INFINI Labs
极限科技 · 让搜索更简单

向量数据库技术实践

# 从国产替代到自主创新

## Easysearch 在搜索引擎领域的信创实践

主讲人

极限科技 INFINI Labs · 曾勇

CONTENTS · 目录

# 演讲内容导览


01 **AI 时代下搜索引擎的新定位**
从工具到 AI Infra

02 **搜索引擎国产化的必要性**
信创浪潮下的必修课

03 **国产化需要考虑的核心问题**
改造、迁移、回退、服务、运维

04 **我们是谁**
极限科技 INFINI Labs

05 **我们都做了什么**
从引擎到平台，从替代到创新

06 **成功案例**
一汽集团 · 移动云

07 **总结与展望**
三阶段演进与未来方向

01

# AI 时代下
# 搜索引擎的新定位

从工具到 AI Infra

搜索引擎正从"找到信息"的工具，变成连接企业数据与 AI 能力的核心基础设施。

# 搜索技术演进：两个时代的分水岭

传统搜索时代

## 帮用户"找到信息"

基于关键词匹配和倒排索引，用户主动输入，系统被动返回。搜索引擎是一个"工具"。

### 核心技术

关键词匹配 | 倒排索引 | 全文检索 | BM25 评分

### 典型场景

企业文档搜索 | 网站搜索 | 日志检索 | 数据查询

AI 搜索时代

## 理解信息并产生价值

从"匹配字面"升级到"理解语义"，从"返回列表"升级到"驱动决策"。搜索引擎成为基础设施。

### 技术演进

- 语义搜索 + 向量检索
- BM25 + kNN 混合检索
- RAG · 检索增强生成
- Agent 自主调用搜索

### 一句话变化

从"找信息"→"用知识"

# AI 搜索典型应用场景

R

## RAG 企业知识库

企业数据 → 搜索引擎索引 → 检索相关知识 → RAG 组装上下文 → 大模型生成答案

典型应用

企业知识助手 · 智能客服

运维诊断助手 · 合规审查

M

## 多模态搜索

搜索对象不再局限于文本：图片、视频、音频、结构化数据都可以被索引、被理解、被检索

典型应用

以图搜图 · 安防视频检索

设计素材库 · 内容审核

A

## Agent 自主搜索

Agent 自主规划多步任务 → 调用搜索引擎获取上下文 → 综合多源信息执行决策

典型应用

自动化运维 Agent

风控决策 Agent · 投研分析

→ 三种场景的共同点：搜索引擎都是 Agent 的"知识来源"和"决策依据"。

# AI Infra 整体架构

!

搜索引擎正从"应用工具"升级为连接企业数据与大模型能力的核心基础设施。没有好的搜索底座，RAG 和 Agent 就失去了高质量的知识来源。而一旦它成为核心，它的自主可控就成了绕不开的前提。

# 当基础设施失控，代价是什么

同样的失控，发生在"工具"和"基础设施"上，代价天差地别。

- 当它只是"工具"

## 失控 = 体验受损

- 搜索慢一点、查不到，影响单点体验
- 随时可以更换，替换成本低
- 影响范围有限，可容忍、可补救

过去：搜索引擎是查东西的"工具"，出事不致命。

→

- **当它成为"基础设施"**

## 失控 = 全线失守

- 承载企业核心数据，敏感信息高度集中
- 支撑 AI 决策与业务连续，牵一发动全身
- 难以轻易更换，替换成本极高

现在：它是数据主权、业务连续、AI 能力的共同底座，失控不可接受。

?

