elastic

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

elastic

在LLM带来的新革命中

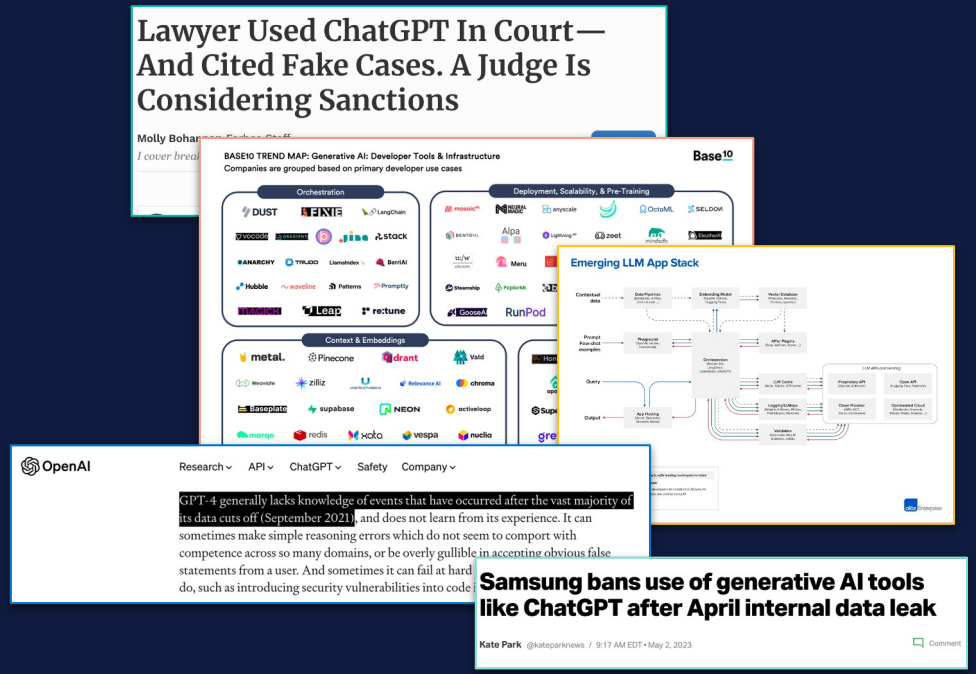
搜索比以往任何时候都更重要！

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
- ...



早期

Retrieval Augmented Generation

向量搜索因此而引爆

elastic

发展 ...ter: Elasticsearch Relevance Engine™ (ESRE)

