全民k歌黑产对抗之路

--安全体系架构与技术

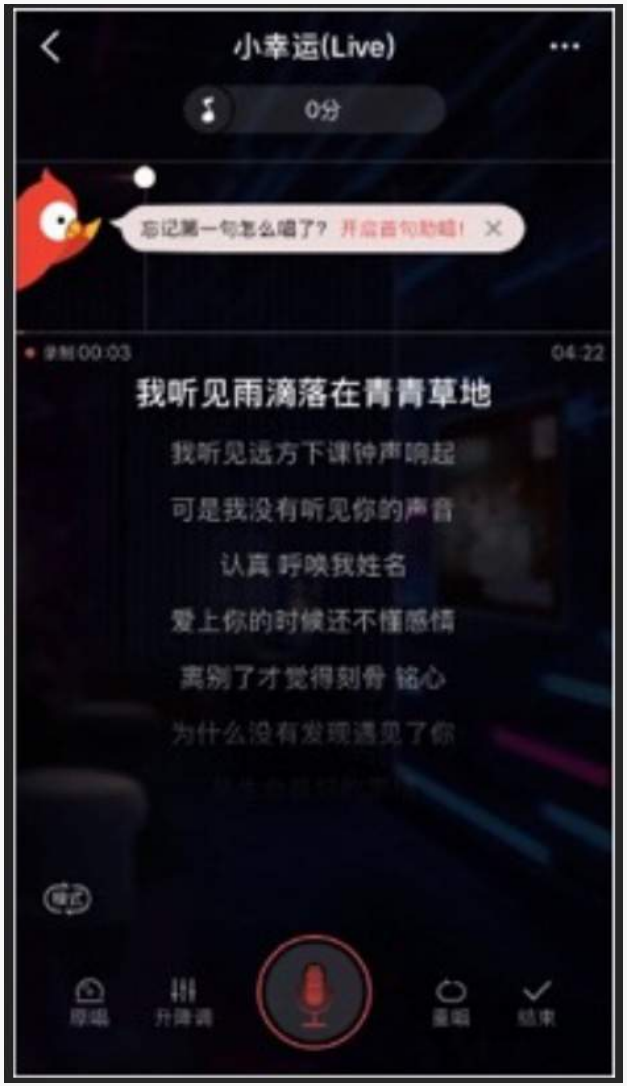
腾讯音乐 chrisluo(罗静)

基础安全-业务安全

全民K歌

全民k歌-唱

全民K歌


小幸运(Live)
0分
忘记第一句怎么唱了？ 开启首句助唱！
04:22
我听见雨滴落在青青草地
我听见远方下课钟声响起
可是我没有听见你的声音
认真 呼唤我姓名
爱上你的时候还不懂感情
离别了才觉得刻骨 铭心
为什么没有发现遇见了你
原唱
升降调
重唱
结束


小幸运(Live)
0分
04:22
我听见雨滴落在青青草地
原唱
升降调
重唱
结束


x14

全民k歌-听

全民K歌


关注 好友 热门 附近
最后一次
送礼 评论 转发


父爱如山，愿岁月长留
全民K歌
父爱如山
愿岁月长留
小时候，他总是我们坚实的后背，保护我们，细心呵护我们。而今天，他们正在老去，我们所希望的仅仅是愿岁月长留...
播放全部
天亮了
K歌
如果可以，我愿为他建造一个美丽花园
单车
K歌
重新编曲的单车，让你重温经典。


