有赞搜索引擎实践

洪斌

背景-电商对信息检索的需求

大数据
- 大规模索引创建
- 大规模搜索校验

背景-电商对信息检索的需求

大数据
- 大规模索引创建
- 大规模搜索校验

时效性
- 数据可靠传输
- 实时索引

背景-电商对信息检索的需求

大数据
- 大规模索引创建
- 大规模搜索校验

时效性
- 数据可靠传输
- 实时索引

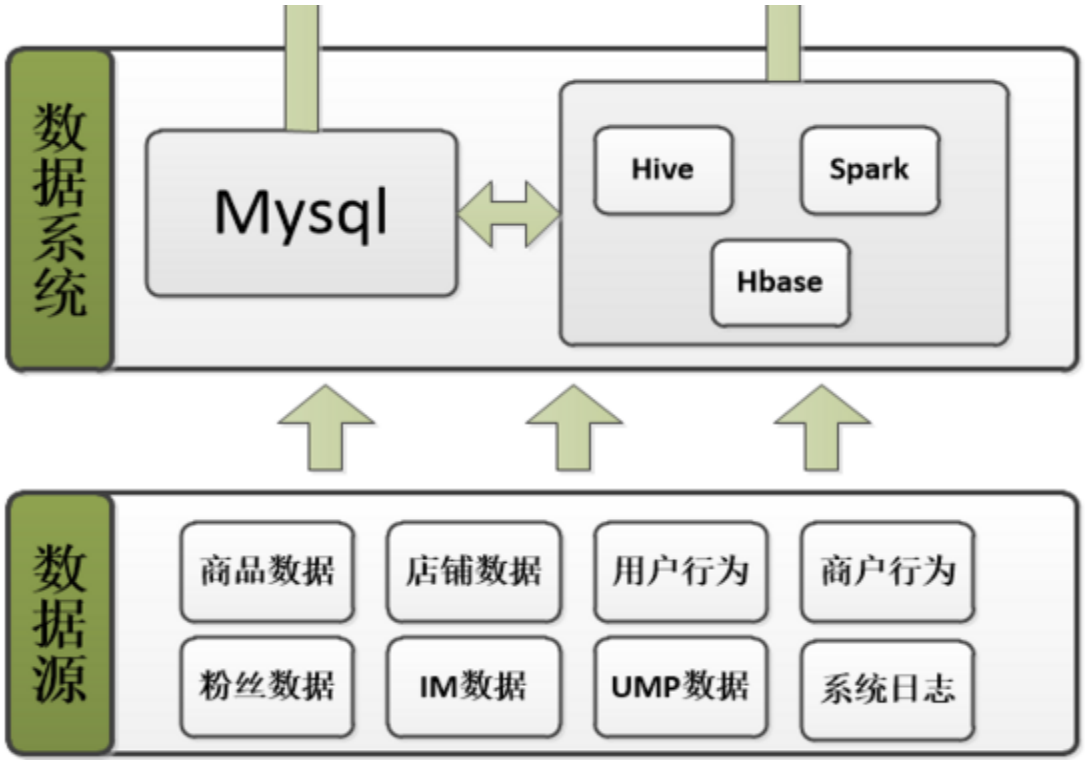
相关性
- 准确性
- 权重性

••去重　　••反作弊

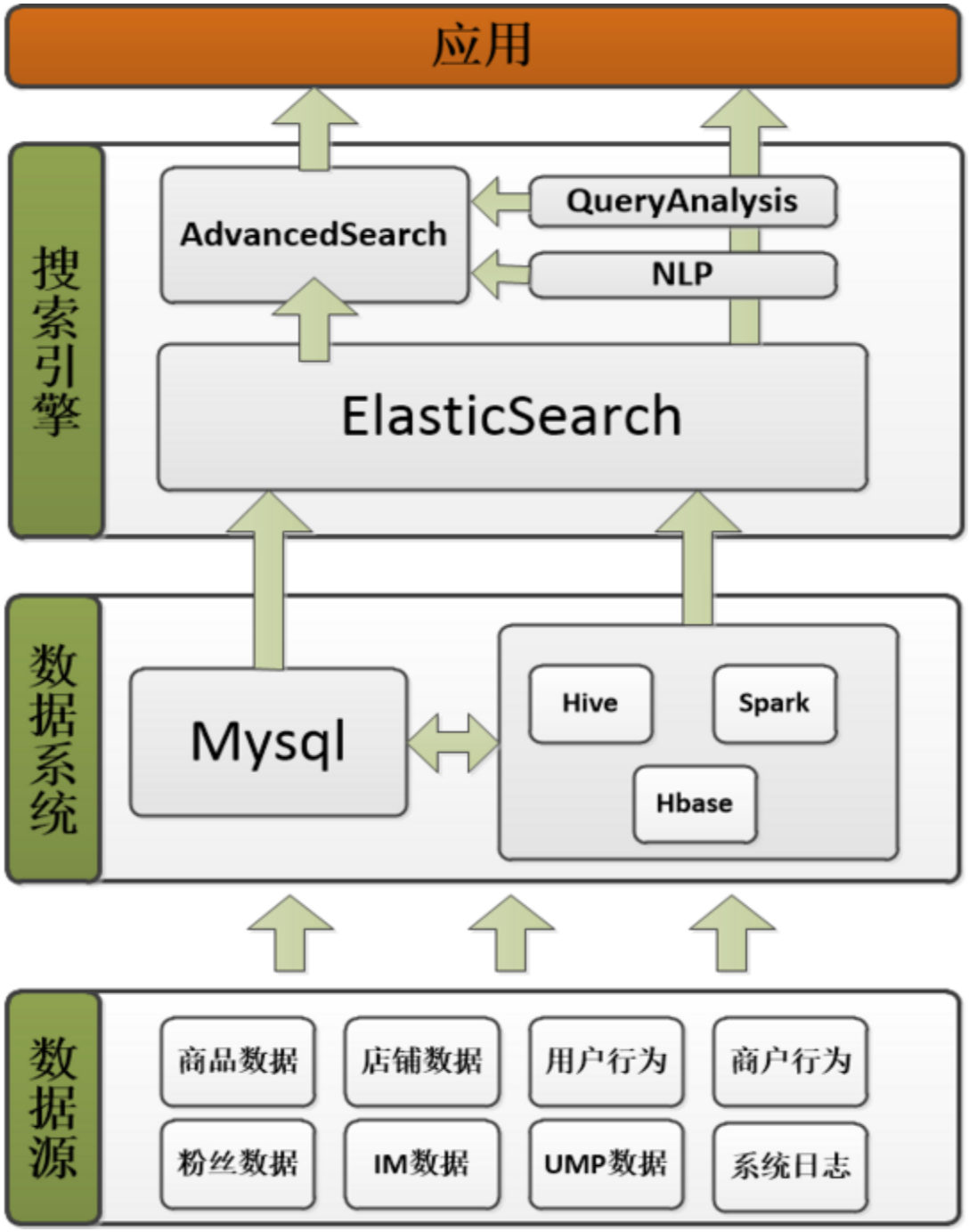
有赞搜索总体架构

有赞搜索总体架构

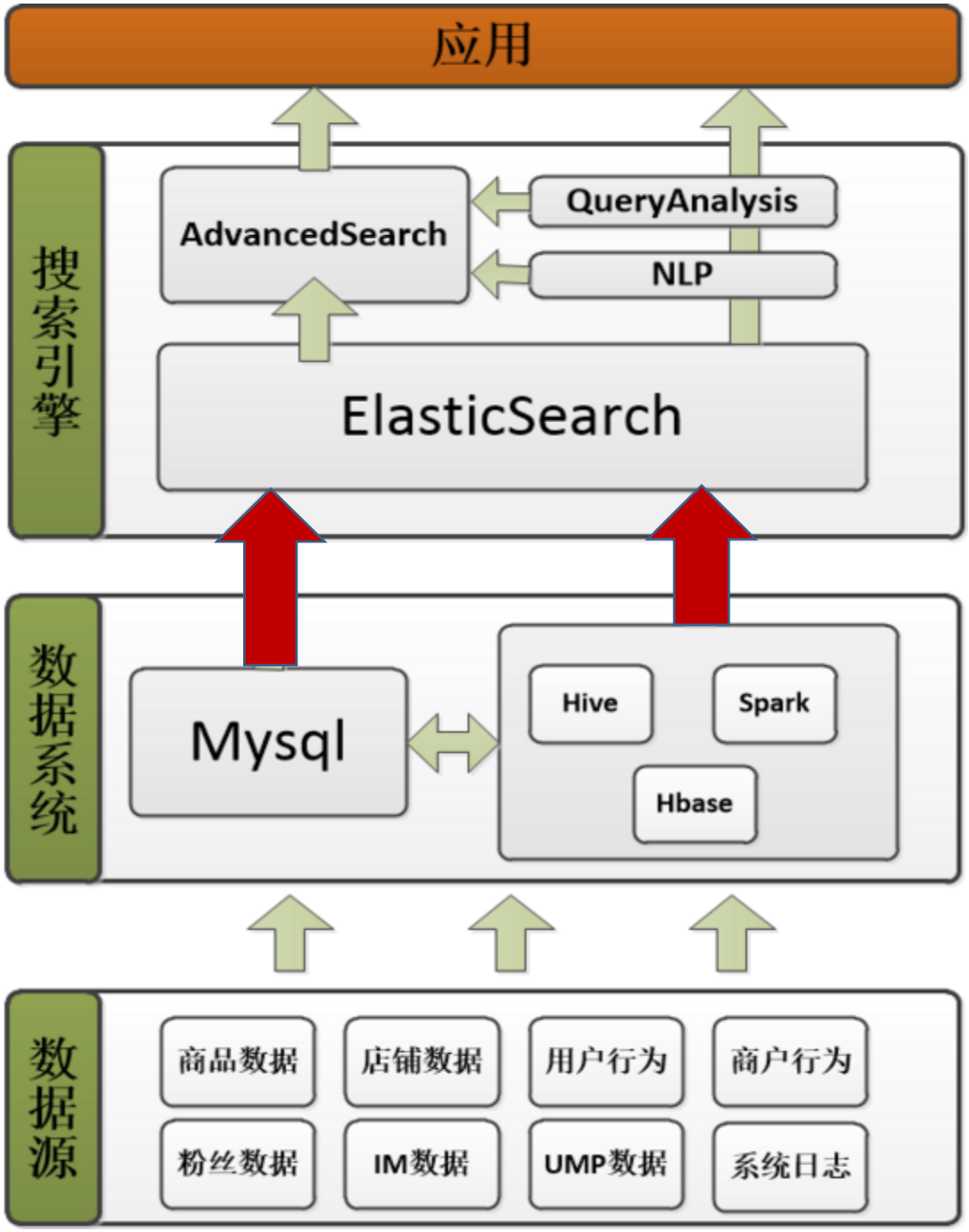
有赞搜索总体架构


数据系统
Mysql
Hive
Spark
Hbase
数据源
商品数据
店铺数据
用户行为
商户行为
粉丝数据
IM数据
UMP数据
系统日志

有赞搜索总体架构


应用
搜索引擎
AdvancedSearch
QueryAnalysis
NLP
ElasticSearch
数据系统
Mysql
Hive
Spark
Hbase
数据源
商品数据
店铺数据
用户行为
商户行为
粉丝数据
IM数据
UMP数据
系统日志

有赞搜索总体架构


应用
搜索引擎
AdvancedSearch