不仅仅是一个向量库

elastic

AI search需要更多的能力

进化

1，LLM进行查询转换

您的问题

转换后的问题

+

上下文窗口

正确答案

4，LLM进行re-ranking

GAI / LLM

5， 整个 RAG 应用生命周期的相关性反馈、校准、评估、全观测与安全防护

您的业务数据

公共互联网数据

WWW

3，Model as Analyzer

2，数据处理与向量转换

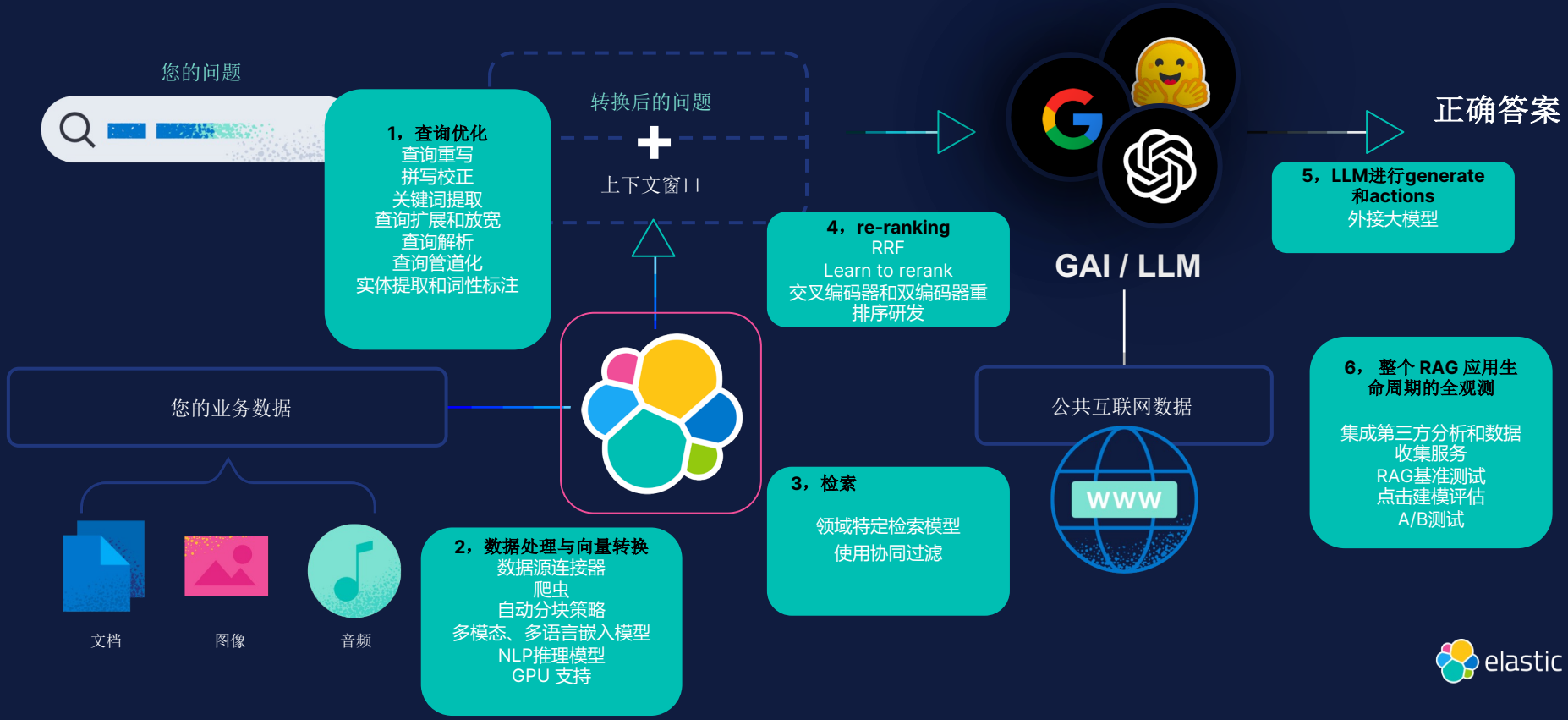
文档

图像

音频

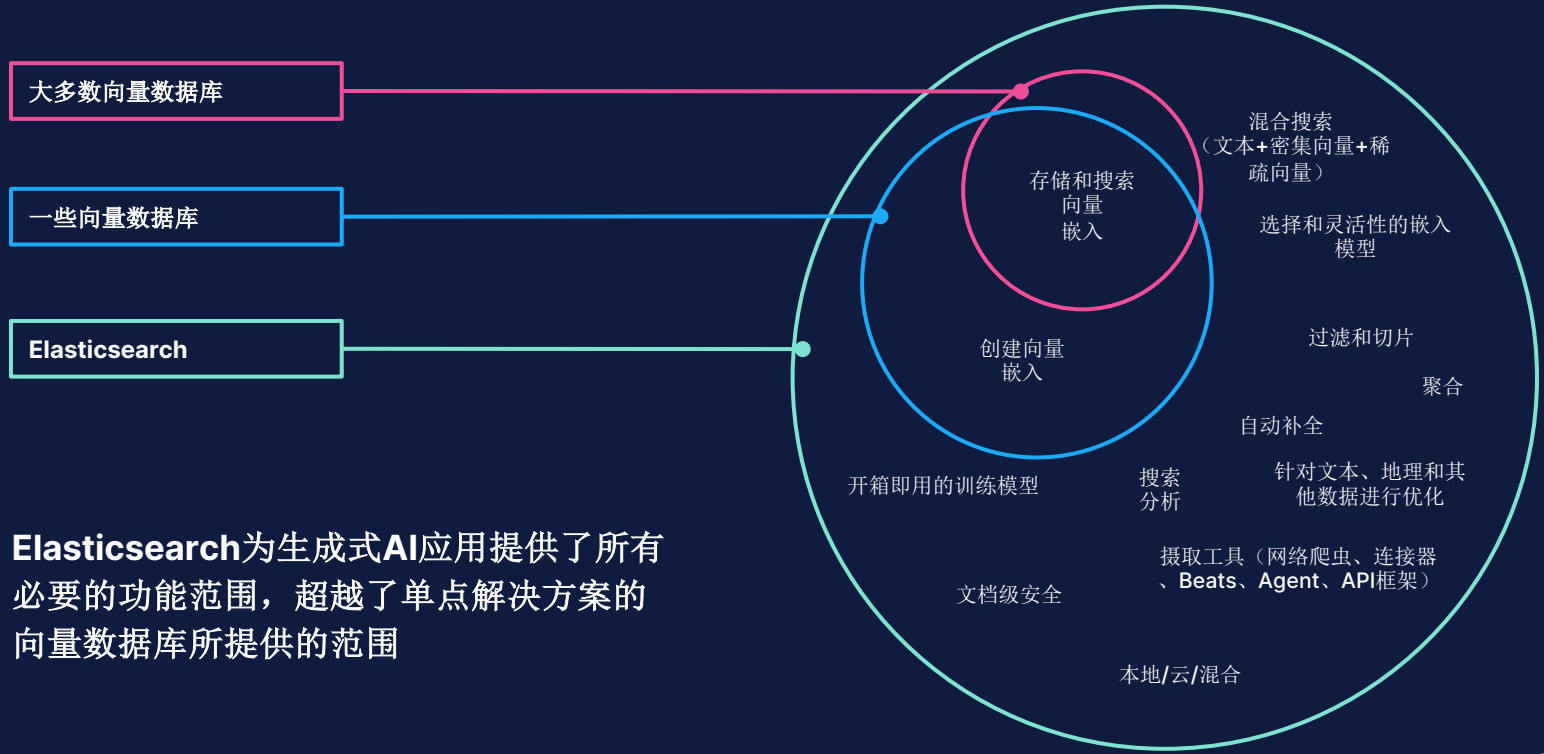