热门歌单
沁人心肺,一秒爱上
欢快追爱曲
众里寻声千百度
欧美|CHANGE
Good night

全民k歌-看

全民K歌


18:01
35%
00:06
Rita黄沙源
每次唱田馥甄的歌都会有特别多的感慨，希望你们能喜欢,比心~
收录此作品的歌单


进入房间
女主角


BELL
4:21 PM
22%
短视频
全部
推荐
附近
与我无关
shakila
每次和你走在路上都会看…
一次就好
叶儿首爱
奇妙能力歌

安全形势

全民K歌

接入
- DDoS攻击
- Xss注入
- Sql注入
- Csrf攻击
- DNS篡改

账号
- 恶意注册
- 恶意刷粉
- 私信骚扰
- 刷等级
- 盗取账号

文字
- 灌水
- 人身攻击
- 广告
- 谩骂

图片
- 色情
- 广告
- 招嫖
- 涉政
- 涉恐

音频
- 涉政
- 涉恐
- 传销
- 涉黄
- 诈骗

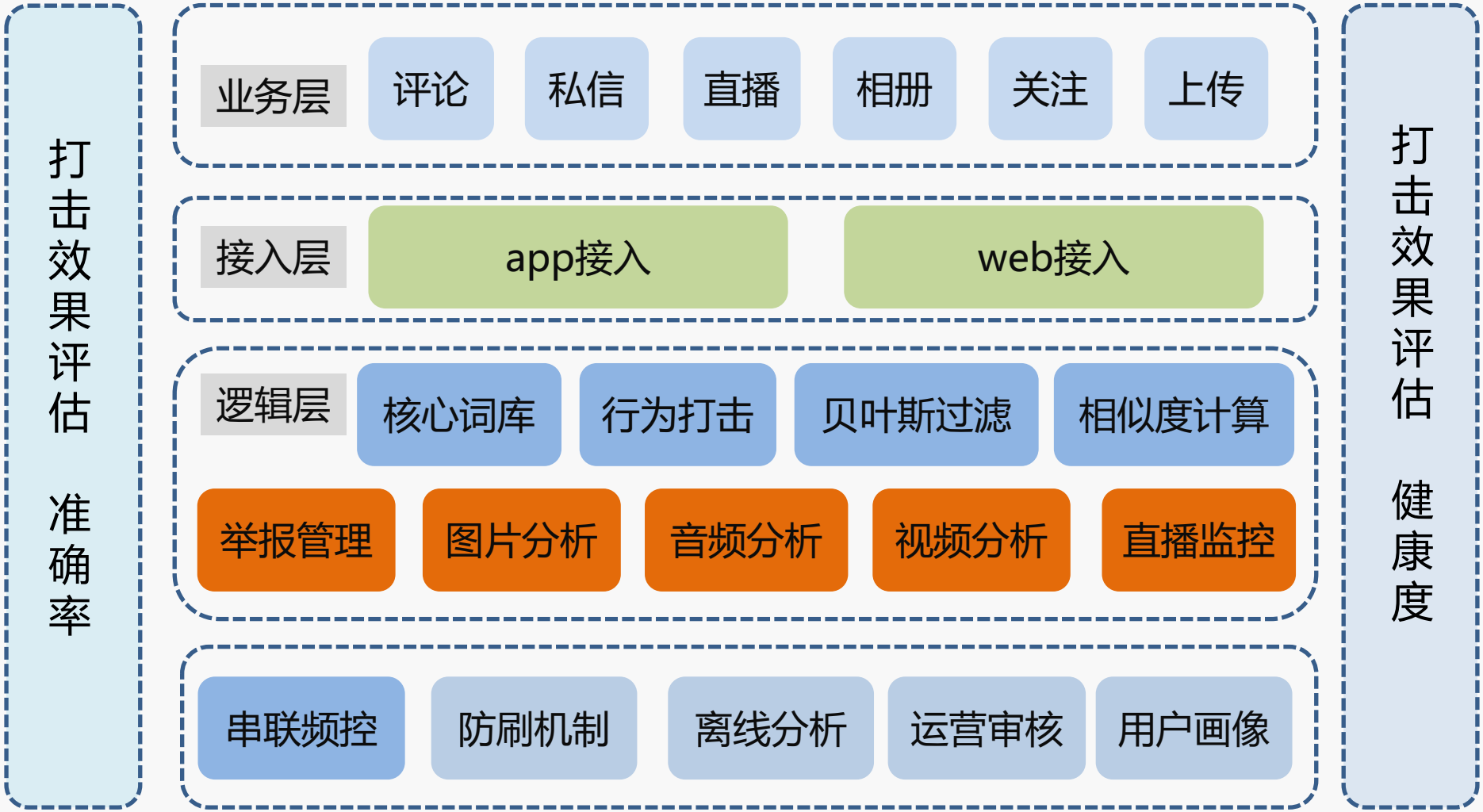
视频
- 涉政
- 涉恐
- 传销
- 涉黄
- 诈骗

直播
- 涉政
- 涉恐
- 传销
- 涉黄
- 人身攻击

活动
- 恶意刷榜
- 薅羊毛
- 盗取作品
- 盗取资料

安全架构

全民K歌

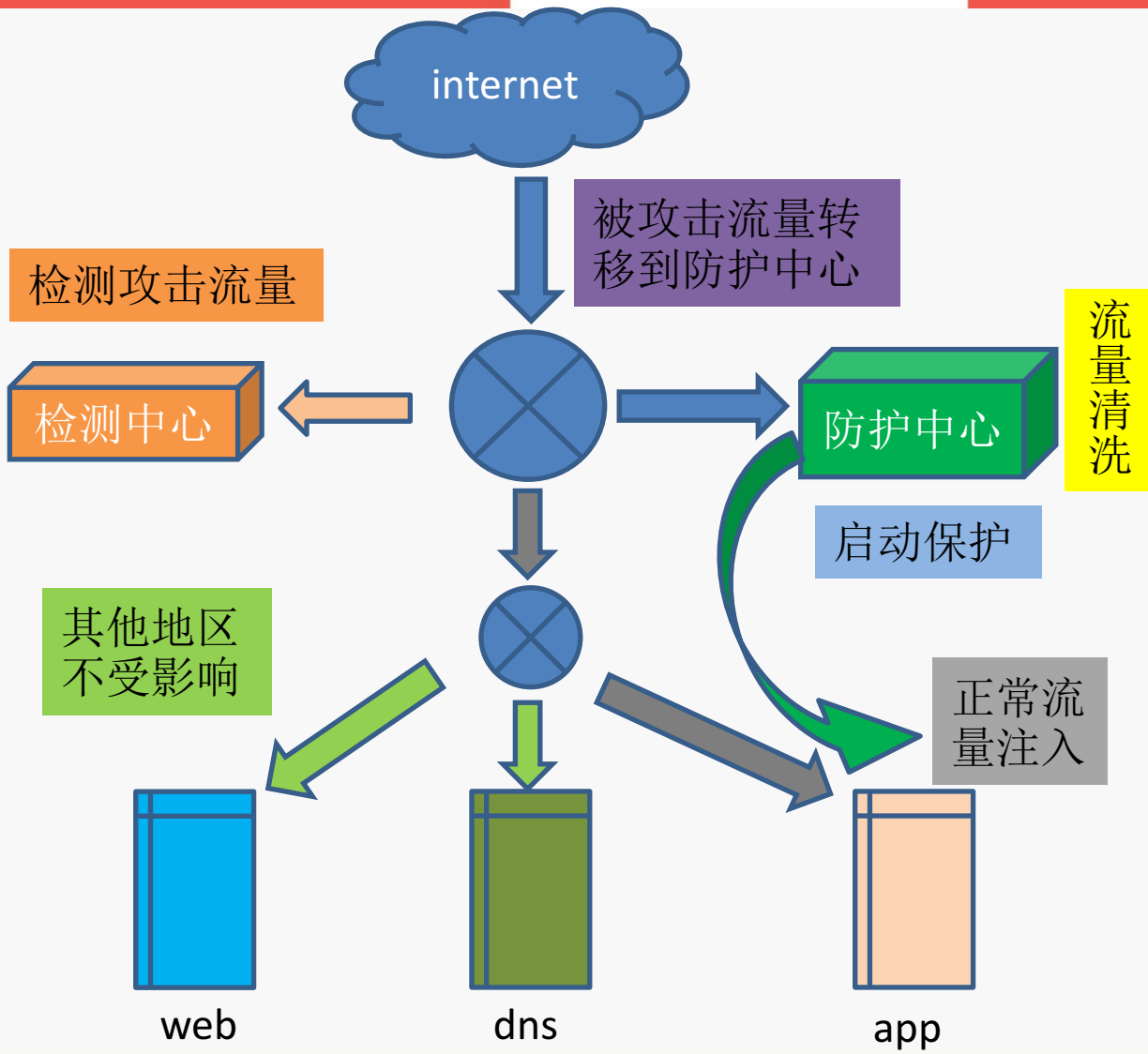

打击效果评估 准确率
业务层
评论
私信
直播
相册
关注
上传
接入层
app接入
web接入
逻辑层
核心词库
行为打击
贝叶斯过滤
相似度计算
举报管理
图片分析
音频分析
