基于Ray的蚂蚁数据构建引擎在搜推、RAG场景的实践

谢 涛

蚂蚁AI数据构建引擎负责人

2025/12/20

蚂蚁开源 ANT OPEN SOURCE

RAY

谢涛

蚂蚁AI数据构建引擎负责人

个人简介：
2015年毕业于上海交通大学
先后供职于IBM、爱奇艺
2021年加入蚂蚁
深耕于分布式计算、特征计算和搜推
广索引数据处理领域

蚂蚁开源 ANT OPEN SOURCE

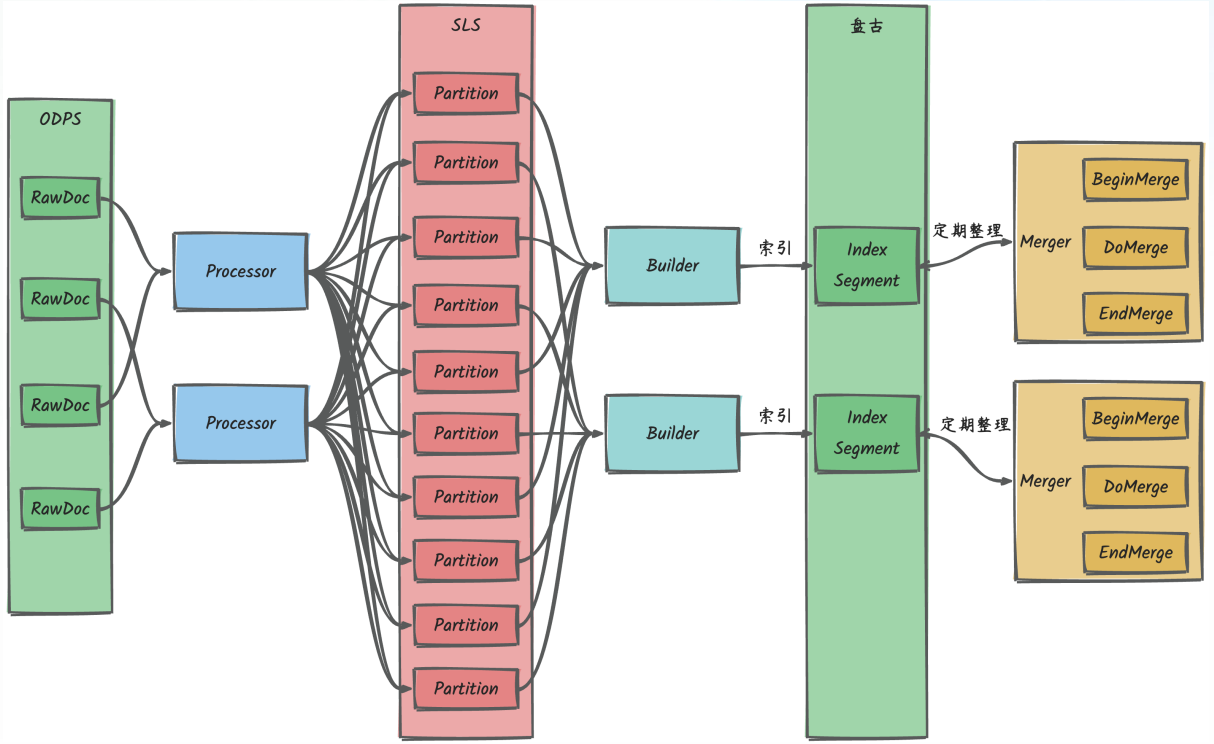
RAY

CONTENT
目录

基于Ray的海量数据构建提效

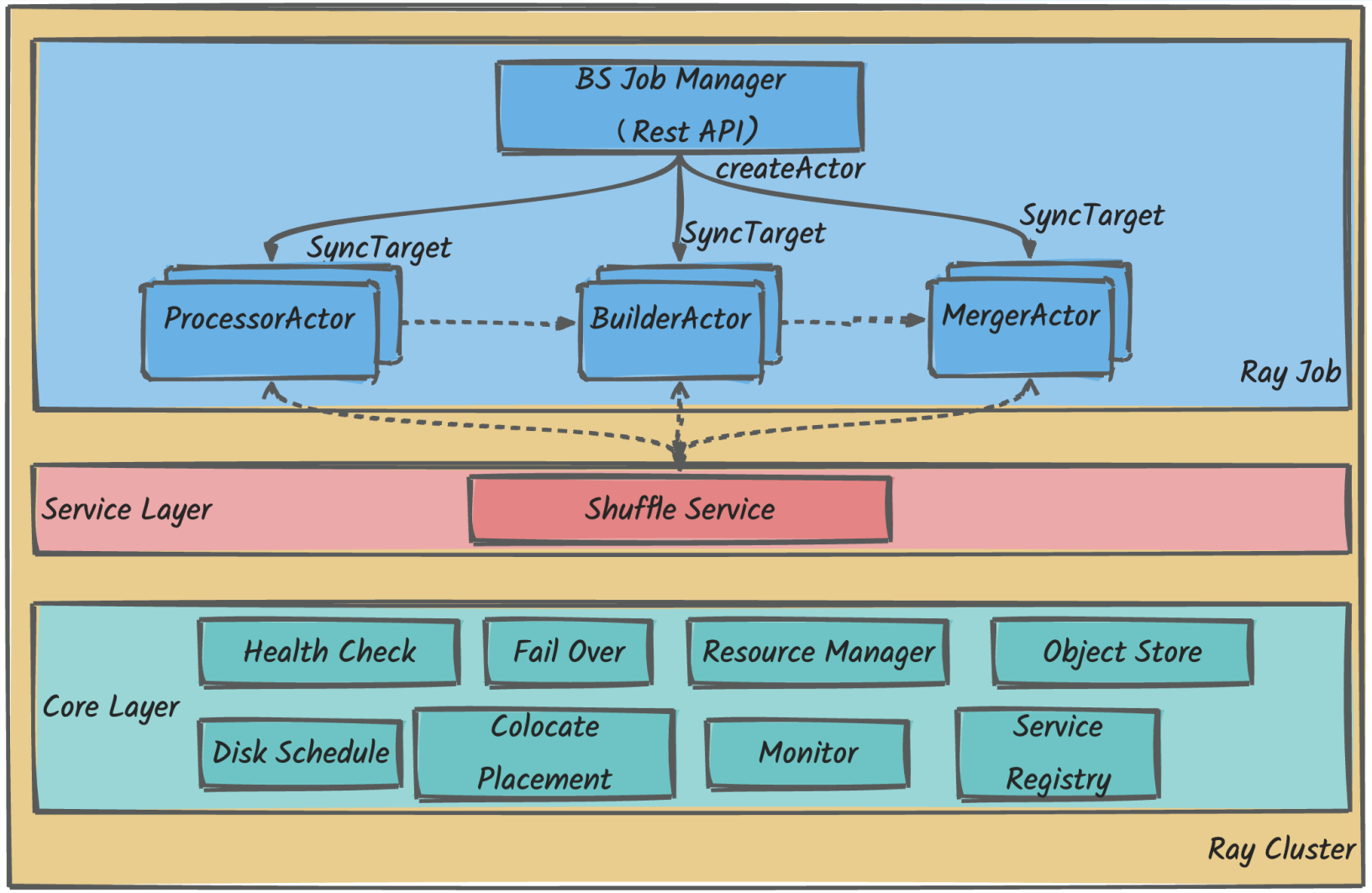