**于是问题变成了——**

**这么重要的基础设施，凭什么还要依赖国外技术？自主可控，从技术选项变成了战略必选。**

02

# 搜索引擎国产化的必要性

## 从可选项到必修课

国产替代正在从"外围系统"走进"深水区"。搜索引擎作为核心数据闸门，已无法继续依赖国外基础软件。

# 信创三阶段演进

信创远不止"软件国产化"，而是国家层面推动的、实现信息技术体系自主可控的系统性产业升级。

01 **基础设施国产化**
CPU · 服务器 · 操作系统

02 **基础软件国产化**
数据库 · 中间件 · 应用软件

03 **数据基础设施自主创新**
搜索引擎 · AI 平台 · 数据智能基础设施

目标：2027 年前完成央国企核心系统改造，工作重点从外围进入承载核心业务和数据的"深水区"。

**必须直面的四大风险**

01 **供应链断供**
俄乌冲突已敲响警钟：商业支持、安全更新、法律许可可能随时终止

02 **法律与协议**
SSPL/ELv2/AGPL 持续收紧；SaaS 场景下合规边界高度复杂

03 **数据安全合规**
国密算法、等保 2.0、审计溯源、数据遥测——先天不足

04 **技术黑盒**
技术路线被国外厂商主导，深度依赖社区与商业支持

# 选型 5 个关键指标

国产化背景下的搜索引擎选型，早已不是单纯的技术决策，而是数据安全、业务连续性、长期技术战略三重决策。合格的国产搜索引擎，需在五个维度同时达标。

| | | |
|---|---|---|
| 1 | 自主可控 | 核心引擎是否自己掌控？关键 Bug 能否自主修复？技术路线是否受外部制约？ |
| 2 | 企业级稳定性 | 能否支撑 TB/PB 级数据、高并发、7×24 运行？能跑 Demo 和能跑核心业务是两回事。 |
| 3 | 信创真实适配 | 在国产 CPU、OS 环境下是否能"即装即用"？还是需要大量定制化补丁？ |
| 4 | 安全治理内建 | 权限控制、审计溯源、字段脱敏是否搜索引擎自身能力？而非依赖外部组件拼接？ |
| 5 | 长期演进能力 | 能否支撑未来 3–5 年的向量检索、AI 搜索、多模态搜索等演进需求？ |

选的是搜索引擎，本质是选长期技术底座。

03

# 国产化需要考虑的核心问题

## 替代之前，先回答五个问题

业务连续、迁移可行、回退可控、服务可靠、运维可控——五个问题不过关，国产化就是冒险；全部过关，才是机会。

# 国产化之前，先回答五个问题

搜索引擎是企业核心基础设施，国产替代不能只关注"能不能跑"，更要回答这五个关键问题。以下逐一抛出，答案见后续章节。

01

## 现有业务改造成本如何?

企业已有的搜索应用，切到国产引擎需要改多少代码？API 是否兼容？存量数据能否平滑迁移?

02

## 如何无缝平滑迁移?

迁移过程中能否保持业务在线？不同场景下有哪些迁移路径可选？是否有零停机的方案?

03

## 是否能够安全降级回退?

迁移后如果出现问题，能否安全退回到原系统？回退窗口多长？原始数据是否完整保留?

# 国产化之前，先回答五个问题

04

## 产品售后服务如何保障?

原厂还是代理商？技术团队能否直接对接？应急响应多快？部署、培训、持续支持谁来负责?

> 搜索引擎不是一次性采购，是长期技术依赖。售后服务决定了国产化之后"能不能用好"。

05

## 运维团队能否支撑新架构?

是否需要重新学习？现有经验和技能能否复用？日常运维复杂度会不会大幅上升?

> 运维团队是国产化落地的"最后一公里"。团队接不住，再好的引擎也白费。

→ 以上五个问题，不只是在问"能不能替"，更是在问"替完之后怎么办"。接下来看我们为此做了哪些工作

04

# 我们是谁

## 极限科技 INFINI Labs

专注于搜索技术、数据分析与 AI 搜索基础设施——构建面向企业的新一代搜索与 AI 基础设施。

# 极限科技，搜索基础设施建设者

https://infinilabs.cn/

INFINI Easysearch

企业级的分布式搜索型数据库

Coco

为现代团队打造的统一搜索与 AI Agent 中心

INFINI Cloud

一站式企业搜索管理与创新平台

INFINI Gateway

面向搜索场景的高性能数据网关

INFINI Console

一站式的数据搜索分析与管理平台

INFINI Pizza

新一代端侧及分布式实时搜索引擎

## 解决方案



- AI & Agent 搜索与企业知识管理解决方案
- Elasticsearch 企业级平替综合解决方案
- 一站式企业搜索基础设施平台建设方案
- 核心业务场景的海量数据实时搜索方案