视频分析
直播监控
串联频控
防刷机制
离线分析
运营审核
用户画像
打击效果评估 健康度

接入安全-DDoS保护

全民K歌

应用策略：

- DPI检测技术，快速准确地发现针对业务的各种DDoS攻击
- 采用运营商黑洞路由、外网核心ACL、专业清洗设备等多种手段，形成多层级的防护架构
- 防护带宽2T，部署CDN100+，全网调度对抗攻击流量

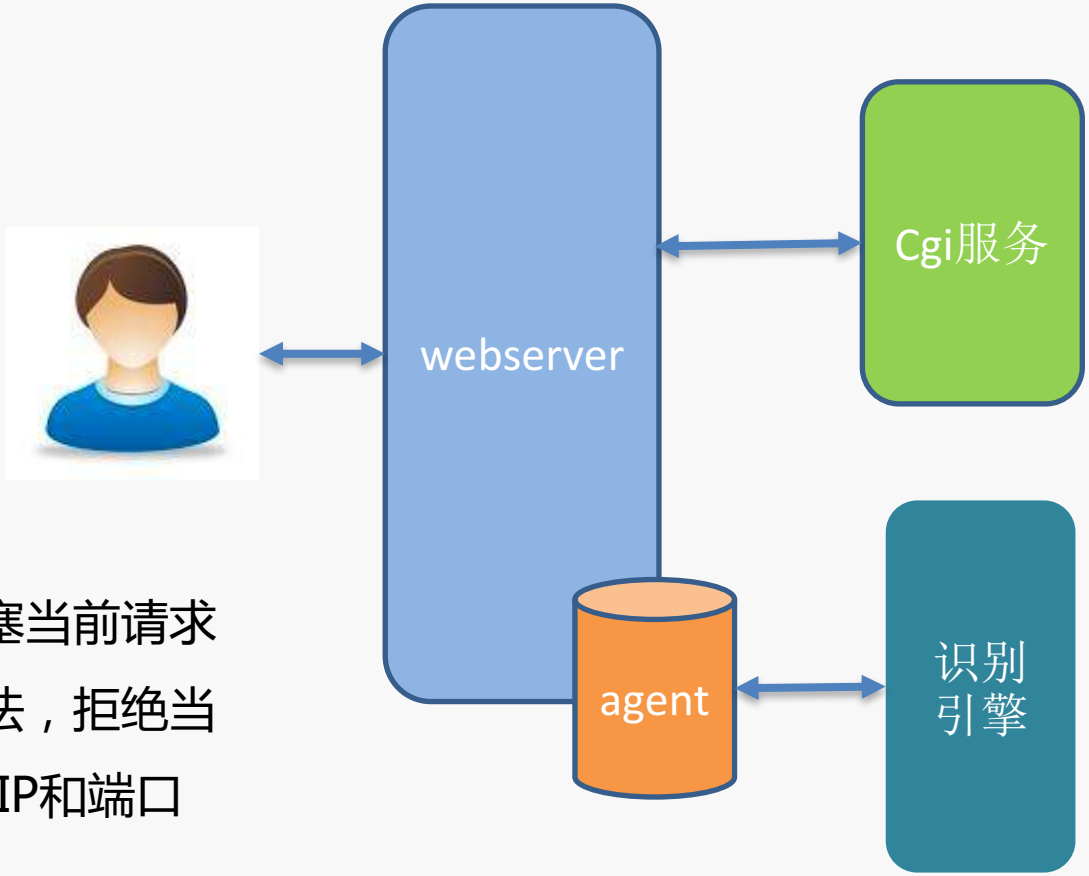


接入安全-web防护

全民K歌

Cgi安全问题扫描：

- 1、开发提交安全扫描
- 2、测试环境自动扫描
- 3、线上服务安全防护
- 4、发现漏洞提交安全工单



sql注入、xss、csrf检测：

- 接受请求，转发到检测服务，阻塞当前请求
- 检测服务分析恶意程度，如果非法，拒绝当前请求，合法则返回后端cgi机器IP和端口
- Cgi处理后返回正常数据

文字策略-关键词

全民K歌

应用策略：

- 组合关键词
- 关键词划分等级
- 过滤转义，把全角、异体转为标准内容再进行匹配
- 小语种识别，对含有高危小语种的内容单独处置
- 中文转拼音，对高危内容进行同音识别

文字策略-相似度

全民K歌

文本的基本元素是词汇

比较算法：

- Jaccard相似度
- Simhash-汉明距离
- 余弦相似度

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{A \cdot B}{|A| \times |B|}$$

