KUBERNETES COMMUNITY DAYS BEIJING 2025

构建基于企业数据的高精度生成式人工智能应用

郑予彬
开发者布道师
亚马逊云科技

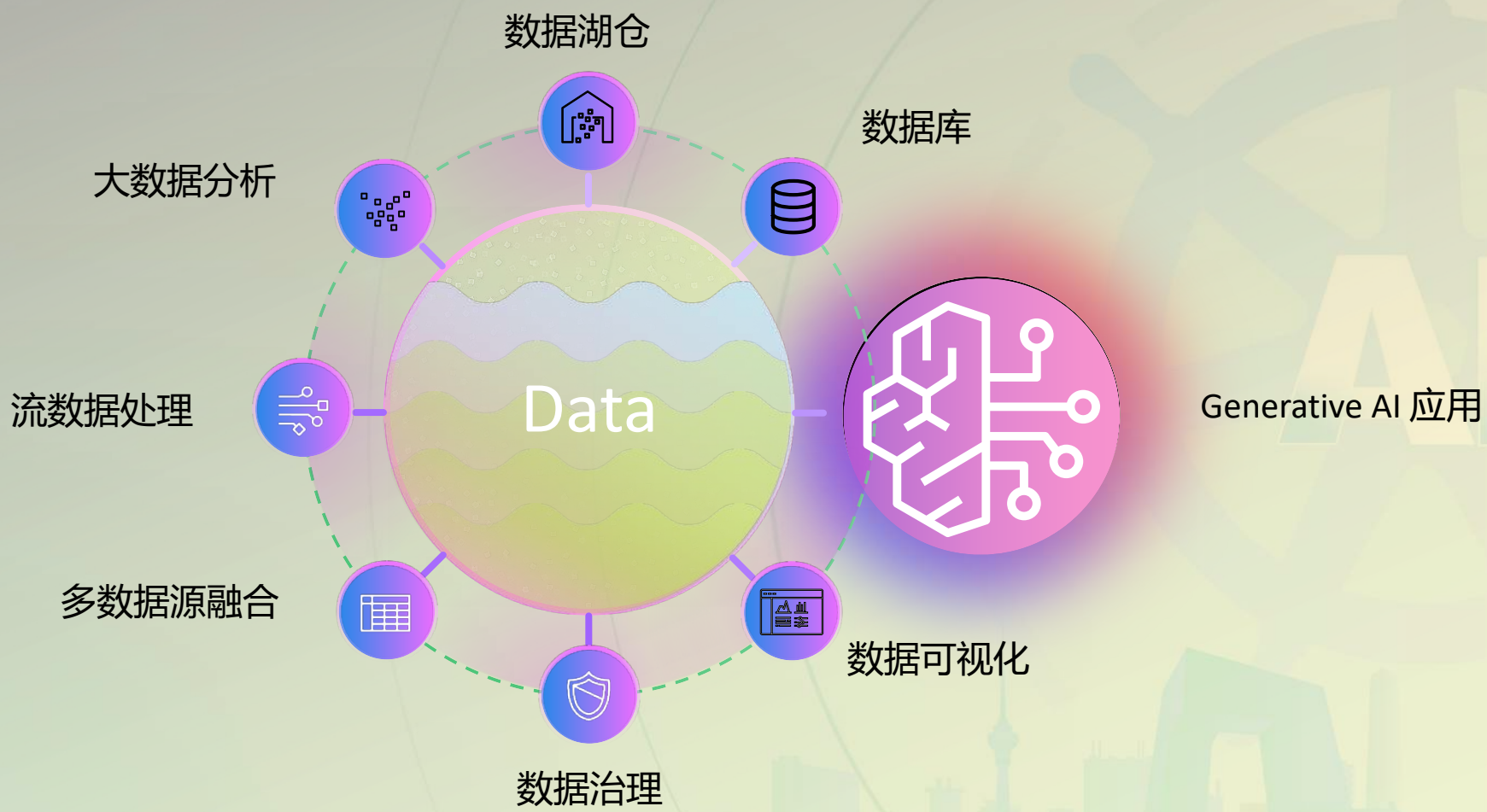
Content

目录

Part 01

生成式AI在企业中的转化潜力

GenAI 应用植根于数据平台



企业迫切需要转变为构建数据驱动型公司

数据驱动型公司可实现**每年30%+的速度增长**

Forrester, “Insights-Driven Businesses Set the Pace for Global Growth,” https://bit.ly/3r4uRiL