05

# 我们都做了什么

## 从引擎到平台，从替代到创新

围绕 Easysearch，构建从引擎内核、企业级特性到 AI 搜索平台的完整能力体系。

# Easysearch–企业级分布式搜索引擎

墨天轮搜索型数据库排行榜长期第一名

INFINI Easysearch

INFINI Easysearch 是一个分布式的搜索型数据库，实现非结构化数据检索、全文检索、向量检索、地理位置信息查询、组合索引查询、多语种支持、聚合分析，提供企业级的安全和加密，并提供完善的周边管理工具。

- 兼容 ES 及其生态
- 完善的企业级功能
- 内核及功能增强
- 易用性
- 国产自主可控

# 向量检索：语义搜索与混合搜索

Easysearch 内置向量检索能力，将全文检索、向量检索与混合搜索统一在同一引擎中，简化 AI 搜索架构与运维。

## 核心能力

**01 kNN 向量检索**

稠密浮点向量字段，近似（LSH）与精确（exact）两种检索模式

**02 混合搜索**

BM25 全文 + 向量相似度双路召回，boost 权重融合排序

**03 向量索引模型**

HNSW、LSH 局部敏感哈希 + exact 精确，支持 cosine 相似度

**04 Embedding 接入**

对接外部 Embedding 模型，实现文本向量化与语义检索

向量数据库能力

一份引擎 = 全文检索 + 向量检索 + 混合搜索

不必再单独部署向量数据库，降低架构复杂度与运维成本。

## 典型应用场景

| | |
|---|---|
| 语义搜索 | 按语义相似度匹配，而非仅依赖关键词 |
| RAG 检索增强 | 为大模型提供高质量上下文 |
| Agent 长期记忆 | Mem0 集成，向量化记忆召回 |
| 企业知识库 | 语义检索与结构化过滤 |
| 多模态检索 | 文本、图像特征等向量数据统一检索 |

# 企业级特性 + 全场景兼容性测试

## A 补齐大量企业级特性

默认 TLS 加密 | RBAC + ABAC 双模权限 | 字段级脱敏 | LDAP / AD 认证 | IP 黑白名单

防暴力破解 | 磁盘文件加密 | 快照加密 | 限流限速阻断 | 国密算法 | 审计日志

文档级权限

> Easysearch 从内核层面补齐了 ES 开源版所缺失的企业级能力，让企业拿到的不只是搜索引擎，而是一个"可治理的基础设施"。

## B 全面的兼容性测试

| | |
|---|---|
| API 兼容 | REST API 全部常用接口 |
| SDK 兼容 | Java · Python · Go · Node.js |
| DSL 兼容 | Query DSL 完整语法 |
| 索引兼容 | ES 6.x / 7.x / 8.x 数据读写 |
| 生态兼容 | Kibana · Logstash · Beats · Filebeat |
| 信创兼容 | 龙芯 · 鲲鹏 · 飞腾 + 麒麟 · 统信 UOS |

