好豆BING需求说明（APP统计分析部分）

2014/05/27

概述

（此处省略宏图大计美好愿景800字。）

常用术语

本文中，“用户”在大多数情况下，与“设备”同义，按设备ID 统计。在谈到“登录用户”时，则理解为网站/应用注册用户，按用户ID 统计。

术语	术语含义
新增用户	首次启动应用的用户（以设备为判断标准）
活跃用户	在统计周期内启动过应用的用户（去重），启动过一次的设备即视为活跃用户，包括新用户与老用户
累计用户	截止统计周期结束时间，启动过应用的所有独立用户（去重，以设备为判断标准）。
留存用户	某段时间内的新增用户，经过一段时间后，仍继续使用应用的被认作是留存用户，这部分用户占当时新增用户的比例既是留存率。
启动次数	在收到用户发起的一次请求时，如果之前5分钟（超时时长）没有发生请求，则该请求被视为开始一个会话（启动）。*注：友盟口径分为安卓设备与iOS 设备。安卓设备启动是通过再所有Activity 中调用MobclickAgent.onResme() 和MobclickAgent.onPause() 方法来监测的。若用户使用过程中进入home 或其他程序，经过一段时间间隔后才返回之前的应用，如果间隔超过x（x可以由开发者自由设定，默认30）秒，则被认为是两次启动。iOS 设备在iOS4.x之后的系统，由于iOS 开始支持后台运行，进入后台即算是当前统计会话结束，当*

	再次进入前台时，算作一次新的启动行为，并开始新的统计会话。
单次使用时长	用户首次发起请求后，被视为开始一个会话。如果在每次请求接下来的5分钟内有下一个请求发生，则该会话持续，直到5分钟内没有请求发生，此时被视为该会话结束。该会话第1个请求发生时间到最后一次请求发生时间之差，即为本次会话时长（单次使用时长）
平均单次使用时长	单次使用时长的均值，即应用的总使用时长/总启动次数。
数据变化率	(本次数据-上次数据)/上次数据
过去N天活跃用户	过去N天内启动过应用的用户（去重），启动过一次的用户即视为活跃用户，包括新用户与老用户。
过去N天活跃占比	过去N天活跃用户占累计用户的比例
IDFA	苹果设备的 Identifier for Advertiser 的缩写。
IDFV	苹果设备的 Identifier for Vendor 的缩写。

数据说明

原始数据

APPLOG

在原BI系统中，该数据被命名为CDA39907，对应的表名为： dw.t_dw_applog
在大数据平台上，该数据HDFS存放路径为： /user/yarn/logs/source-log.php.CDA39907
该数据的shark访问路径为： logs.log_php_app_log

字段信息

字段名	字段说明
request_time	请求时间。该请求时间格式为Unix 时间戳。
device_id	设备ID。iOS 设备ID 格式：设备MAC 地址的MD5+IDFA+IDFV。安卓设备ID 格式：haodou+设备IMEI。
channel_id	渠道ID。格式：渠道编码+版本。
userip	用户访问IP
appid	应用ID。取值，1:去哪吃iphone/2:菜谱安卓/3:去哪吃安卓/4:菜谱iphone/5:华为机顶盒/6:菜谱ipad
version_id	版本。
userid	用户ID。未登录用户为0。
function_id	请求调用的函数
parameter_desc	请求传递的参数
logdate	日志日期。格式：yyyy-mm-dd。

中间数据

活跃设备日表（分应用）

字段信息

字段名	字段说明
statis_date	统计日期（PK）
device_id	好豆设备ID（PK）
channel_id	渠道ID。格式：渠道编码+版本。（PK）
version_id	版本ID
userip	当天首次访问IP
userid	当天首次登录用户ID。
dev_imei	安卓设备IMEI
dev_uuid	安卓设备UUID
mac_md5	苹果设备MAC 的MD5 值
idfa	苹果设备IDFA
idfv	苹果设备IDFV
req_cnt	访问请求次数
effect_cnt	有效访问次数
isvirtual	虚拟机标识。取值：1，是，0，否（缺省）
isfake	虚假标识。取值：1，是，0，否（缺省）

历史设备总表（分应用）

字段信息

字段名	字段说明
device_id	好豆设备ID（PK）
first_day	首次活跃日期。格式：yyyy-mm-dd。
first_channel	首次活跃渠道ID。格式：渠道编码+版本。
first_version	首次活跃版本ID
first_userip	首次活跃访问IP
first_userid	首次登录用户ID。
last_day	最近活跃日期。格式：yyyy-mm-dd。
last_channel	最近活跃渠道ID。格式：渠道编码+版本。
last_version	最近活跃版本ID
last_userip	最近活跃访问IP
last_userid	最近登录用户ID。
dev_imei	安卓设备IMEI
dev_uuid	安卓设备UUID
mac_md5	苹果设备MAC 的MD5 值
idfa	苹果设备IDFA
idfv	苹果设备IDFV
virtual	虚拟机判断信息
issilent	沉默用户标识。取值：0，否（缺省）／1，是
isvirtual	虚拟机用户标识。取值：0，否（缺省）／1，是
isfake	虚假标识。取值：0，否（缺省）／1，是
uninstall_time	卸载日期。格式：yyyy-mm-dd hh24:mi:ss