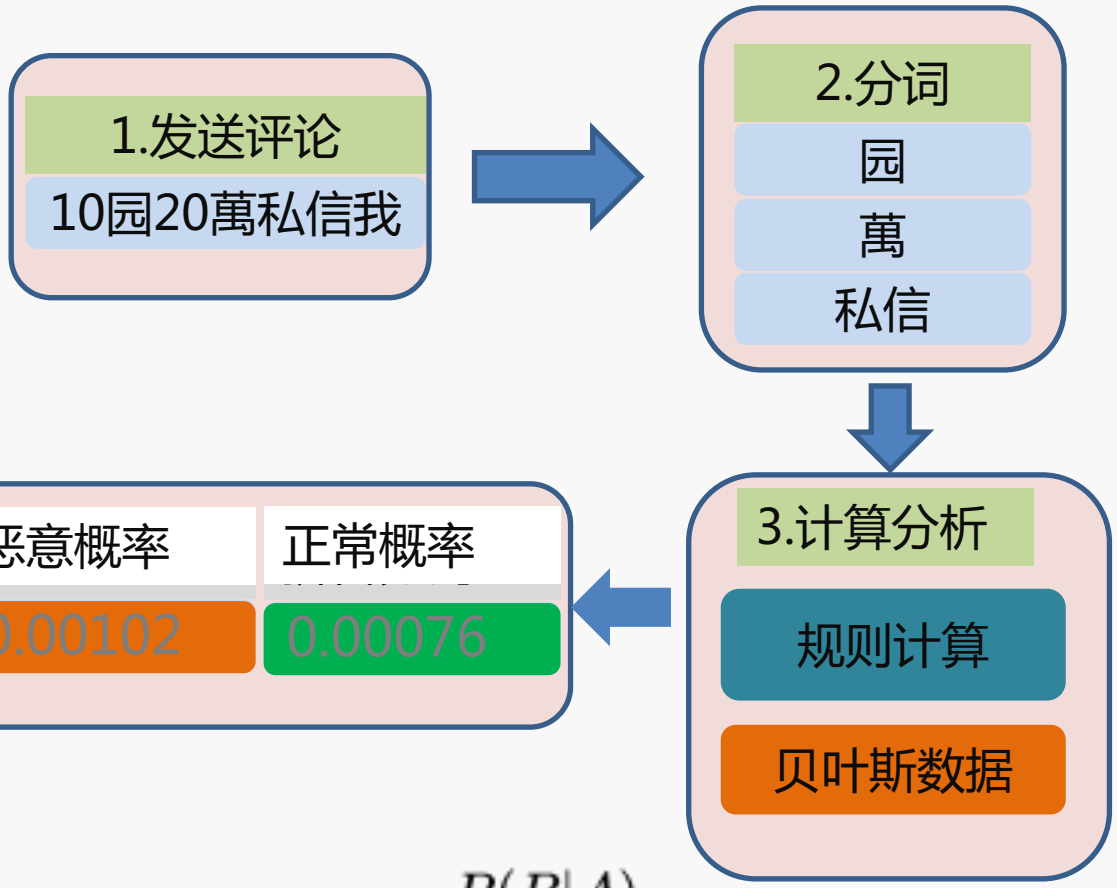
策略：

- 单个作品下的评论内容分析
- 同一个用户发出的评论分析

文字策略-贝叶斯

全民K歌

文本	分类
喜欢 唱歌 私信	正常
聊天 找我 私信	正常
元 萬 加 私信	恶意
10园20萬私信	?