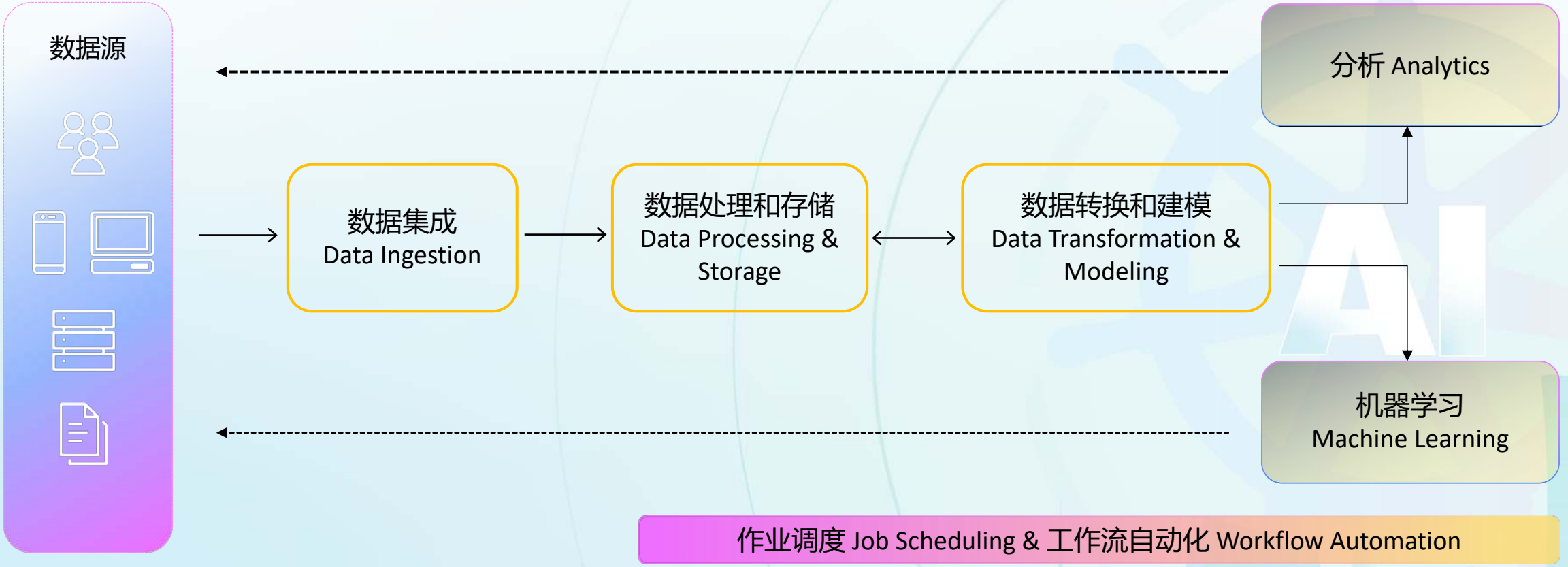
68%的企业表示仍然无法从数据中获取高价值

Accenture, “Closing the Data-value Gap,” https://accntu.re/33V6sU3

只有 **28%的企业反馈**已建立数据文化

Accenture, “Closing the Data-value Gap,” https://accntu.re/33V6sU3

构建数据平台是实现数据驱动型公司的关键



Part 02

RAG架构-连接企业数据与大语言模型

企业的数据是关键的差异化因素

企业数据

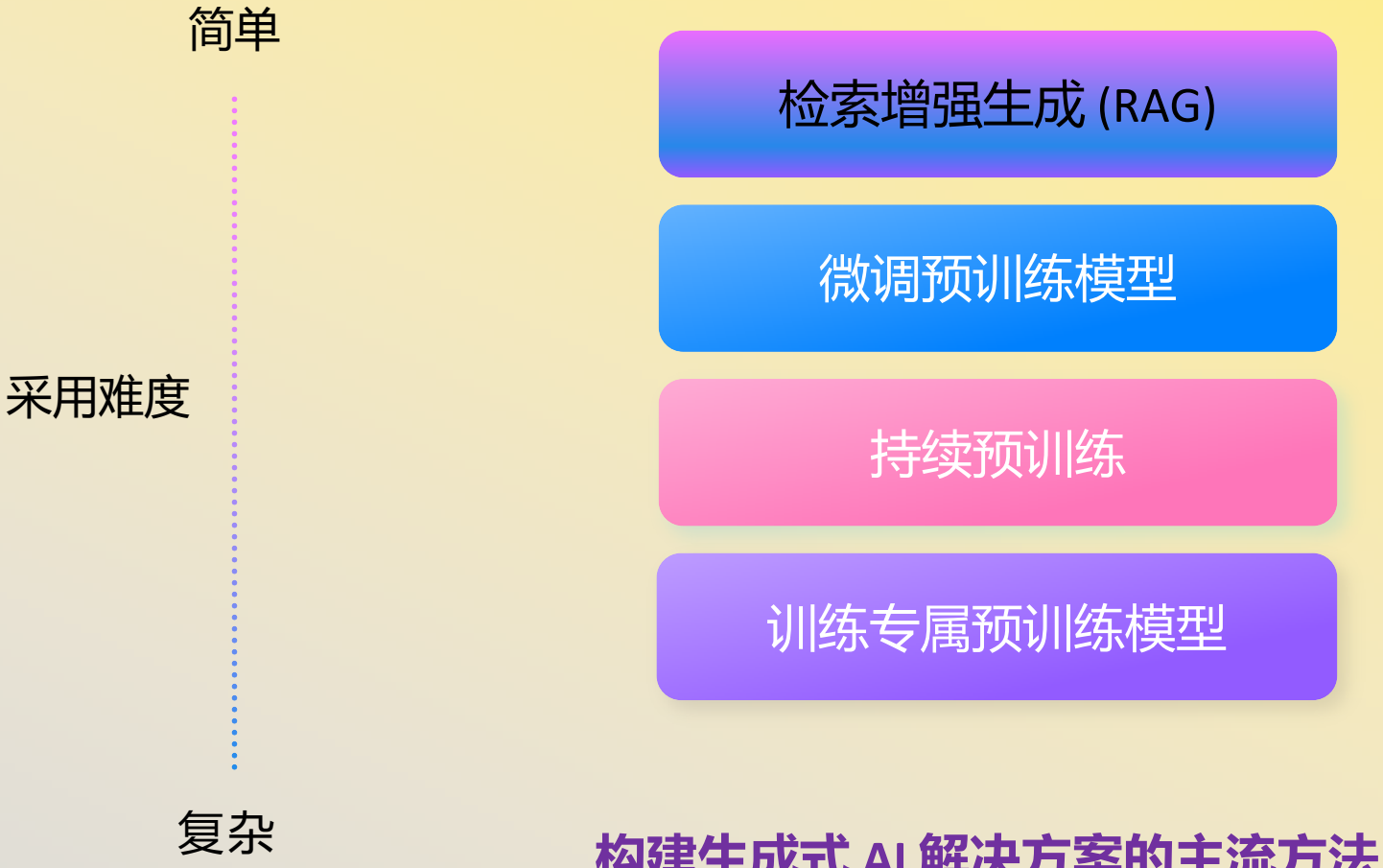
检索增强生成 (RAG)