# 国密集成 + 安全备份容灾

合规要求越来越高，Easysearch 从算法、备份、容灾三个层面给出端到端的安全保障。

## 国密算法集成

SM2　SM3　SM4

通信加密 + 数据加密，满足等保 2.0 与密评要求，适用于金融、政务等高合规要求行业。

## 安全备份

- 全量 + 增量 Snapshot 备份
- 备份文件支持加密
- 鼎甲 DBackup 深度兼容认证

## 异地容灾

- Gateway 跨机房双活部署
- CCR 跨集群复制 (v2.3.0+)
- 主备切换 · 自动 + 手动

# 存储优化：压缩、分层与快照搜索降本增效

海量数据场景下，存储成本是搜索引擎长期运营的最大开销。Easysearch 通过一系列内核级优化，让数据"存得更省、查得更快"。

## 存储优化技术

**01 Rollup 汇总**
历史数据降采样，保留关键聚合指标，大幅降低存储占用

**02 冷热分离架构**
热数据 SSD 高吞吐，冷数据低成本存储 / 归档，按需分层

**03 快照搜索**
对象存储直接挂载快照索引，归档数据即时可查

**04 ZSTD 压缩**
高压缩率压缩算法，兼顾存储成本与查询吞吐

**05 时序合并策略**
基于时间范围的合并，优化时序数据索引与查询效率

降本增效

### 存储与吞吐的双重优化

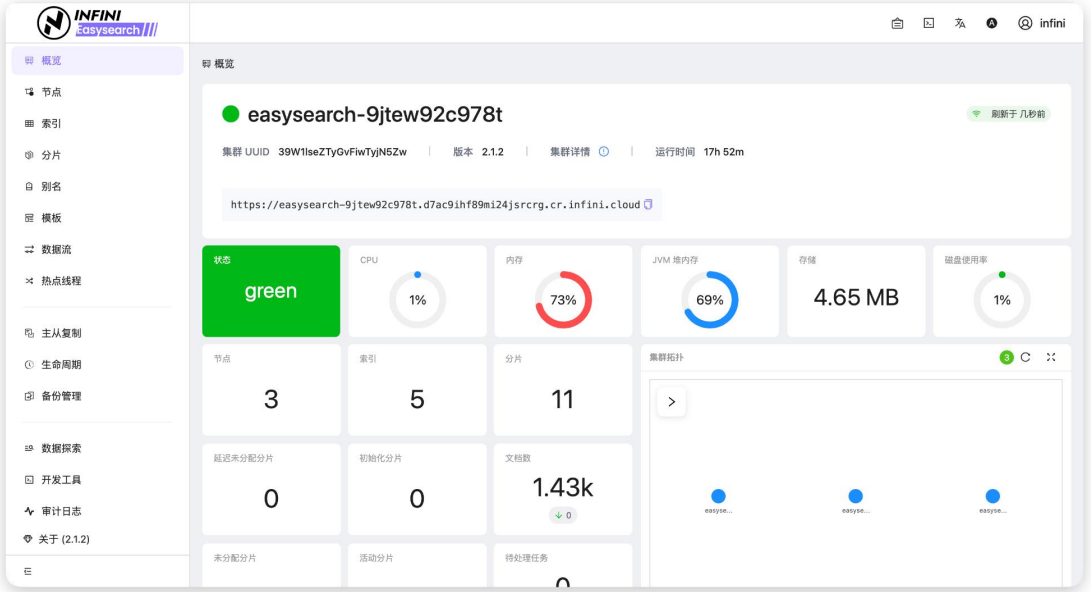
| 84% | 5× |
| --- | --- |
| 压缩率 8.7GB → 1.4GB | ZSTD 查询吞吐提升 |

数据来源：INFINI Labs 官方基准测试 —— 8.7GB Nginx 日志综合压缩（ZSTD + source_reuse）；吞吐为 Easysearch(ZSTD) 对比 Elasticsearch 7.10.2(best_compression) 的对等环境实测。

