



未来运维之道: Elastic AI助手的高效运维与数据智能

李捷：首席解决方案架构师

Jul 27, 2024



Agenda



我们将在今天分享:

- 什么是 AI search? AI search的发展
- Elastic Solution中的GenAI
- AI 智能助手与elasticsearch维护

在LLM带来的新革命中
搜索比以往任何时候都更重要！
信息整合是大模型最 **Solid** 的场景！

大模型和搜索的互相成就 造就了 AI Search

生成式人工智能应用程序
具有特定的挑战：

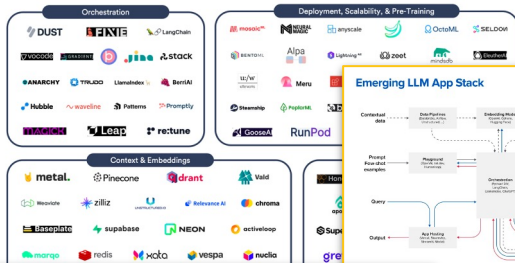
- 幻觉，错误的答案
- 复杂的技术堆栈
- 实时访问私人数据
- 安全与隐私
- ...

Lawyer Used ChatGPT In Court—
And Cited Fake Cases. A Judge Is
Considering Sanctions

Molly Bohan
I cover bread

BASE10 TREND MAP: Generative AI: Developer Tools & Infrastructure
Companies are grouped based on primary developer use cases

Base10



OpenAI

Research API ChatGPT Safety Company

GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its data cuts off (September 2021), and does not learn from its experience. It can sometimes make simple reasoning errors which do not seem to comport with competence across so many domains, or be overly gullible in accepting obvious false statements from a user. And sometimes it can fail at hard do, such as introducing security vulnerabilities into code.

**Samsung bans use of generative AI tools
like ChatGPT after April internal data leak**

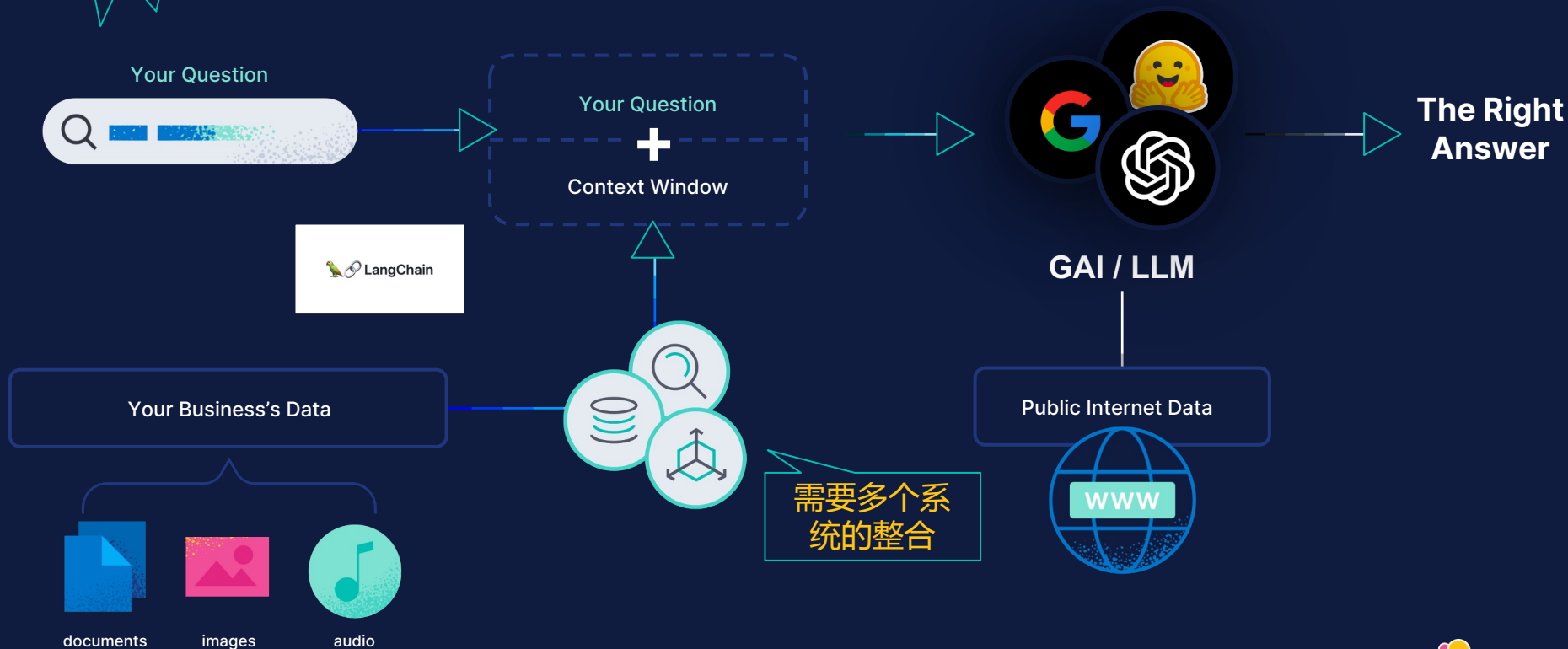
Kate Park @kateparknews / 9:17 AM EDT • May 2, 2023

