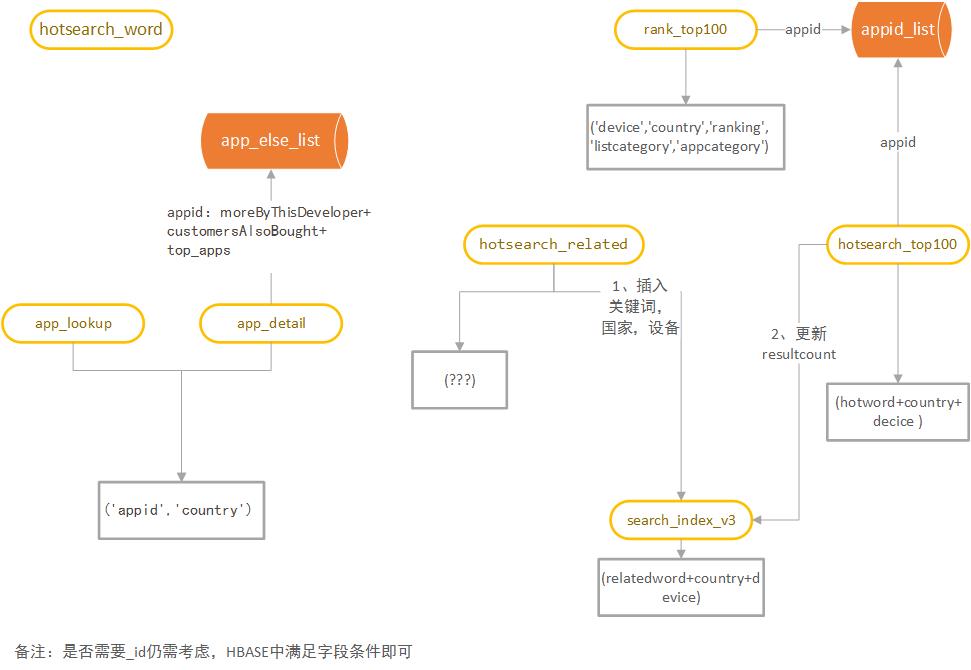
# App\_store 存储逻辑

## 总体逻辑流程图



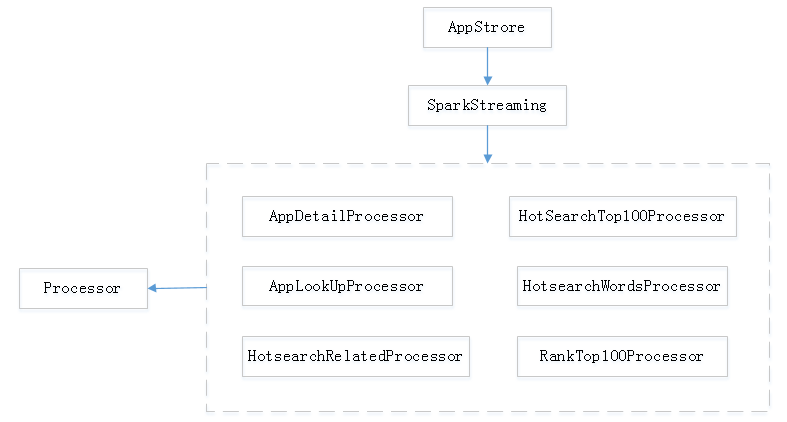
## 代码说明

1、AppStrore是SparkStreaming的入口

2、获取的DStream中kv.\_1为Topic，kv.\_2为值

3、根据Topic做一个模式匹配，通过processRDD方法调用几个主类

4、主类分别为AppDetailProcessor，AppLookUpProcessor，HotsearchRelatedProcessor，HotsearchRelatedProcessor，HotsearchWordsProcessor，RankTop100Processor