elastic

Elastic在AI search上的投资与Roadmap



elastic

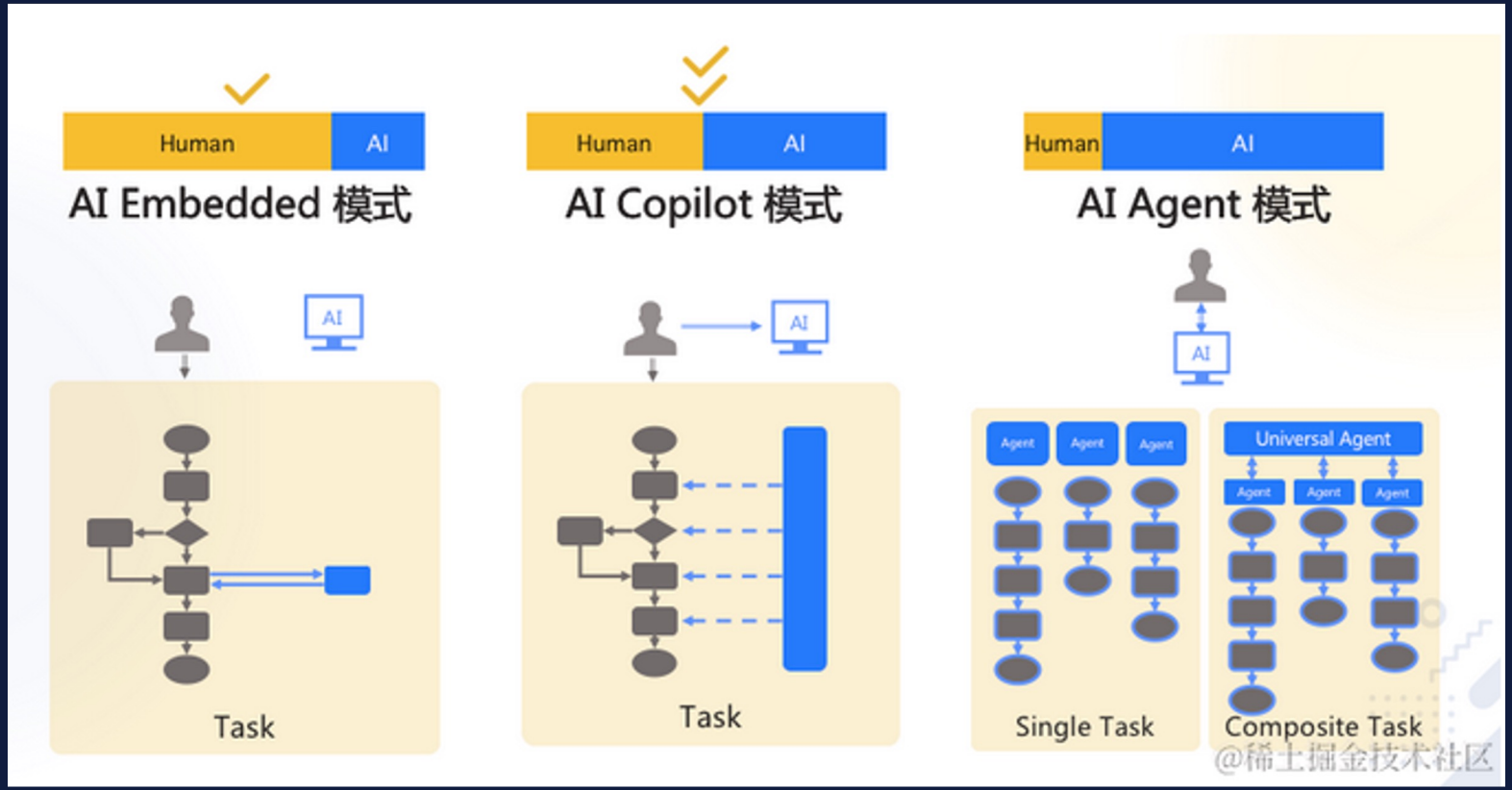
Elastic拥有您所需的所有功能



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

elastic

RAG - Copilot - AI Agent



elastic

RAG + AI Agent

Natural Language Operations

Logs and Events