Comment

早期

Retrieval Augmented Generation

向量搜索因此而引爆





发展

Enter: Elasticsearch Relevance Engine™ (ESRE)

不仅仅是一个向量库

Your Question



Your Question



Context Window

LangChain

Your Business's Data



documents



images

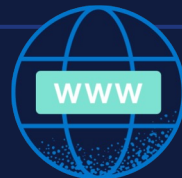


audio



GAI / LLM

Public Internet Data

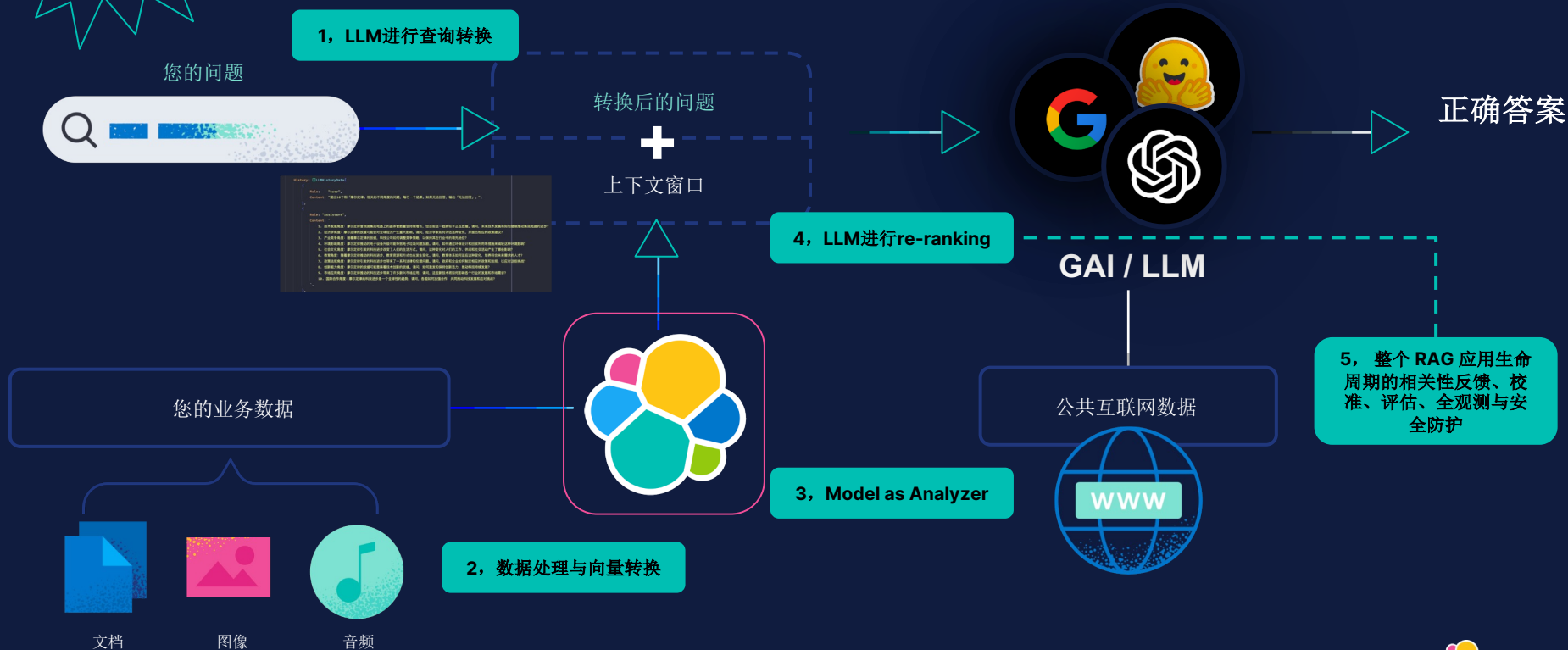


The Right Answer

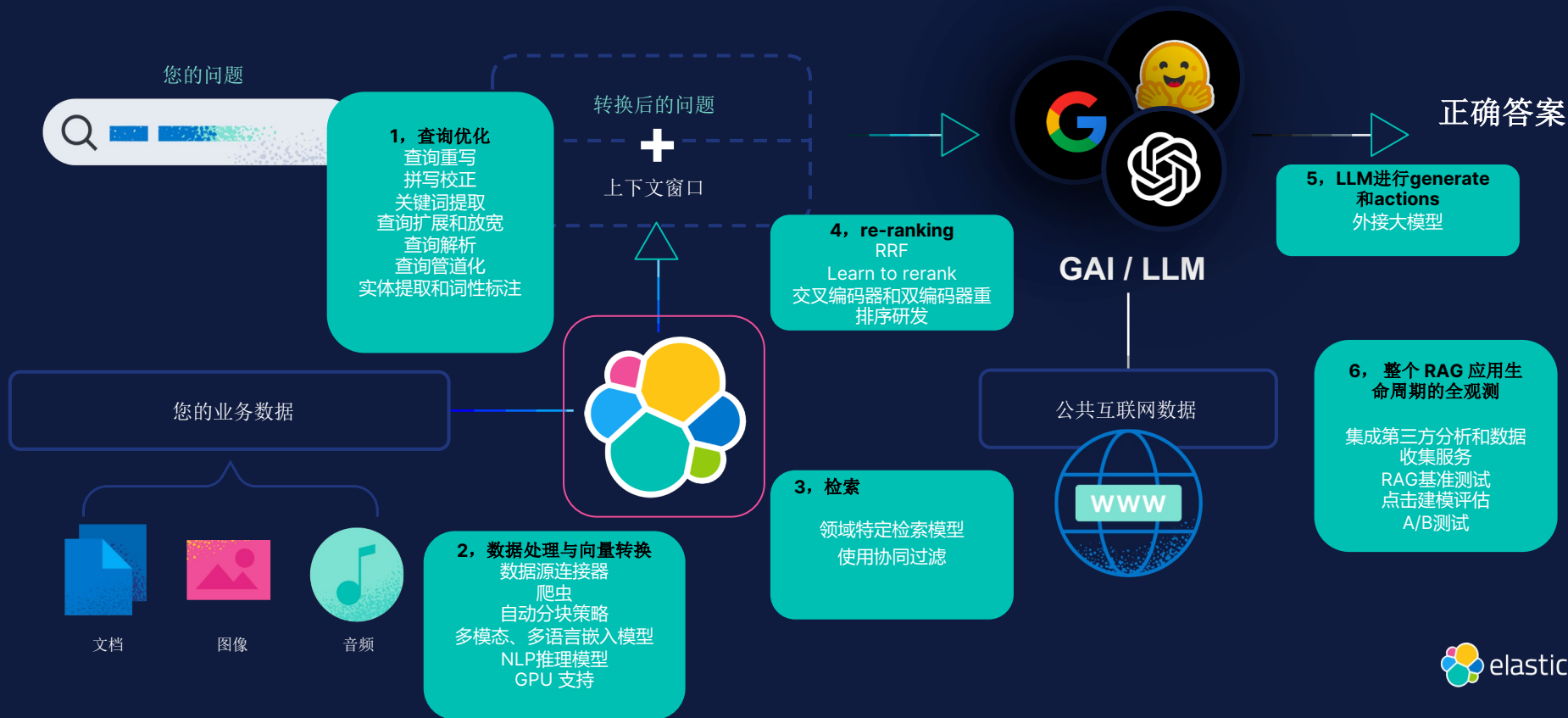


进化