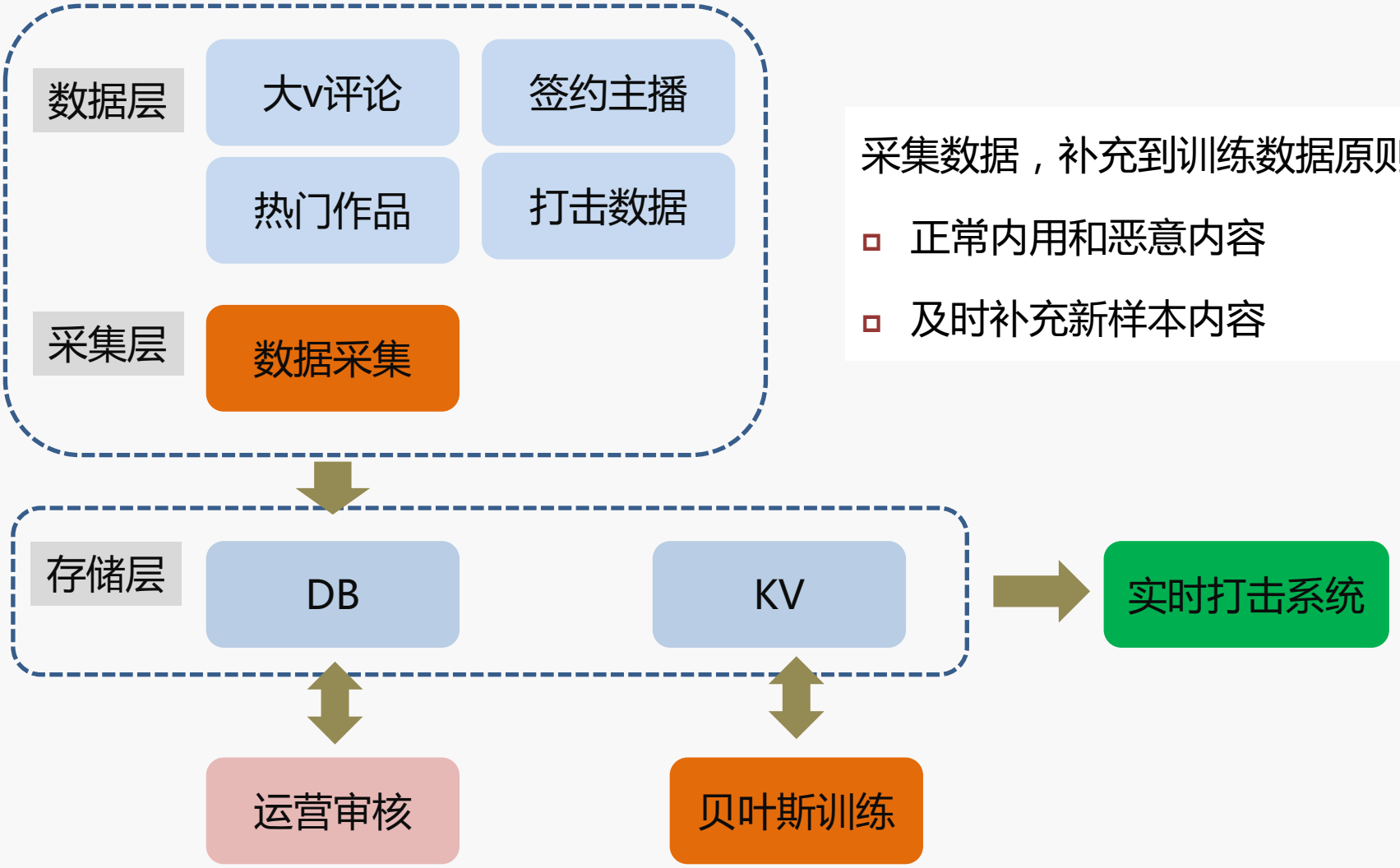
应用场景：

- 支持针对具体业务的训练库
- 算法不仅支持文本文类，还可以用于其他分类场景
- 可以结合业务加入其他纬度数据

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

文字策略-贝叶斯

全民K歌



采集数据，补充到训练数据原则：

- 正常内用和恶意内容
- 及时补充新样本内容

图片策略

全民K歌

行为策略：

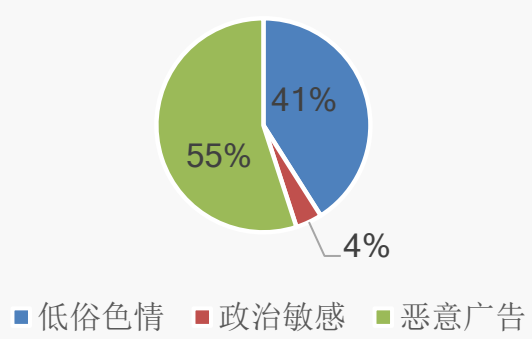
账号体系、行为分析

图像识别：

图像DNA、OCR、色情识别

人工审核：

图片审核、种子库运营



图片策略

全民K歌

热门黄图特点：

- 有组织的人肉作案，量相对小
- 容易对抗和绕过，不汇聚
- 图片自动检测容易误打击

打击方案：

- 人工确认加历史数据用作决策树训练
- 行为特征分析加上帐号特征
- 图片出现频率和相似度计算
- 图片文字率以及图片ocr识别
- 昵称贝叶斯聚类分析
- 曝光率异常分析
- 自动打击加人工审核

