

# 标签交流分享会

数据运营部

2017.6.19

2 打标签流程

3 标签结果处理

4 标签运营优化迭代情况

#### 维度

#### 5维

- **▶ URL标签**
- **➢ APP标签**
- **➢ SITE标签**
- ▶ 临时标签
- > 关键字标签

#### 行业

#### 8大类

- ▶ 基础(国信)
- 金融(亚信)
- > 房产(亚信)
- 汽车(泰岳)
  - 电商(数据研发)
- 区域房产
- ▶ 微信
- > 旅游

# 统计

#### URL:

- 23个一级分类;
- **188个二级标签**;
- 1277个三级标签;
- 1944个四级标签;
- > 299个五级标签;
- ▶ 52个六级标签;

#### APP:

3648个APP;

#### SITE:

> 84260网站

## 体系建设参考

- 市场部红头文件;
- > 各厂家标签体系;



兴趣标签体系及标 码--内部使用(5月



APP标签体系及标 码--内部使用(5月



Site标签体系及编 内部使用版(5月份

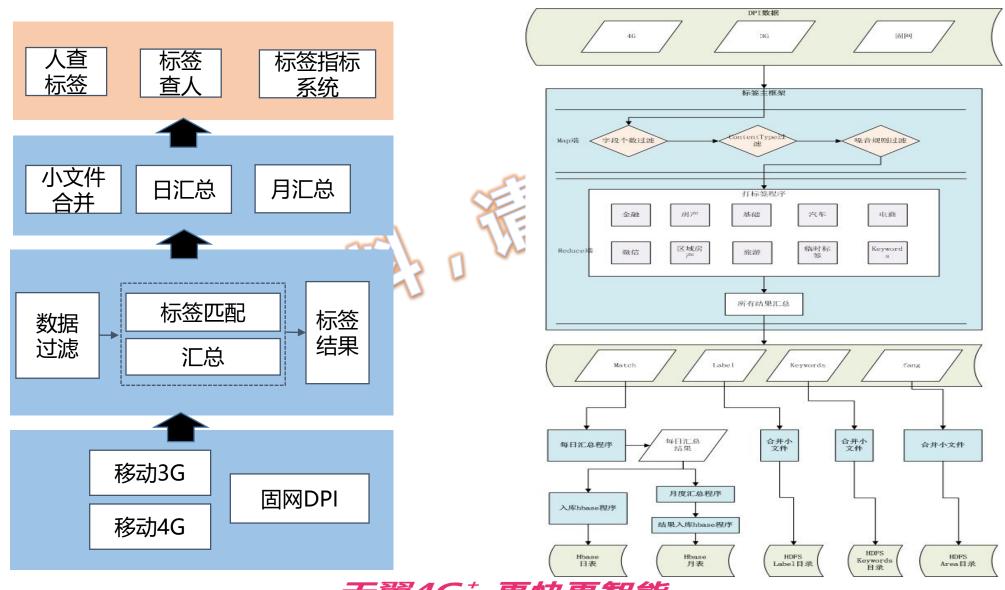
#### 天翼4G<sup>+</sup> 更快更智能

2 打标签流程

3 标签结果处理

4 标签运营优化迭代情况

# 标签架构&流程图

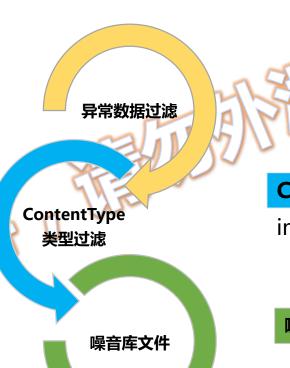


天翼4G<sup>+</sup> 更快更智能

# 数据过滤准则

#### 异常数据过滤

- 口 字段不够
  - ·移动DPI:38、40字段;
  - ·固网DPI:12字段
- □ URL长度异常
  - ·长度小于9
- 口 主标识为空
- · 移动DPI: mdn、imei/meid、imsi字 段全部为空;
  - ·固网DPI: useraccount字段为空;



#### ContentType类型过滤

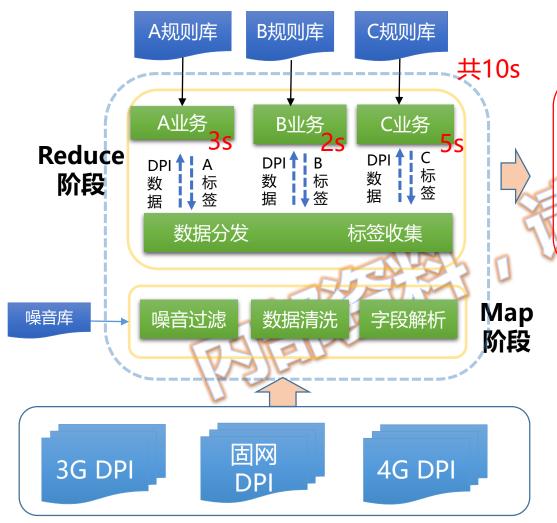
image、application等7种类型;