QueryAnalysis
NLP
ElasticSearch
数据系统
Mysql
Hive
Spark
Hbase
数据源
商品数据
店铺数据
用户行为
商户行为
粉丝数据
IM数据
UMP数据
系统日志

索引构建

索引构建

- 全量构建
- 增量构建
  - 快

索引构建

- 全量构建
  - 大　，数据量巨大如何在短时间构建
  - 全　，数据不小心丢了如何弥补
- 增量构建
  - 快

索引构建

- 全量构建
  - 大
  - 全
- 增量构建
  - 快

索引构建-全量

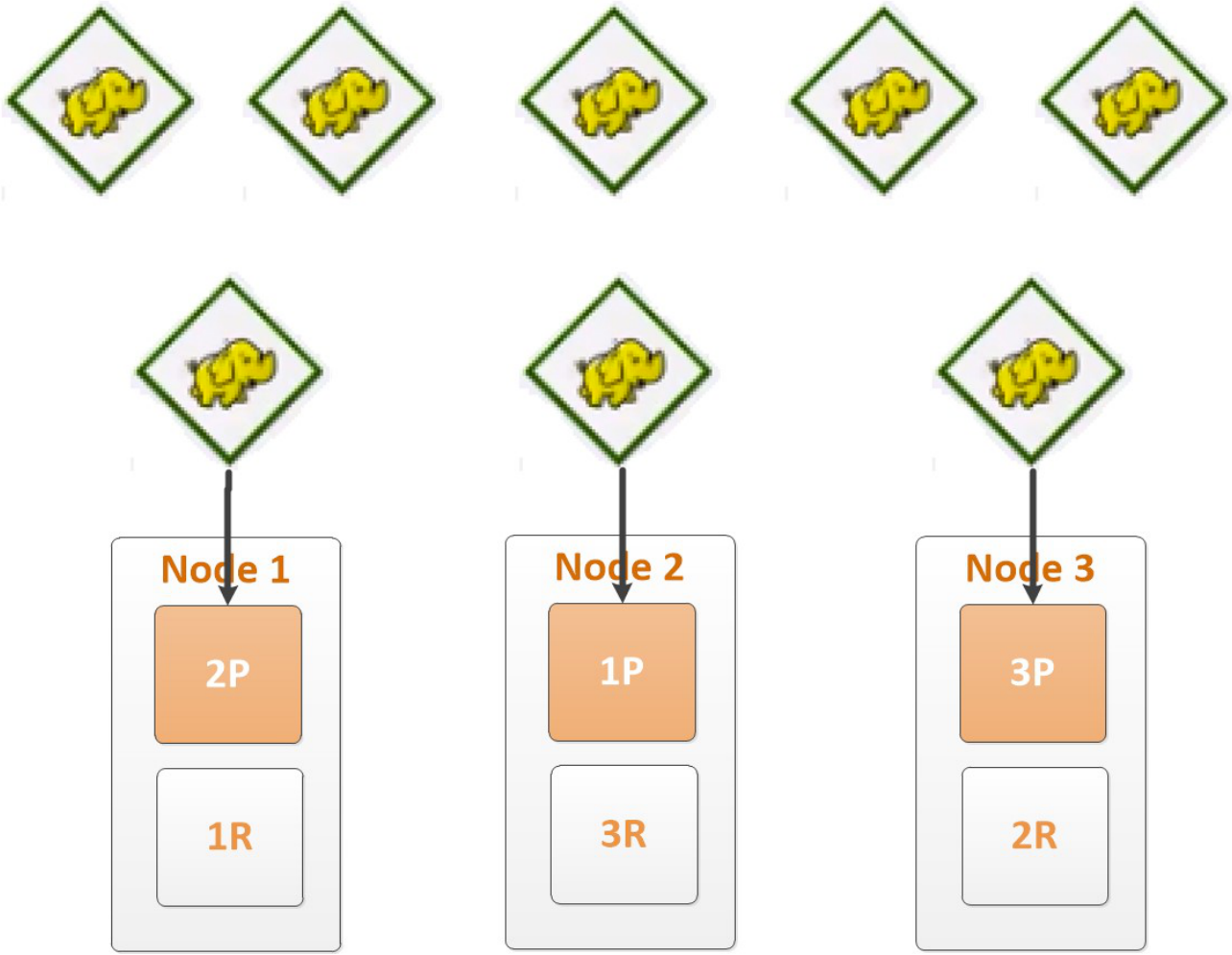
索引构建-全量

- Hadoop构建优势
  - 离线计算环境
  - 数据源无关

索引构建-全量

- Hadoop构建优势
  - 离线计算环境
  - 数据源无关
- 解决方案
  - 方案一: Hadoop构建lucence每个node的索引分发到每个es node上

索引构建-全量



索引构建-全量

- Hadoop构建优势
  - 离线计算环境
  - 数据源无关
- 解决方案
  - 方案一: Hadoop构建lucence每个node的索引分发到每个es node上
  - 方案二: 讲es集群映射成Hive外部表. 通过并行Restful技术进行并行更新（es-hadoop）