音频策略

全民K歌

恶意检测算法：

- 重点监控人物提取音频指纹，进行指纹比对分析
- 声音场景识别，分析在唱歌或说话
- 声音内容识别，是否小语种，特征片段匹配等

天明先生的真情分享！

0 28

天明先生触动心灵的经典语

1 18

富人与穷人的思维观念

3 21

WXB使云家人享天伦之乐

0 30

ZJZB发展春天正式到来

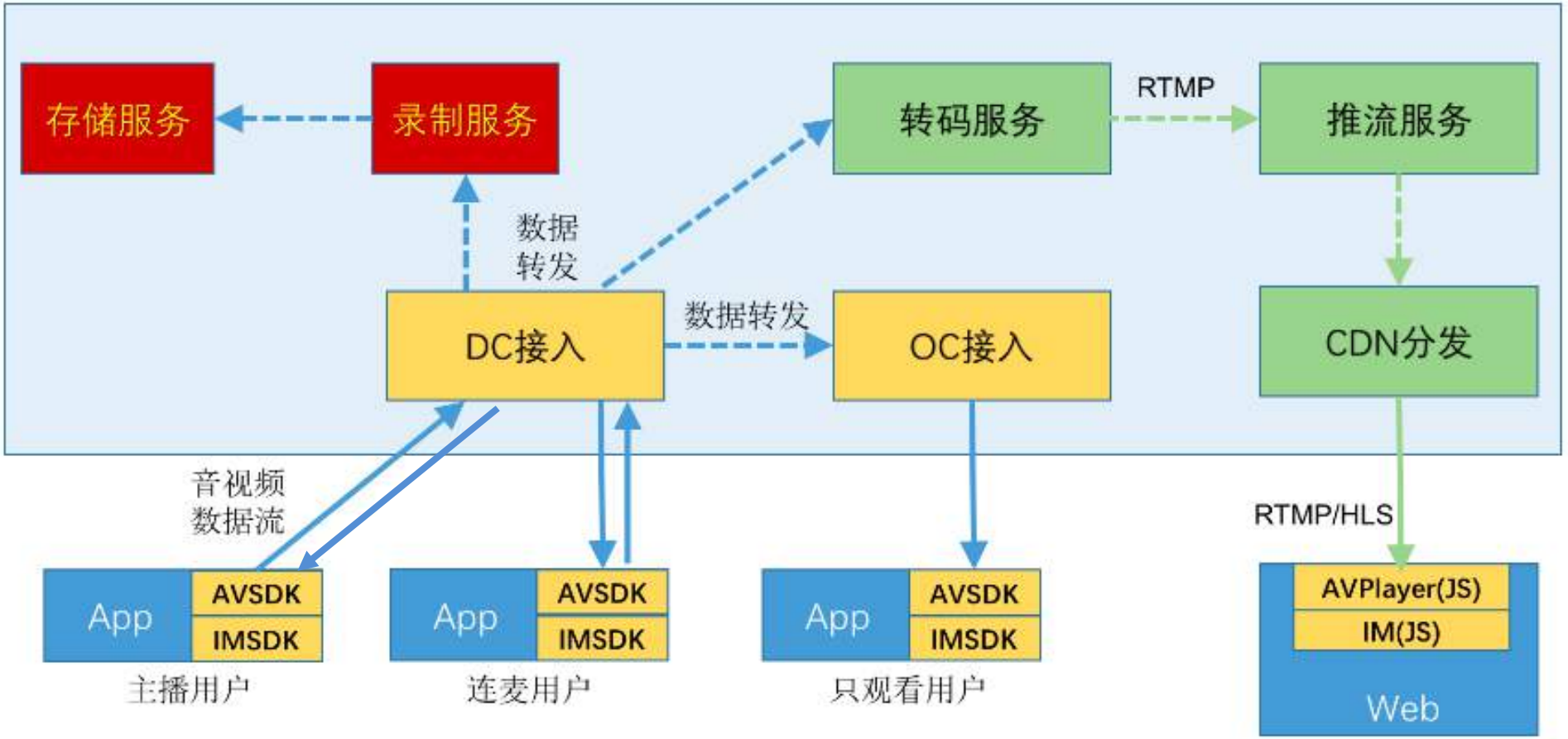
0 48

17跟老大疯一年什么都有

0 104

视频直播

全民K歌


存储服务
录制服务
转码服务
RTMP
推流服务
数据
转发
DC接入
数据转发
OC接入
CDN分发
音视频
数据流
RTMP/HLS
App
AVSDK
IMSDK
主播用户
App
AVSDK
IMSDK
连麦用户
App
AVSDK
IMSDK
只观看用户
AVPlayer(JS)
IM(JS)
Web

直播监控

全民K歌

1. 直播流旁路推流，可以实时查看直播情况
2. 对直播流进行截图，按图片的敏感度打分提交人工审核
3. 对主播进行实名认证
4. 对问题主播进行警告，随时切断直播信号，严重者进行封号

防刷-频控策略

全民K歌

实现方案：

- 1、结合账号特征和用户画像，精细化管理
- 2、实时监控，报表输出，动态调整
- 3、命中频控用户需要进行短信或图片码验证
- 4、流水查询，可回溯分析

防刷-设备IP分析

全民K歌

IP画像：

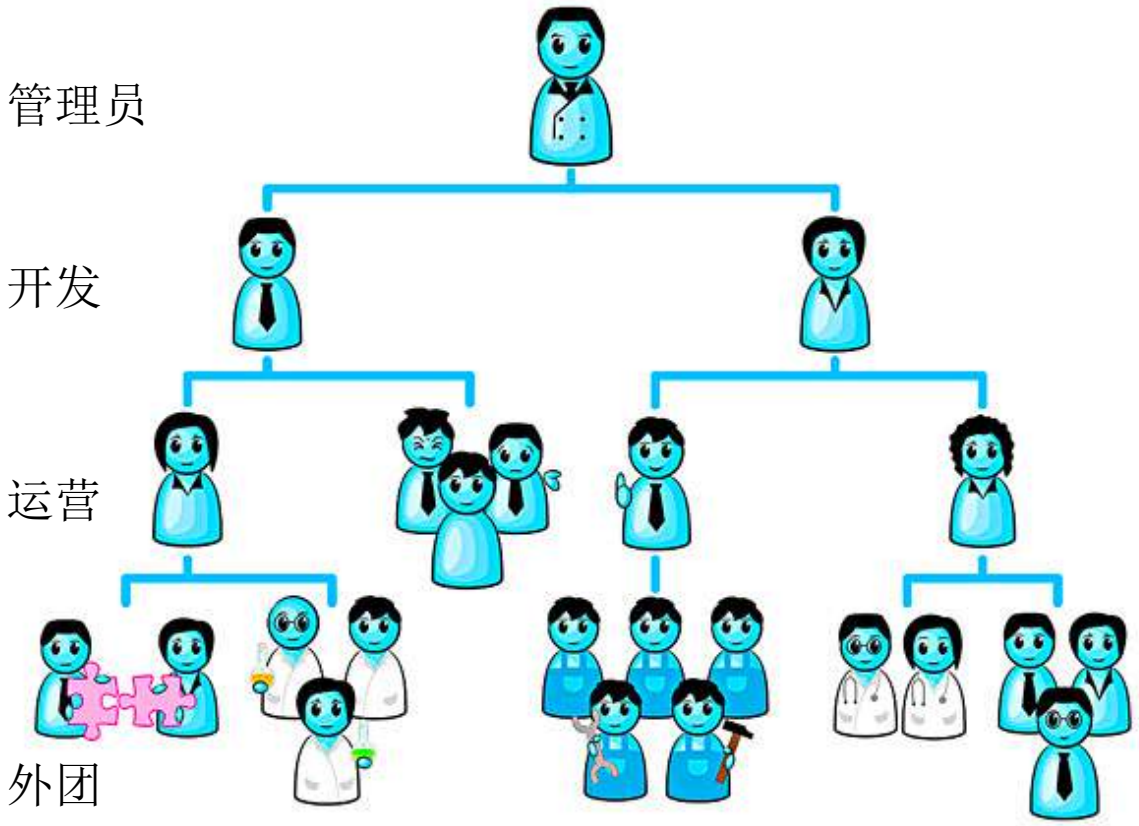
- 基于海量用户分析用户行为，提供ip信用评级
- Ip信息分类：代理/vpn，idc服务器、网关、腾讯用户IP、运营商、局域网
- 应用场景：恶意爬虫、恶意注册/登录、机器人识别、恶意刷单、刷量

设备画像：

- 后台计算确认移动终端唯一身份
- 基于海量用户从设备活跃度、模拟设备、新增设备、常用设备、黑产设备对移动设备进行安全画像
- 基于画像实时判断移动设备风险等级

