

# ES的垂直搜索平台架构思考

吴英昊 丰坚



# 垂直搜索平台架构

- **Elasticsearch**与垂直搜索平台
- 离线索引和实时索引
- **query**分析
- 打分和排序
- 其他问题
- 总结



# Elasticsearch和垂直搜索平台



# 垂直搜索

- 电商类搜索引擎
- 旅游类搜索引擎
- 文档类搜索引擎



# 垂直搜索特点

- 一般没有爬虫系统
- 数据比较结构化
- 支持各种维度的排序
- 数据实时性要求非常高
- 与推荐系统和广告系统的融合

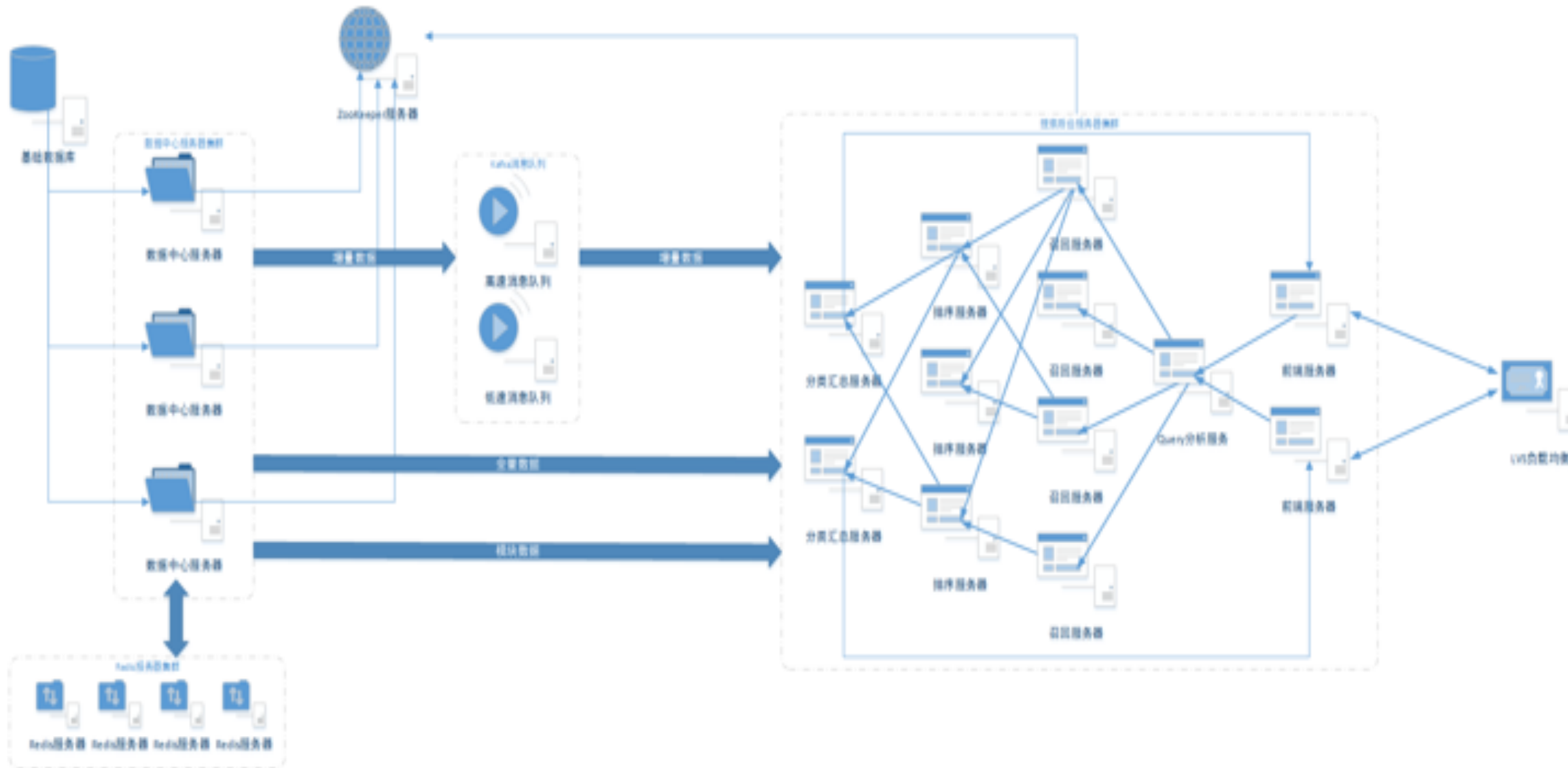


# 垂直搜索引擎模块

- 索引器 (**Elasticsearch**)
- 检索器 (**Elasticsearch**)
- **query**分析 (**plugin**)
- 离线打分排序
- 在线打分排序 (**plugin**)
- 商业因素排序 (**plugin**)
- 人工干预 (**plugin**)

