Smarter observability with AIOps, generative AI, and machine learning

Table of contents

The case for AIOps

Artificial Intelligence for IT Operations (AIOps) is the application of artificial intelligence (AI), machine learning (ML), and analytics to improve the day-to-day operational work for IT operations teams. Simply put, AIOps is the ability of software systems to ease and assist IT operations via the use of AI and ML and related analytical technologies. AIOps capabilities can be applied to the ingestion and processing of a variety of operational and business data, such as logs, traces, metrics, and much more.

With the increasing complexity of distributed applications along with the adoption of cloud-native technologies, teams have been dealing with three major changes in the application environments they observe and manage: data volume, complexity, and pace of change. AIOps can play a key role and, when implemented and used properly, help them navigate these challenges effectively, freeing up operations teams to focus on more important work.

Integrating AIOps into your observability solution with ML and generative AI can optimize your operations and give you even more visibility into your systems.

Understanding AIOps for observability

AIOps continues to be a hot topic among developers, site reliability engineers (SREs), and DevOps professionals. The case for AIOps is especially crucial given the expansive nature of today's observability efforts across hybrid and multi-cloud environments. As with most observability platforms, it all starts with your telemetry data: metrics, logs, traces, and events.

Once IT operations teams collect and begin analyzing that data, the benefit of AIOps becomes rapidly clear. AIOps aims to accurately and proactively identify areas that need attention and assist IT teams in solving issues faster. It's simply not feasible for the human mind to absorb and analyze petabytes of raw observability data — but a machine can. Adding AIOps delivers a layer of intelligence via analytics and automation to help reduce overhead for a team. Let's dive in to answer common questions on this critical topic!

How can AIOps help me?

AIOps can significantly reduce the time and effort required to detect, understand, investigate, determine root causes, and remediate issues and incidents. In turn, saving time during troubleshooting can help IT personnel focus more of their energy on higher-value tasks and projects.

Defining the undefinable

Definitions and explanations by analyst groups and vendors seek to clarify the often murky and confusing world of AIOps. Despite the complexity, it's clear that AIOps will be an essential tool for navigating today's hybrid and multi-cloud environments.

Why do you need AIOps as part of your observability strategy?

From digital transformation initiatives to cloud migration to distributed, hybrid, or cloud-native application deployments, evolving technologies are dramatically changing the IT operations landscape.

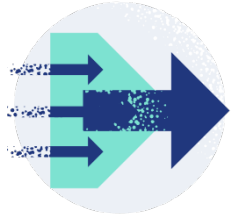
The landscape changes have the following three characteristics:

Data volume

The volume of data for observability continues to increase exponentially.

Complexity

Applications, workloads, and deployments continue to become more complex, ephemeral, and distributed.



Pace of change

Rate at which changes (application and infrastructure) occur is faster than ever before.

These are not mutually exclusive. In some ways, quite the opposite. For example, high rates of change and complex deployments utilizing auto-scaling mean an even higher volume of data. More data means parsing, analyzing, and extracting value out of this data becomes more difficult.

Leveraging AI and ML to summarize and roll up data and to intelligently tier the data for storage can help alleviate some of the challenges around telemetry data volume. Clear visual depictions of an application environment, via infrastructure and service dependency maps for example, and contextual navigation help align troubleshooting efforts with how users naturally think of their deployment. Furthermore, auto-surfacing of problems, anomalies, and root causes will address some of the other complexity challenges.

Observability platforms will need to keep track of all application and infrastructure changes and correlate those changes with system behavior and user experience because those changes are often the root cause of acute, anomalous behavior.

Tech Tidbit

A software upgrade or patch for a new feature may have unintended consequences. Enabling AIOps helps teams be more agile and adept at keeping pace with those frequent changes, which ultimately helps sustain service performance.

How does AIOps drive business value for an organization?

Given the volume, complexity, and pace of change in today's cloud-native and hybrid application environments, AIOps is increasingly moving from being a nice-to-have capability to a mission-critical competency for IT operations teams.