微调预训练模型

持续预训练

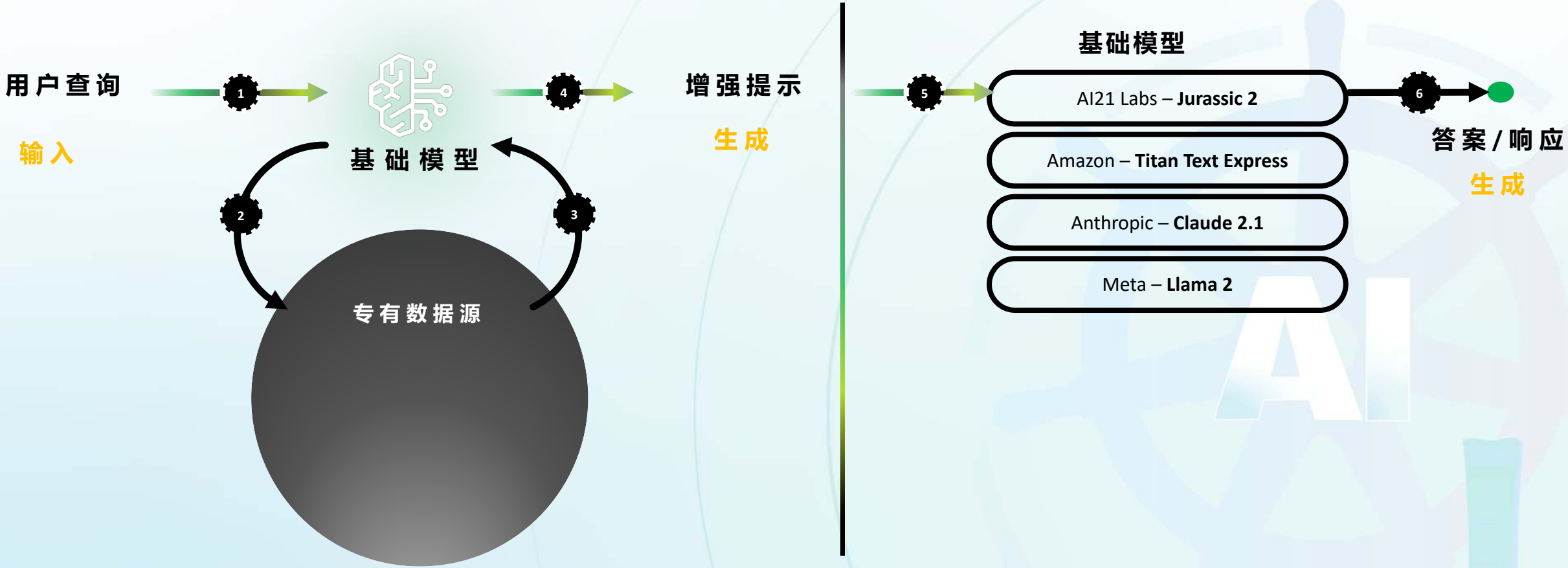
训练专属预训练模型

企业的数据是关键的差异化因素



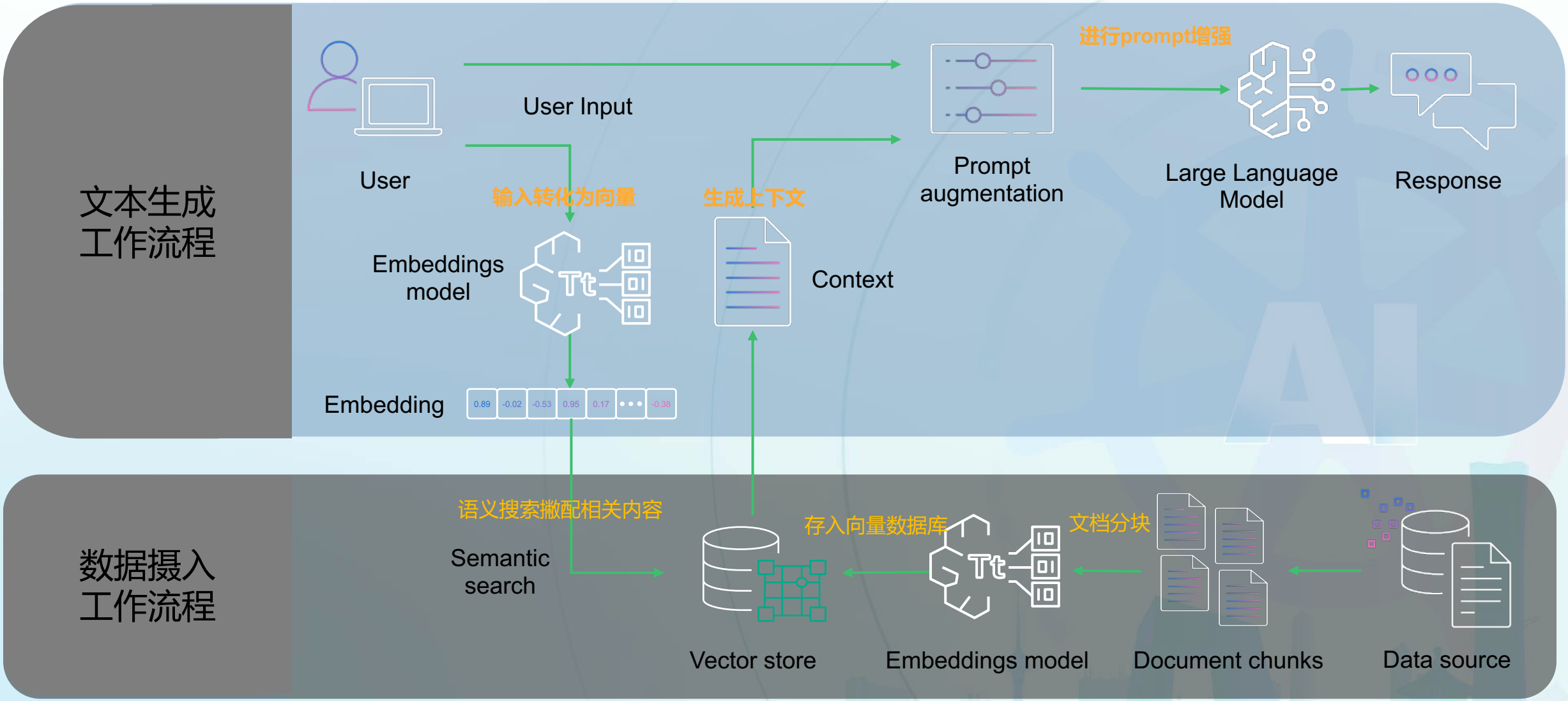
构建生成式 AI 解决方案的主流方法

Retrieval augmented generation(RAG)