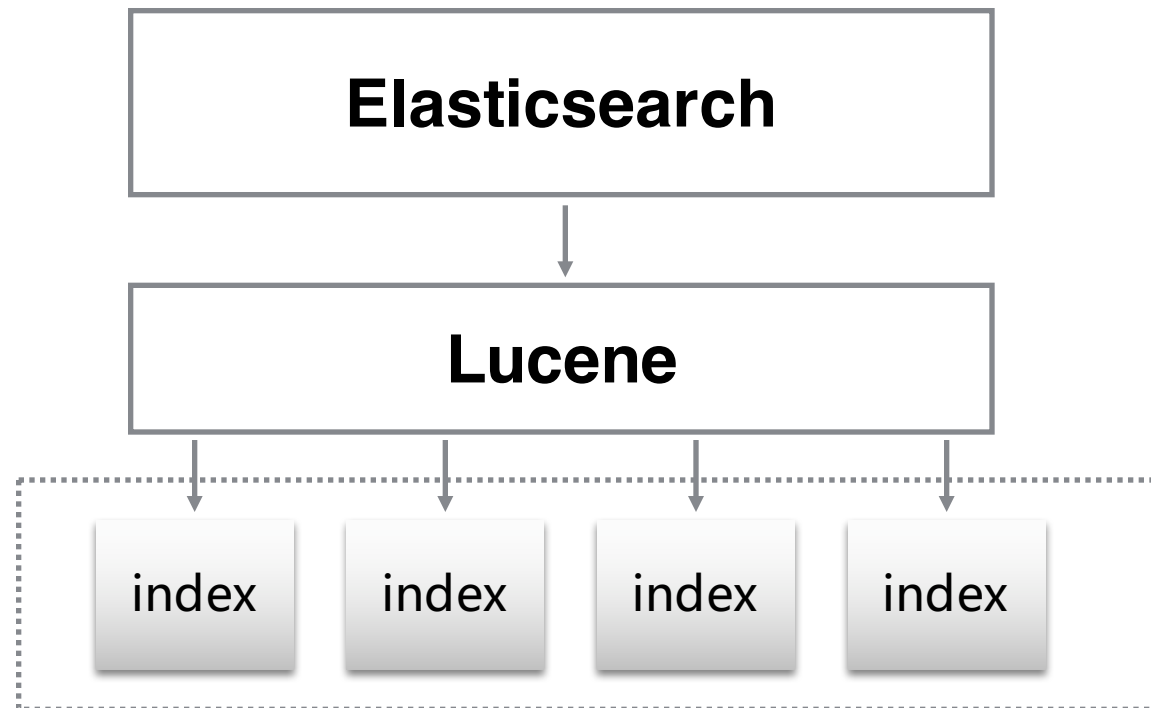


# 复杂的搜索平台



# Elasticsearch

- 集成度高
- 部署简单
- 功能完整且强大
- 伸缩性和可靠性高
- 功能扩展能力强
- **Lucene**性能全面



围绕Elasticsearch进行扩展和改造





