MARKET PERSPECTIVE

Elastic AI Assistant Shows What an AI Assistant Can Become

Christopher Kissel Michelle Abraham

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: Elastic AI Assistant Shows What an AI Assistant Can Become

On June 9, 2023, Elastic presented Elastic AI Assistant. The AI Assistant comes after the announcement of the Elasticsearch Relevance Engine (ESRE). In the past four months, several companies have launched AI assistants focusing on various aspects of threat detection, triage, and threat hunting, which are capabilities that Elastic AI Assistant provides. However, there are aspects of what an analyst does pertaining to security beyond incident detection and response (IDR) that Elastic AI uniquely addresses.

Key Takeaways

- Generative AI has been applied to the first steps of the IDR, and this greatly simplifies the life of the security operations analyst. However, a true AI assistant would replicate the tasks that an IT/security analyst would do, and Elastic AI Assistant offers more than just IDR processes.
- The biggest fear about generative AI is that it will misunderstand signals and produce "false negatives." Generative AI is about as good as the underlying data structure. A significant feature of Elastic AI Assistant is that it will provide instructions on how to format data for ingest and then check to see if the data is properly configured.
- Elastic is all about search, and its AI Assistant enhances what a company can do with its search engine.

Recommended Actions

- Elastic has created the Elastic AI Assistant so that it can leverage existing detection and response rules from other appliances and turn it into an Elastic Query (and does so with 95% accuracy) or create rules as observed in a new threat environment. This means that an IT/security operations team can spend appreciably less time tuning its tools and actually use them.
- Data structure and IDR are inextricably tied. That the Elastic AI Assistant, using Elastic Common Schema, helps structure data at the time of ingest is an important differentiator and should go on T-shirts.
- Elastic has philosophy that its platform (ultimately) is easily configurable for customers. Elastic is designing its assistant to work within tightly defined domains. In turn, this allows customers to "bring your own use case" or work with what has been prepackaged.

Source: IDC, 2023

August 2023, IDC #US51056523

NEW MARKET DEVELOPMENTS AND DYNAMICS

It is fair to say that the basic conflict in cybersecurity has been much the same over the past decade. There are simply too few analysts. The analysts that are manning security operations are overburdened, and both burnout and competitive swiping of personnel make it difficult to develop a team or institute universal processes. IDC believes that for every three entry-level analysts (tier 1), there is one advanced analyst (tier 2/3); a more sustainable ratio might be 5:1. It is funny because on balance, cybersecurity vendors are doing an excellent job of creating point products, platforms, and cybersecurity services but are not gaining ground relative to the adversary. Yes, the adversary is improving, but the sheer amount of data, coupled with new surfaces (cloud architecture, operational technology [OT], and wireless and mobile), makes it that cybersecurity solution providers are running in quicksand. Until now.

Let's be clear that the application of generative AI represents an inflection point in cybersecurity. There is truth in that we can define generative AI by the sum of its parts; that being artificial intelligence and machine learning (AI/ML), natural language processing (NLP), and large language models have been a part of it and cybersecurity applications for some time now. Paying respect to the OpenAI concept, it was the release of ChatGPT specifically that seems to have unified and crystallized these disparate capabilities.

IDC is developing a comprehensive profile of the pros and cons of generative AI, but for now, a quick understanding suffices. This is a quick review of the pros of generative AI:

- If the domains are well defined, generative AI can create content (readable English or native language), source code, and even images and voice context to complete an application. People (currently just people) who can build parameters for generative AI to work for line of business, marketing, and IT/cybersecurity are fast becoming a sought-after skill by digital transforming companies. (Elastic notes that well-defined domains do helps with high-relevance searches but necessarily a requirement of its platform.)
- In theory, if content created by generative AI is comparable with what humans can create, the idea is that the human being can fine-tune generative AI code/content. As we will see, that may be easier said than done.
- With the proper safeguards, generative AI allows a company to fail fast.
- Generative AI is exceptionally versatile. It can be used to write code, explain reports in readable English or other languages, refine queries, and find patterns in data lakes.

There are also limitations:

- This generation of generative AI does not quite have the "spark of genius." It cannot write a RESTful API or OpenAPI connecting applications by itself. (However, if a connector was broken up thoughtfully into 20 different stages, it might write the application that way.)
- Try as it might, generative AI may not be able to obfuscate proprietary data. Realize too that this is not all generative AI's fault; end users may not be curating data properly for generative AI to discriminate between allowable and impermissible information. In certain environments, generative AI access to data will be impermissible, period.
- Generative AI hallucinates. Often a generative AI will create false conclusions when it is at capacity or if data is improperly structured.