举报管理

全民K歌



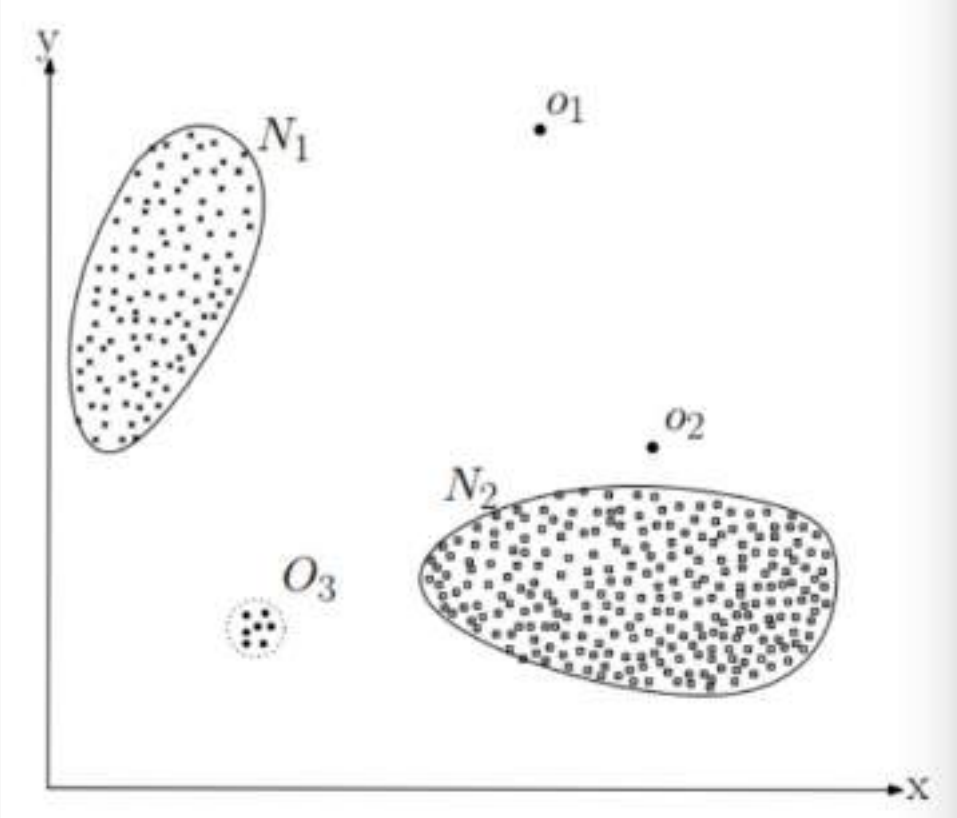
- 分等级授权、不同等级对应不同操作权限
- 综合用户等级、画像、举报次数等排序后审核