## 数据结构介绍

数据源将存储在kafka的一个topic中，通过type区分。共6种type。

（红字加粗为和其他表有关联，需要注意的字段）

### hotsearch\_words

* 数据内容

appstore的热搜词。

* 数据格式

{'ranking': 3, 'updatetime': '2017-03-06T14:15:08.501000', 'country': 'us', 'hotword': 'feeld', 'device': 'iphone', 'type': 'hotsearch\_words'}

* 存储位置

hotsearch\_words表

* 处理逻辑

无。直接存储

### hotsearch\_related

* 数据内容

appstore的热搜词对应的相关关键词。

* 数据格式

{'ranking': 3, 'updatetime': '2017-03-06T14:15:09.214000', **'searchindex'**: 4605, **'country'**: 'us', 'hotword': 'poker - texas holdem', **'relatedword'**: 'poker texas holdem online', **'device'**: 'iphone', 'type': 'hotsearch\_related'}

* 存储位置

hotsearch\_related ，search\_index

* 处理逻辑

1. 数据存入hotsearch\_related表中
2. Relatedword的值及searchindex更新到search\_index\_v3表中

### 3.hotsearch\_top100

* 数据内容：

appstore的热搜词的APP搜索结果。

* 数据格式：

{'updatetime': '2017-03-06T16:36:53.388000', 'country': 'cn', 'results': [u'1110614468', u'1110654670', u'117678340'], 'hotword': '\xe5\x8d\xa1\xe7\x89\x8c\xe7\xb1\xbb', 'device': 'iphone', **'resultcount'**: 4683, 'type': 'hotsearch\_top100'}

* 存储位置：

hotsearch\_top100，search\_index

* 处理逻辑：
  1. 存入hotsearch\_top00
  2. 更新对应的resultcount到search\_index表中

### rank\_top100

* 数据内容：

appstore的榜单信息。

* 数据格式：

{'updatetime': '2017-03-06T16:42:01.273000', 'country': 'cn', 'results': [395680819, 507056600, 1184187247, 1126403961, 907828692, 1190810262, 1193597129, 981372916, 344784375, 372047576, 621561671, 1114659672, 960436333, 662090840, 539397400, 992815164, 1060085514, 1156325747, 380090605, 1195317699, 561396899, 536108352, 1106374233, 916366645, 1100139790, 1155918748, 899441948, 1162675136, 1163603649, 627446343, 1070068355, 1012274888, 1107501843, 1168560959], 'type': 'rank\_top100', 'device': 'iphone', 'listcategory': '\xe4\xbb\x98\xe8\xb4\xb9', 'appcategory': 6016}