In life, atrophy and time conspire against humans, but generative AI is diametrically different. Generative AI improves and becomes more accurate with more data and self-teaches and learns as it goes. In cybersecurity, generative AI may not be sufficient in and of itself to write whole applications, but it can assemble sections of codes beautifully and at hyperspeed. Again, the term *safeguards* is doing a lot of work in these concepts, but the productivity gains far outweigh the risks of cybersecurity applications/products bypassing generative AI altogether.

In the beginning of March 2023, Microsoft announced Microsoft Security Copilot using ChatGPT (see *Is Microsoft Security Copilot Disruptive Technology?* IDC #IcUS50532123, March 2023). At RSA 2023, SentinelOne Purple AI, CrowdStrike CharlotteAI, and Google Cloud Security AI Workbench were product announcements also featuring natural language processing. It is also fair to mention StrikeReady CARA and Darktrace as companies offering technologies with strong AI-assisted automation in their security operations center (SOC) assistant tools.

What these tools have in common is that generative AI has been developed as a "digital assistant." The platforms focus on the first steps an analyst would take when an alert is triggered. The analyst would create a timeline, gather artifacts, develop a relational graph of applications accessed by suspected infected machines, pull reports about available threat intelligence, and determine the overall risk environment. These manually intensive processes that may take as few as 5 minutes or as many as 30 minutes are now available instantaneously. From there, an analyst can ask the platform in natural language what to do next — the platform answers in kind and then can guide the search. It happened so fast, and these are no small achievements.

This type of triage, risk assessment, and guided search is essential to Elastic AI Assistant, but this serves as the starting point. Elastic AI Assistant does more.

Committed to Open Frameworks — How Elastic Common Schema, Eland, Elasticsearch Relevance Engine, and Elastic AI Assistant Are Connected

There are arguably three guiding principles that Elastic remains committed to. The first is Elasticsearch has to provide an optimized (fast and accurate) search function. Second, the Elastic platform has to empower the customer; that is, to the degree that it is possible, Elastic technology has to help the customer with data migration, allow customers to import its own machine learning, and incorporate open frameworks while not insisting on techniques that create a "vendor lock" situation. Last, the Elastic platform has to provide continuous observability of a heterogeneous network while providing context for problem solving.

To begin, in a bid to offer an open schema for all to use, Elastic contributes the Elastic Common Schema (ECS) to OpenTelemetry (it is both an answer and an alternative to the Open Common Schema Framework [OCSF]). The goal of both endeavors is to create a common log format that can be used by IT/cybersecurity appliances that would negate the need for translators.

The first technology developed was Eland. Eland is a library that allows customers to upload compatible machine learnings to Elastic to be managed and used at ingest and at retrieval (ML) models.

Eland is interesting. ML is a term and a concept that is ubiquitous in cybersecurity. However, not all ML engines are the same. Often, monitoring data is only needed locally and does not feed into a data lake; Eland helps with this. In addition, the end user may want to create specific use cases for ML or isolate data. There are three ways that Elastic allows a customer to maximize ML:

- An end user can access PyTorch models. Developed by Meta AI, PyTorch is an open source ML framework that businesses can access license free. Eland over PyTorch allows uploads for command-line, Docker, and self-created Python code. Docker is the least complicated to work with because it does not require a local installation of Eland and all of its dependencies.
- The customer can load the model it has already trained. This seems optimal, but ultimately, the Eland will understand the NLP transformers.
- Let Elastic create specific vector embeddings to maximize NLP. Elastic has thought this problem out as well. The Elastic Learned Sparse EncodeR handles semantic searches across domains and provides language identification. The Elastic Agent is used to ingest and index data in real time from cybersecurity appliances. For better fidelity, it is possible to integrate with third-party transformers models such as OpenAI's GPT-3 or Microsoft Azure OpenAI. Eland supports NLP and embedding models, supervised learning models, and generative AI.

All of this starts to funnel back into search. In marketing collateral, Elastic will refer to its application as "Elastic AI Assistant powered by Elasticsearch Relevance Engine (ESRE)." Strong ML algorithms help establish relationships enriching search. Aptly named a "relevance search engine," ESRE builds upon what Eland is doing with transformers and data ingest. At the heart of ESRE is an algorithm based on the BM25, which is a technique used on the premise of "best matching." The approach is algorithmic based where entity frequency such as names (not fields), information context, and recency of use is weighted and scored for relevance. The BM25F version used by Elastic is a multifield version of what would be done in a singular data lake.