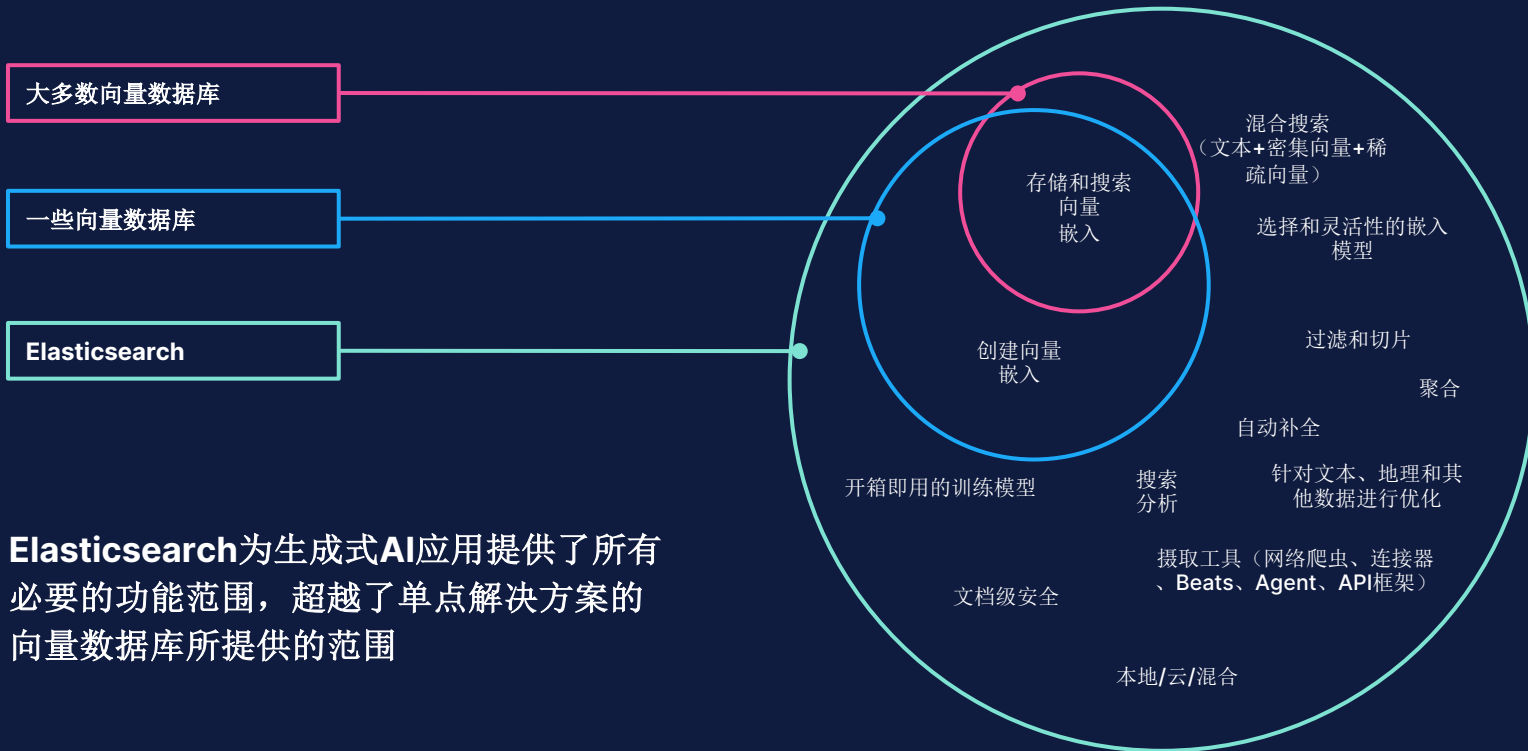
AI search需要更多的能力



Elastic在AI search上的投资与Roadmap

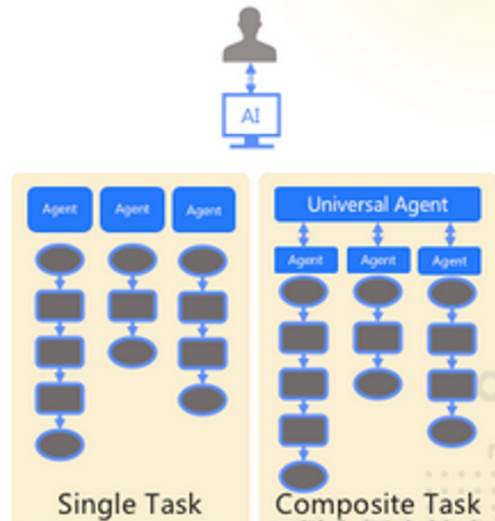
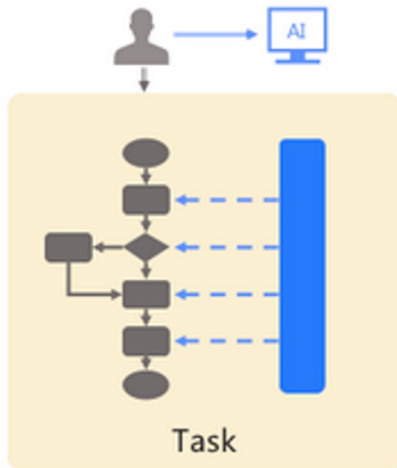
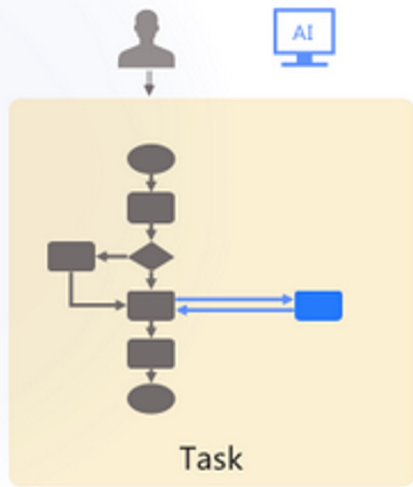


Elastic拥有您所需的所有功能



Elasticsearch为生成式AI应用提供了所有必要的功能范围，超越了单点解决方案的向量数据库所提供的范围

RAG - Copilot - AI Agent



@稀土掘金技术社区

RAG + AI Agent

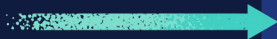
Natural Language
Operations

Logs and Events

Cluster information

Domain specific
knowledge

Alerts



Elastic AI Assistant

构建你所需的知识库

获取集群中的有效上下文

语义搜索

调用ES和Kibana的API

协助你获得深层次的洞见

提供有效的建议或者解决方案

帮助你进行必要的操作

RAG for Search

Powered by **ESRE** Elasticsearch Relevance Engine™

- 快速验证与调试
- 代码样例
- 性能与成本分析

The screenshot displays the Elastic Playground interface, which is used for testing and debugging RAG applications. The interface is divided into several sections:

- Left Sidebar:** Contains navigation links for Search, Build, and Playground. Under Search, there are links for Home, Content, Indices, Connectors, and Web crawlers. Under Build, there are links for Playground (highlighted), Search Applications, and Behavioral Analytics. Under Getting started, there are links for Elasticsearch, Vector Search, and AI Search. Under Enterprise Search, there are links for App Search and Workplace Search.
- Top Bar:** Features the Elastic logo, a search bar with the placeholder text "Find apps, content, and more.", and links for Playground Docs and Endpoints & API keys.
- Main Content Area:**
 - Playground Header:** Includes the title "Playground" and a "TECH PREVIEW" badge.
 - Chat Interface:** Shows a conversation history. The first message is from the system: "Welcome! Ask a question to get started." The second message is from the user: "You asked at 09:27 ES | QL 支持 CCS 吗?". The third message is from the AI: "AI Grounding answer based on 3 document sources". Below this, the AI response is shown: "AI responded at 09:27 3,523 tokens sent". The response includes a summary and a list of document sources: "是的, ESQL支持查询本地和远程集群索引, 支持CCQ, 支持协调器模式和远程模式。具体信息可以参考文献 [Foundation - [x] ESQL now supports CCQ: https://github.com/elastic/elasticsearch/pull/101640], [Enrich - [x] Resolve policies across clusters: https://github.com/elastic/elasticsearch/pull/104355 - [x] Support any mode: https://github.com/elastic/elasticsearch/pull/104840 - [x] Support coordinator mode: https://github.com/elastic/elasticsearch/pull/104936 - [x] Support remote mode: https://github.com/elastic/elasticsearch/pull/104993 - [] Improve for remote remote: https://github.com/elastic/elasticsearch/issues/105095]和[RCS 2.0]".
 - Model settings:** Includes a dropdown for the model (currently set to "Azure OpenAI" with "ai-assistant" selected) and a section for instructions: "You are an assistant for question-answering tasks. 并用中文回答". There is also a checkbox for "Include citations" which is checked.
 - Bottom Bar:** Contains buttons for "Regenerate" and "Clear chat", and a text input field for "Ask a question" with a send button.

AI Assistant for Observability

Powered by **ESRE** Elasticsearch Relevance Engine™

- 提供警报 RCA
- 解释日志信息
- 解释主机进程
- 解释 APM 错误
- 剖析分析库和函数

The screenshot displays the Elastic Observability interface. On the left is a sidebar with navigation options: Overview, Alerts, SLOs, Cases, Logs, Stream (selected), Anomalies, Categories, Infrastructure, Inventory, Metrics Explorer, Hosts (BETA), APM, and Services. The main area is titled 'Stream' and shows a search bar with the text 'Search for log entries... (e.g. hostname:host-1)'. Below the search bar, there are tabs for 'Customize' and 'Highlights'. The log entries are displayed in a table with columns for date, dataset, and message. The right-hand panel, titled 'Details for log entry BG5UTYoBueMvUUqGcTBC', shows the log entry details and a search bar. A red box highlights the AI Assistant panel, which contains two questions: 'What's this message?' and 'How do I find similar log messages?'. Both questions are answered by the Elastic AI Assistant, providing helpful insights from the Elastic AI Assistant. The connector used is 'Azure OpenAI GPT-4'.

Observability Logs Stream Settings Alerts and rules Add data AI Assistant

Observability

Overview Alerts SLOs Cases Logs Stream Anomalies Categories Infrastructure Inventory Metrics Explorer Hosts BETA APM Services

Stream

Search for log entries... (e.g. hostname:host-1)

Customize Highlights

Aug 31, 2023 event.dataset Message

Showing entries from Aug 31, 15:41:07

15:41:07.272	frontend-node.log	[fronte
15:41:07.272	frontend-node.log	fronte
15:41:07.272	frontend-node.log	fronte
15:41:07.275	frontend-node.log	fronte
15:41:07.275	frontend-node.log	fronte
15:41:07.277	postgresql.log	[postgr

Details for log entry BG5UTYoBueMvUUqGcTBC

From index .ds-filebeat-8.5.0-2023.08.31-000043

Investigate

What's this message?
Get helpful insights from our Elastic AI Assistant. Connector: Azure OpenAI GPT-4

How do I find similar log messages?
Get helpful insights from our Elastic AI Assistant. Connector: Azure OpenAI GPT-4

Search...

Field	Value
@timestamp	2023-08-31T20:41:07.272Z
agent.ephemeral_id	f4d0276d-7e75-402b-ae64-39f927f41789