# 离线索引和实时索引

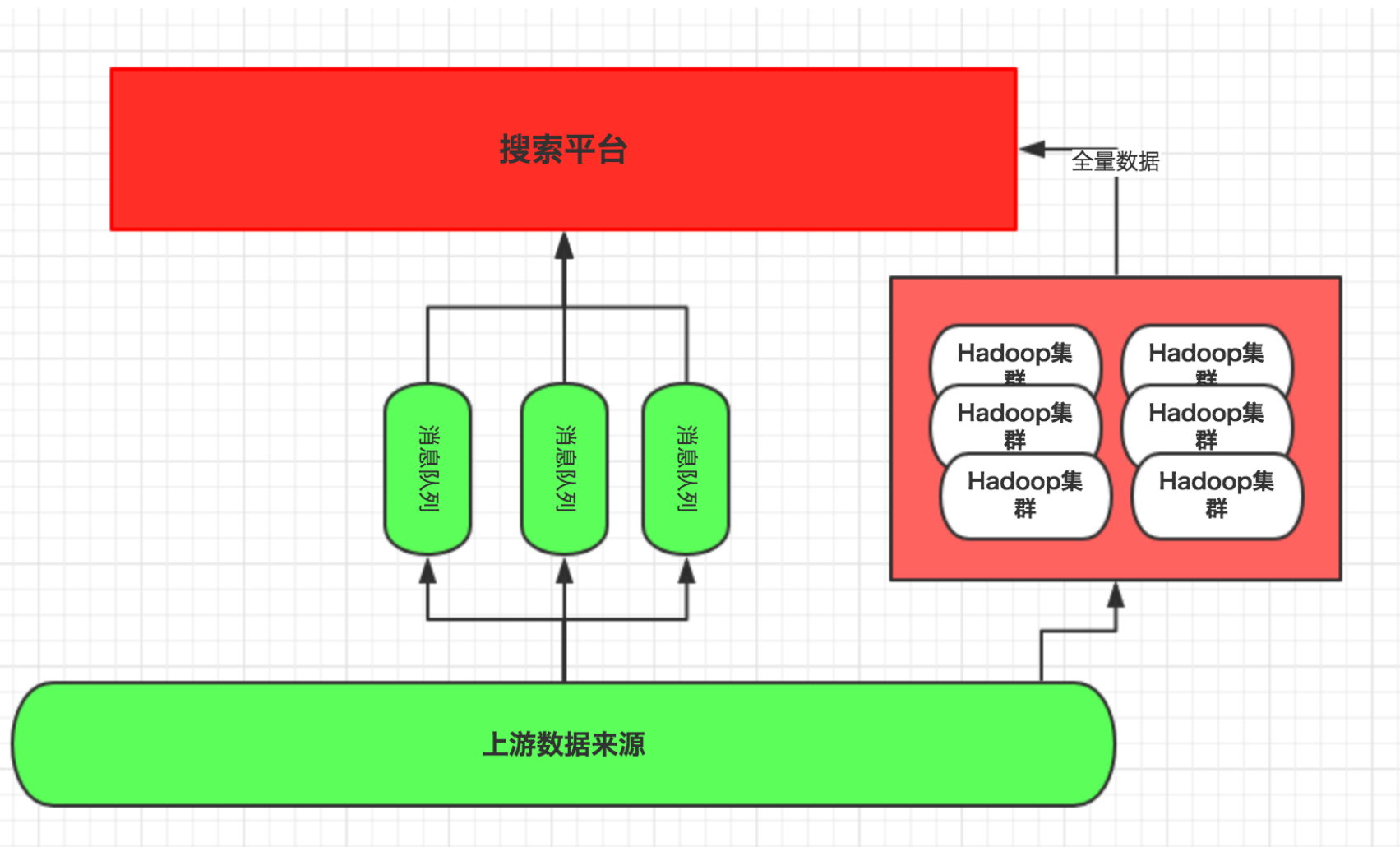


# 离线索引和计算

- 必不可少的离线索引部分
- 精确的离线打分
- 精确的文本打分模型
- 减轻在线计算压力
- 个性化推荐标签

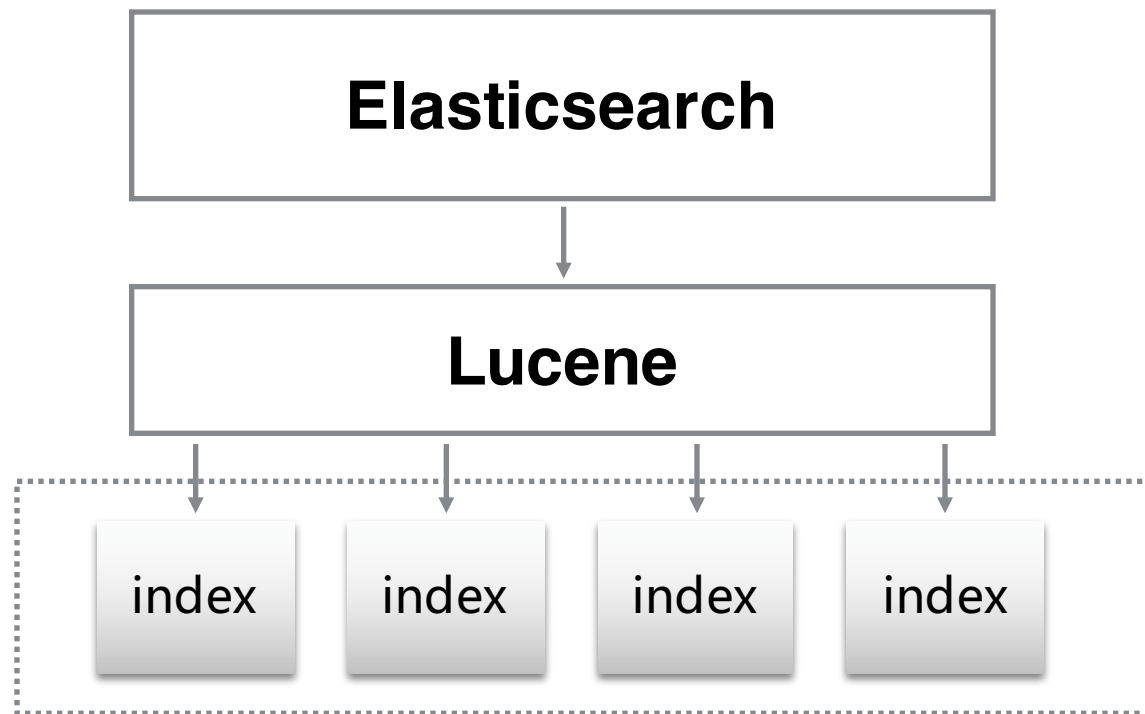