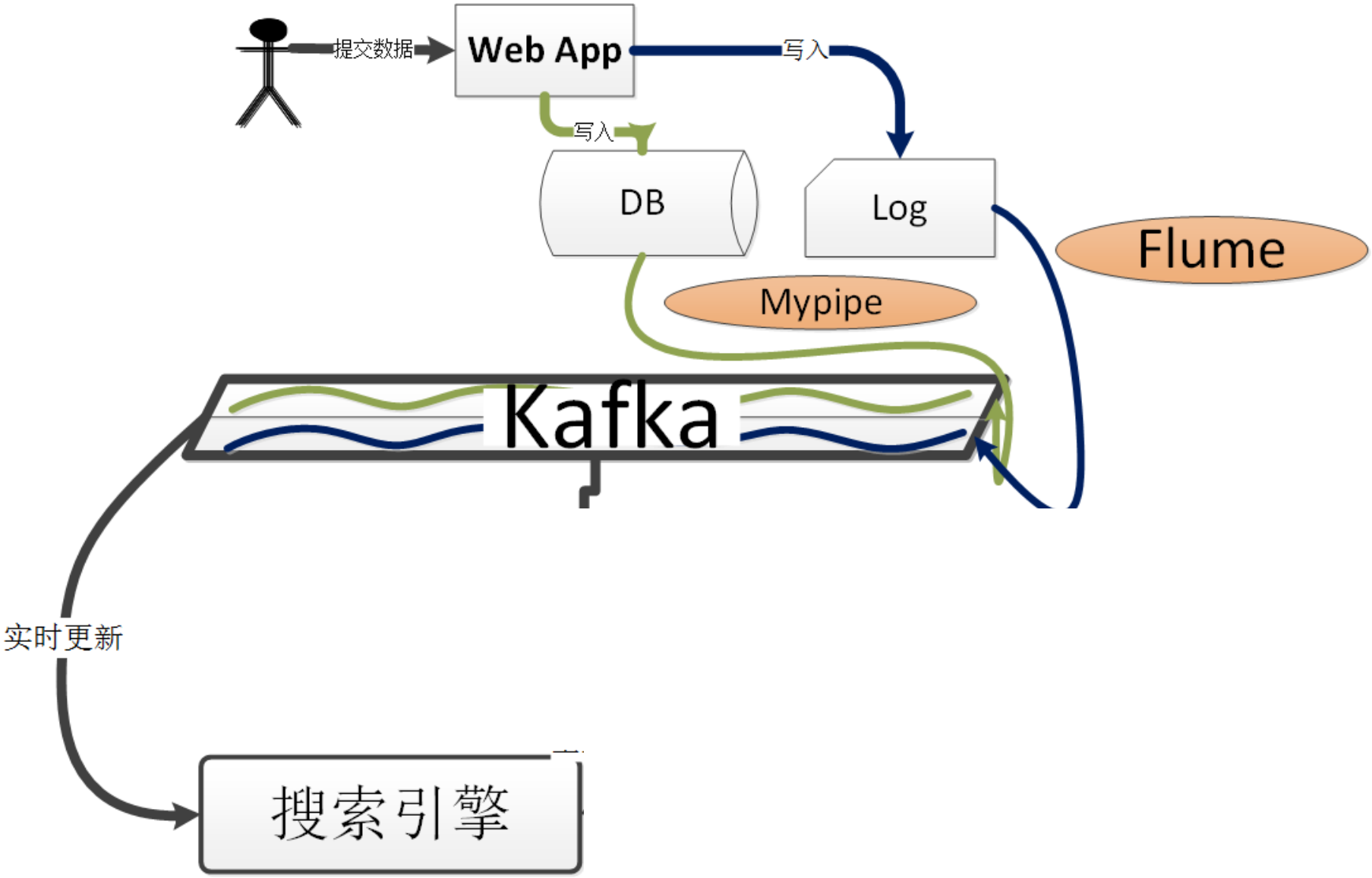
两种方案都可以解决全量更新的单点问题

索引构建-全量

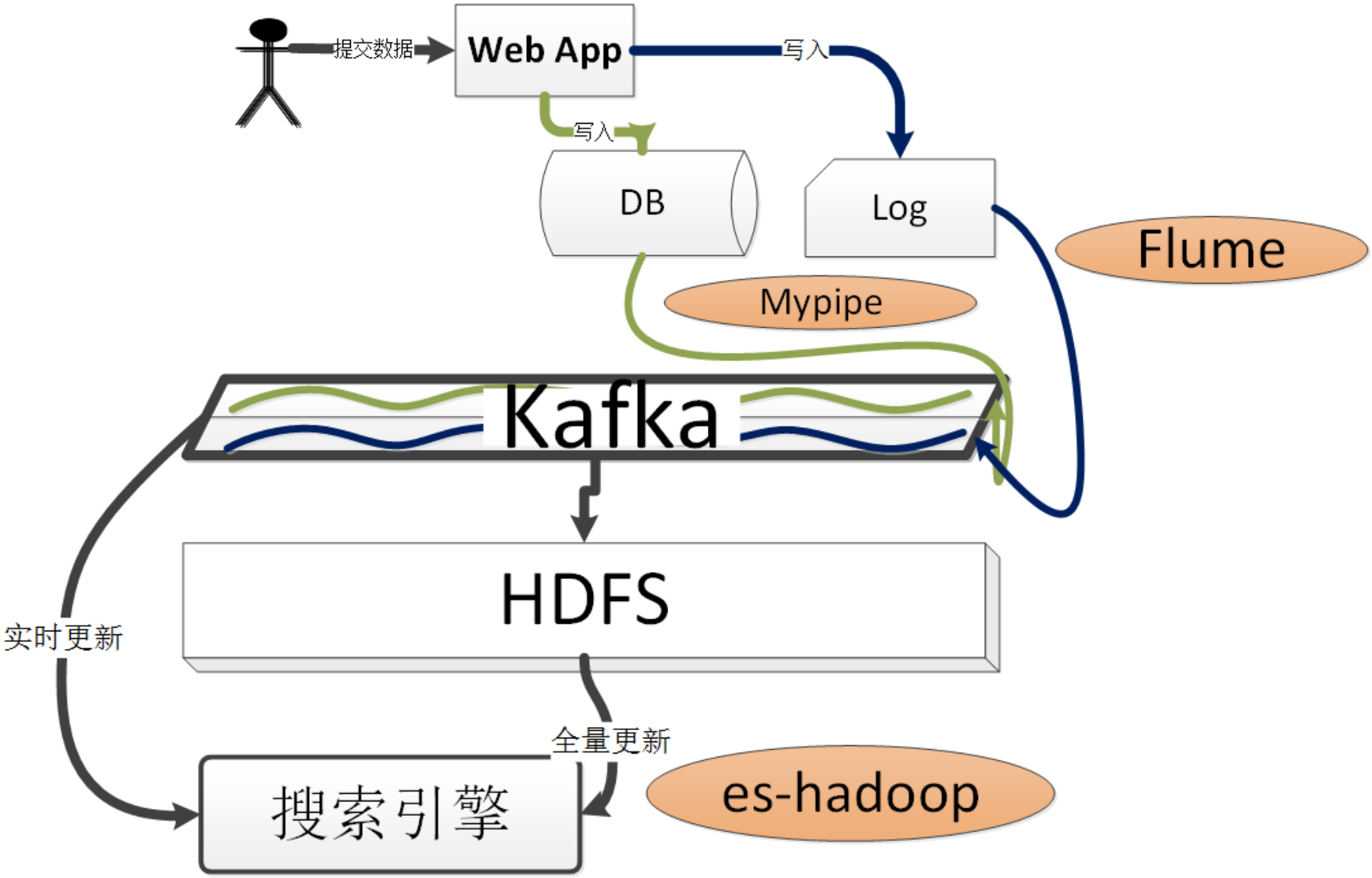
- Hadoop构建优势
  - 离线计算环境
  - 数据源无关
- 解决方案
  - 方案一: Hadoop构建lucence每个node的索引分发到每个es node上
  - 方案二: 讲es集群映射成Hive外部表. 通过并行Restful技术进行并行更新（es-hadoop）
  - 两种方案都可以解决全量更新的单点问题