RAG的主要优势在于它能够将外部知识库的信息与语言模型结合，从而生成更准确、更相关的回答。这种方法特别适合需要访问专有数据或最新信息的应用场景

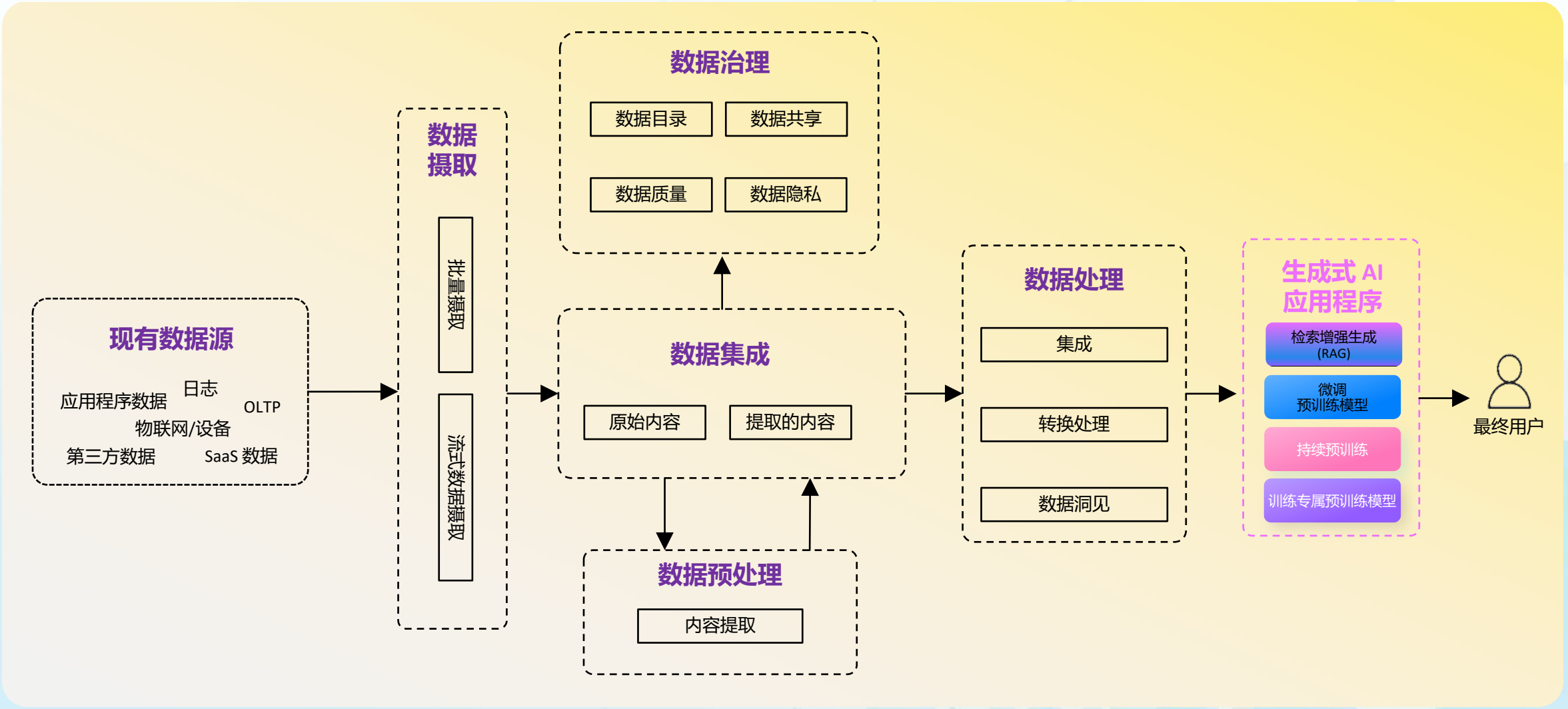
RAG 工作流程



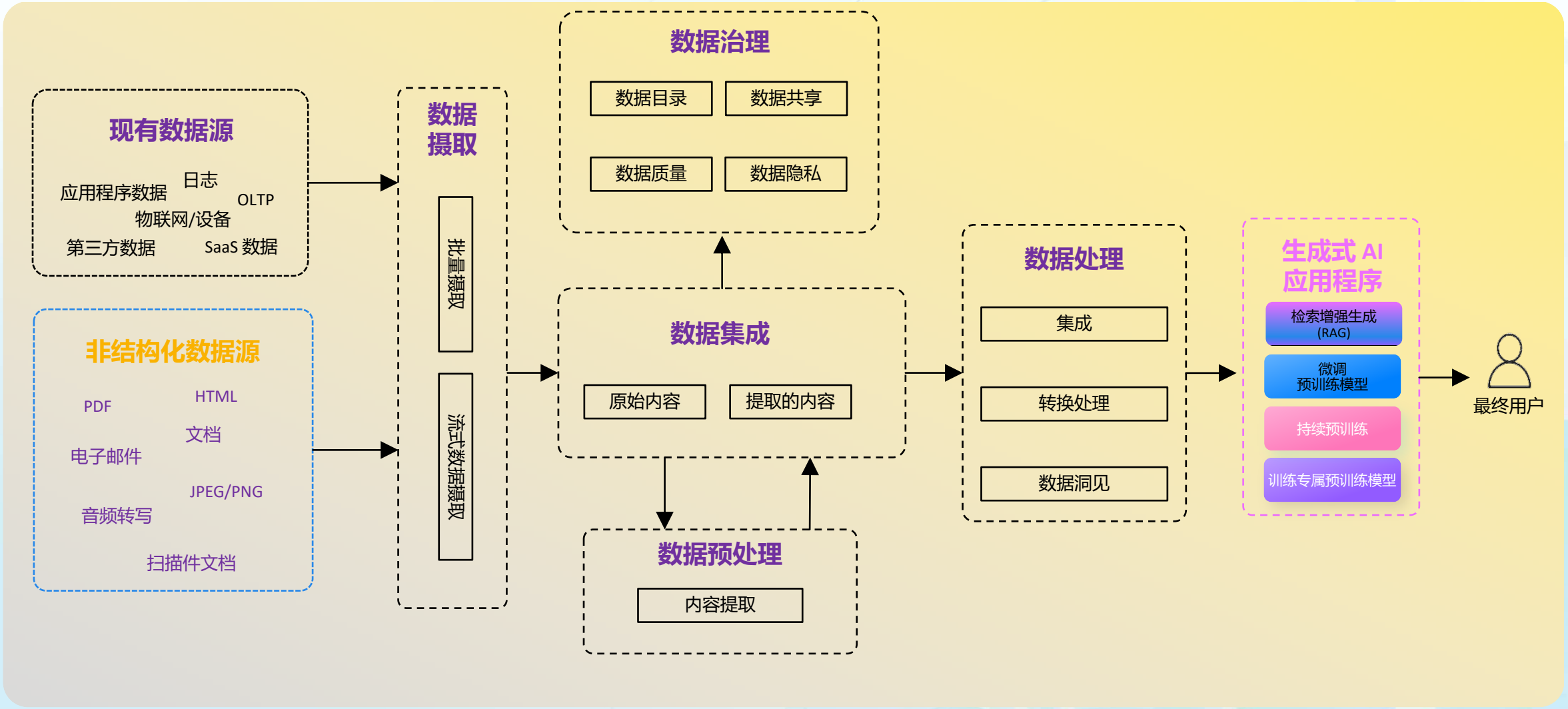
结合外部知识，提高回答准确性和相关性，支持实时数据更新，处理私有或定制数据

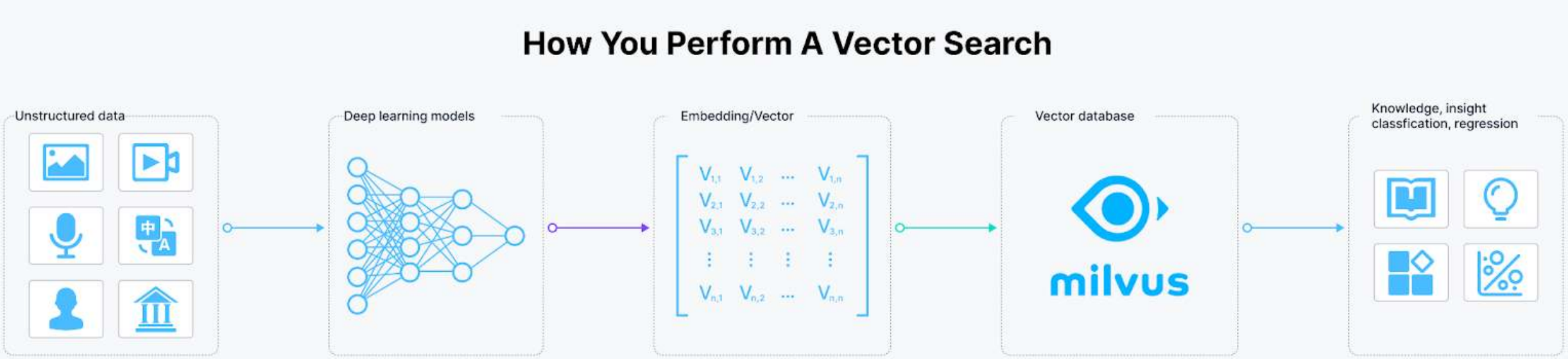
数据基础

为生成式 AI 应用程序扩展



为生成式 AI 应用程序扩展





https://zilliz.com.cn/blog/how-to-deploy-open-source-milvus-vector-database-on-amazon-eks.md

Part 03

如何优化云上Gen AI 工作负载

越来越多客户选择基于容器构建数据平台

现代化的容器平台可以高效地管理调度计算资源，有效支撑数据业务对敏捷性和可靠性的严苛要求。

- **客户容器化改造后获得的收益**

成本节省

01100
10110
11110 **23% 计算资源利用率提升**

业务敏捷

28% 营收增长

45% 更多的应用部署

36% 开发速度增加

可靠运维

40% 宕机时间减少

13% 提升新业务部署成功率

生产力提升

80% 增加运维效率# VMs per admin

72%

of surveyed Enterprises adopted Containers

*The goal is to put more and more on containerized platforms. It really gives us the **ability to scale**.*

IT Director, Wholesale Trade

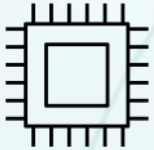
Prepared by

Known

成熟度：Data on Kubernetes 是如何走到这一步的?