Cluster information

Domain specific knowledge

Alerts

协助你获得深层次的洞见

提供有效的建议或者解决方案

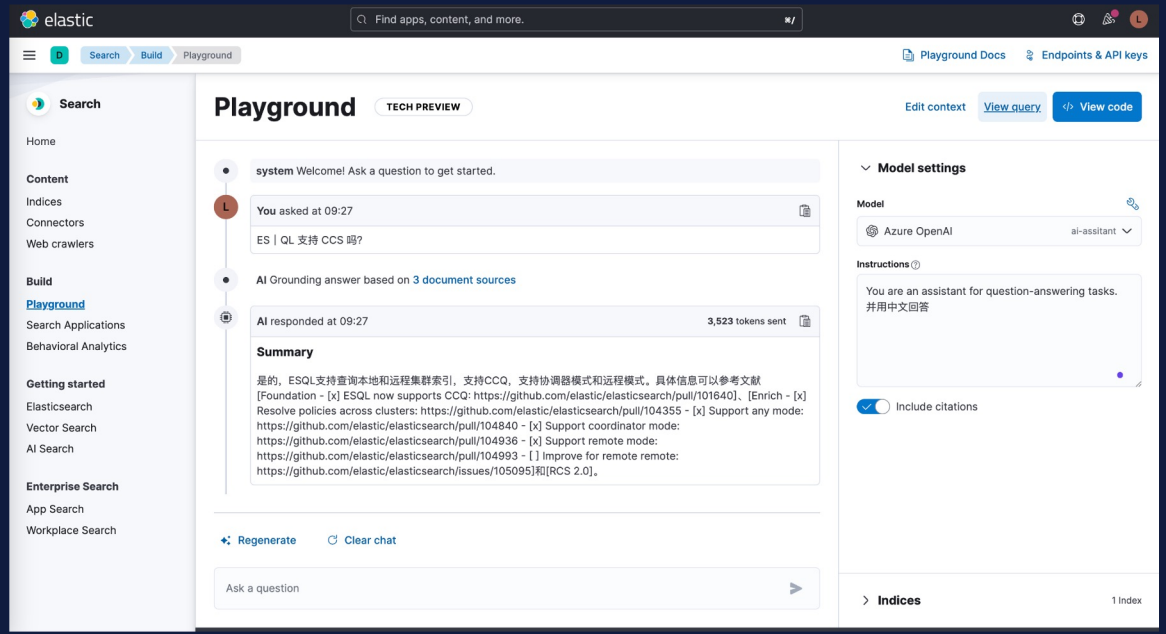
帮助你进行必要的操作

elastic

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

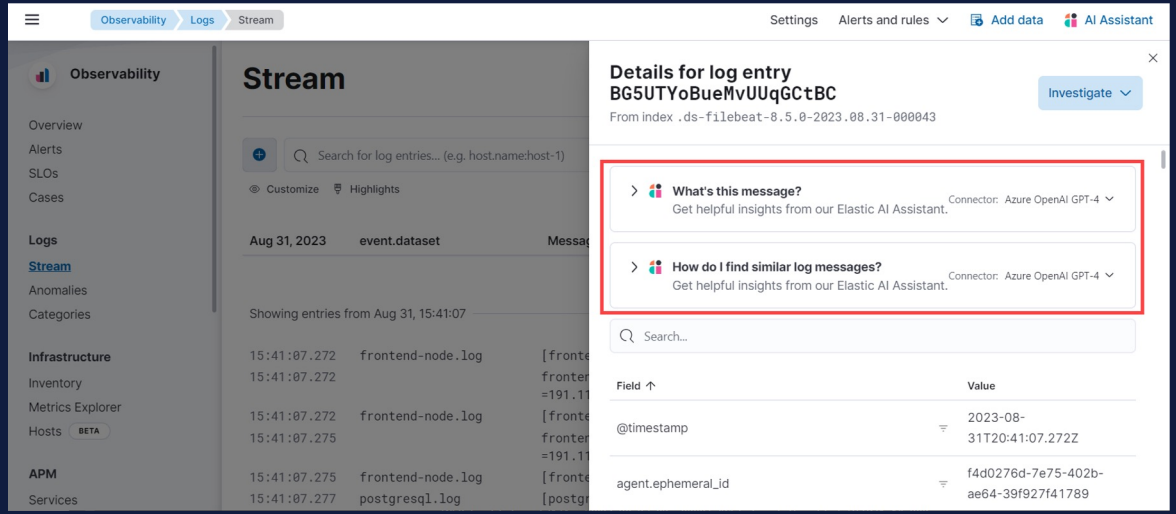


elastic

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