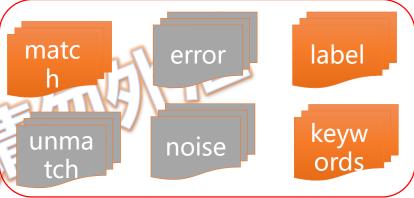
#### 噪音库文件过滤

.jpg、.gif、.jpeg等51种后缀;



# 打标签匹配流程







#### 匹配结果

框架程序通过读取DPI数据并对数据进行过滤,去噪音处理,再把数据依次传给A、B、C标签程序,A、B、C标签程序通过对传进来的DPI数据的URL分别进行比对匹配,最终将结果发给框架程序,由框架程序汇总后输出到HDFS;

~== WINKING UM

2 打标签流程

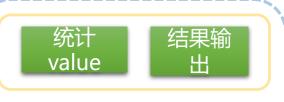
THE THE

- 3 标签结果处理
- 4 标签运营优化迭代情况

# 标签汇总流程

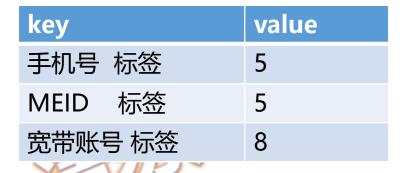
Reduce 阶段

Map阶段



合规性 判断







# 日汇总

标签汇总程序通过读取框架程序输出的标签结果,并对结果数据进行汇总处理,汇总每个用户的每个标签数,汇总完后输出到HDFS。

月汇总

牛顿冷却算法



月度标签权值算法 模型.docx

**3G**: 手机号|MEID|......|标签A|标签B|标签C|标签D|...... **4G**: 手机号|IMEI|......|标签A|标签B|标签C|标签D|......

**固网**: 宽带账号|.....|标签A|标签B|标签C|标签D|.....

# 标签入库流程



| key     | column | value |
|---------|--------|-------|
| 手机号     | 标签     | 5     |
| MEID    | 标签     | 5     |
| 宽带账号    | 标签     | 8     |
| M+省份+标签 | 手机号    | 5     |
| D+省份+标签 | MEID   | 5     |
| 省份+标签   | 宽带账号   | 8     |
|         |        |       |

用户查标签

标签查用户



| key     | value |
|---------|-------|
| 手机号 标签  | 5     |
| MEID 标签 | 5     |
| 宽带账号 标签 | 8     |

标签汇总结果

标签入库程序对汇总后的数据进行处理,然后按照:

- 手机号、标签、标签次数;
- meid号、标签、标签次数;
- 标签、手机号、标签次数;
- 标签、meid号、标签次数;

方式存储到Hbase中供其他用户查询使用。

天翼4G<sup>+</sup> 更快更智能

~== WINKING UM

- 2 打标签流程
- 3 标签结果处理

4 标签运营优化迭代情况

# 标签运营优化迭代情况

#### 口 迭代优化

业务部门需 关键字标签 主框架数据过滤 标签指标系 临时标签多 标签体系覆 承接标 求承接; bug多轮修改; 轮优化; 优化; 盖优化; 统构建; 签项目 **2016**0620 20170310 热门APP优 各家打标签 补数、汇总合 APP标签体 ·标签项目了解; jar包分离; 化; ·业务部门问题、 流程梳理; 并bug修正 系优化思考; 规则库常规更新; 需求梳理;

#### 口 主要成果

- 数据过滤占比
  - · 90% -> 75% -> 70%;
- **卜 标签体系覆盖率** 
  - · Site: 4%~99.8%;
- > 临时标签优化
  - ·对比测试,提高5倍匹配率;

#### > 关键字标签

- ·分隔符bug发现,修改,供客户正常使用;
- **▶** 热门APP优化
  - ·按照类别进行top20的持续优化;
- > 标签指标体系
  - ·多维度评估打标签质量;



天翼4G \* 更快更智能

# 标签现有主要问题及下一步方向

#### 口 现有主要问题及下一步方向

> 资源不足

·数据量增加,业务增加,但资源不变;

解决方案:资源申请;主框架业务剥离;

> 金融表标签APP规则不准

解决方案:协调厂商重新维护;

➢ 金融APP打标签逻辑错误

解决方案:修正逻辑错误;

> 房产打标签不准,无优化

·抓包工具不好用,所打标签基本不可用,需要重建;

解决方案:与区域房产合并,厂商手工维护;

> APP标签体系不统一

·历史原因,APP无体系,各厂商自行维护,纵观全局,非常凌乱;

解决方案:制定APP标签体系,将现有APP重新整理;

气车标签黑盒解不开,且无APP

· 厂家自己维护的标签库与提供给云公司的标签库不一致;

· 实现逻辑与实际业务对不上;

解决方案:问题不解决,不验收或者作为遗留问题;

电商标签基本无人维护

· 项目结束,无运营合同;

解决方案:电商https率高,不予维护;或可安排专人学习和维护或者重建;

# 训训