# Elasticsearch+Hbase标签体系方案

目录

**[1. 整体方案](#_Toc23679_WPSOffice_Level1)** **[2](#_Toc23679_WPSOffice_Level1)**

[1.1. 架构图](#_Toc3514_WPSOffice_Level2) [2](#_Toc3514_WPSOffice_Level2)

[1.2. 表字段概要设计](#_Toc28933_WPSOffice_Level2) [2](#_Toc28933_WPSOffice_Level2)

**[2. 技术可行性分析](#_Toc3514_WPSOffice_Level1)** **[4](#_Toc3514_WPSOffice_Level1)**

[2.1. Es特性](#_Toc303_WPSOffice_Level2) [4](#_Toc303_WPSOffice_Level2)

[2.2. 实时数据同步](#_Toc15747_WPSOffice_Level2) [4](#_Toc15747_WPSOffice_Level2)

[2.3. 离线数据同步](#_Toc12609_WPSOffice_Level2) [6](#_Toc12609_WPSOffice_Level2)

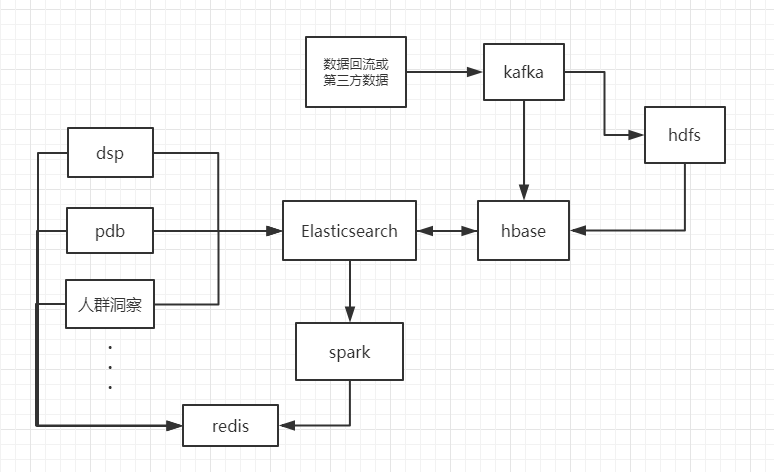
[2.4. 与solr对比](#_Toc20040_WPSOffice_Level2) [6](#_Toc20040_WPSOffice_Level2)

**[3. 注意事项](#_Toc28933_WPSOffice_Level1)** **[6](#_Toc28933_WPSOffice_Level1)**

说明：Elasticsearch简称为es

# 整体方案

## 架构图



## 表字段概要设计

### Es

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| gid | 性别 | 年龄段 | 省份 | 城市 | 操作系统 | 媒体 |  | ..... |
| 233o | 男 | 25-35 | 广东,福建 | 广州,深圳 | ios | 今日头条，腾讯视频... |  | ..... |

逗号（,）分隔多值的列按逗号分词并建索引。

### Hbase

|  |  |  |  |
| --- | --- | --- | --- |
| rowkey(deviceid) | cf:gid | cf:realValue | cf:type |
| 34455f | 233o | hhjj455b | mac |

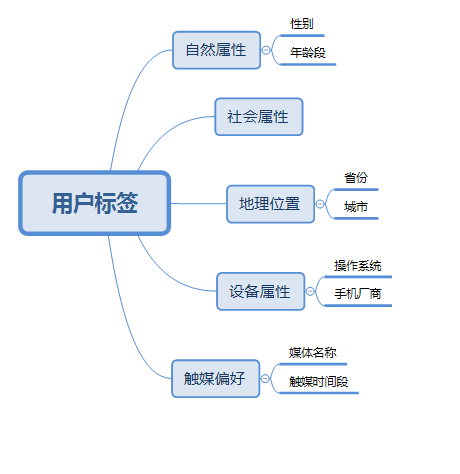
这张表是deviceId和gid的对应关系表。其中的rowkey为deviceId的加密后的值，gid保存的是gid的值，其中的type为加密用的方式，realValue列为原始值aes加密后的值，可以解密出原始值的。

|  |  |  |  |
| --- | --- | --- | --- |
| rowkey(gid) | cf:性别 | cf:年龄段 | ...... |
| 233o | 男：0.5,女：0.4 | 15-20:0.3，25-35:0.5 | ...... |

存储权重信息主表，权重低于阈值的值就不同步到es

### mysql

|  |  |
| --- | --- |
| id | 自增长ID |
| tag\_id | 标签id |
| tag\_name | 标签名 |
| pid | 父级id |
| pids | 所有的父级id |
| multiValue | 这个标签单值还是多值 |
| weight\_threshold | 权重阈值，权重超过这个阈值的值才会同步到es |
| is\_es | 是否同步到es,对叶子节点有用 |



# 技术可行性分析

## Es特性

标签系统对外暴露的服务与es支持情况

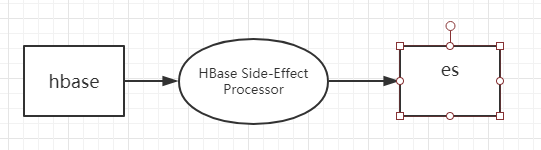
|  |  |
| --- | --- |
| **标签系统** | **es特性** |
| 标签动态添加 | 支持（动态添加列） |
| 任意字段组合查询 | 支持 |
| 个性化分词 | 支持（自定义分词器） |
| 根据id快速数据查询 | 支持 |
| 数据分析 | 支持(整合spark) |

## 实时数据同步

满足增量数据更新

### Hbase同步到es

利用HBase的Relication机制，同步数据到es。



HBase Side-Effect Processor 是Lily Hbase Indexer的一个子项目。Lily Hbase Indexer是我们当前正在使用的，用来同步hbase数据到solr。

HBase Side-Effect Processor可编程实现同步逻辑，满足当前的设计。

需要注意：HBase Side-Effect Processor这个项目需要评估可靠性。

### Kafka同步到es（推荐）

kafka里的数据写一份到hbase,再写一份到es。

行业内常规的做法，可靠且可控。

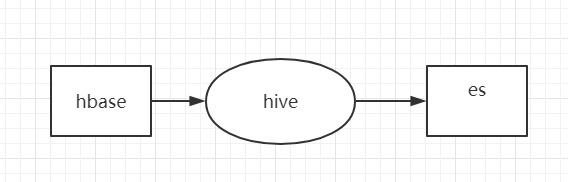
**推荐这个方法。**

### 其他方法

Hbase特性原生支持,通过Hbase协处理器可以把hbase数据实时同步到es。

缺点：风险很大。耦合太强，不好调试，同步逻辑有bug，不但会影响同步数据，还会影响写数据到hbase。

## 离线数据同步



Es支持与hive整合，方便日常维护和开发。

## 与solr对比

单纯的只用来做二级索引,Es与solr区别不大。

这里针对我们的标签系统做针对性的比较。

|  |  |  |
| --- | --- | --- |
|  | solr | es |
| 标签动态添加 | 支持，但不够灵活 | 支持,而且很灵活 |
| 与spark整合 | 官方不支持 | 官方支持 |
| 与hive整合 | 官方不支持 | 官方支持 |

可以说es为与hadoop生态整合提供完美支持。

Es与spark整合可以很高效生成目标人群包。

# 注意事项

es索引库分片数初始化问题，要根据数据增长情况分配够分片数。