In addition, Elastic also uses an embedded vector search. Creating search context is not wholly intuitive. Vector search finds related data using approximate nearest neighbor (ANN) algorithms. For example, the type of context around the term *chief executive officer* might be CEO, boss, president, and final decision maker. You can see in cybersecurity where there are similarities in malware signatures, rules, and detections and closely aligned playbooks where a vectorized search function can improve efficiency.

So should you use BM25F or vector search? Search engines allow for weighting, prompting, and other ways to manipulate the score and the relevance of results. Elasticsearch has added Reciprocal Rank Fusion (RRF), which is an improved alternative algorithm using linear hybrid scoring across query types that may include those utilizing vector search. RRF can be used to combine results from BM25 queries with ELSER queries or BM25 with vector search queries.

In June 2023, Elastic announced Elastic AI Assistant to offer security practitioners a chat interface to ask questions using natural language. The assistant is based on the Elasticsearch Relevance Engine, which helps customers build search AI applications without requiring them to run their own LLM models. (Note that a customer may build ML models using the Eland library; however, the Elastic AI Assistant is not dependent on Eland usage. It connects to OpenAI or Azure OpenAI with properly generated API keys.) Prebuilt prompts are currently available including summaries, queries, and suggested actions with all of the context saved in the chat framework and also able to be added to a case during an investigation. Prompts also allow the security analyst to anonymize the data being sent to the AI model to meet data privacy and security concerns.

One advantage for Elastic in training the model is that Elastic is an open platform so all information about its code and platform is easily available to all. Elastic is currently connecting to OpenAI and Azure AI models to provide AI output though in the future this will be expanded to other models including allowing the customers to connect their own models. Elastic tests the data efficacy of generally available models to ensure responses to prompts are accurate before deciding to integrate with them though customers would need to assume the risks of connecting to internal models. Customers are able to see a dashboard view of how many tokens their searches and queries are using so usage can be monitored.

Elastic expects the Elastic AI Assistant to be particularly helpful for customers that are migrating their SIEM to Elastic from a previously used vendor. The Elastic AI Assistant can convert detections, queries, and data mapping from the customer's previously used language and schema to the EQL and ECS used by Elastic in much less time than it would take humans to do so. This feature is more than a checklist item. If a company wants to migrate from an AWS environment to an Azure environment, the detection rules can come with them. Similarly, open source detection rules from a Shodan or Sigma can be translated and ported onto appliances such as AV, firewalls, or SIEM filters. These rules seem to be adequately converted about 95% of the time.

As mentioned previously, the first uses of generative AI have been in the first steps of alert investigation. This is a subset of the fields that Elastic provides on the home screen and in the first prompt include:

- Alerts by name
- Severity level of the alerts
- host.name
- Agent.status
- user name
- rule.name
- process.executable
- Rule.type
- Process.name
- Rule description

These fields can also be changed, and the pre-populated fields greatly reduce what analysts have to go through in assembling artifacts and timelines, a process affectionately called triage. The Elastic AI Assistant will then present an alert summary such as which process hash was used, provide specific recommended actions (active steps toward remediation if you will), and then provide triage steps, which is guided threat hunting (included in this is the guided threat hunting is what tools the analyst should use). These functions are all impressive, and generally where AI digital assistants are headed. Note that IDC does not have a lab environment where we can run digital AI assistants through its paces, and we can formally and objectively score the results, but the Elastic AI Assistant does seem to be in the upper echelon of this new and rising tool class. However, there is a little more to the story.

Mapping data is going to be massively important, and there are several reasons. First, companies have to make sure that personally identifiable data, intellectual property, and data that is subject to local regulations is not exposed and, in many cases, explicitly not used in training machine learning models. This will be a daunting challenge. Second, without proper data discipline, the aforementioned

hallucinations from generative AI become increasingly likely. Last, if data is misaligned, it cannot be properly used.

One of the key aspects of Elastic AI Assistant is that it monitors data at the time of ingestion to see that it is properly aligned with the Elastic Common Schema. Conforming to this schema brings all of the previous elements into play. ESRE becomes more efficient, and third-party transformers and third-party ML models can be used.