## 价值体现

- 降低存储成本　冷数据归档 + 高压缩率
- 提升查询性能　Rollup 降采样，秒级聚合
- 支撑海量时序　日志 / 监控数据长期保留

# 部署向导 + 可视化管理后台




### 1 Easysearch 安装部署向导

一键拉起集群，自动检测环境、推荐最优参数。支持单节点快速体验与多节点生产集群部署。信创环境（统信 UOS、麒麟 V10）一键部署。

### 2 Easysearch 内置管理界面

引擎自带 Web UI，无需额外安装。提供集群概览、索引管理、数据搜索、SQL 查询、安全配置等能力。随引擎运行，轻量零依赖。

# 多集群纳管 + K8s 私有搜索云

从单集群到多集群，从物理机到云原生——Easysearch 提供两种面向规模化的运维形态。

## INFINI Console · 多集群统一纳管

同时纳管多套 Easysearch / Elasticsearch 集群：跨版本集中监控、统一告警、统一权限、统一运维。

50+ 节点 | 跨版本兼容 | 统一告警 | 统一权限

典型场景 → 平安银行　200+ 集群统一管理

## INFINI Cloud · K8s 私有搜索云

**免基础设施运维**
一键创建集群

**云原生弹性扩展**
按需扩缩容

**全链路性能监控**
内置告警

**多租户隔离**
权限可配

私有部署 · 数据不出企业内网

# 轻量压测方案 + 面向搜索的网关

围绕搜索引擎的使用场景，Easysearch 提供了两款配套工具：一个负责"测"，一个负责"管"，覆盖从性能验证到生产流量治理的完整链路。

INFINI Loadgen

## INFINI Loadgen · 轻量级搜索压测解决方案

专为搜索场景设计的压测工具，开源免费，无需复杂配置即可模拟真实查询负载。

**核心能力**

- 模拟真实业务数据与查询模式
- 模块参数随机化，贴近生产场景
- 详细性能报告：吞吐 / 延迟 / 错误率
- 轻量易用，完美替代 Rally

迁移场景价值：迁移前压测对比 ES 与 Easysearch，用数据验证性能，让决策有据可依。

INFINI Gateway

## INFINI Gateway · 面向搜索场景的智能网关

不同于通用 API 网关，专为搜索流量设计，是连接应用与搜索引擎之间的"智能管道"。

**核心能力**

- 流量管理：路由、限流限速、熔断降级
- 查询加速：缓存、请求合并、结果预取
- 安全审计：请求日志、行为追踪、异常检测
- 无缝迁移：双写双读、灰度切流、安全回退

生产场景价值：扛住日增百 TB 流量（好未来），查询延迟从 100–1000ms 降至 0.3–5m。

# Coco AI 为现代团队打造的统一搜索与 AI Agent 中心

Connect + Collaborate

https://coco.rs

### 打通异构数据，跨平台统一搜索

Coco AI 统一连接飞书、语雀、Wiki、Jira、GitLab、MySQL 等企业应用，实现跨平台文档、对话与文件的一站式搜索。

### 智能数据消费，无缝 AI 能力访问

基于语义的 AI 搜索与生成式对话，基于 LLM Wiki 设计理念，通过智能体帮助团队快速获取内部知识，打造企业专属 AI 工具箱，桌面端一键调用，提升工作效率。

### 安全第一，隐私优先

支持自托管，数据不出环境；企业级权限控制与数据脱敏，满足严苛的安全合规需求。

### 开源开放，扩展性强

支持连接器、Webhook、MCP、Skills、Deep Research 等多维扩展，提供 API 对接与 Widget 集成，在线市场一键安装。

# INFINI Pizza · 下一代实时搜索引擎

INFINI Pizza

RUST **基于 Rust 的新一代新一代分布式纯实时搜索引擎**

INFINI Pizza 是用 Rust 从零构建的搜索引擎，面向实时检索与嵌入式搜索场景。Rust 的内存安全与零成本抽象，让它天然具备高性能与高可靠的双重优势。