* 存储位置：

Rank\_top100

* 处理逻辑：

1. 存入rank\_top100的数据：除了key为data以外的所有

### search\_index\_v3

* 数据内容：

搜集所有关键词的去重列表

* 数据格式：

数据源来源 type为：hotsearch\_related，hotsearch\_top100

* 存储位置：

Search\_index\_v3表

* 处理逻辑：

1. 去重逻辑：relatedword+country+device
2. Hotsearch\_related中：

{'updatetime': item['updatetime'],

"searchindex": item['searchindex'],

"keyword": item['relatedword'],

"country": item['country'],

"device": item['device'],

"resultcount": 0}

1. hotsearch\_top100中：更新resultcount的值。

### app\_detail

* 数据内容：

Appstore每个app的具体信息

* 数据格式：

见附件：app\_detail.json（1,2,3都有可能）type：app\_detail

* 存储位置：

App\_detail表

* 处理逻辑：

1. 去重生成逻辑：appid+country
2. Json数据源data中，data['data']下的数据，清洗后插入/更新到app\_detail

### app\_lookup

* 数据内容：

Appstore每个app的信息（与app\_detail相似）

* 数据格式：

见附件：app\_detail\_lookup type：app\_lookup

* 存储位置：

App\_detail表

* 处理逻辑：

1. 去重逻辑：appid+country
2. Json数据源data中，data['data']下的数据，清洗后插入/更新到app\_detail

（事实上app\_lookup内容和app\_detail内容一致，单列出来是因为两者相同内容字段名key有变化，方便处理）

## 表结构列表及定义

见附件：aso项目中基础表-V1- （或者待完成）

## 字段说明

|  |  |
| --- | --- |
| 字段名 | 字段意义 |
| updatetime | 更新时间 |
| hotword | 实时热搜词 |
| device | 设备（iphone,ipad) |
| country | 国家 |
| listcategory | 榜单类型，如免费，付费等 |
| appcategory | 应用类别，如游戏等 |
| keyword | 关键词 |
| searchindex | 搜索指数 priority |
| resultcount | 搜索结果数 |
| ranking | 排名 |
| searchtime | 查询日期 |
| relatedword | 相关搜索词 |
| appname | app名字 |
| app\_url | APP 的url |
| appimg | applogo |
| companyname | app开发商 |
| appid | appid |
| category | 分类 |
| version | 版本 |
| language | 语言 |
| appintro | app简介 |
| versionHistory | app历史版本 |
| rid | elasticsearch唯一uid标识id,搜索用 |
| art\_url | app详情页面 |
| artistName | 作者名 |
| valueCurrentVersion | 当前版本评分 |
| value | 总体评分 |
| ratingCountCurrentVersion | 当前版本评分数 |
| ratingCount | 总体评分数 |
| screenshots | app界面截图 |
| genreNames | 类别（文字表示） |
| price | 价格 |
| copyright | 版权 |
| purchases | 热门app内购买项目 |
| requirements | 兼容 |
| contentRatingsBySystem | 评级 |
| customersAlsoBought | 顾客还买了 |
| familyShareEnabledDate | 是否家庭共享 |
| moreByThisDeveloper | 部分开发人员的其他app |
| top\_apps | 相关app排名 |

## 类别映射关系

GENRE = {36: u'总榜', 6000: u'商务', 6001: u'天气', 6002: u'工具', 6003: u'旅游', 6004: u'体育', 6005: u'社交', 6006: u'参考', 6007: u'效率', 6008: u'摄影与录像', 6009: u'新闻', 6010: u'导航', 6011: u'音乐', 6012: u'生活', 6013: u'健康健美', 6014: u'游戏', 6015: u'财务', 6016: u'娱乐', 6017: u'教育', 6018: u'图书', 6020: u'医疗', 6021: u'报刊杂志', 6022: u'商品指南', 6023: u'美食佳饮', 6024: u'购物', 7001: u"动作游戏", 7002: u"探险游戏", 7003: u"街机游戏", 7004: u"桌面游戏", 7005: u"扑克牌游戏", 7006: u"娱乐场游戏", 7007: u"骰子游戏",7008: u"教育游戏", 7009: u"家庭游戏", 7011: u"音乐", 7012: u"智力游戏", 7013: u"赛车游戏", 7014: u"角色扮演游戏", 7015: u"模拟游戏",7016: u"体育", 7017: u"策略游戏", 7018: u"小游戏", 7019: u"文字游戏",13001: u"新闻与政治", 13002: u"流行与时尚", 13003: u"家居与园艺", 13004: u"户外与自然", 13005: u"运动与休闲", 13006: u"汽车",13007: u"艺术与摄影", 13008: u"新娘与婚礼", 13009: u"商务与投资", 13010: u"儿童杂志", 13011: u"电脑与网络", 13012: u"烹饪与饮食",13013: u"手工艺与爱好", 13014: u"电子产品与音响", 13015: u"娱乐", 13017: u"心理与生理", 13018: u"历史", 13019: u"文学杂志与期刊",13020: u"男士兴趣", 13021: u"电影与音乐", 13023: u"子女教养与家庭", 13024: u"宠物", 13025: u"职业与技能", 13026: u"地方新闻",13027: u"科学", 13028: u"青少年", 13029: u"旅游与地域", 13030: u"女士兴趣"}

# 25个总类别。游戏子类：18个，报刊杂志子类：28个