# 离线索引和计算

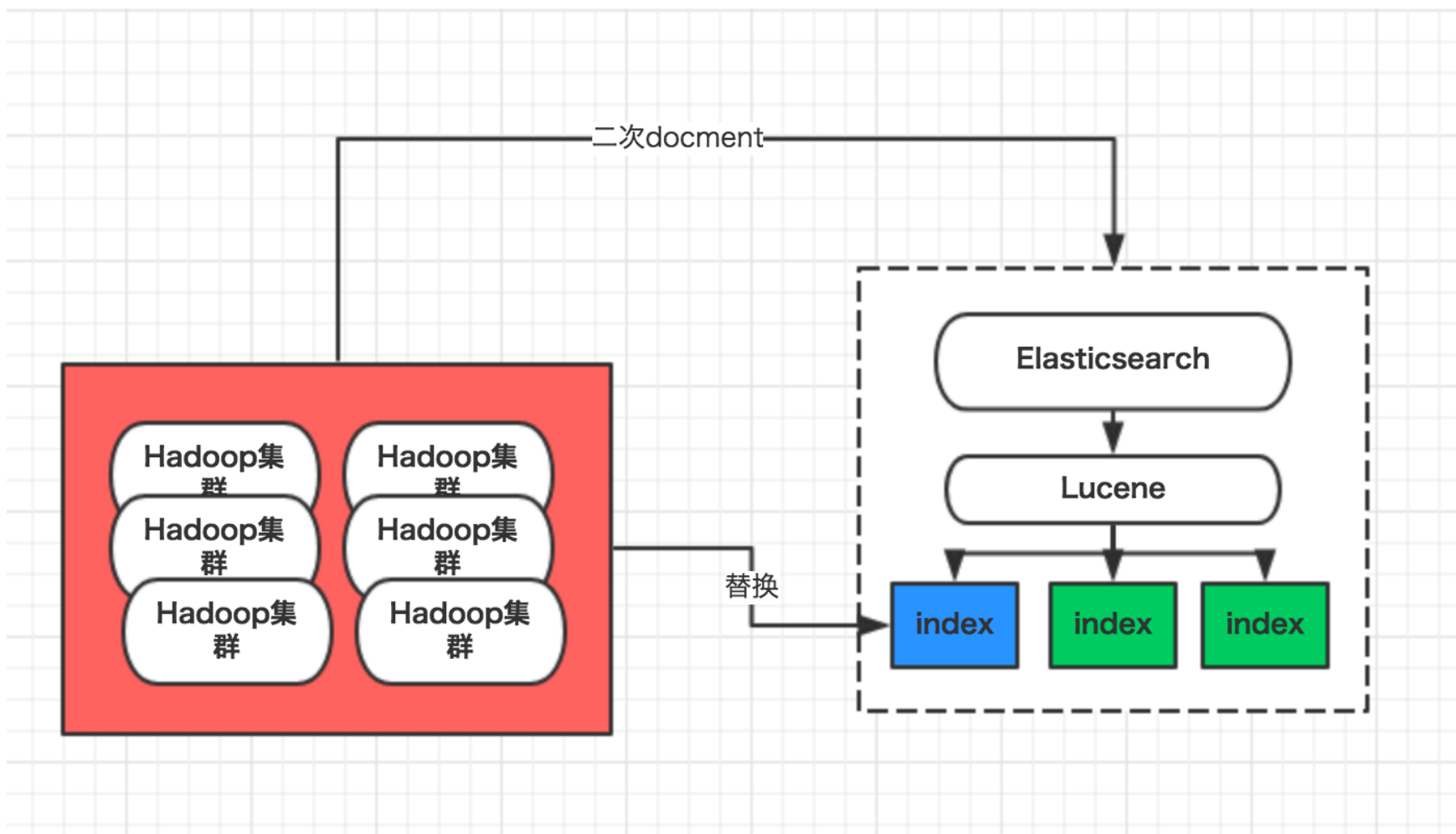


# ES离线索引和计算

- 彻底的索引构建
  - 最直接
  - 问题太多
- 折中生成二次**document**
  - 较简单
  - 控制度小
- 分词器问题
- 分片问题等等