内存安全 | 无 GC 停顿 | 高性能 | 低写入延迟 | 插件化架构 | 嵌入式友好

### 为什么是 Rust?

- 内存安全：编译期杜绝悬垂指针、越界访问
- 无 GC 停顿：避免 JVM 卡顿，延迟更稳定
- 极致性能：接近 C 的运行时开销

### 应用场景

- 静态站点嵌入式搜索（Hugo）
- 边缘计算场景的轻量检索
- 低延迟实时索引写入

技术定位

01 **面向未来**
Easysearch 的下一代内核技术储备

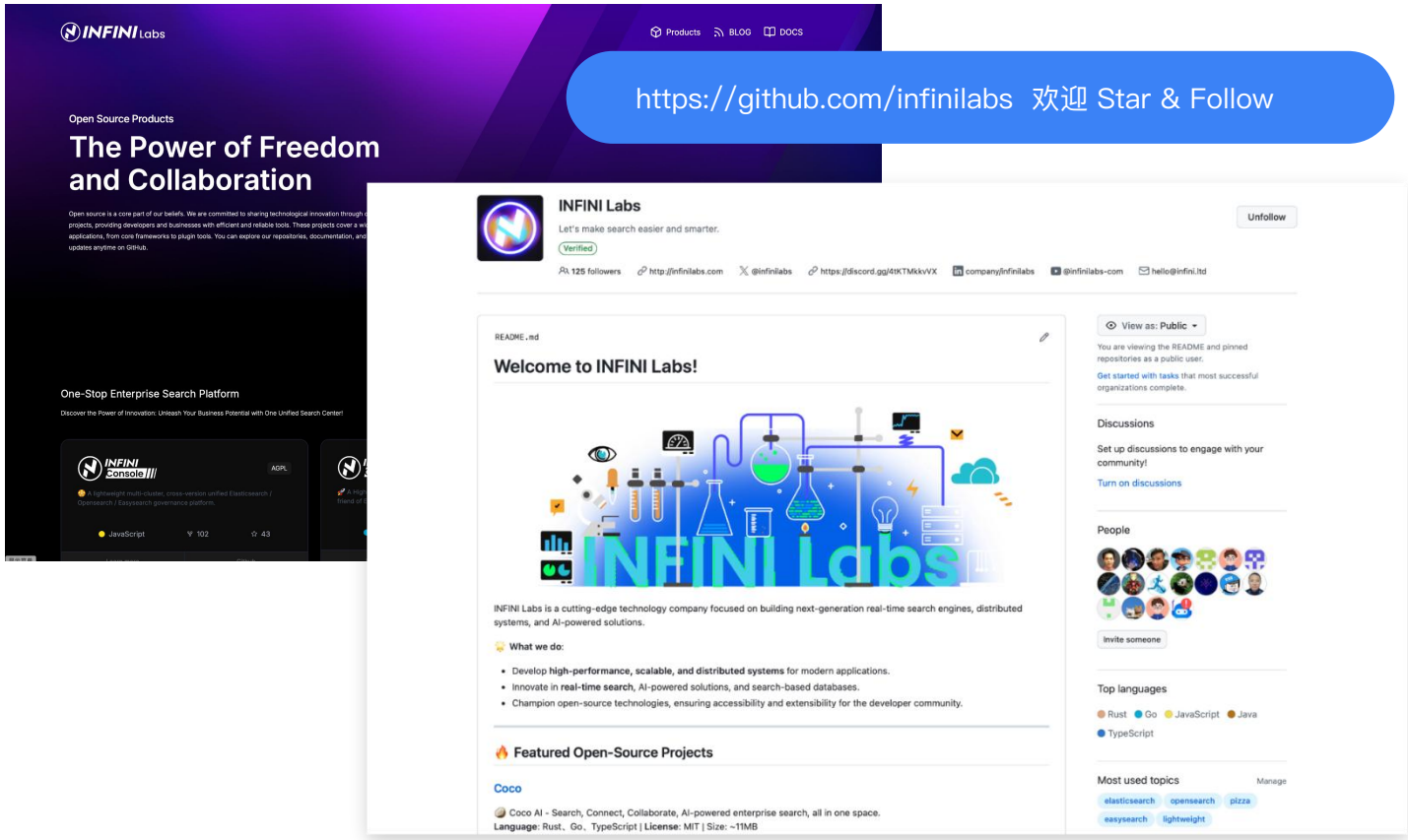
02 **双内核并行**
Lucene 内核 + Rust 内核共同演进

03 **技术领先**
从"跟跑"走向"领跑"的关键一步

# 开源工具与生态

## 开源工具生态

| | |
|---|---|
| Analysis-IK 中文分词 | IK 分词 + 自定义词典 |
| Analysis-STConvert | 简体 / 繁体字符转换 |
| Analysis-Pinyin | 中文转拼音检索 |
| INFINI Gateway | 为搜索场景设计的高性能网关 |
| INFINI Console | 一站式的数据搜索分析与管理平台 |
| INFINI Loadgen | API 压测 / 性能基准 |
| INFINI Agent | 多实例监控采集 |
| INFINI Framework | 通用开发框架 |
| INFINI Pizza | 新一代端侧及分布式实时搜索引擎 |



06

# 成功案例

## 从金融到制造，从通信到政务

Easysearch 已在大型汽车制造集团、头部云服务商、保险集团、银行、券商等关键行业落地。