While AIOps can significantly reduce the mundane and repetitive work required by IT operations (ITOps), SRE, and DevOps teams, there are also significant business benefits:

- **Reducing MTTD (mean time to detection)** and MTTR (mean time to resolution), which means less service downtime, improved SLAs, and better customer experience
- **Helping organizations deal with rapidly growing data volumes** intelligently, reducing total cost of ownership (TCO), and alleviating scale challenges
- **Reducing signal and alert noise** and implementing better automation, freeing operations teams to take on higher-value initiatives
- **Improving organizations' ability to handle ever-increasing IT complexity**, allowing them to innovate and bring new features to market more quickly and frequently

Modern hybrid and cloud-native environments continue to push the boundaries of what operations folks can manage for their enterprise. Cost analytics, tracking business metrics, and alignment of business impact with observability data are just a few examples of the more recent challenges facing ops teams.

The good news is that the same AIOps concepts and analytics capabilities such as baselining, anomaly detection, and correlations that help observability are equally adept at solving the newer business challenges. AI and ML capabilities can go even further and help make sense of any new signals and data, allowing users to extract useful, actionable insights that contribute to business success.

The role of machine learning in AIOps

Machine learning (ML) is a branch of artificial intelligence (AI) that focuses on the use of data and algorithms to imitate the way humans learn, gradually improving accuracy over time. ML involves feeding large amounts of data into computer algorithms so they can learn to identify patterns and relationships within that data set. As the algorithms receive new data, they continue to refine the underlying model and improve performance over time.

ML is important because it learns to perform complex analyses using examples (model training), without programming specialized algorithms. Compared to traditional algorithmic approaches, ML enables you to automate more, improve customer experiences, and create innovative applications that were not feasible before.

Examples include:

- **Predicting** trends to improve business decisions
- **Personalizing** recommendations that increase revenue and customer satisfaction
- **Automating** the monitoring of complex applications and IT infrastructure
- **Identifying** spam and spotting security breaches

What are the advantages of machine learning?

ML can help empower your teams to get to the next level of performance in the following categories:

Automation

Cognitive tasks that are challenging to humans — due to repetitiveness or data volume — can be automated with machine learning. Examples include monitoring complex networked systems, identifying suspicious activity in complex systems, and predicting when equipment needs maintenance.

Customer experience

Intelligence delivered by machine learning models can elevate user experiences with proactive anomaly detection and alerting, and faster root cause analysis of issues, catching (and resolving) problems before your users do.

Innovation

Machine learning solves complex problems that weren't possible with purpose-built algorithms, freeing up your teams from labor-intensive data analysis and manual troubleshooting so they can focus on innovative projects that are strategic to your business.

The importance of a unified observability platform for AI and ML

The more comprehensive and rich the data that's available to analyze, the more that can be done with that data through the application of AI and ML technologies. A modern, unified observability platform with all your operational data will be the foundation of any AIOps efforts for the future.

Advanced application of AI and ML can help drive additional use cases such as extracting business insights, deriving predictive or leading indicators across multiple signals, or defining and deploying entirely customized AI-driven workflows when the need arises. Observability systems will start to form more closed loops; collecting, storing, and analyzing data and detecting and remediating more incidents automatically with increasingly less human intervention.

Using machine learning for anomaly detection

In AIOps, machine learning is most useful for anomaly detection. Anomaly detection is the process of using algorithms to identify unusual patterns or outliers in data that might indicate a problem. Anomaly detection is used to monitor IT infrastructure, applications, and networks and to identify activity that signals a potential impact on application performance or could lead to a network outage later. Anomaly detection can also be used to detect security breaches and fraudulent bank transactions. Learn more about AIOps and machine learning now.