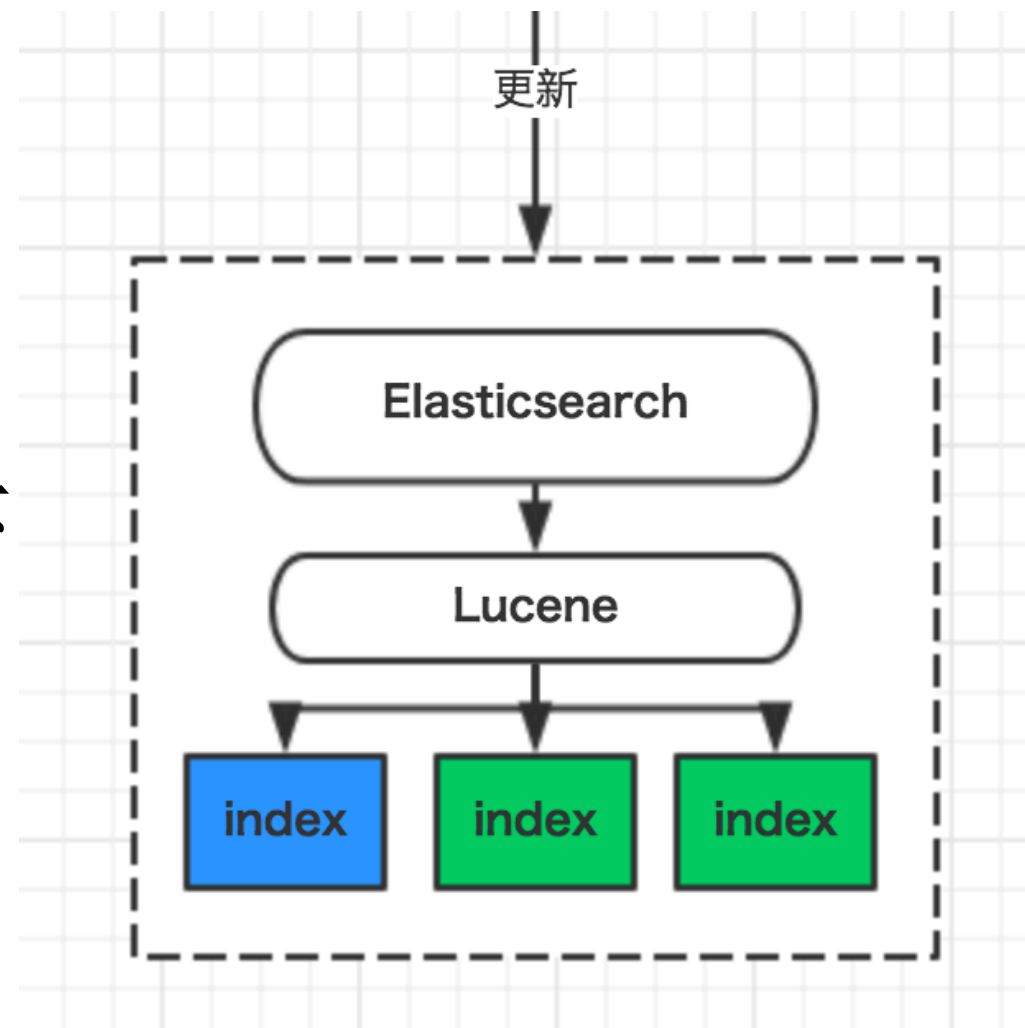


# ES离线索引和计算



# 实时更新文档

- 合并 → 删除 → 重建
- 解决频繁，大量局部更新的场景
- doc value ?