索引构建-增量更新

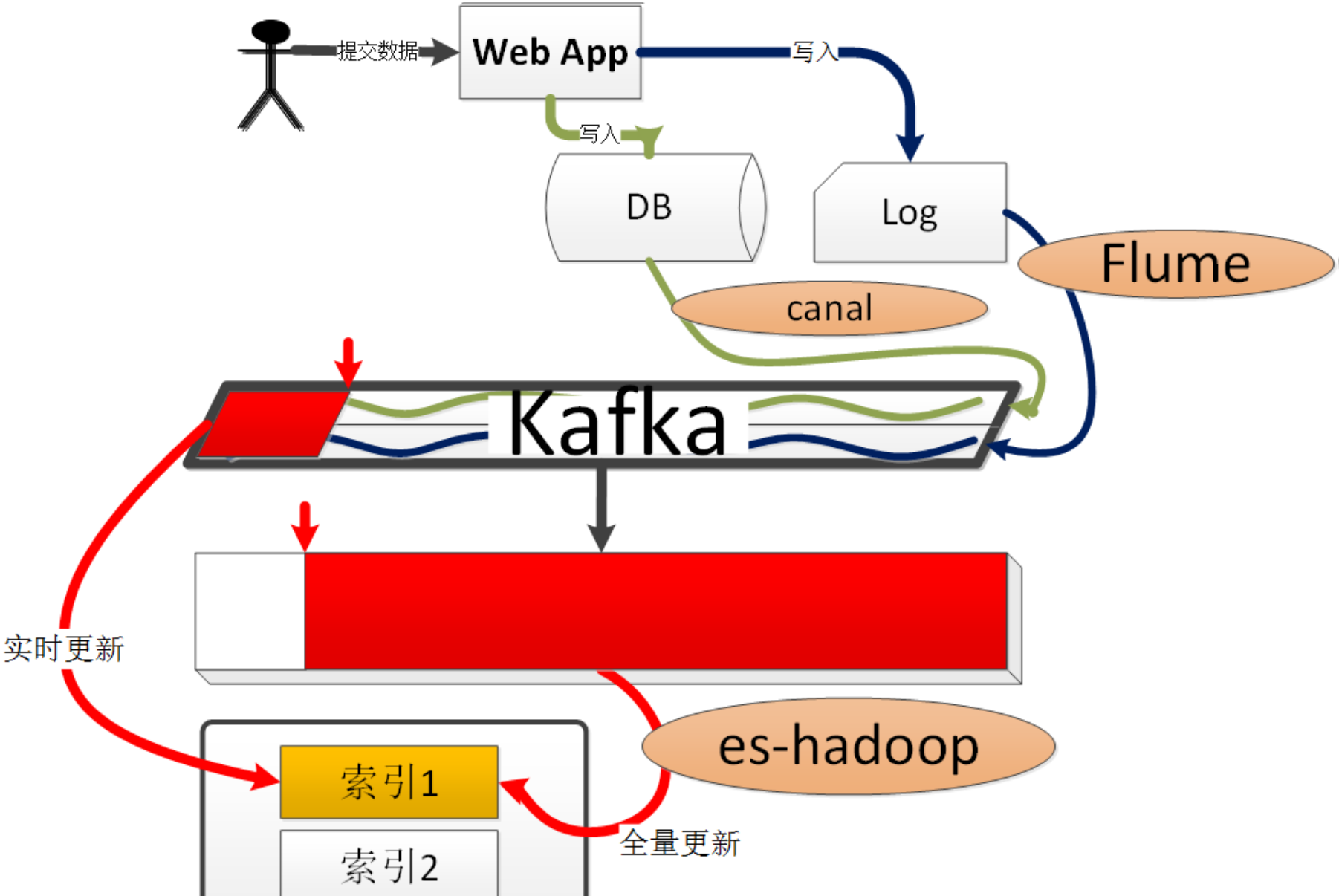
索引构建-增量更新



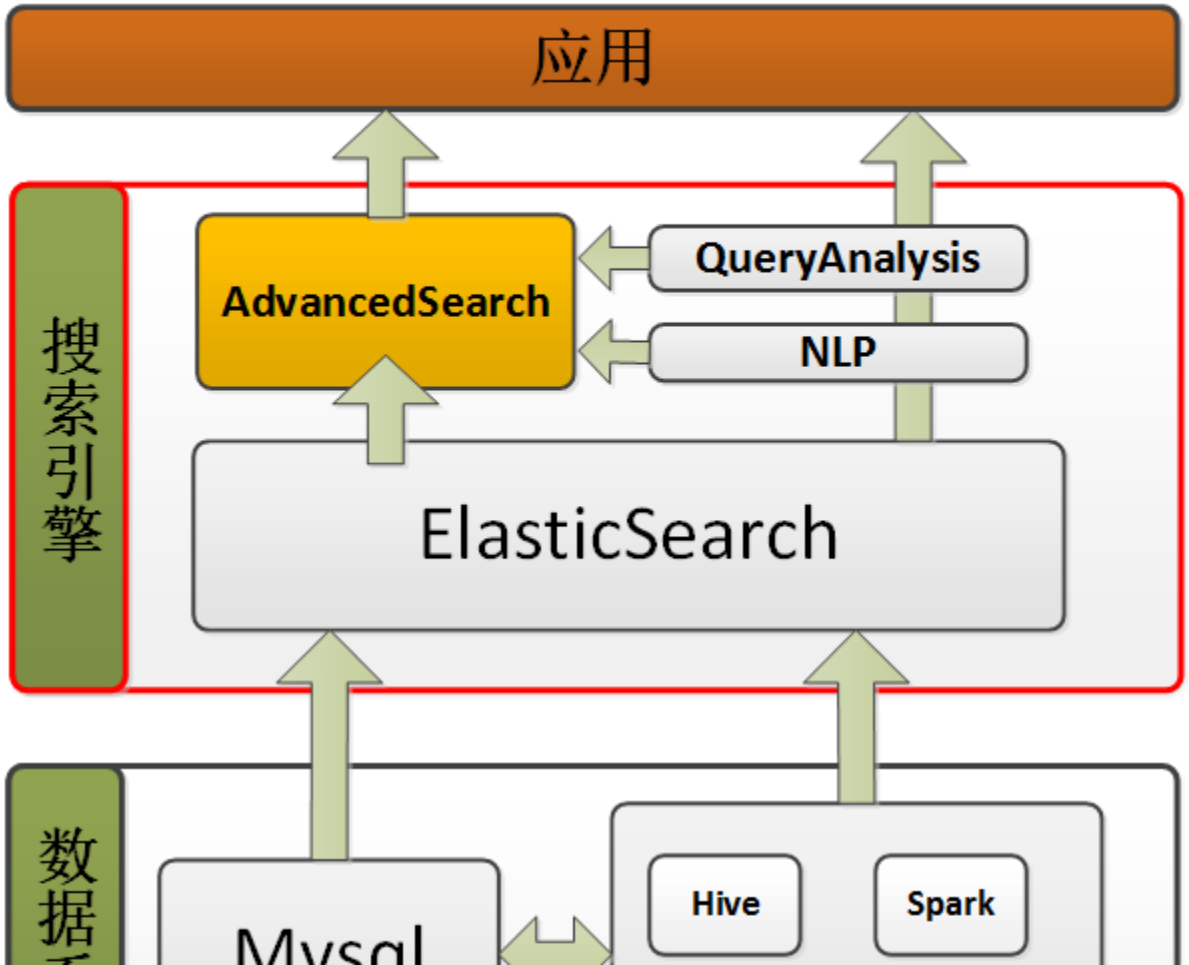
索引构建-增量更新


提交数据
Web App
写入
写入
DB
Log
Flume
Mypipe
Kafka
HDFS
实时更新
全量更新
搜索引擎
es-hadoop

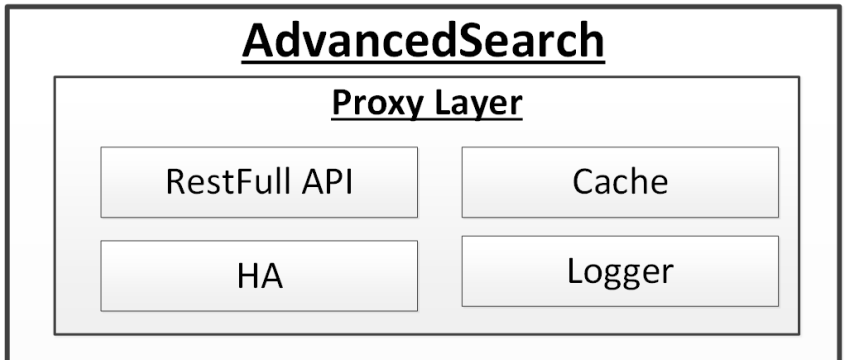
索引构建-增量更新


提交数据
Web App
写入
写入
DB
Log
Flume
canal
Kafka
实时更新
es-hadoop
索引1
索引2
全量更新

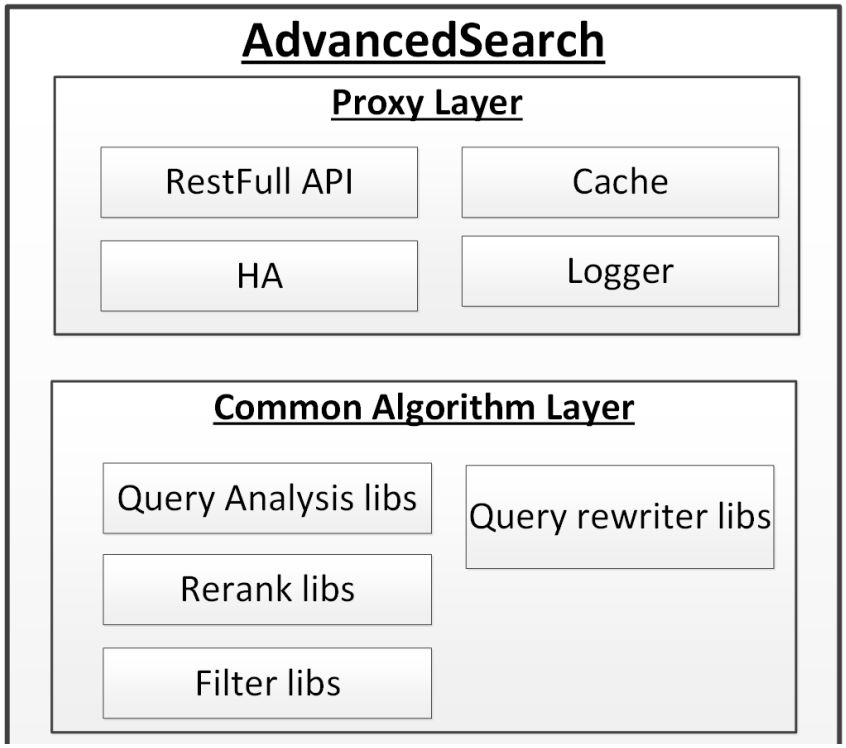
高级搜索


应用
搜索引擎
AdvancedSearch
QueryAnalysis
NLP
ElasticSearch
Hive
Spark

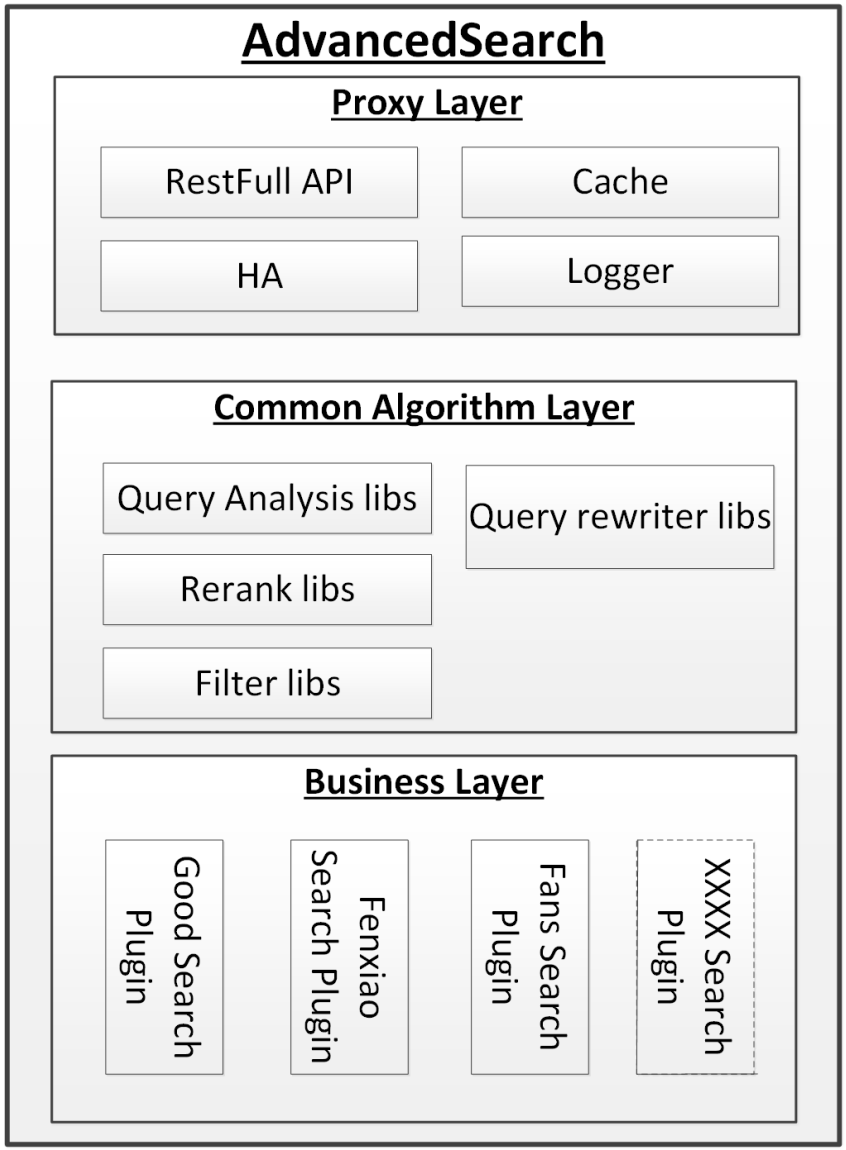
高级搜索



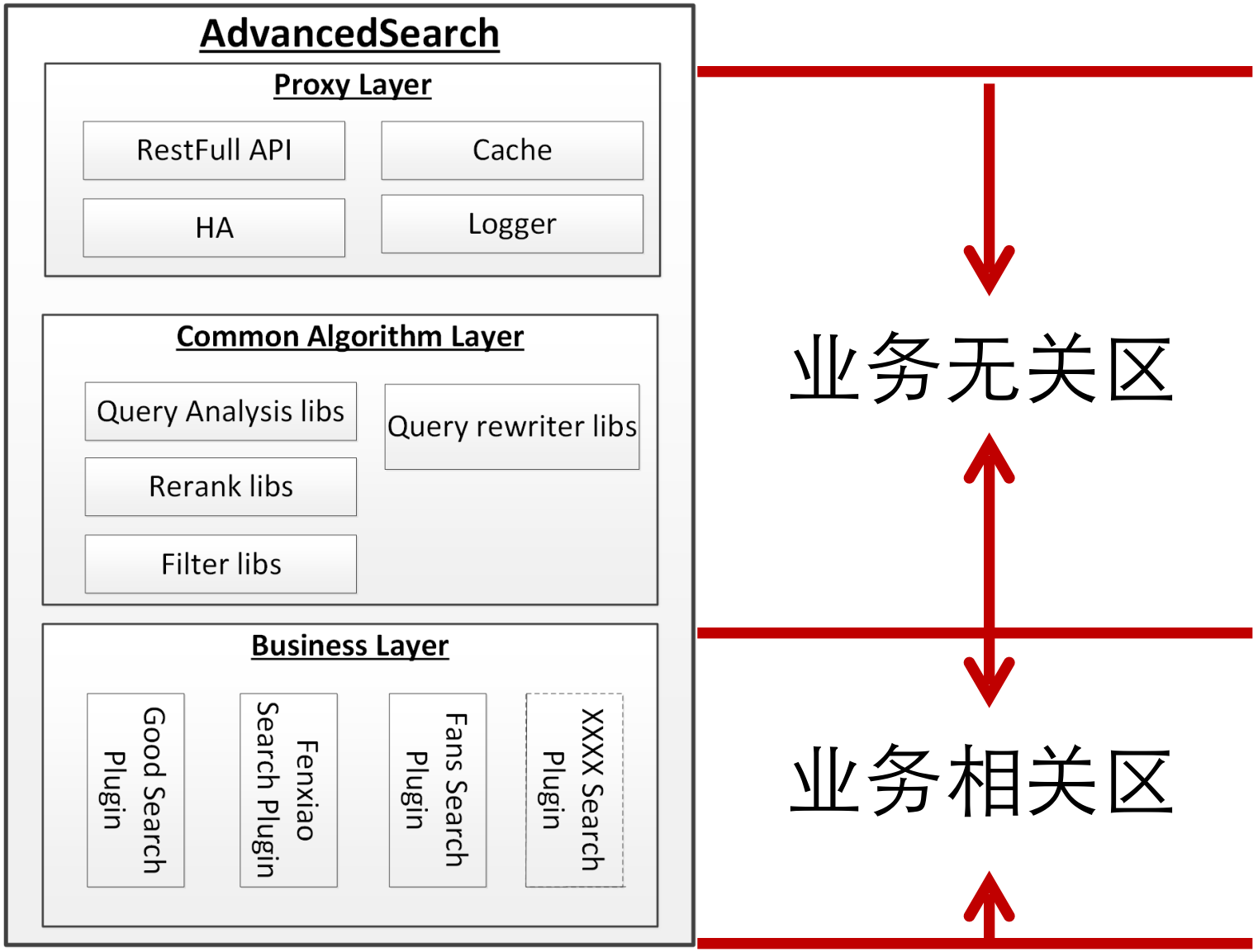