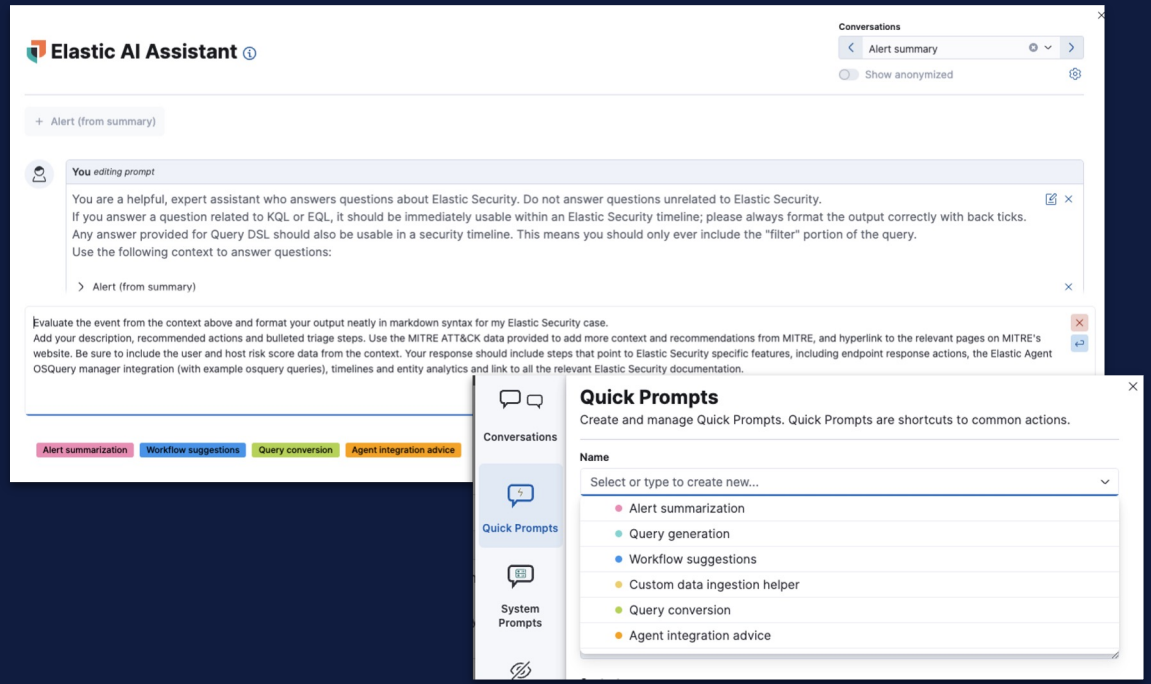


elastic

AI Assistant for **Security**

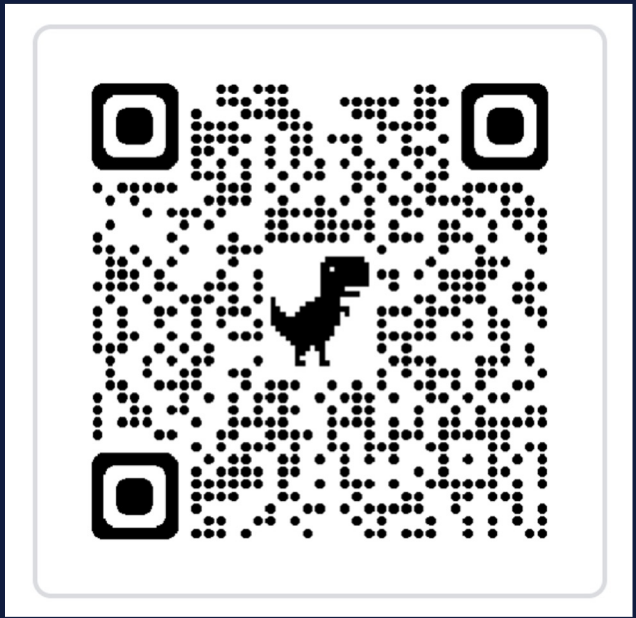
Powered by ESRE Elasticsearch Relevance Engine™

- 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



elastic

DEMO



elastic