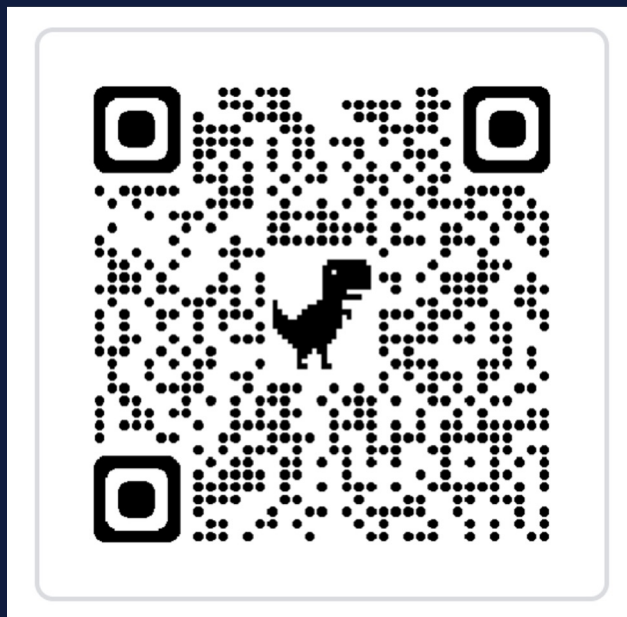
AI Assistant for Security

Powered by **ESRE** Elasticsearch Relevance Engine™

- 警报汇总
- 查询生成
- 工作流程建议
- 数据摄入助手
- 查询转换
- 代理集成建议

The screenshot displays the Elastic AI Assistant interface. At the top, the title "Elastic AI Assistant" is shown with a help icon. Below the title, there's a "Conversations" section with a dropdown menu set to "Alert summary" and a "Show anonymized" toggle. The main chat area shows a conversation with the assistant. The user's prompt is: "Alert (from summary)". The assistant's response is: "You are a helpful, expert assistant who answers questions about Elastic Security. Do not answer questions unrelated to Elastic Security. If you answer a question related to KQL or EQL, it should be immediately usable within an Elastic Security timeline; please always format the output correctly with back ticks. Any answer provided for Query DSL should also be usable in a security timeline. This means you should only ever include the 'filter' portion of the query. Use the following context to answer questions: > Alert (from summary)". Below the chat area, there's a section for "Evaluate the event from the context above and format your output neatly in markdown syntax for my Elastic Security case." with a list of suggested actions: "Alert summarization", "Workflow suggestions", "Query conversion", and "Agent integration advice". On the right side, there's a "Quick Prompts" sidebar with a search bar and a list of prompts: "Alert summarization", "Query generation", "Workflow suggestions", "Custom data ingestion helper", "Query conversion", and "Agent integration advice".

DEMO



不仅于此

AI Assistant 能够帮我们完成更智能，更规范，更轻松高效的运维



无脑接入
积累技术
债的开始

解决更多的问题

我们如何开始日常的Elasticsearch工作？从日志场景开始

我们需要接入一个新的登录日志：

```
2023-12-29T17:24:03, srv-  
80.rogers.com, rogers,  
219.118.103.238, CN, Beijing,  
172.16.0.3, success, demo, log_on,  
authentication, Mozilla/5.0  
(Linux; Android 4.1)  
AppleWebKit/536.2 (KHTML, like  
Gecko) Chrome/27.0.889.0  
Safari/536.2
```

或

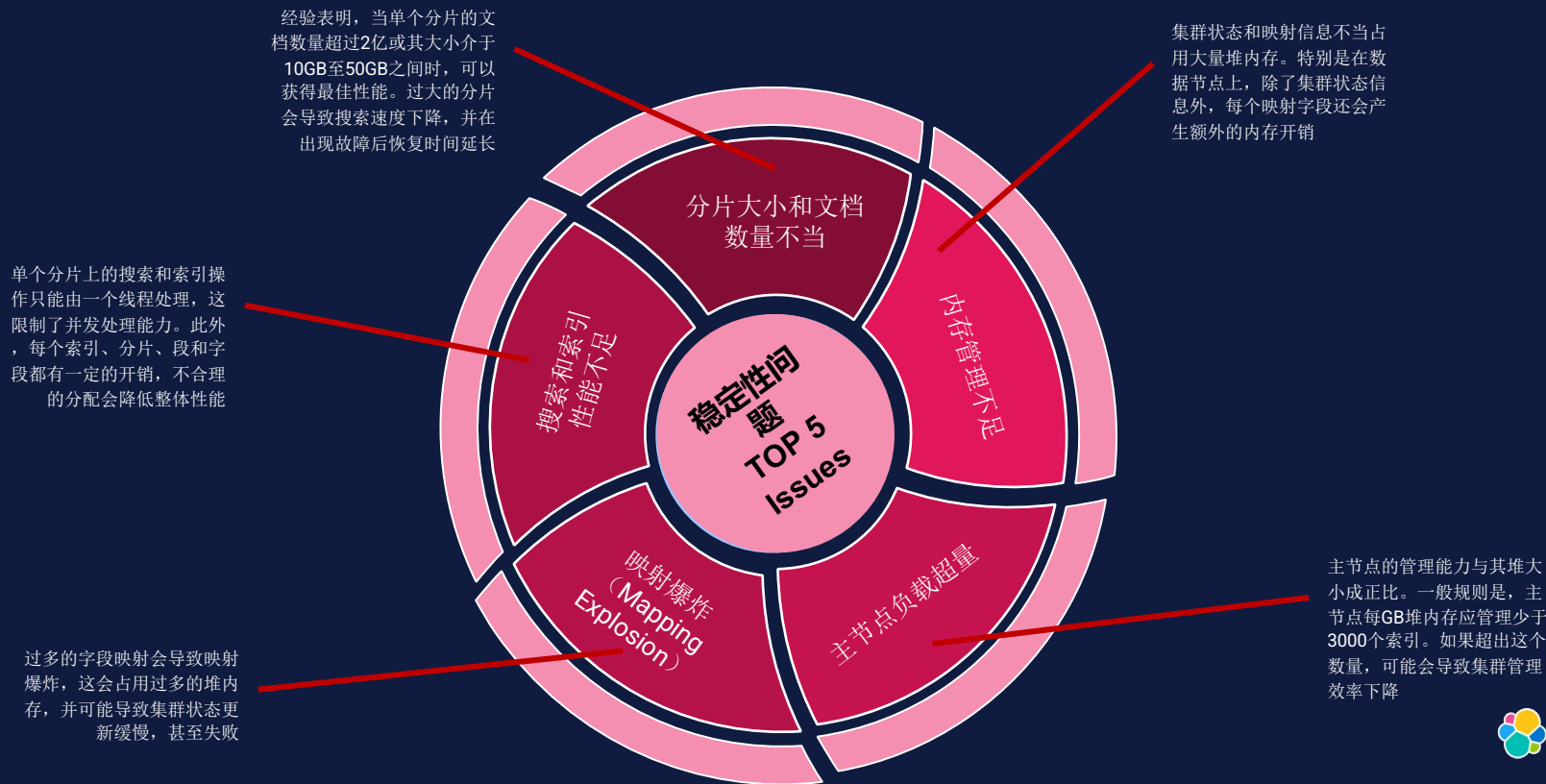
```
{  
  "@timestamp": "2023-12-29T17:24:03",  
  "host": {  
    "name": "srv-80.rogers.com"  
  },  
  "user": {  
    "name": "rogers"  
  },  
  "source": {  
    "ip": "219.118.103.238",  
    "geo": {  
      "country_iso_code": "CN",  
      "city_name": "Beijing"  
    }  
  },  
  "host.ip": "172.16.0.3",  
  "event": {  
    "outcome": "success",  
    "dataset": "demo",  
    "action": "log_on",  
    "category": "authentication"  
  },  
  "user_agent": {  
    "original": "Mozilla/5.0 (Linux; Android 4.1) AppleWebKit/536.2  
(KHTML, like Gecko) Chrome/27.0.889.0 Safari/536.2"  
  }  
}
```