# 中国一汽集团

大型汽车制造集团 · 搜索基础设施国产化升级

2022 年 9 月，国资委发布 79 号文，针对 98 家央企提出信创要求，标志行业信创全面开展。
2023 年开始进行 ES 国产化替代工作。

| | |
|---|---|
| 几十个<br>业务场景 | 700+<br>节点 |
| 15 TB<br>日均写入 | 3年+<br>稳定运行 |

**业务属性划分**

| | |
|---|---|
| IT基础应用 | 24 |
| 大营销 | 109 |
| 大财务 | 7 |
| 大数据 | 12 |
| 大研发 | 5 |
| 分子公司 | 7 |
| 管理与支持 | 6 |

**各领域超大系统**

| | | |
|---|---|---|
| IT基础应用 | 统一监控平台 | API网关 |
| 大营销 | 智能网联云平台2.0 | 红旗营销管理系统 |
| 大财务 | 企业业务中台（EIP） | 财务共享中心 |
| 大数据 | CDP平台 | 奔腾UBI |
| 大研发 | 一汽云仿真平台 | D-life6.0 |
| 分子公司 | 旗享智投平台 | 资产交易电商平台 |
| 管理与支持 | 协同办公系统 | 管理者工作台 |

# 移动云

国内领先云服务商 · Elasticsearch 国产化替代与云原生升级

在公有云上对外提供托管 ES 服务，涉及多个使用区、机房。
早期使用 Elastic 产品支撑该业务，目前以完成国产化改造，使用 Easysearch 提供服务。
该项目获得"2023年中国信通院大数据星河标杆案例"。