Common observability challenges and use cases for AIOps today

Reducing noise for improved issue detection

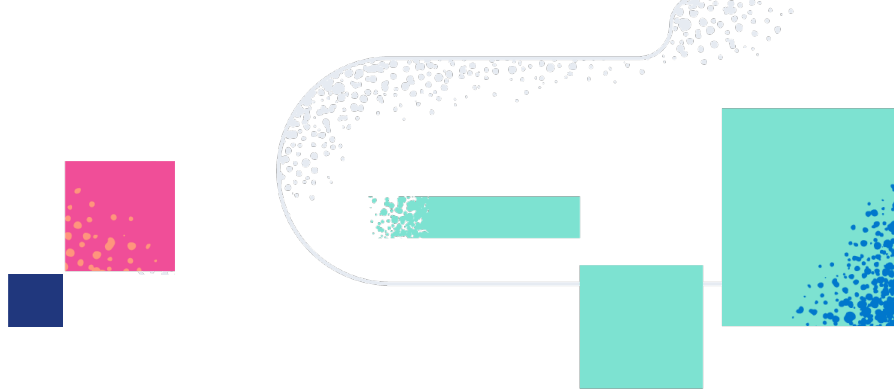
Providing context for faster root cause analysis

Democratizing data and analytics across the organization

Reducing noise for improved issue detection

Observability platforms can ingest and analyze massive amounts of data from multiple sources in real time, allowing SREs to get a comprehensive view of the system's behavior and identify potential issues as they arise. AIOps functions can be used to automatically identify patterns in diverse data and highlight relationships and correlations that are not easily evident through basic dashboards and data visualizations.

This can be particularly useful for detecting and resolving problems that are transient, difficult to predict, or hidden within the system's normal operating range. For example, when an application is running slow, AIOps can be used to automatically identify probable causes of slow or failed transactions.



With so much data being generated by modern systems, it can be overwhelming for SREs to sift through all the noise and determine which alerts are the most important. Observability platforms can use AIOps techniques and machine learning algorithms to identify patterns and correlations between different alerts, allowing SREs to prioritize their efforts and focus on the most pressing issues. Many types of noisy data can be reduced by AIOps automation, for example:

> **Multiple sets** of similar or duplicate information
>
> **Too many detected** issues and alerts (both manual and automatic), some of which might have the same underlying root cause
>
> **Informational notification** events

These all contribute to varying levels of noise in the observability data and workflows. Alert fatigue has never been more likely for SRE or IT Operations teams than when observing modern application deployments. AIOps helps reduce noise and surface important insights with the right context, making IT operations teams more efficient.

By automatically prioritizing entities and information based on business and user impact, AIOps helps focus on what's most critical. AIOps can also help detect and de-duplicate information based on data characteristics and can cluster or group similar information, presenting them together, further reducing noise when troubleshooting. As new types of observability signals and data are ingested, time series baselining via unsupervised machine learning and anomaly detection significantly reduces the manual effort needed to monitor and track that data.

Providing context for faster root cause analysis

Root cause analysis (RCA) is a proven troubleshooting technique used by teams to identify and resolve problems at their core, rather than attempting to treat symptoms. Root cause analysis is a structured, step-by-step process designed to seek out primary, underlying causes by gathering and analyzing relevant data and testing solutions that address them.

When an issue does arise, AIOps can help SREs and developers identify the root cause more quickly. By analyzing data from multiple sources, AIOps can identify the underlying cause of a problem, even if it is not immediately apparent. These insights can help SREs resolve issues more efficiently and prevent them from reoccurring in the future.

Automatically surfacing contextual information surrounding the issue helps speed up the investigation by presenting relevant information inline and in workflows. AIOps can correlate multiple events and behaviors around an issue, aiding more holistic investigations and cutting down MTTD and MTTR. For a smaller set of specific, well-understood symptoms, AIOps can fully automate the journey from symptom to root cause, removing the need to manually iterate through the investigation.

Democratizing data and analytics across the organization

AIOps aims to ease the lives of IT operations teams, reducing the amount of manual work needed, especially for routine and repetitive tasks, and helps find the needle in the haystack. This allows operations users to focus on higher-level activities like platform architecture, platform engineering, automation, security, and more.

Ideally, your AIOps platform will allow you to democratize ML and analytics for the non-data scientists in your organization, such as SRE teams and business users. With pre-configured models for common use cases and easy-to-use workflows to facilitate customization, your entire organization can work to run your data-driven business more efficiently.

Democratizing your observability data for everyone in your organization? Now that's a north star that I can look up to!

Building trust for AIOps in observability

IT personnel, SREs, and DevOps engineers have a couple of adoption hurdles they must cross to successfully adopt and use AIOps for their observability use cases.