离线分析

全民K歌

异常检测算法：

- 基于已经标记的大数据特征样本
- 适合大数据，并行处理方便
- 可解释，方便问题回溯



实现方案：

- 标记设备、用户画像等数据
- 行为数据上报到HDFS，通过Hive/Sparksql跑出可疑用户
- 分析数据的聚集度，排序后抽样确认

审核平台

全民K歌

直播监控平台 | 举报审核平台 | 黄图审核

Mv审核平台 | Top行为审核 | 小语种内容审核

清唱作品审核 | 直播监控黑名单 | 实名审核

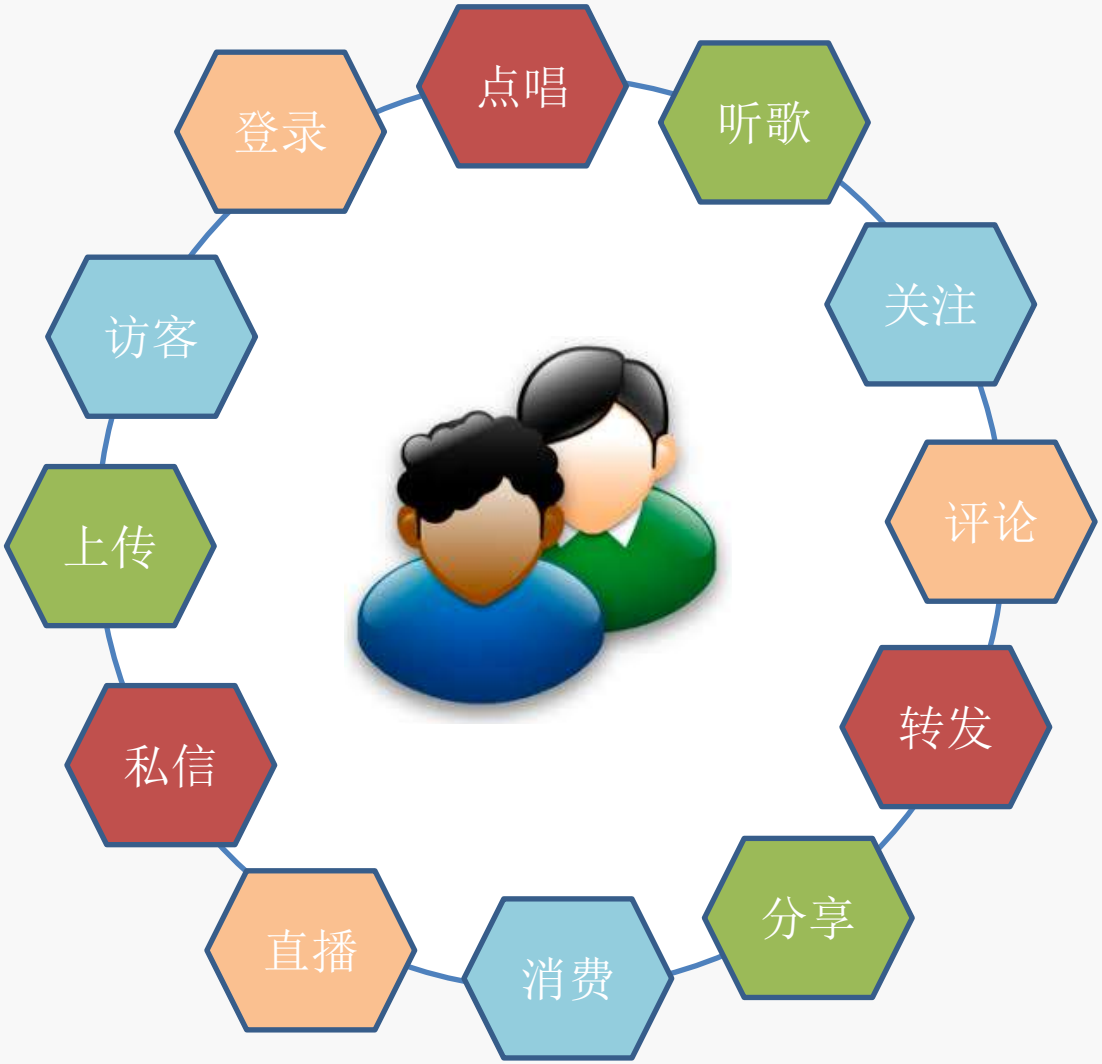
安全健康度审核 | 打击准确率审核 | 封号复核

运营审核：

- 可疑数据top排序，人工审核确认
- 只要投入少数人力，审核数据进入恶意样本进行训练

用户行为画像

全民K歌



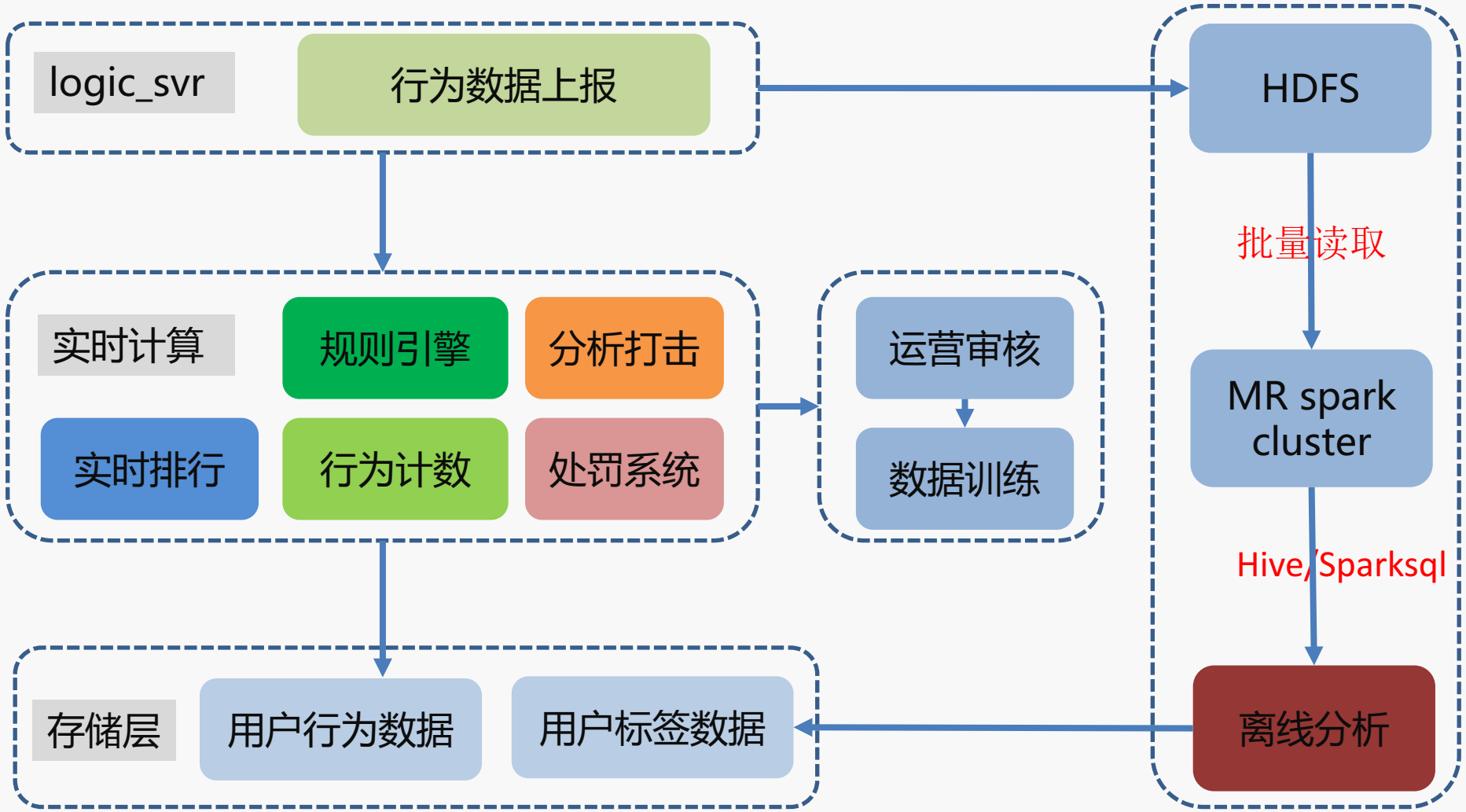
信用度评估：

- 用户画像
- 历史行为累积
- 用户分类预测

实现方案：

- 分维度计算，综合考量
- 实时累积、离线计算、分析预测

用户行为分析


logic_svr
行为数据上报
HDFS
批量读取
实时计算
规则引擎
分析打击
实时排行
行为计数
处罚系统
运营审核
数据训练
MR spark cluster
Hive/Sparksql
存储层
用户行为数据
用户标签数据
离线分析

安全策略和效果评估

全民K歌

- **效果评估：**

1. 平台内容健康度
2. 打击内容准确率

- **安全策略评估：**

1. 用户数据和打击效果的平衡
2. 打击效果和投入成本的平衡
3. 内容分析结合场景和账号行为
4. 新技术的引入-人工智能
5. 安全对抗长期存在，需review改进


THANKS