支撑全站万亿级正倒排，KV，KKV索引构建

问题：
1. 海量索引构建场景的资源弹性调度
2. 长尾场景的构建提效
3. 稳定性与成功率



Ray底座迁移

1. Ray C++ API
2. PBM worker Actor化
3. 集群与作业模式
4. HC，FO，资源调度基于Ray原生能力
5. Auto Scale



Ray迁移收益

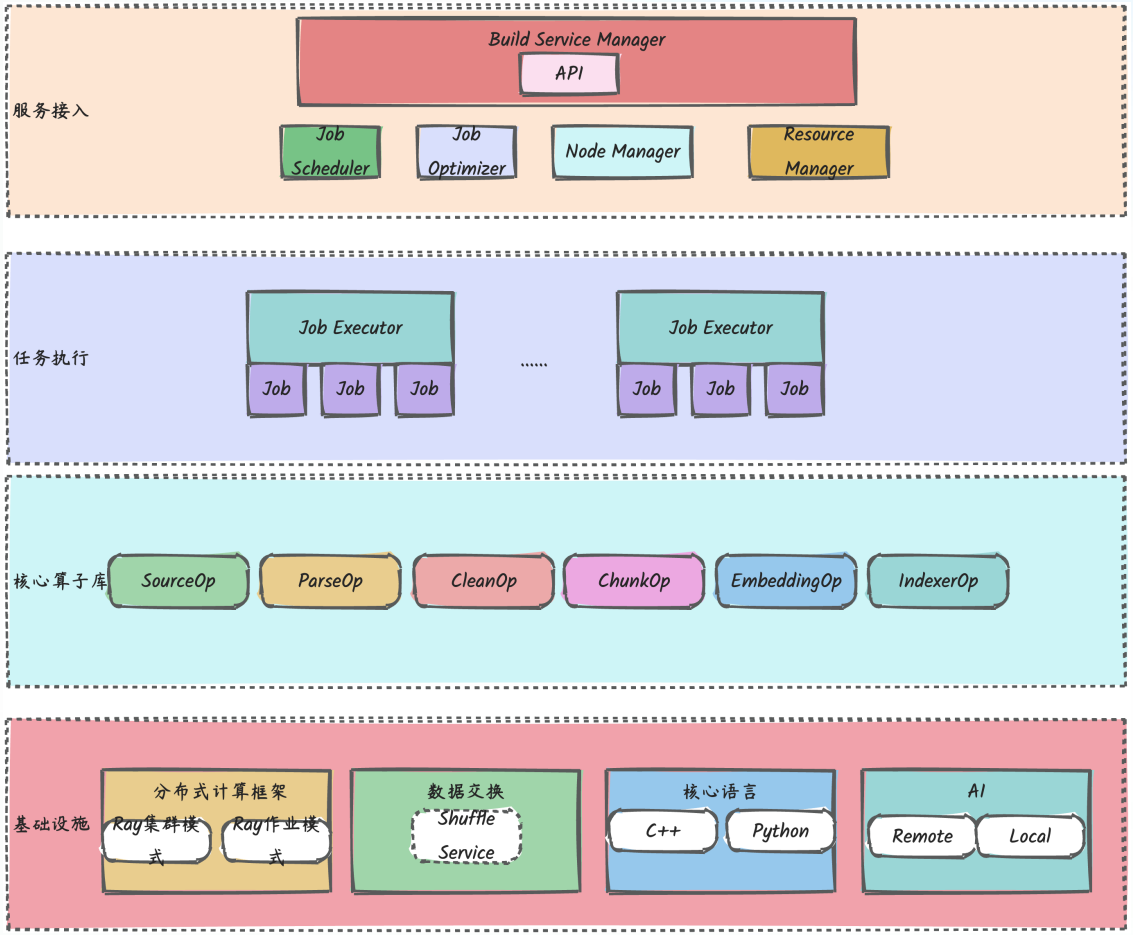
1. 容器规格从10+种减少至2种，资源弹性瓶颈消除
2. Actor轻量化改造，异构worker按需加载资源，调度overhead减少，小表构建耗时减少1倍+
3. 构建成功率和稳定性提升