Users are faced with questions such as what is the true business value beyond the hype and will AIOps actually help them detect and remediate problems better and more efficiently than their current monitoring or observability setup. Beyond the hype, users may not always know if they will truly benefit from AI and ML for their specific use cases.

And then there are trust hurdles:

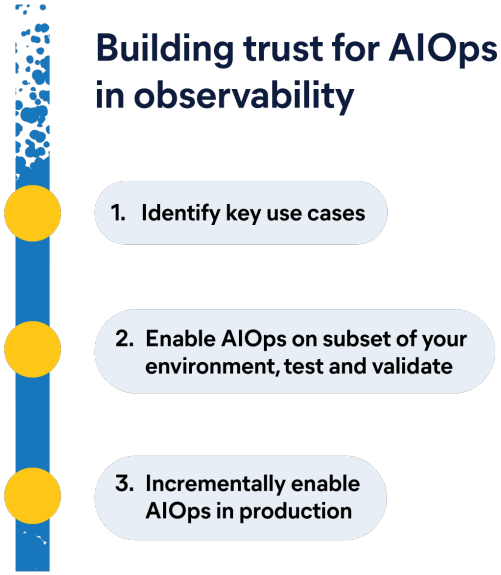
- Users find it difficult to tell whether the AIOps-based insights are accurate.
- Users might not fully understand how comprehensive the analysis is, the information used, and how the algorithms work.
- Users are unsure how conclusions are arrived at or if those conclusions are relevant to their current investigation.

The result: general distrust of black box AIOps systems. In some cases, organizational pressures or policies motivated by a lack of trust may also present barriers to AIOps adoption.

Our experience has shown that the best way for AIOps to provide its value is through its slow and steady adoption. First, identify specific, time-tested, and proven use cases to start adopting AIOps as a proof of concept (POC). Next, enable AIOps functionality on a smaller subset of your application environment while validating and socializing benefits and outcomes at each stage. Once you've seen some success, incrementally enable more AIOps functionality with a move toward production environments. This deliberate deployment path alleviates some of the typical challenges associated with deploying new technology that can otherwise deter widespread AIOps adoption.

Testing and proving technology effectiveness in a smaller lab or non-production environment and measuring and showcasing results to management can help increase confidence and get buy-in before deploying AIOps in a real-world production environment. Such testing might unearth other gaps and requirements, such as missing or inconsistent data, shallow coverage, or insufficient storage or computing.

As you deploy AIOps in production, check to see if your observability solution can scale its features appropriately and handle your enterprise workloads. Certain AIOps features that work well in lab or POC environments may struggle to keep up with larger-scale requirements typically encountered in production environments.



The future of AIOps with generative AI

How generative AI will impact observability, today and tomorrow

If you've interacted with ChatGPT, Open AI's natural language processing tool, then you've interacted with generative AI technology and large language models (LLMs).

A large language model, such as ChatGPT, Amazon Bedrock, or Google Bard, is a specific type of generative AI model that generates information based on a variety of inputs: the data it is pre-trained on and the query submitted by the user. When queried, it will search the breadth of the data it is trained on, match data to the query, synthesize it, and provide that synthesis as a natural language response to the user. In other words, it responds in plain English. As it happens, these capabilities are well-suited to some current observability problems.

THE PRESENT

Using generative AI and search for observability

Your observability platform gives you visibility into your logs, metrics, traces, functions, libraries, and a host of other systems and data-related information. Generative AI can help you navigate your observability platform and provide further insights and guidance into what you're seeing with a simple query. Consider these use cases:

Explain x

Not sure what the function, log, or trace you're seeing is? You could query the generative AI tool to provide you with more information about the data.

Synthesize information

The generative AI tool can even go one step further and synthesize the information you see on your observability platform to generate a neat report or visualization for you.

Improve efficiency

With its ability to explain and synthesize, the generative AI tool can help bolster your team's level of expertise and efficiency. For example, if your code is eating up a lot of CPU, you could query the generative AI tool to use code profiling data to identify resource-intensive functions to optimize to improve resource usage, and down the line, reduce costs.