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

elastic

解决更多的问题

无脑接入
积累技术
债的开始

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

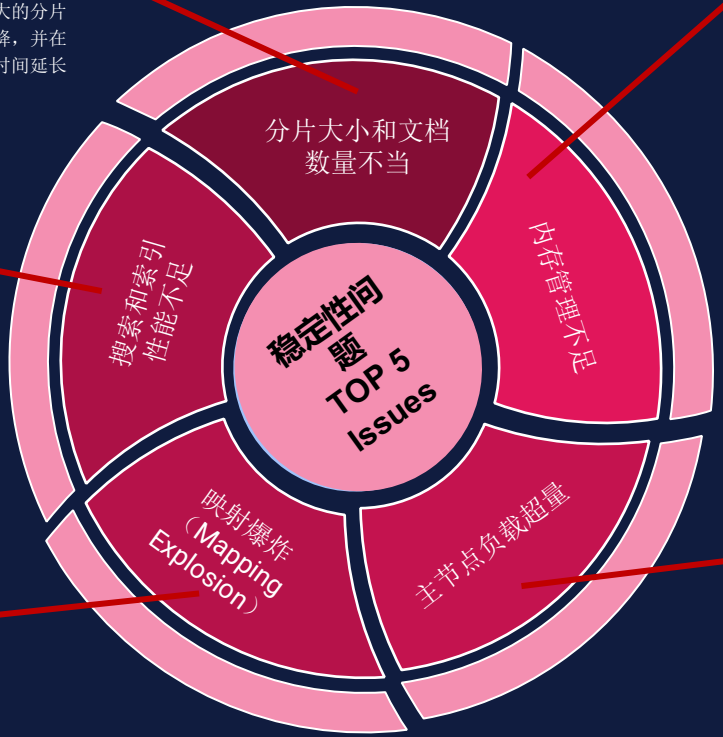
- 没有索引模板
  - 没有统一的数据schema规范，总是依赖动态mapping
  - 没有别名
  - 索引不滚动切分
- 没有数据生命周期管理
- 不进行数据分片的管控
- 不使用Datastream处理日志和指标数据