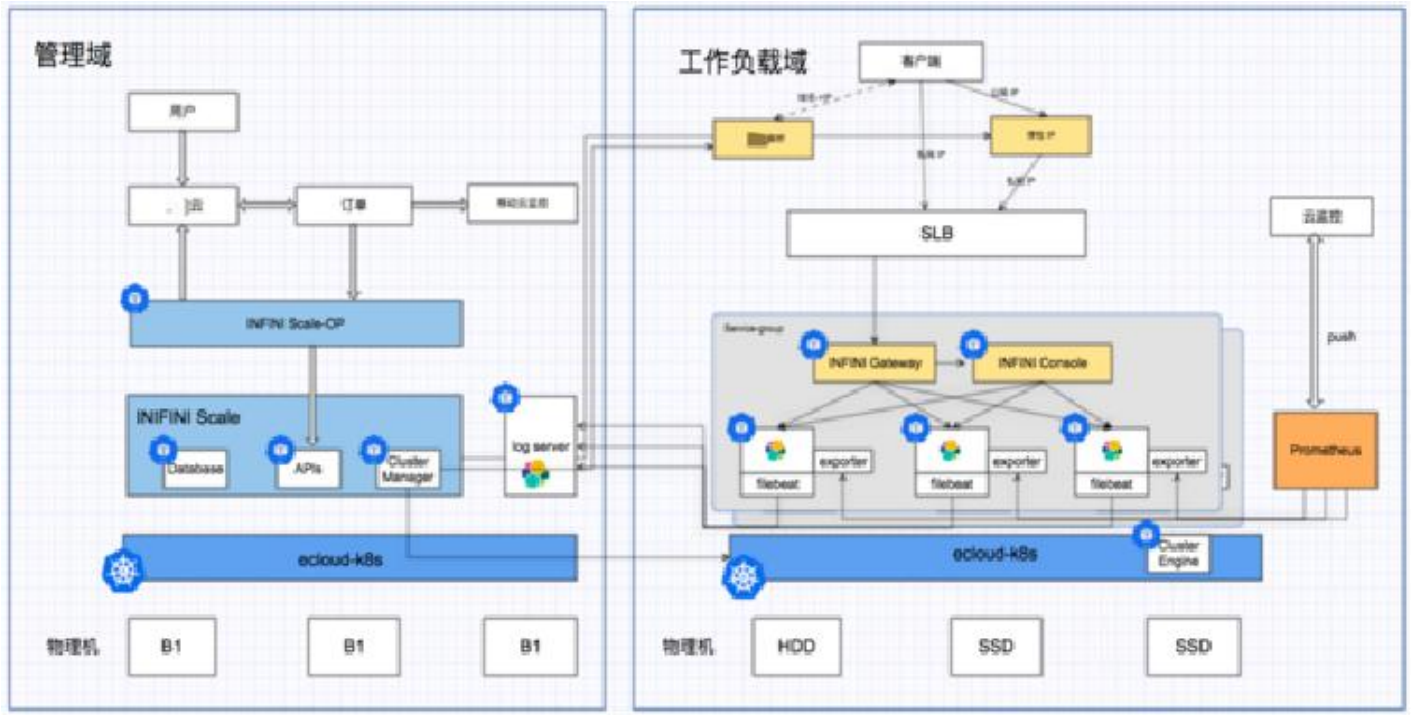
**13 个**
可用区

**3000+**
节点

**PB级**
最大集群

**3年+**
稳定运行

## 2023年度中国信通院大数据"星河"标杆案例



07

# 总结与展望

## 从国产替代到自主创新

搜索引擎国产化不是终点，而是起点。从替代走向自主创新，从工具走向 AI Infra，每一步都是面向未来的战略选择。

# 从国产替代到自主创新：三阶段演进

已完成

## 01 国产替代

ES 兼容 → 功能补齐 → 业务连续迁移，让企业"敢替"。

- API / DSL 全兼容
- 零改造业务迁移
- 信创全栈适配

进行中

## 02 企业级增强

内核优化 → 性能突破 → 安全内建，让企业"好用"。

- Rules / ZSTD / IK 性能突破
- 国密 + 备份容灾
- Coco AI 智能助手

未来

## 03 AI Search Infra

RAG 底座 → Agent 支撑 → 多模态 → Rust 内核，让企业"领先"。

- RAG / Agent / 知识库
- 多模态搜索
- INFINI Pizza Rust 内核

国产化替代 ≠ 简单换芯 · 而是从替代到创新的战略升级

# 核心主张与行动建议

## 搜索引擎的国产化，不是换个引擎，
## 而是建个底座。

主张 1

### 把数据主权拿回手里

核心索引数据不再依赖国外基础软件，搜索能力自主可控

主张 2

### 把技术路线握在手中

性能优化、内核演进、AI 集成完全可控，可按业务节奏演进

主张 3

### 为 AI 时代打底座

搜索是 RAG、Agent、知识库的核心基础设施，先打底座再谈 AI

行动建议

1 早评估

2 分步走

3 借 AI 之势

INFINI Labs
极限科技 · 让搜索更简单

# 感谢聆听


W 官网 easysearch.cn

@ 邮箱 hello@infini.ltd

T 服务热线 400-139-9200


极限科技 INFINI Labs · 让企业数据被快速搜索，直接服务于 AI 与业务决策