As a natural language processing platform, a large language model (LLM) can easily translate languages from Javascript to Python or JSON. It is one of the reasons an LLM is incredibly useful in the context of observability.

Bottom line: there are specific observability functions, such as interpretations of log messages and errors, script conversion, and report generation which align with the current capabilities of generative AI. By making observability “problems” search problems, you can use generative AI capabilities to your advantage.

Generative AI and LLMs: The upcoming societal shift

The arrival of ChatGPT (March 2023) marked the dawn of a new era. For many, this was the first introduction to directly interacting with an LLM through a web browser interface. But this seemingly simple interaction marked a seismic shift in the relationship between humans and AI. There was excitement and glee from technology enthusiasts, but there was also skepticism and outright fear. Could this be just a fad? Would LLMs replace jobs?

But the excitement around LLMs is growing. Google, Facebook, and other tech giants are releasing their own LLMs and chatbots. Open source LLMs are rapidly evolving with some insiders suggesting they might eventually surpass Google and OpenAI.

The increasing interest in LLMs has already led to transformations across various industries. In the coming years, the norm for people to interact with data will be with search boxes, chatbots, and prompts built to execute workflows. LLMs are poised to change our lives in ways we can barely begin to imagine.

The generative AI societal shift", June 2023, https://www.elastic.co/search-labs/blog/articles/generative-ai-societal-shift

THE CAVEAT

It’s all about the training for LLMs

It’s important to note that generative AI capabilities are only as good as the data the LLM is trained on. If it has never been fed logs or a given set of libraries, it will not provide complete or accurate responses to queries looking for explanations on those topics. That said, there are ways to draw value from the tool without time and resource-intensive training. How?

What is RAG?

Retrieval augmented generation (RAG) is a framework that enables users to ‘feed’ an LLM private or proprietary data so it has the most up-to-date information. This increases the LLM’s efficiency and accuracy and ensures that the user has additional data sources for the LLM to generate more useful query responses.

There are two components to RAG: retrieval and augmented generation. The latter implies that the query is augmented by additional data or information. You might recognize this component as prompt engineering. By augmenting the prompt, the user is ‘prepping’ the LLM backend for the most up-to-date information retrieval, and is, therefore, able to draw the most value from the tool.

The thing about privacy

If you use LLMs with your private data, there is a risk they could potentially train themselves on that data leading to concerns that this information could be regurgitated publicly. This, of course, is less than ideal for most businesses. Sensitive information leaks are one of several reasons that organizations are hesitant to adopt LLMs internally. RAG can circumvent some of these privacy concerns, while the current foolproof solution is investing in private commercial LLMs.

THE FUTURE

Will AI go autonomous?

So where does generative AI in the realm of observability go from here? It's all speculative, but the industry seems primed to develop autonomous agents.

But before the technology is ready to provide a reliable autonomous agent, there are several leaps it must make:

Language-driven interfaces

One of the current challenges any observability platform faces is how it presents information to the user. There are currently only two modes: pre-built custom dashboards, and signal-type dashboards. Pre-built custom dashboards offer high-level, single-pane-of-glass visibility. While very useful for providing context and correlation across data sets, they still require a deeper manual dive into investigating issues.

Single-type dashboards offer a granular — but separate — view of your logs, your traces, your service dependency maps, and so on. A language-driven interface is the next step to bridging the gap between holistic and granular visibility — it would enable a dynamic single-pane-of-glass mode which provides you with a relevant set of signal dashboards in a single view. A language-driven interface would enable the user to dialogue with the system to pull up the necessary dashboards or views, enabling a conversational problem-solving process that corresponds to the intricate and dynamic nature of observability issues.

With current technologies, language-driven interfaces aren't such a far stretch. The next step is for generative AI to take on an assistant role.

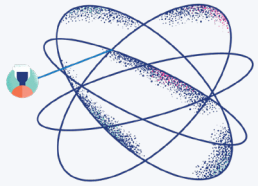
Tech Tidbit: Observability versus security usage of LLMs

It's important to note that while the field of security has established, publicly available frameworks for troubleshooting, observability does not. In the realm of observability, every problem is rather unique, which furthers the need for sophisticated planning and reasoning capabilities in its generative AI tools.