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

经验表明，当单个分片的文档数量超过2亿或其大小介于10GB至50GB之间时，可以获得最佳性能。过大的分片会导致搜索速度下降，并在出现故障后恢复时间延长

集群状态和映射信息不当占用大量堆内存。特别是在数据节点上，除了集群状态信息外，每个映射字段还会产生额外的内存开销

单个分片上的搜索和索引操作只能由一个线程处理，这限制了并发处理能力。此外，每个索引、分片、段和字段都有一定的开销，不合理的分配会降低整体性能



过多的字段映射会导致映射爆炸，这会占用过多的堆内存，并可能导致集群状态更新缓慢，甚至失败

主节点的管理能力与其堆大小成正比。一般规则是，主节点每GB堆内存应管理少于3000个索引。如果超出这个数量，可能会导致集群管理效率下降

elastic

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?

elastic

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

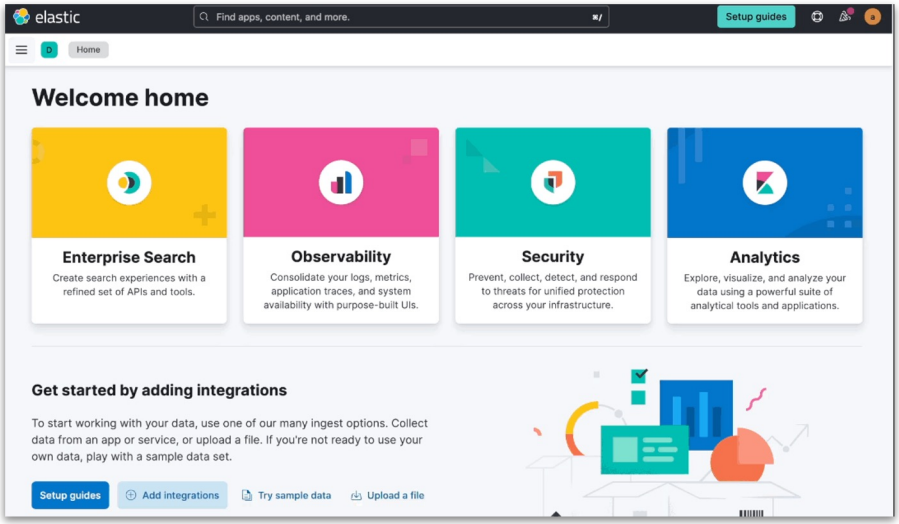
AI驱动的搜索: Elastic Learned Sparse Encoder

驱动

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

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

elastic

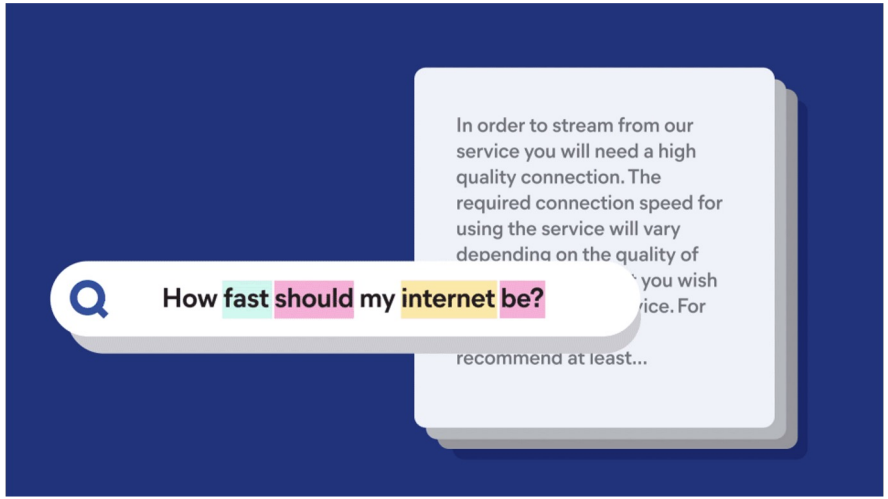
AI-powered Search: Elastic Learned Sparse Encoder