应用设备留存表

字段信息

字段名	字段说明
statis_date	统计日期
device_id	好豆设备ID
first_day	首次活跃日期。格式：yyyy-mm-dd。
first_channel	首次活跃渠道ID。格式：渠道编码+版本。
first_version	首次活跃版本ID
actlog	活跃日志串。缺省：000_{000}（60个0）。对应天数活跃，相应天位置改为1。

注：目前保留两个月内的留存情况。

渠道虚假活跃IP表

字段信息

字段名	字段说明
statis_date	统计日期
appid	应用ID
channel_id	渠道ID。格式：渠道编码+版本。
userip	虚假活跃IP地址
dev_num	该IP 下的活跃设备数
req_cnt	该IP 下的访问请求次数

数据处理过程

1．从当天APPLOG 中，取出设备ID ，以及相应的应用ID，与相应应用的历史设

备库中的设备ID 比较。如果该设备ID 不存在历史设备库中，则该设备为新增设备，并将该设备ID 添加到历史设备库中；如果该设备ID 存在历史设备库中，则该设备不记为新增设备（用户）。

2. 基于当天APPLOG 以及历史设备库，计算相关指标。

3. 将计算出的相关指标数据，导入MySQL 数据库。 注意：由于历史原因，过去的iOS 设备ID 只有第1段，在后来的版本中才增加了IDFA 与IDFV，需要避免由于扩展造成的新增重复计算。

关于虚假判定 目前根据单个渠道发行版本在单IP 的活跃用户数进行虚假判定。如果单个渠道发行版本在单IP 的活跃用户数大于12，则该IP 被视为该渠道发行版本的虚假活动IP，该IP 上的的所有该渠道发行版本的活跃用户都被视为虚假用户。

注意：虚假用户的判断策略需要维护，以保证虚假用户判断的有效性（查全性与查准性）。

功能性需求

应用概况

展现以下指标： - 日新增用户数 在当天首次启动应用的用户数（以设备为判断标准） - 日活跃用户数 在当天启动过应用的用户数（去重），启动过一次的设备即视为活跃用户，包括新用户与老用户。 - 日留存用户数 昨日新增用户，在当天依旧使用应用的用户数。 - 日平均单次使用时长 当天单次使用时长的均值，即应用的总使用时长/总启动次数。 - 日启动次数 在收到用户发起的一次请求时，如果之前5分钟（超时时长）没有发生请求，则该请求被视为开始一个会话（启动）。

注：友盟口径分为安卓设备与iOS 设备。安卓设备启动是通过再所有Activity 中调用MobclickAgent.onResme() 和MobclickAgent.onPause() 方法来监测的。若用户使用过程中进入home 或其他程序，经过一段时间间隔后才返回之前的应用，如果间隔超过x（x可以由开发者自由设定，默认30）秒，则被认为是两次启动。iOS 设备在iOS4.x之后的系统，由于iOS 开始支持后台运行，进入后台即算是当前统计会话结束，当再次进入前台时，算作一次新的启动行为，并开始新的统计会话。 - 累计用户 截止统计周期结束时间，启动过应用的所有独立用户

（去重，以设备为判断标准）。 – 过去7天活跃用户 过去7天内启动过应用的用户（去重），启动过一次的用户即视为活跃用户，包括新用户与老用户。 – 过去30天活跃用户 过去30天内启动过应用的用户（去重），启动过一次的用户即视为活跃用户，包括新用户与老用户。

应用趋势

留存分析

包括日留存分析，周留存分析。

日留存分析 分析维度：日期，渠道版本 分析指标：新增用户数，1$_{7天后留存率}$，14天后留存率，30天后留存率等指标。

周留存分析 分析维度：周数，渠道版本 分析指标：新增用户数，1$_{7周后留存率}$等指标。

非功能性需求

日志

处理过程日志文件 后台处理过程应将处理过程的启动，运行期间以及运行结果在日志文件中予以记录。 并支持DEBUG启动方式，将更明细的过程过程内部处理情况输出到日志文件。

界面操作日志文件 用户在操作界面上的活动，应该予以记录。用户活动包括但不限于：用户登录，用户注销，菜单项的选择，查询活动的发起等。

数据质量

性能

安全

- 未登录用户不能访问任何功能页面，访问时，跳到登录提示页面。
- 用户不能访问未对其授权的功能页面。