常见的数据治理问题

- 没有索引模板
 - 没有统一的数据schema规范，总是依赖动态mapping
 - 没有别名
 - 索引不滚动切分
- 没有数据生命周期管理
- 不进行数据分片的管控
- 不使用Datastream处理日志和指标数据

因为缺乏数据治理规范，导致数据治理水平较低，进而表现出集群性能不高，稳定性欠佳，数据安全性不强等问题



66

开放生态的优势：为什么 GPT-4 能在不进行 fine tuning 的情况下精准的完成 elasticsearch 的基础操作？

自带搜索的优势：AI 助手还需要哪些增强才能完成更加高级的操作？



需要企业内部知识库才能执行的高级功能

- 配置连接器，如企业微信，钉钉
- 配置快照仓库，使用snapshot和restore进行数据备份和恢复
- 配置远程集群
- 配置 SSO, LDAP
- 根据企业架构信息，进行推理与决策
- 根据 playbook, runbook 执行修复特定任务的修复指令

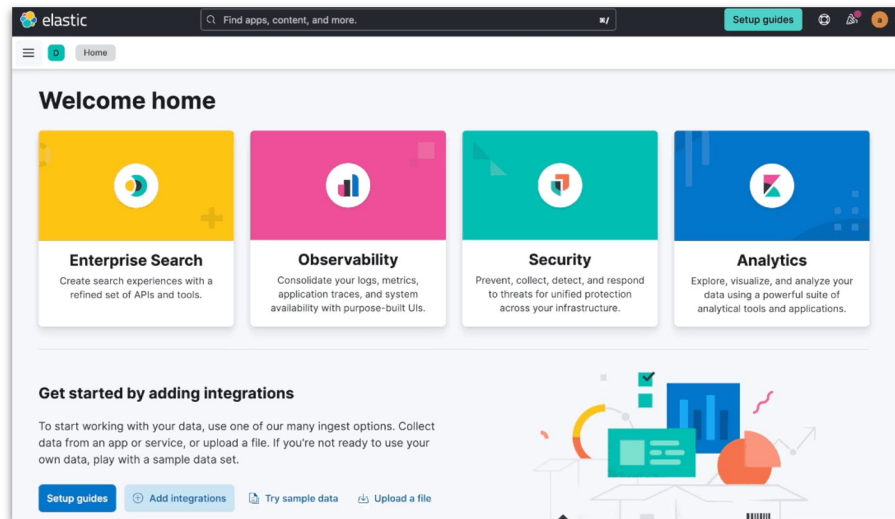
AI驱动搜索: Elastic Learned Sparse Encoder

驱动

- 用于增强语义搜索
- Elastic的新相关性引擎
- 开箱即用的AI驱动搜索

价值

- 卓越的语义相关性：基于意义的搜索，而不仅仅是分词
- 最先进的技术，在行业基准测试中优于竞争对手
- 通过我们精简的_search API，一键开启



原生集成

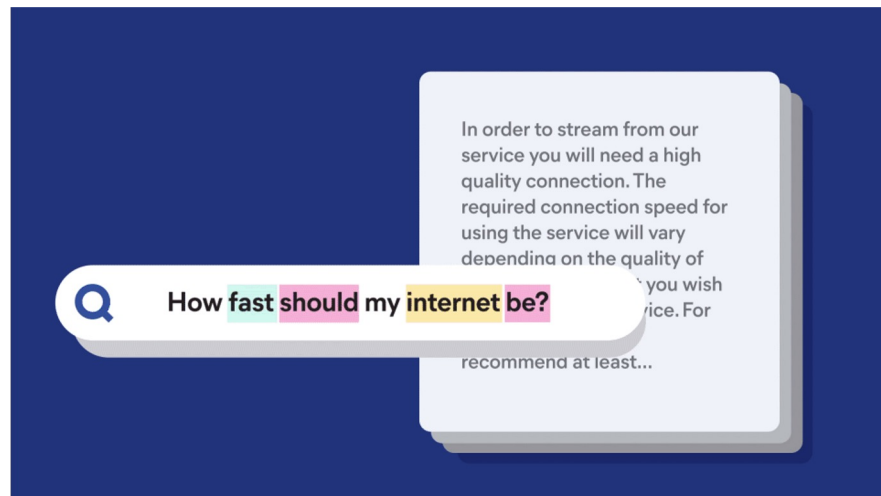
AI-powered Search: Elastic Learned Sparse Encoder