Three types of AI assistants

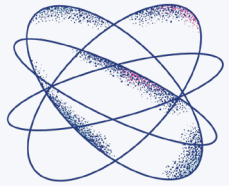
AI assistant with a human *in* the loop

In this mode of operation, the observability platform triggers an alert, the AI assistant identifies it, explains it to the human, and then asks whether the human would like to perform a follow-up task (e.g. look at *x* set of logs). In this mode, the LLM is analyzing and making suggestions, while in constant conversation with a human, and only prompted to act upon human input.



AI assistant with a human *on* the loop

Here, the LLM is empowered to do more steps on its own. It'll perform an analysis on its own, provide you with a report, and then ask you to authorize the next step, which may be taking remedial action.



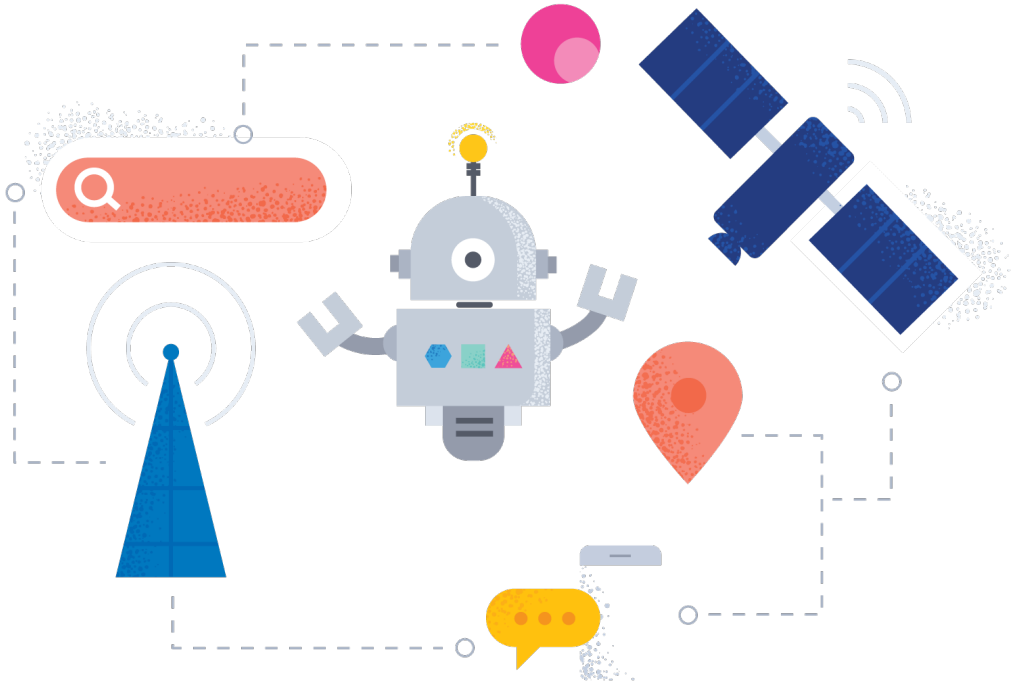
Autonomous agent

The human is removed from the loop and total autonomy is granted to the machine. The LLM does all the analysis and debugging or patching on its own, based on its analysis of the issue. It might restart hosts, change hosts, change configurations, and so on.

In the first two assistant modes, the SRE is called to their screen when an alert goes off (just as they would be now), and they have a say in what happens by micromanaging tasks (in the loop) or simply managing them (on the loop).

Imagine this future scenario: your observability platform gets an alert, the assistant analyzes the alert in real time and offers commentary, and it could also make remediation suggestions. It's important to understand that for the assistant to do this, the AI needs to have a solid understanding of observability workflows and react to new information. You get new intel, you have to change course in accordance with the new information. This particularity of observability explains why the autonomous agent is still out of reach.

A technological reality in which organizations employ autonomous agents is still out of grasp: How do you trust the technology to make the correct decisions? Privacy and perceived lack of control are concerns that many businesses express — and they are valid. The technology cannot currently perform complex reasoning and planning tasks, and before it can be autonomous, it must learn these skills.