Stateful-sets
管理有状态应用

GPU支持



2015 2017 2019 2021 2023

Custom Resource Definition
实现可扩展的 k8s 资源管理

EKS 提供托管的 Kubernetes 服务

Amazon **EKS**

- EKS 与上游 Kubernetes 100%兼容，并积极参与代码贡献
- EKS 紧跟社区发布新版本的 Kubernetes，并且保留足够的时间让客户测试和升级版本
- EKS 提供高性能，可靠和安全的托管Kubernetes服务
- EKS 使 Kubernetes 的操作和管理变得简单，并且与其他亚马逊云科技托管服务深度集成

Amazon EKS 帮助客户更专注于构建可靠、稳定和安全的现代化应用

集成开源方案 FOR ML & DATA-SPECIFIC ORCHESTRATION

可观测性　工作流调度　持续集成和发布　批处理调度

自动化调度 AUTOMATE COST-OPTIMIZED ORCHESTRATION ON EKS

多租户　极致弹性　高效调度　GPU 管理

Karpenter

丰富灵活的选择FLEXIBLE INFRASTRUCTURE CHOICES

Trainium　Inferentia　GPUs

EC2 instances (Spot/ On-Demand/Reserved)	Trn1(n)	Inf2	Inf1	P5	P4de	P4d	P3	G5	G5g	G4dn	G4ad	DL1	VT1	F1

基于 EKS 构建数据和机器学习平台需要考虑的问题

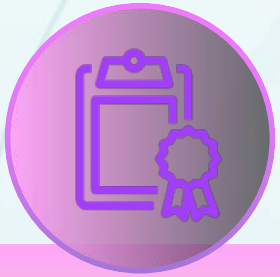
- 大规模扩展到上千节点
- 网络配置
- 高可用设计
- 多租户和安全管理
- 批处理调度
- 选择合适的计算和存储资源
- 日志和监控
- 集群运维

Data & AI on EKS (DoEKS) 项目

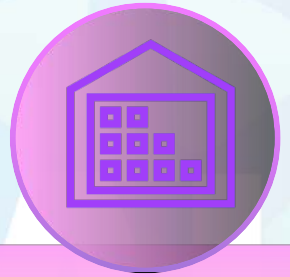
通过开源方式，帮助客户在 EKS 上更好地构建数据平台和 GenAI 应用

基础设施即代码
Infrastructure as Code (IaC) templates

性能指标参考报告
Performance benchmark reports



基于最佳实践
Amazon best practices for data workloads (e.g. Spark, Kafka, Ray)

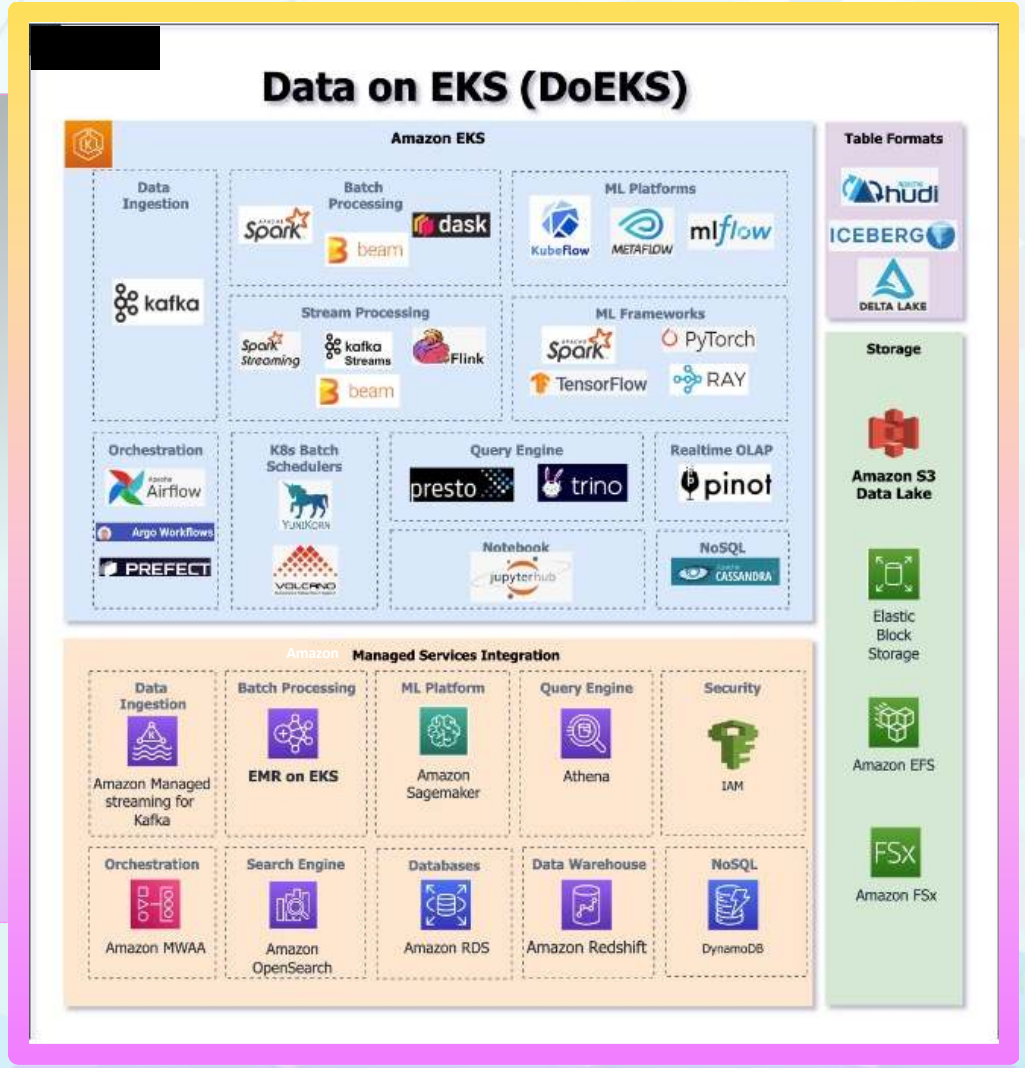


丰富的部署架构模板选择
Deployment examples and architectures

Visit Data on EKS portal to learn more:

Data on EKS 涉及的应用

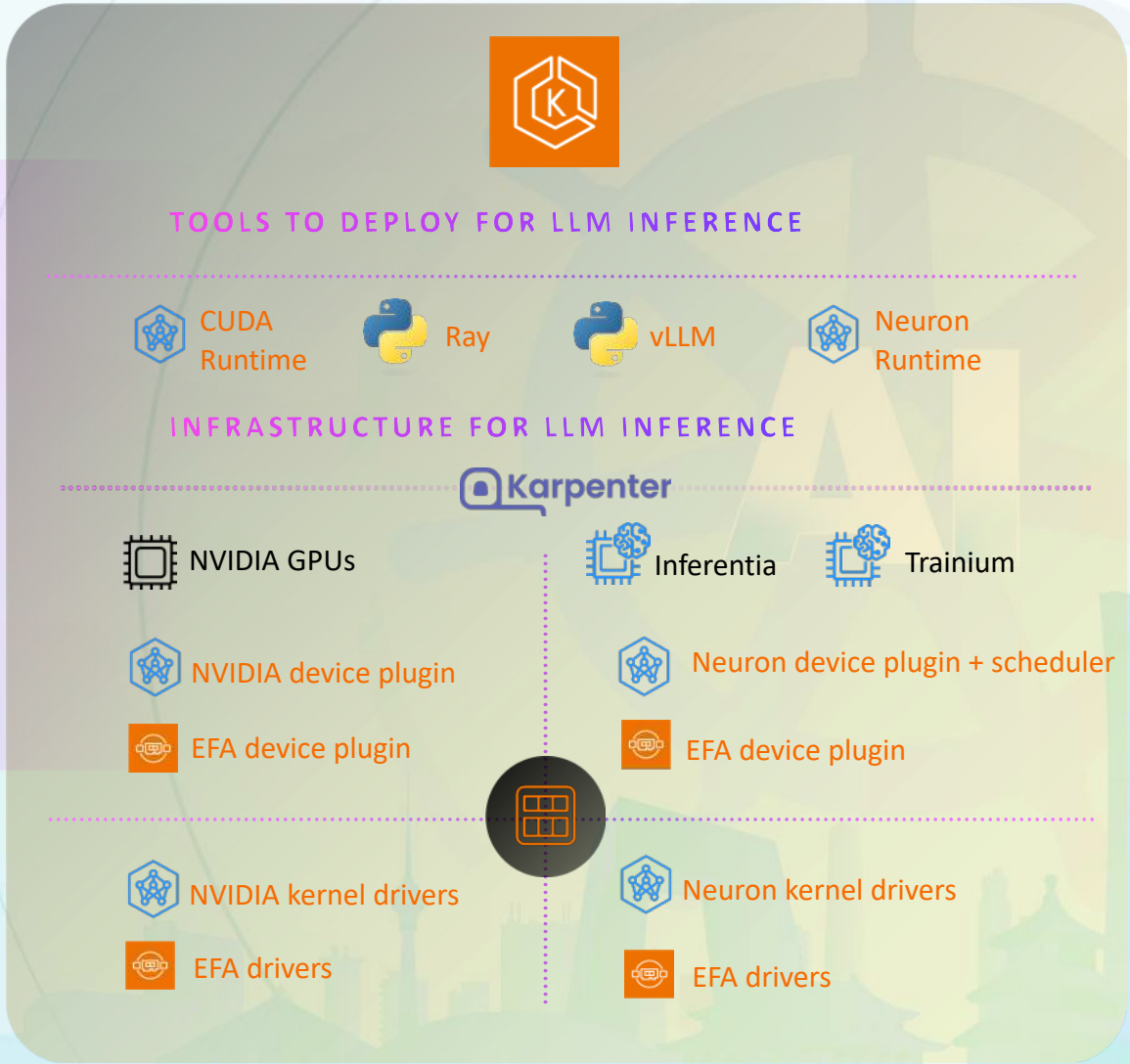
- 分布式数据库
- 数据分析
- 流数据
- 分布式查询
- 作业调度
- AI 与机器学习



DoEKS: 机器学习与生成式 AI

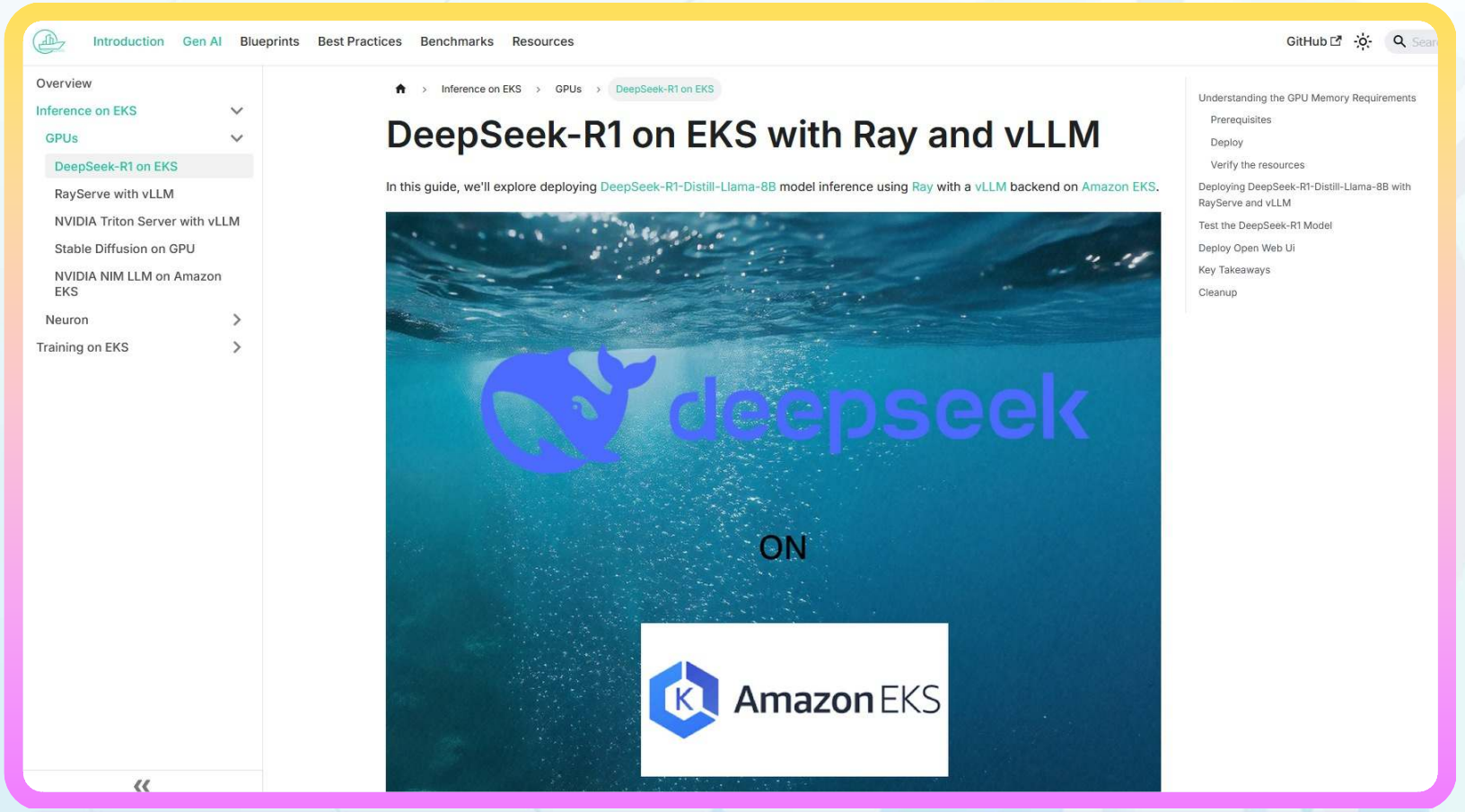
基于 EKS 部署模型推理

- 极致的 GPU 利用率和弹性
- 通过使用 Spot 实例降低成本
- 高效开发部署
- 基于开源生态，便于灵活扩展和个性化定制



一键部署 DeepSeek-R1 on EKS with Ray and vLLM

- 开源
- 简化部署



https://awslabs.github.io/data-on-eks/docs/gen-ai/inference/GPUs/ray-vllm-deepseek