# Query分析模块



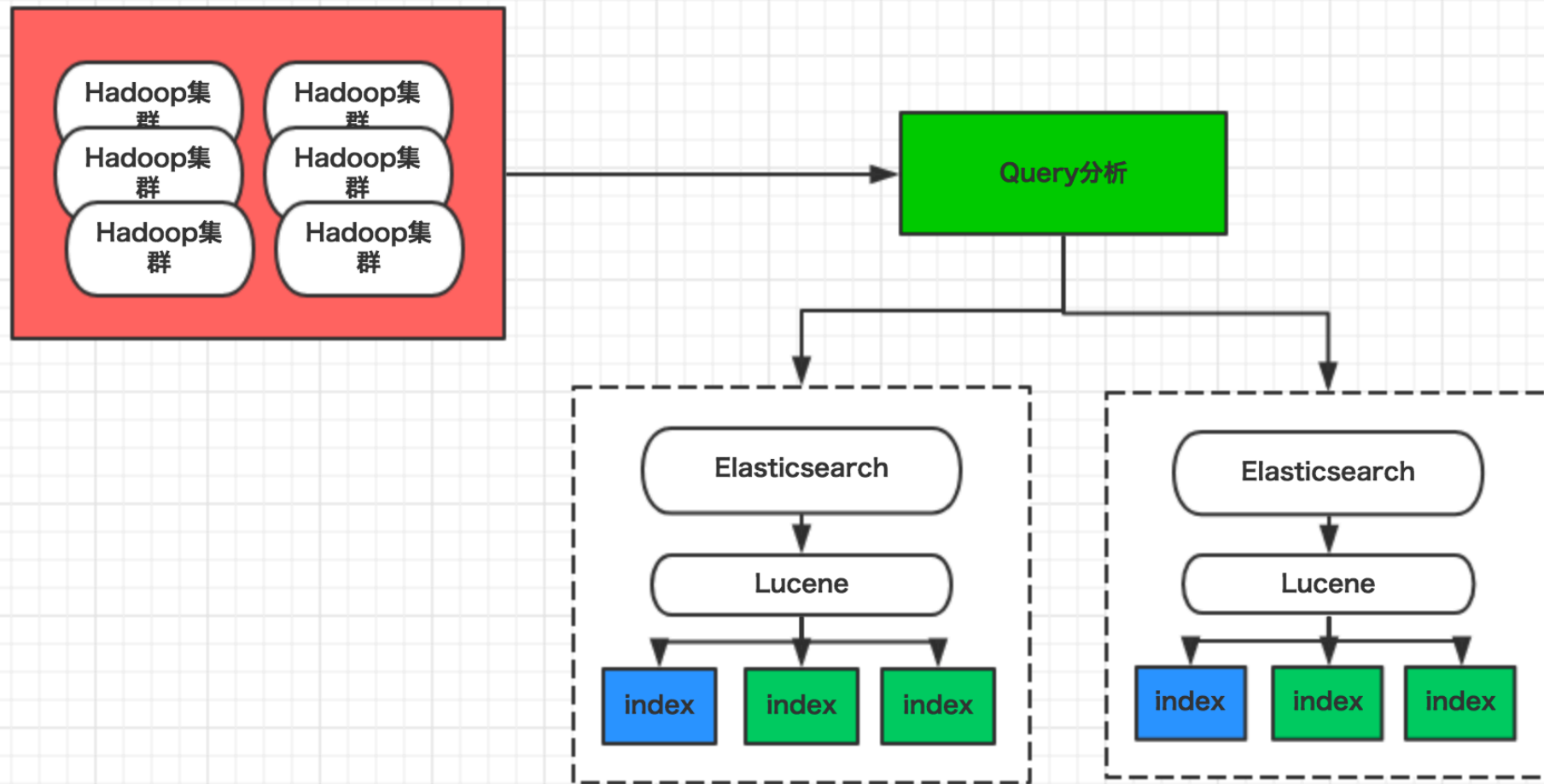
# Query分析

- 搜索质量的保证
- 大量离线训练数据
- 分类，语义识别
- 对应多业务多集群
- 插件？独立服务？





# Query分析



# 打分和排序



# 离线打分和在线排序

## 离线计算

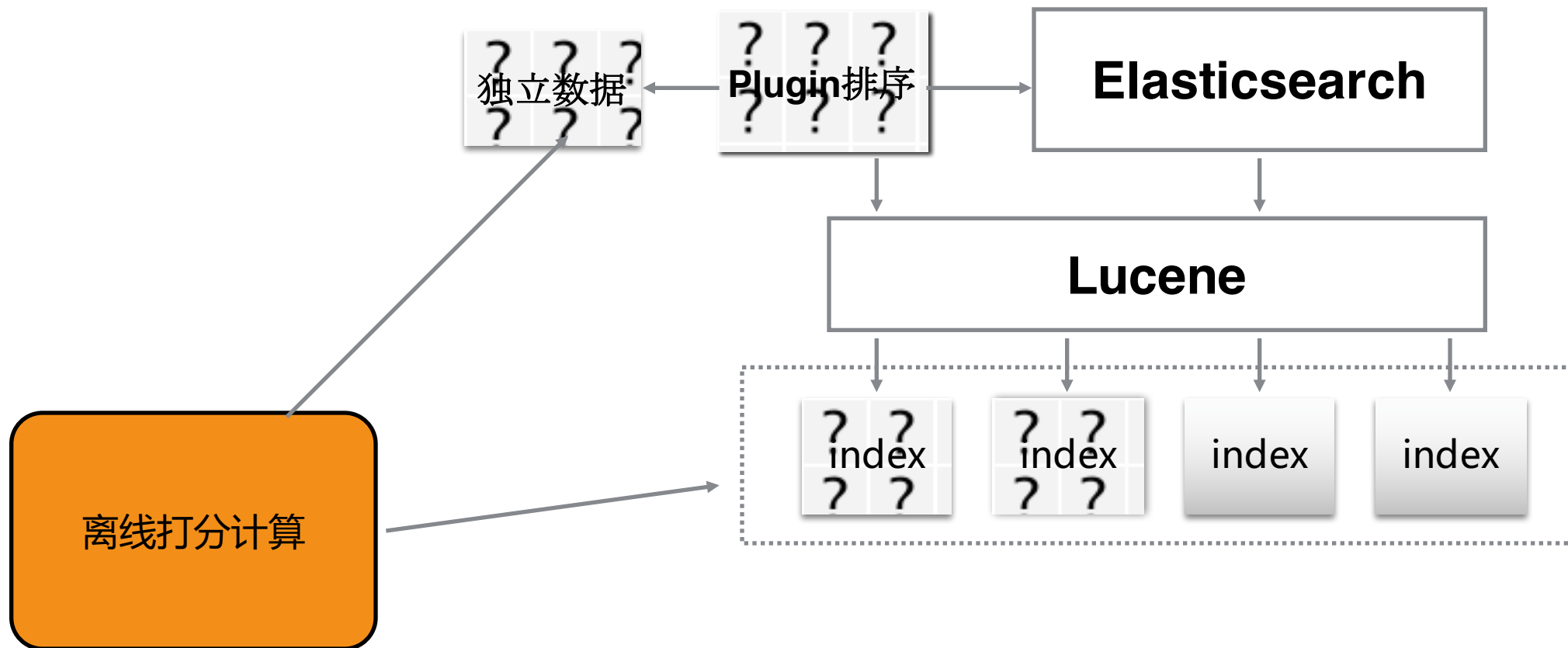
- 文本打分，辅助索引
- 生成辅助索引
- 权重计算
- 机器学习