AI搜索的不可逾越的障碍

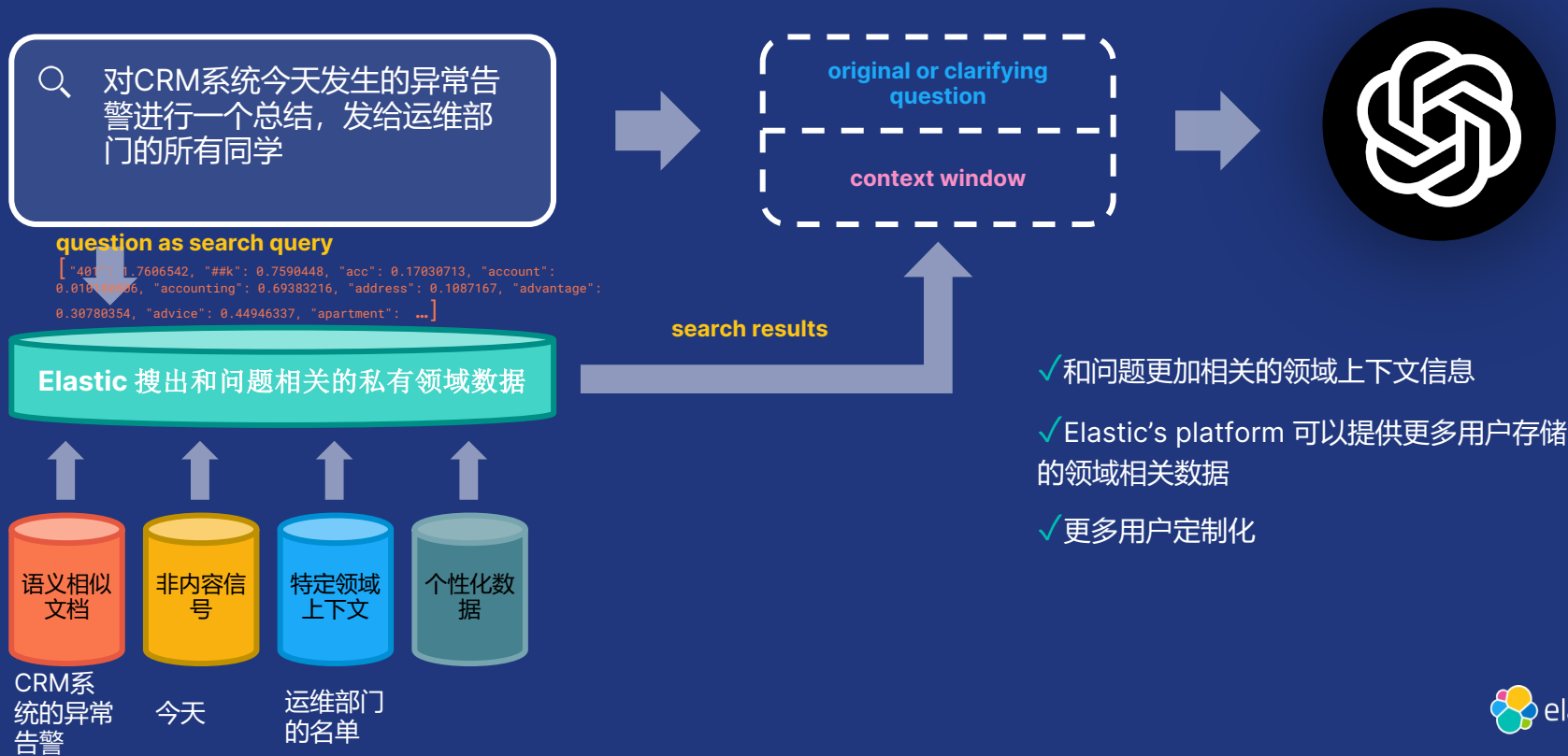
- 大量的专业知识和资源
- 流程明显不同于传统的搜索

Enter Elastic Learned Sparse Encoder

- 良好适配不同知识领域(out of domain)
- 开箱即用
- 解决词汇不匹配的问题
- 比向量搜索结果更具可解释性
- 最先进的技术：优于SPLADE
- 合适倒排索引，在ES上运行飞快



上下文窗口工作流程示意图





AI助手也有不好的地方

- 需要访问外部网络，用户体验稍差
- 只要推理过程超过10秒，就出现异常
- 不可预测性，同一条指令，可能每次执行的方式都不一样！

66

谢谢