Hugging Face

Hugging Face Hub - The leading open platform for AI builders

客户挑战

- Serverless inference for 超过一百万个模型
- 为每个模型提供专属推理endpoints
- 700k+ spaces (AI demo apps)
- 免费增值模式 Freemium pricing

解决方案

✓ 超过 20 个 EKS 集群, 超过 **2 千个节点** (CPU, GPU, Inf2)

✓ **Bin-packing** 基于实际对 GPU 的需求

✓ **Space Free tier**: 在1个 EKS 集群内, 使用 150 个节点实现 3 万个AI 应用 demo 空间

✓ **Time-shared GPUs**, 需要**间隔数秒**就加载模型

DoEKS AI/ML Blueprints

Ray on EKS

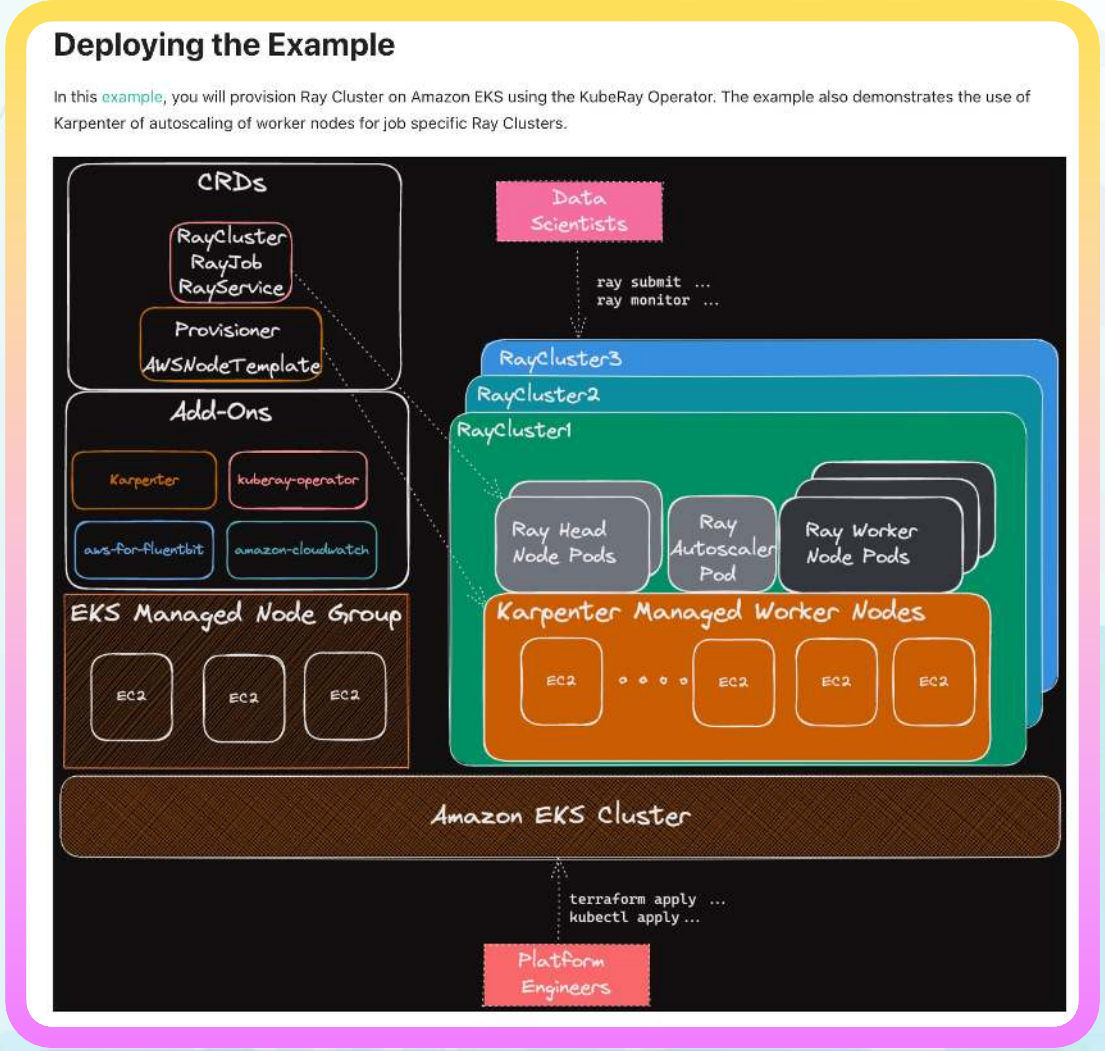
Trainium

Jupyterhub on EKS

Inferentia

Stable Diffusion on Inferentia
Stable Diffusion on GPU

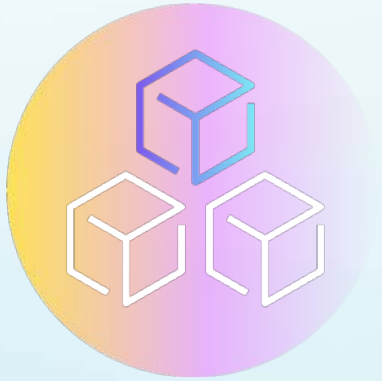
NVIDIA Triton with vLLM
NVIDIA NIM LLM on EKS



DoEKS: Data Processing (Apache Spark)