## 在线排序

- 插件实现
- 脚本实现
- 辅助索引
- 独立数据
- 必须集成



# 离线和在线排序



# 商业因素排序

- 快速响应
- 随时调整
- 开发周期短，随时需要**AB**
- 与架构无关



# ES打分排序

- 插件和脚本完美契合
- 辅助索引
- **ES**的模块化和插件化设计
- 对开发人员友好

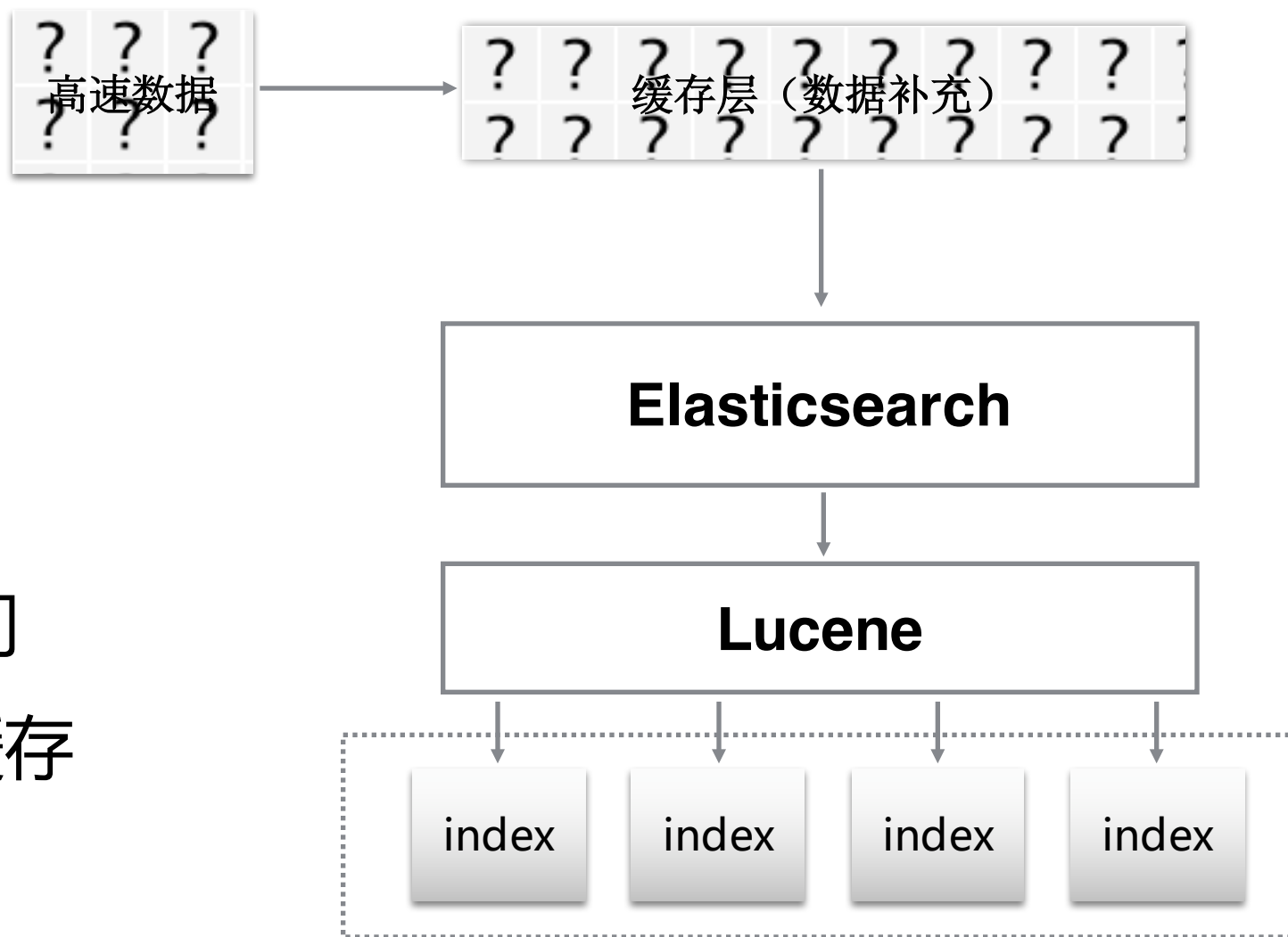


# 其他问题



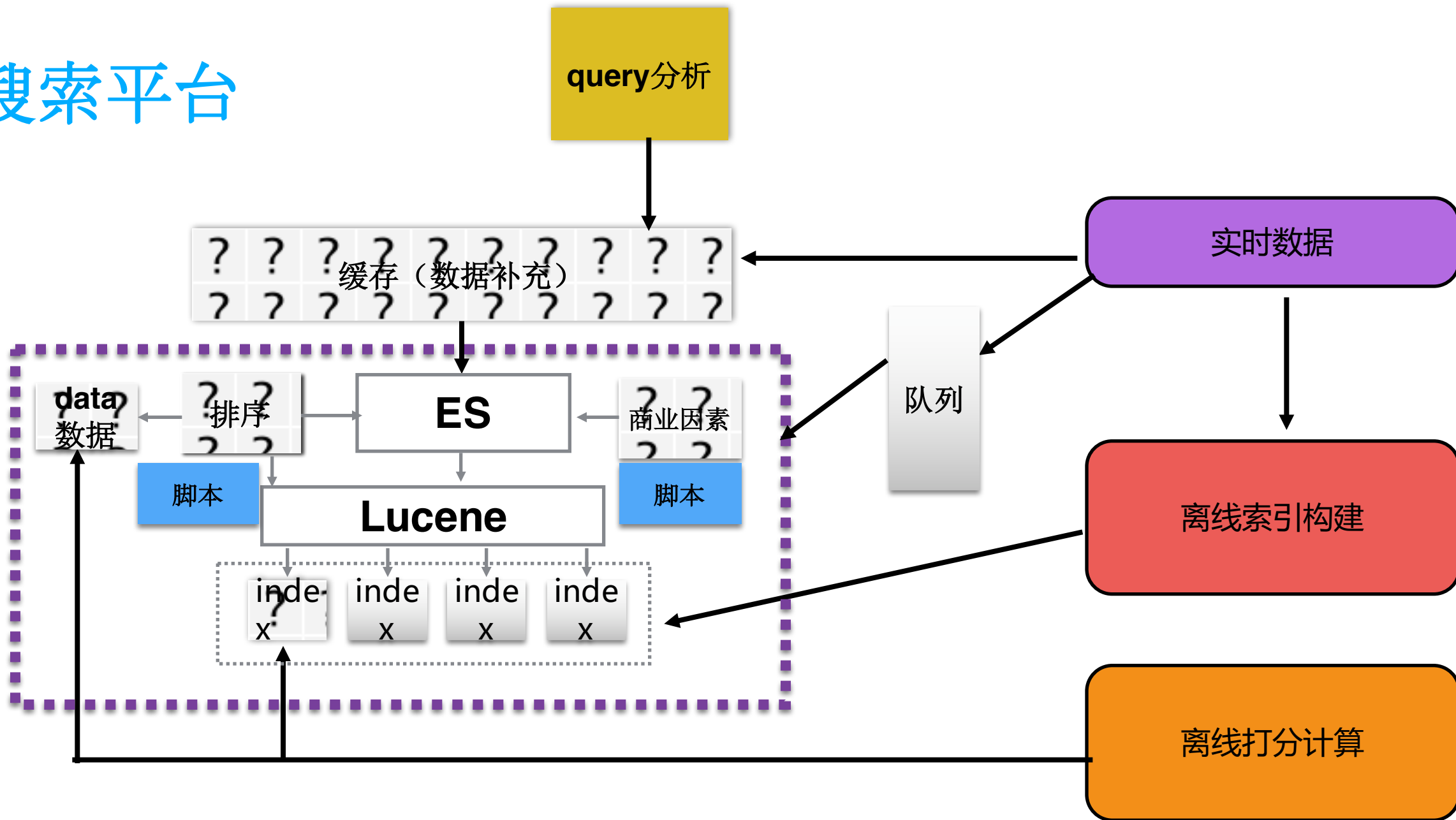
# 其他问题

- 分片和数据量级
- 热搜词和长尾词
- 集群稳定性
- 服务治理，命名空间
- doc缓存和结果集缓存
- 其他外围服务





# 搜索平台



# 总结

- 充分利用**Elasticsearch**的优势
- 该造的轮子还是要造
- **Elasticsearch**的适用场景
- 代码的复用
- 成本考虑



# 谢谢

## THANK YOU

——  
丰坚



[www.antgroup.com](http://www.antgroup.com)