview of the data scientists building and testing foundational models, but it is likely that the ongoing monitoring and management will be transferred to operations in the future, especially as model behaviors become better understood. This is an area where observability vendors that are accustomed to creating visualizations for IT operations will have an advantage in creating context-specific analyses and visualizations of AI applications and infrastructure.

Observability vendors are taking two approaches to observability for AI. The first is treating AI platforms like applications, reporting on the application's performance, and either inferring the status of the AI platform or service, or using the platform's APIs to query model and application performance. Many observability vendors, such as Cisco Systems Inc. (App Dynamics), Splunk, Dynatrace and New Relic, can report on application and AI performance, usage reporting and cost analysis.

The use of collection agents, instrumenting the application itself (along with the underlying software environment), using OpenLLMetry via OpenTelemetry, and the aforementioned APIs and data services from cloud platforms allow these observability platforms to collect, correlate and present AI application performance metrics to IT.

However, this is more than just application statistics. AI-specific metrics are collecting token usage, query times and number of queries, which can be aligned to cost. Since the observability platform also knows who is making queries, the costs can be assigned users, groups and roles, providing IT with cost monitoring and management capabilities. In addition, the content of queries including uploaded files as well as responses can be monitored for sensitive data like personally identifiable information, financial information and other proprietary data.

Advanced capabilities are coming to market. More AI applications are supporting retrieval augmented generation, where the AI platform can make queries to external systems. Those queries can be monitored for cost and content, and sensitive information can be flagged and potentially acted upon before it leaves the AI platform, or before queries are made to data stores the user should not have access to. Cost and performance management can be enhanced as well. New Relic, for example, can perform A/B testing on AI foundational models doing comparative analysis on performance, cost and data quality. Helping IT compare across models enables better workload placement decisions. The analysis could also automate workload placement as foundational models change over time.

Observing the AI platforms

New capabilities for observability vendors include observing and analyzing a foundational model's performance by instrumenting and collecting data from it. These are capabilities that AI and ML operations software (Arize, Fiddler.ai, LangChain and Traceloop) also provide that are likely to be more familiar to data scientist roles and IT. Not many observability vendors offer AI model observability, but that will soon change. Dynatrace was early to market with model observability, which it announced in January 2024. Cisco has announced model observability plans for AppDynamics, part of Splunk Observability Cloud, and for Splunk Observability Cloud itself.

AI model observability will become a critical component of AI observability as more enterprises create AI applications using their own infrastructure or a service. Where observability emphasizes end-to-end analysis and visualizations of an application stack, AI observability focuses on pulling out performance metrics from the model itself, and presenting that data to IT in a manner that is actionable. For example, model drift occurs when AI predictions veer away from reality. This can be detected by the foundational model itself or by an observability platform using various statistical analyses to assess the accuracy of the model's output compared with its input. AI observability can also explain how answers are arrived at in deeper levels of detail, which data scientists and skilled IT staff can use to ensure the AI output is still rational.

Observability for AI is on its way

Expect observability vendors to bring a lot of new products, services and enhancements to market in the next 12 to 24 months. Cisco, Dynatrace, HPE, NewRelic and other vendors are executing product road maps specifically to enhance the observability of AI platforms. Road maps include capabilities for visualizing model performance in terms that IT operations can use and act upon; enhancing model performance and performance predictions based on existing and historical trends; improving workload placement based on desired job completion time, data placement requirements or cost; and providing recommendations for application or model optimizations to improve performance, reduce errors and hallucinations, and manage costs. We can also expect an expansion of support for more AI infrastructure, such as cloud-based and on-premises foundational models, orchestration platforms, and databases and storage platforms.

One motivator for observability vendors is to provide tools that help IT operations and data scientists collaborate through the AI application life cycle. A challenge for IT observability vendors will be to tailor their products to IT operations employees who are not experts in AI systems but need to understand key aspects of the platform, while providing analytics and interrogation tools that data scientists will understand and use. AI and ML operations vendors can also tailor their products for IT operations roles, creating a very competitive landscape for all participants.

CONTACTS

Americas: +1 800 447 2273
Japan: +81 3 6262 1887
Asia-Pacific: +60 4 291 3600
Europe, Middle East, Africa: +44 (0) 134 432 8300

www.spglobal.com/marketintelligence
www.spglobal.com/en/enterprise/about/contact-us.html