在 Kubernetes 上运行 Spark

敏捷 Agility



- 降低失败风险
- 更短的发布时间
- 多版本支持

成本 Cost-Efficiency

- 多租户
- 自动扩展
- 精细控制资源

在亚马逊云科技上运行 Spark 的多种方式

基于 EKS 自建 Spark

Flexible options for running open source Spark on EKS

Wide selection of open source integrations *

Portability and versioning *

适合希望在 EKS 上构建可跨平台一致的标准化数据平台，并且愿意维护开源组件以满足自身定制化需求的客户。

EMR on EKS

Low TCO and fast performance

Secure by default

Ease of use

适合希望在 EKS 上构建可跨平台一致的标准化数据平台，但在数据平台层面仍希望使用托管服务以简化管理的客户。

EMR Serverless

Automatic and fine-grained scaling

Resilience to Availability Zone failures *

Share applications with IAM roles *

适合想要避免管理和操作集群，而只想使用开源框架运行应用程序的客户

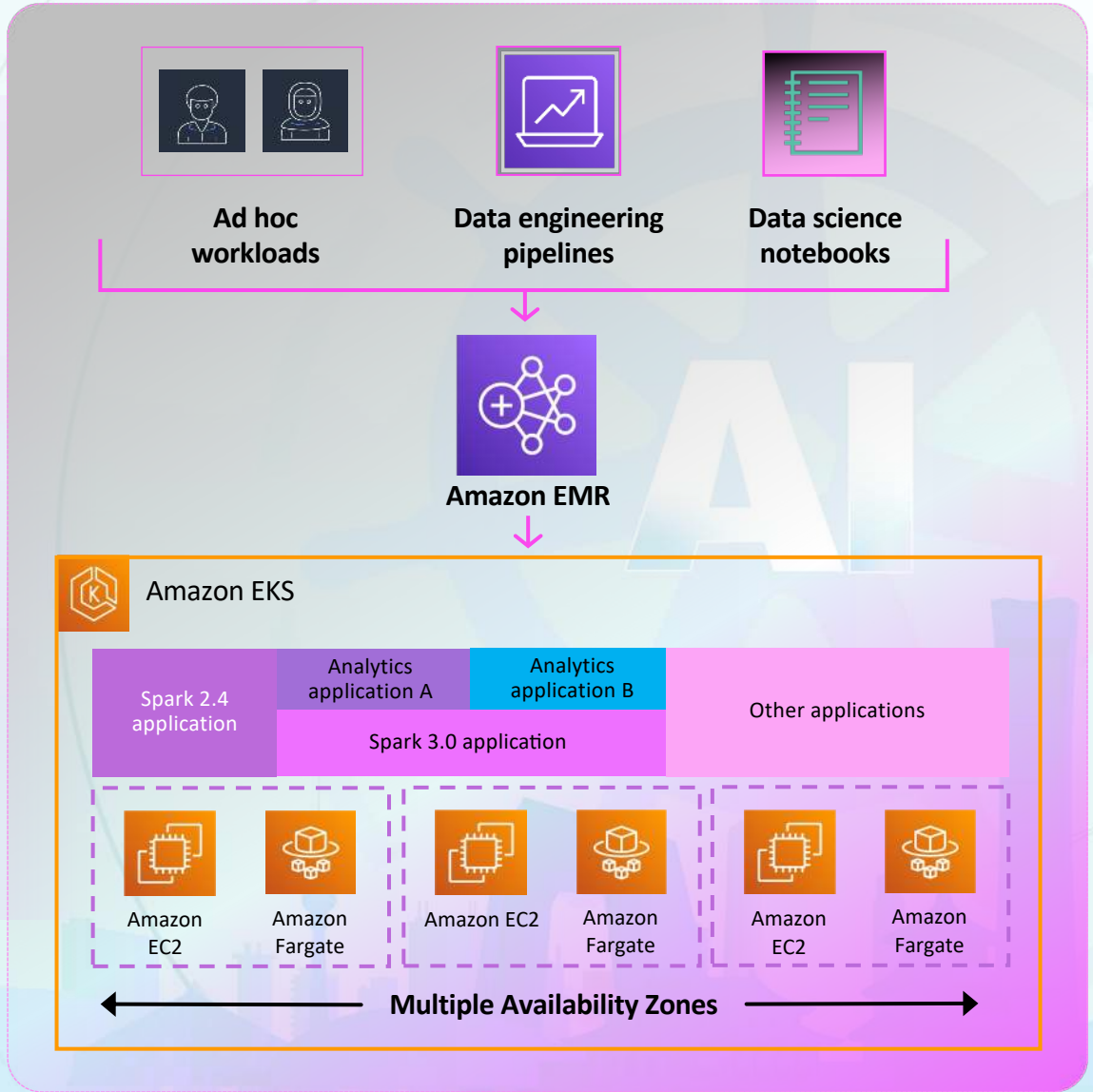
* also supported by EMR on EKS

Most Customer-Managed 客户倾向自建

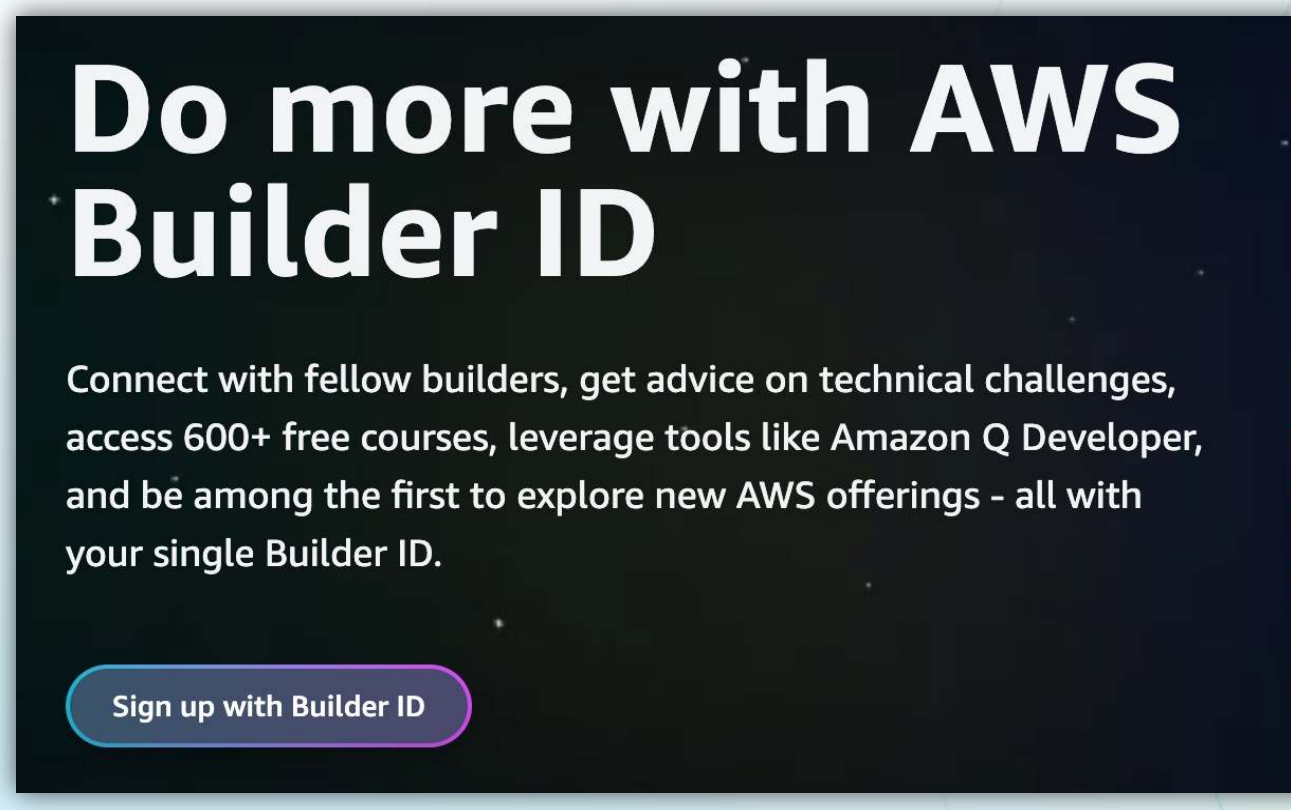
Most Amazon-managed 客户倾向使用托管服务

EMR on EKS

- 支持 Spark, Flink
- 提升资源利用率 Consolidate infrastructure across organization
- 多租户 Manage resource limits by teams and workload
- **性能优异** Spark runtime on EMR on EKS runs 5.4x faster + costs 4.3x less than open source Spark
- 高可用设计 Run application on single AZ or across multiple AZs
- 通过 Fargate 进一步简化资源管理 Choose serverless with Amazon Fargate on Amazon EKS



Sign up to Amazon Web Services Builder ID



Stay Connected with Amazon Web Services

AWS News Blog

Category: AWS re:Invent

Community.AWS

@awsdevelopers

@aws-developers

@awsdevelopers

collectives/aws

Thanks.