高级搜索

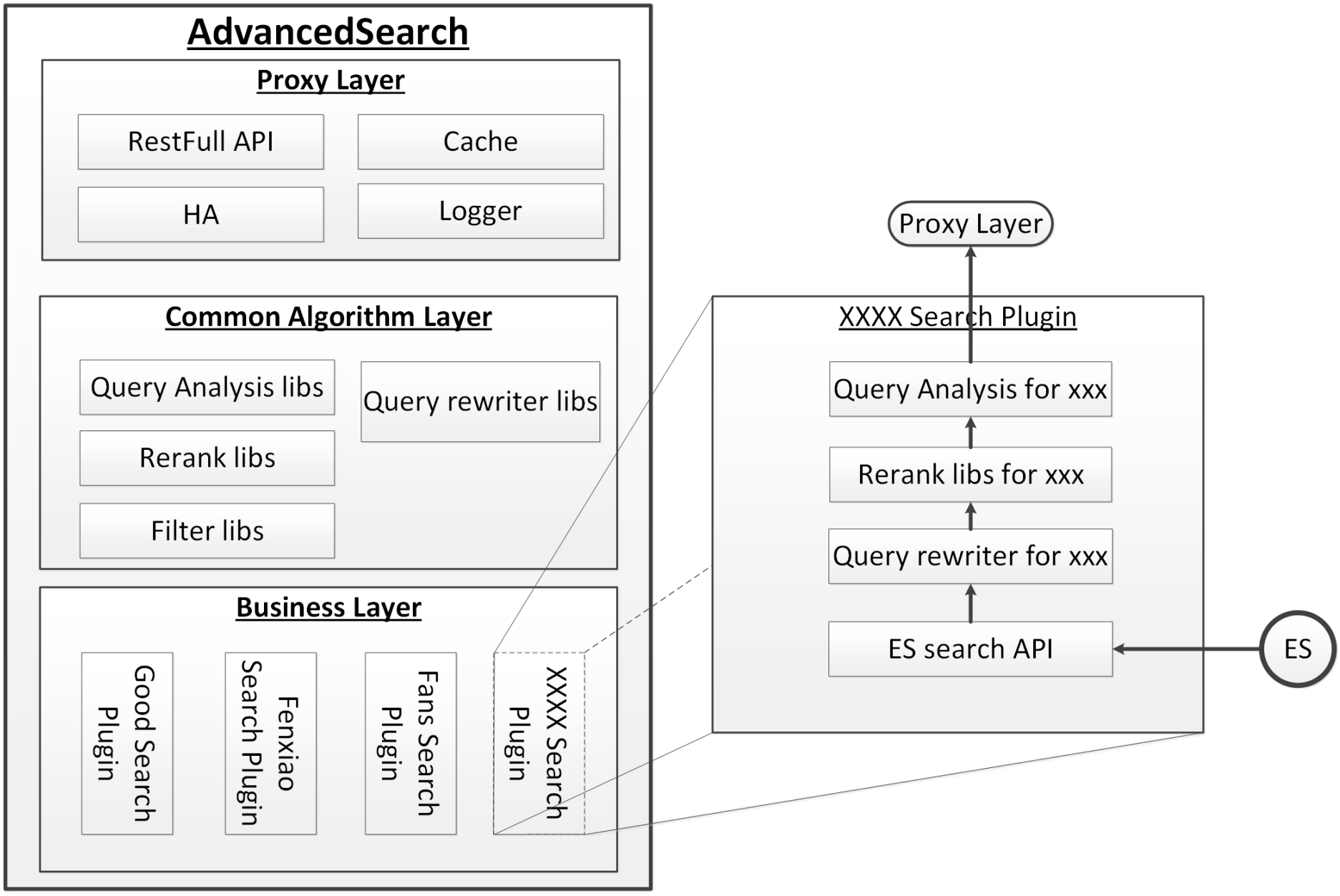

AdvancedSearch
Proxy Layer
RestFull API
Cache
HA
Logger
Common Algorithm Layer
Query Analysis libs
Query rewriter libs
Rerank libs
Filter libs

高级搜索


AdvancedSearch
Proxy Layer
RestFull API
Cache
HA
Logger
Common Algorithm Layer
Query Analysis libs
Query rewriter libs
Rerank libs
Filter libs
Business Layer
Good Search Plugin
Fenxiao Search Plugin
Fans Search Plugin
XXXX Search Plugin

高级搜索


AdvancedSearch
Proxy Layer
RestFull API
Cache
HA
Logger
Common Algorithm Layer
Query Analysis libs
Query rewriter libs
Rerank libs
Filter libs
Business Layer
Good Search Plugin
Fenxiao Search Plugin
Fans Search Plugin
XXXX Search Plugin
业务无关区
业务相关区


AdvancedSearch
Proxy Layer
RestFull API
Cache
HA
Logger
Common Algorithm Layer
Query Analysis libs
Query rewriter libs
Rerank libs
Filter libs
Business Layer
Good Search Plugin
Fenxiao Search Plugin
Fans Search Plugin
XXXX Search Plugin
Proxy Layer
XXXX Search Plugin
Query Analysis for xxx
Rerank libs for xxx
Query rewriter for xxx
ES search API
ES