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

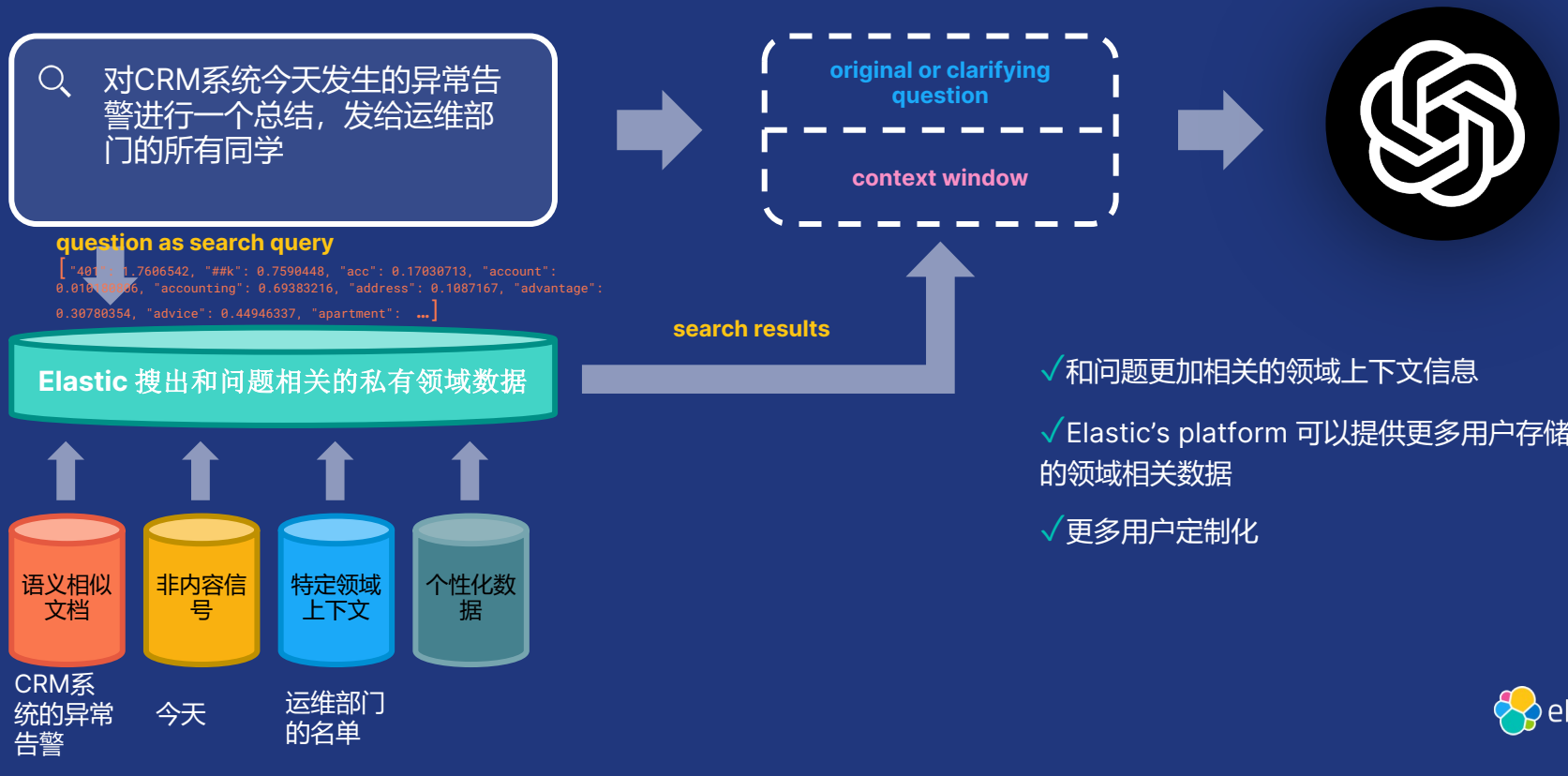
Enter Elastic Learned Sparse Encoder

- 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



elastic

上下文窗口工作流程示意图



✓和问题更加相关的领域上下文信息

✓Elastic's platform 可以提供更多用户存储的领域相关数据

✓更多用户定制化

elastic

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！

elastic

谢谢

elastic