In talking with Elastic, it admits to being early in the journey of how to develop its AI assistant, but express optimism. Elastic intimates that if it can find the right domain controls, it can expand Elastic AI Assistant to help customers with line-of-business and customer insights as well as IT and cybersecurity. As it has been for the longest time, the data is there. The key is unlocking the secrets within the data.

ADVICE FOR THE TECHNOLOGY SUPPLIER

This is the early stages of what generative AI in IT/cybersecurity, especially pertaining to how the technology is used in security operations. While the new norms are still to be established, there are still a few things that can be advised uniquely to Elastic:

- **Continue to develop more use cases for search.** Elastic is a company that is literally conceived from the search function, and this advice is tantamount to preaching to the choir. Cybersecurity analysts often think of search in the context of indexing that enables a team to look for artifacts or indicators of compromise in historical logs or search applied to user behavioral analytics. Search is more than that. It can enrich customer and employee experiences. Search can help scope problems with applications and network performance. Granted, it would be almost impossible to prepackage responses to every search query, but adding chat to ordinary IT functions is a time saver.
- **Note that as Elasticsearch is built on open platforms, Elastic has a special obligation to disclose its methodology if not its algorithms**. Customers will have a growing fear of how answers are derived — in other words a black box approach. There are two problems that AI will not solve. First, it will not always be right. The network is fluid, and the adversary is always looking for ways to pick at the seams. Second, there is value in the process. While the triage process may solve a set of alerts, understanding the logic may help create new detection rules preemptively or create insights about a company's cybersecurity posture.
- **Build use cases specifically around identity.** Many sources including the Verizon Breach report attribute roughly 80% of breaches to poor identity and access management practices. Special attention should be paid to Active Directory.

Pertaining to the Elastic AI Assistant, Elastic already does two smart things. First, Elastic monitors OpenAI usage; that is, it tracks how many tokens each user utilizes in Elastic and then how many tokens are used in other ChatGPT applications. Second, Elastic offers a freemium model of the AI assistant.

IDC sees real potential in the Elastic AI Assistant and expanding its capabilities beyond IT, and cybersecurity can only add to Elastic customer satisfaction.

LEARN MORE

Related Research

- *Can We Trust It? Challenges to Maximizing the Value of AI* (IDC #US50732423, June 2023)
- *Achieve Cyber-Resilience and Trust Through Better Vulnerability Management* (IDC #US50717223, May 2023)
- *Security Maturity and Vulnerability Management* (IDC #US50587123, April 2023)
- *Engendering Trust with Proactive Cybersecurity Using Continuous Risked-Based Posture Assessment* (IDC #US50456123, March 2023)
- "No Turning Back: AI Everywhere Causes a Seminal Shift in the Tech Market," **blogs.idc.com/2023/06/28/no-turning-back-ai-everywhere/**

Synopsis

This IDC Market Perspective looks at the announcement of Elastic AI Assistant, which offers security practitioners a chat interface to ask questions using natural language. Generative AI and its close kin ChatGPT are not "like to have" capabilities in the cybersecurity stack as much as these are seminal requirements. The first iterations of generative AI have been seen in the incident detection and response (IDR) stack. Analysts use natural language processing (NLP) to assemble timelines, refine threat intelligence research, determine risk, find artifacts on networks, and then begin guided threat hunting. These tasks seem remedial, but these are exceptionally time consuming and prone to user error. The first formal general availability announcement of a generative AI, ChatGPT digital assist was made by Microsoft and its announcement of Security Copilot, but several companies have followed suit. (It should be noted that many companies would dispute Security Copilot was the first, although its technology was the first ChatGPT specific. Indeed, generative AI has been a part of different aspects of SIEM, threat intelligence, and security automation for some time now.)

Applying generative AI to the seminal stages of IDR should not be glazed over. This is a massive achievement that helps and will help cybersecurity generalists as well as more refined cybersecurity teams. But cybersecurity is not limited to IDR: truly it entails the proper prepping of the network environment (understand configurations and data structure), ties into line of business, and integrates search proactively into observation and new use cases. Elastic Assistant AI is a positive realization in that direction.

"In truth, we don't fully realize the potential gains and possible hazards coming from the use of widespread generative AI," said Chris Kissel, research vice president, Security and Trust at IDC. "The recent platform announcements of what can be deemed as digital assistants are revelatory. Elastic AI Assistant joins this new class of tools. But Elastic realizes that cybersecurity only exists within the context of the network and the business at large. Several capabilities in AI Assistant such as helping customers properly format data for ingest make its platform safer and more accurate than competing generative AI platforms."

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com