高级搜索

- 反向代理
- 提供丰富的相关性库
- 管理不同的搜索业务
- 屏蔽内部复杂性

评分体系

评分体系

静态分 * 动态分

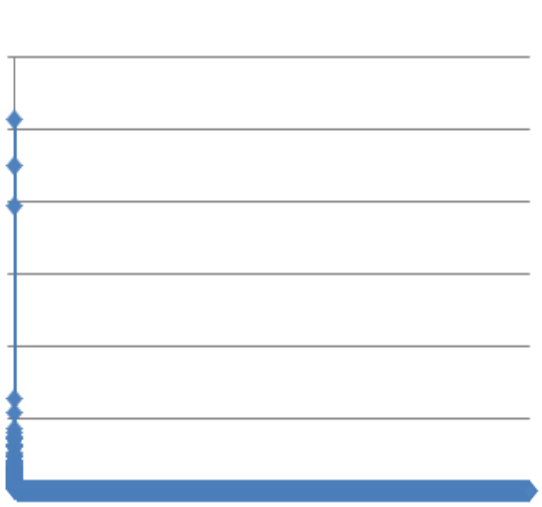
- 静态分体现商品的重要性.
- 动态分体现商品和query的相关性

评分系统-静态分

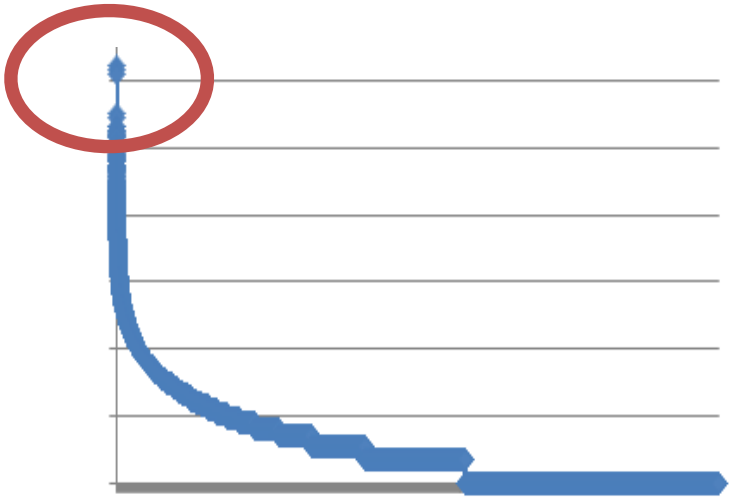
- 目标
  - 稳定性
  - 连续性
  - 区分度


指标集
订单
浏览
……
售后
组合并归一化
指标库
订单分
评价分
……
售后分
参数选择
静态分

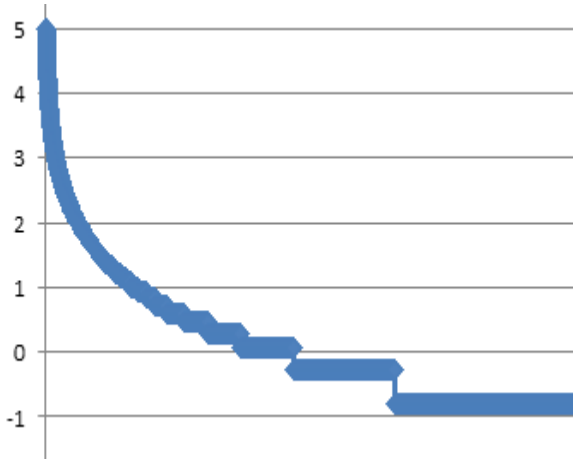
归一化方法



min-max归一化



log归一化



log-zscore归一化

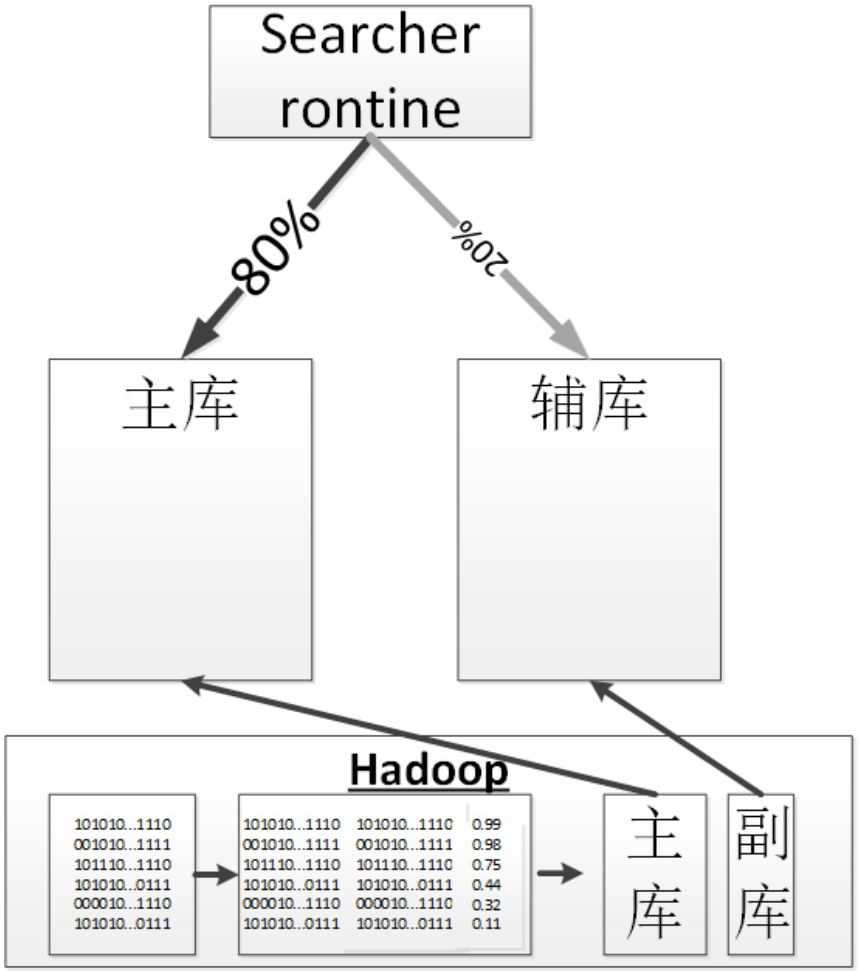
几个例子

去重-商品去重

- 商品去重转化为计算两个向量的相似度
- 著名的All Pair Similarity 问题
- Spark提供技术支持

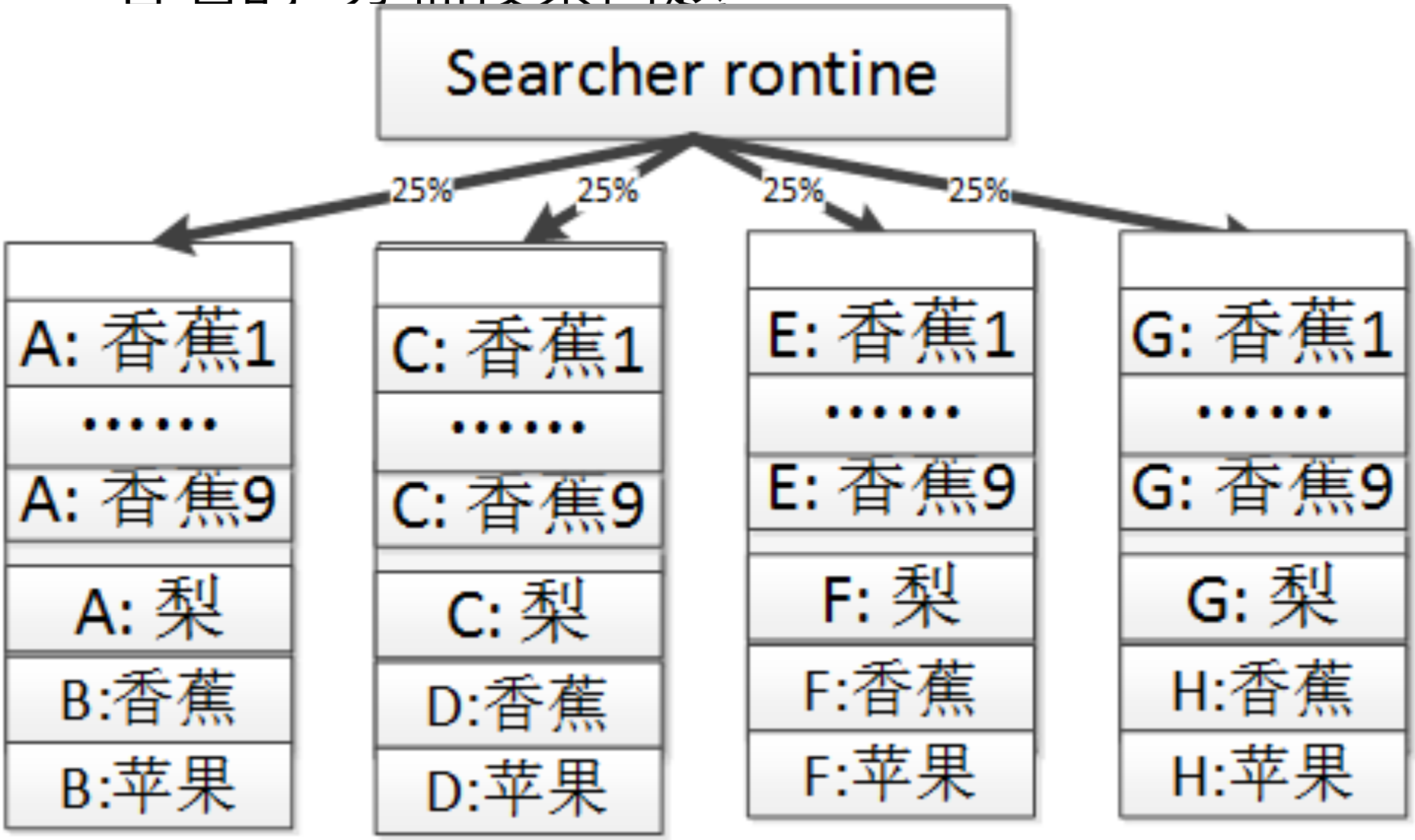
  Matrix.