基于Ray的RAG算子体系建设

算子市场：租户、计费、动态注册

算子编程范式：基于注解的算子执行与输入、输出约束

算子服务：基于Ray作业与集群模式，支撑不同场景的算子服务SLA保障



RAG算子服务- API

调用计量

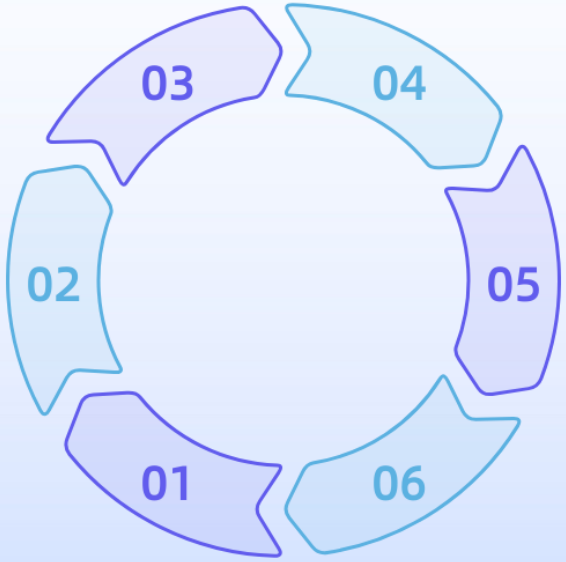
按API调用次数和资源消耗进行精确计量，为计费提供数据支持，提升透明度。

权限隔离

通过租户维度的权限控制，防止越权访问，保障系统资源的安全与独立。

OAuth鉴权

基于OAuth 2.0实现多租户身份验证，确保算子调用的安全性，隔离不同租户的权限。



灵活计费

支持多种计费模式，结合计量数据实现成本分摊，满足多租户商业需求。

限流熔断

集成限流与熔断机制，防止系统过载，保障高并发下的服务稳定性。

智能路由

通过智能路由算法优化请求分发路径，提高API响应效率与系统吞吐能力。

Rag 算子开发约束

01

节点定义

通过#node注解定义算子逻辑单元，明确节点类型及其职责，形成可识别的处理节点。

02

接口契约

声明输入输出接口规范，确保数据流动的兼容性，支持模块化集成与调用。

03

初始化资源

在#setup中完成依赖加载，支持异步模型初始化，建立外部服务连接。

04

异步加载

支持延迟或按需加载大型模型，提升启动效率，避免阻塞主流程执行。

05

执行逻辑

#execute实现核心处理逻辑，要求上下文隔离，保证各次执行互不干扰。

06

幂等性要求

多次执行结果一致，避免副作用，适用于重试与容错机制场景。

07

资源释放

#teardown负责清理内存与句柄，防止资源泄漏，确保环境干净退出。

08

生命周期管理

完整覆盖从初始化到销毁的全过程，保障组件稳定与系统可靠性。

Rag 算子执行层

Code Gen模块确保算子一次编写，即可运行在多个场景

1. Ray 作业模式：支撑日常海量非结构化数据处理场景
2. Ray 集群模式：支撑响应要求较高的异步场景
3. 在线服务模式：支撑响应要求极高的在线服务场景

下一步展望

建设高性能AI数据构建引擎

1. AI Native能力集成与高性能服务保障
2. Remote shuffle service集成 支持海量结构化与非结构化数据的统一处理
3. Embedding能力

蚂蚁开源 ANT OPEN SOURCE

RAY

诚招英才

欢迎加入蚂蚁智能引擎技术部

一起建设下一代AI Native数据引擎

长按识别二维码

欢迎投递