Conclusion

As with many newer technologies, AIOps continues to evolve in response to ever-increasing data, complexity, and pace of change.

In a complicated, cloud-native landscape, just having any AIOps system isn't enough. Choosing the right observability platform is crucial in preparing for the future of AIOps-driven observability and remediation that is just around the corner. With the right platform, organizations can stay ahead of the curve and leverage AIOps to optimize their operations, gain valuable insights, and make data-driven decisions that drive growth and success.

Is there a single platform that can help you use the power of AIOps and generative AI to transform the way you do business? Today?

Yes. Meet Elastic Observability.

Elastic Observability is a comprehensive, full-stack observability solution with a great foundation for AIOps. Ingest all your data across metrics, logs, traces, and even business data — *in a unified platform*.

Elastic and its AI-powered observability

With Elastic Observability, you can consume and process large observability data sets at scale to quickly zero in on the most relevant information to the business. Elastic Observability uses context-aware generative AI and advanced ML to reduce labor-intensive troubleshooting and streamline triage activities to accelerate root cause analysis so teams can focus on innovation. All this and more with our AI-powered technology:

Elastic AI Assistant for observability: Enhances the understanding of application errors, log messages, and alert analysis and provides suggestions for optimal code efficiency through an interactive chat interface. The AI Assistant can integrate with large language models (LLMs) of your choice, while also leveraging proprietary data and runbooks for additional context.

Elasticsearch Relevance Engine™ (ESRE): Designed to power artificial intelligence–based search applications. ESRE is used to apply semantic search with superior relevance out of the box (without domain adaptation), integrate with external LLMs, implement hybrid search, and use third-party or your own transformer models.

Elastic Learned Sparse Encoder (ELSER): This retrieval model trained by Elastic enables you to perform a semantic search to retrieve more relevant search results. This search type provides you search results based on contextual meaning and user intent, rather than exact keyword matches.

Learn more about Elastic Observability's AIOps position ▶

AIOps with Elastic delivers the AI-powered insights you need so you can:

- Empower SREs with generative AI
- Automate anomaly detection
- Proactively detect outliers and trends
- Accelerate problem resolution
- Streamline alerts with your incident management workflows

Ready to take the next step? ▶

BONUS

AIOps and generative AI cheat sheet

Artificial Intelligence (AI)

AIOps: Artificial Intelligence for IT Operations (AIOps) is the application of AI, machine learning (ML), and analytics to improve the day-to-day operational work for IT operations teams.

Machine learning: Machine learning (ML) is a branch of AI that focuses on the use of data and algorithms to imitate the way humans learn, gradually improving accuracy over time. One way they do this is with neural networks that utilize interconnected nodes in a layered structure that resembles the human brain.

There are four types of machine learning:

Supervised machine learning

The algorithm learns from labeled training data sets and improves its accuracy over time to detect learned patterns when it receives new data.

Semi-supervised learning

The algorithm trains on both labeled and unlabeled data to make predictions or decisions based on the available information, then refines its responses by finding patterns in the data.

Unsupervised machine learning

The algorithm analyzes data that has not been labeled and has no reference for identifying patterns.

Reinforcement learning

The algorithm learns through trial and error by getting feedback in the form of rewards or penalties for its actions.

Deep learning: A subfield of neural networks that has many layers, allowing it to learn significantly more complex relationships than other machine learning algorithms.

Generative AI

LLM: A large language model (LLM) is a deep learning algorithm that can perform a variety of natural language processing (NLP) tasks.

Prompting: A prompt is an instruction given to an LLM. Few-shot prompting teaches the model to predict outputs through the use of examples.

RAG: Retrieval augmented generation (RAG) is a framework that enables users to "feed" an LLM private or proprietary, external data so it has the most up-to-date information.

Hallucinations: A hallucination is when an LLM produces a false or nonsensical output or one that does not match the user's intent. Because large language models are not search engines or databases — they only predict the next syntactically correct word or phrase — they can appear to produce results that are factually incorrect or contradictory, especially if the data set they are trained on contains contradictory information.

Thank you.