  columnSimilarities

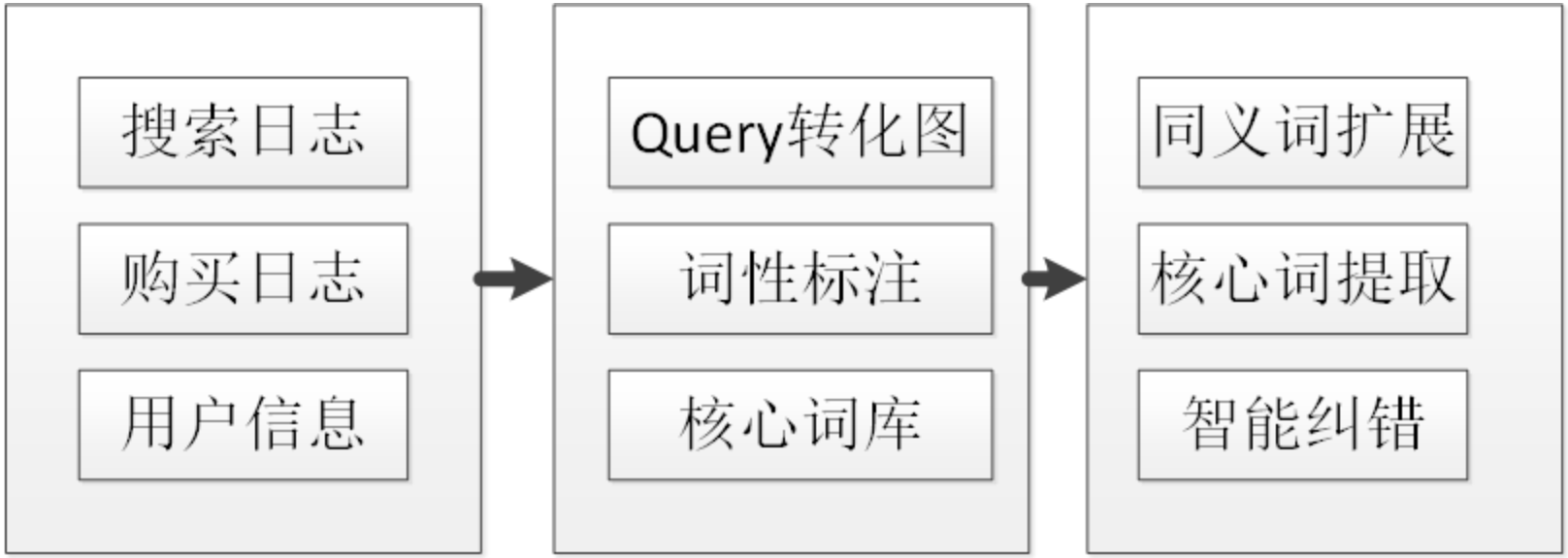


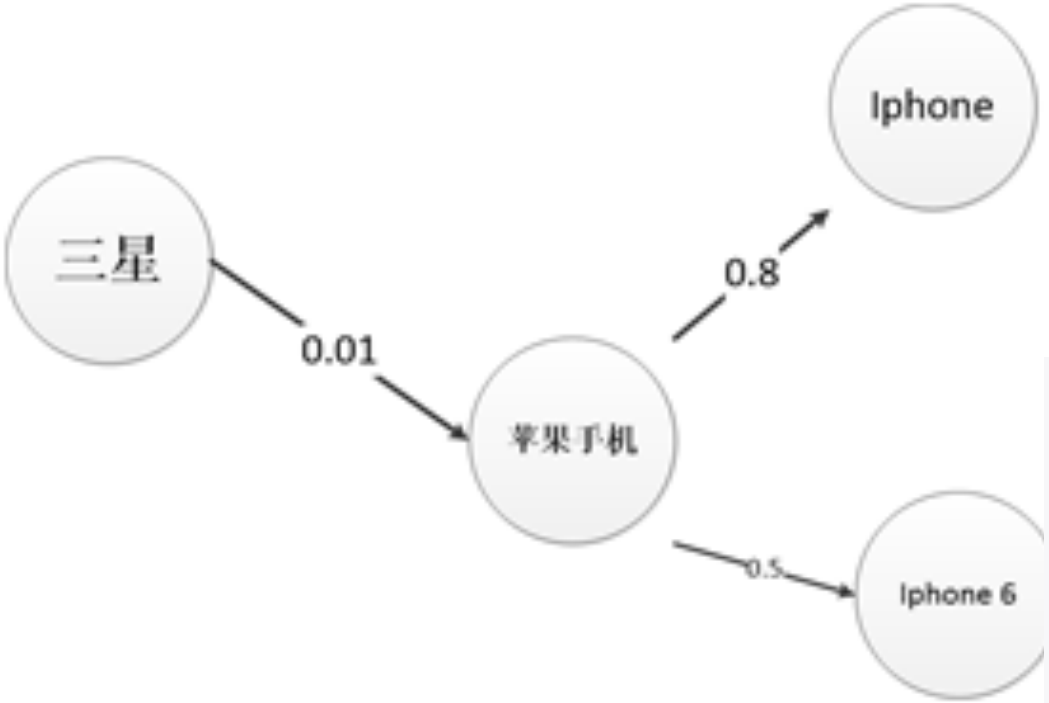
去重-店铺去重

- 著名的”分桶搜索问题”



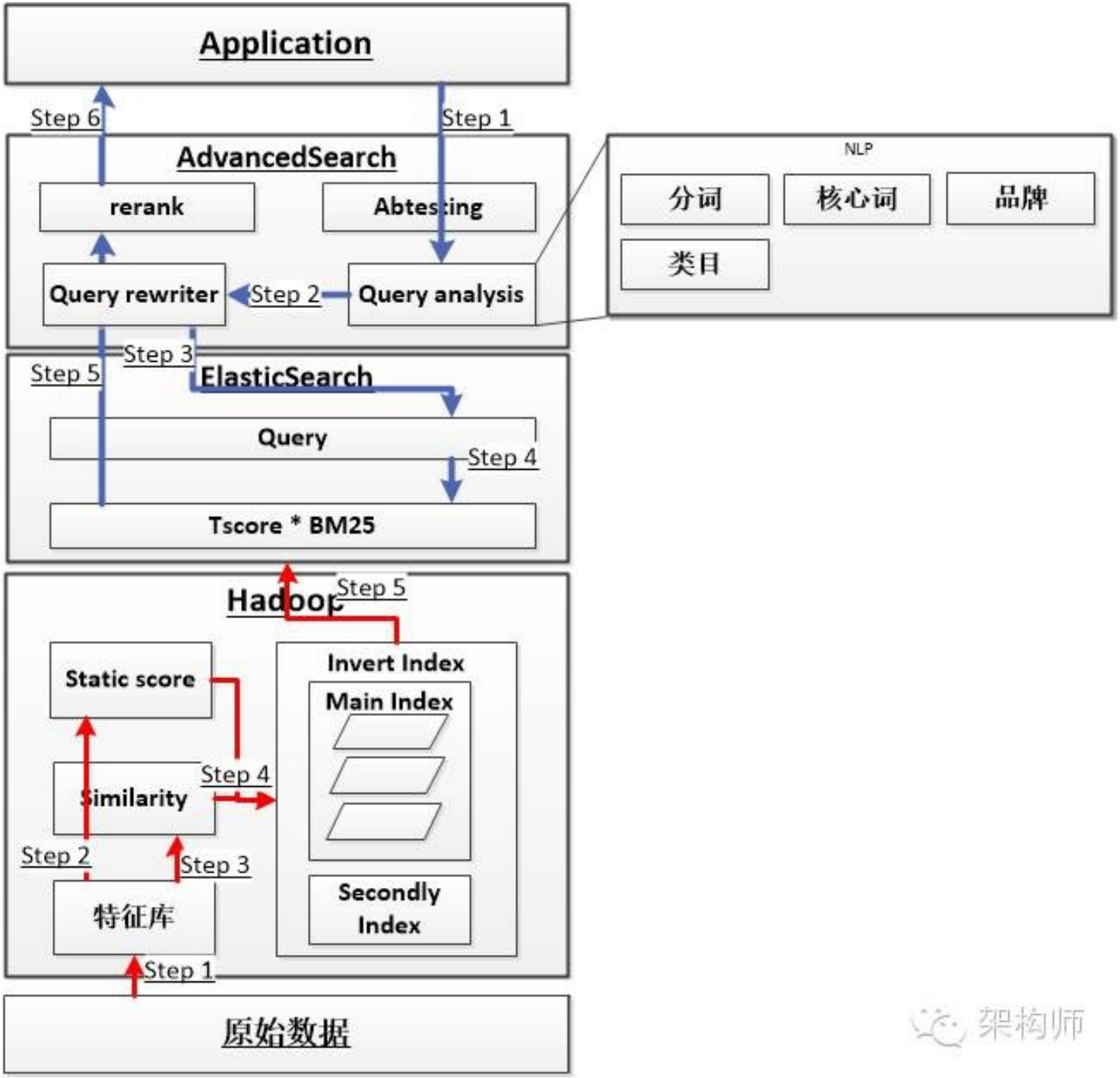
Query分析






```
{
    "query" {
        "match": {
            "query":"苹果手机"
        }
    }
}
```

```
{
  "query": {
    "should": [
        { "match": {
            "content": {
                "query": "苹果手机",
                "boost": 10
            }
        }},
        { "match": {
            "content": {
                "query": "iphone",
                "boost": 8
            }
        }},
        { "match": {
            "content": {
                "query": "iphone6",
                "boost": 5
            }
        }}
    ]
```


Application
Step 6
Step 1
AdvancedSearch
rerank
Abtesting
Query rewriter
Step 2
Query analysis
NLP
分词
核心词
品牌
类目
Step 3
Step 5
ElasticSearch
Query
Step 4
Tscore * BM25
Hadoop
Step 5
Static score
Invert Index
Main Index
Similarity
Step 4
Step 2
Step 3
特征库
Secondly Index
Step 1
原始数据

性能优化

- 应用级队列
- 自动降级
- 善用filtered query
- 其他
  - 关闭分片自动均衡
  - 尽可能延长refresh
  - 尽可能使用bulk
  - 善用rolling技术
  - 物理分离

展望

- 搜索平台化

QA&3Q