埋点日志数据处理

埋点日志

埋点就是在应用中特定的流程收集一些信息，用来跟踪应用使用的状况，后续用来进一步优化产品或是提供运营的数据支撑。比如用于作为用于实现个性化推荐的数据支撑。

埋点方式主流的无非两种方式：

- 自行研发：在研发的产品中注入代码进行统计，并搭建起相应的后台查询和处理
- 第三方平台：第三方统计工具，如友盟、百度移动等

埋点日志主要分为类型：

- 曝光日志：商品被展示到页面被称为曝光，曝光日志也就是指商品一旦被展示出来，则记录一条曝光日志
  - 曝光时间
  - 曝光场景
  - 用户唯一标识
  - 商品ID
  - 商品类别ID
- 点击流日志：用户浏览、收藏、加购物车、购买、评论、搜索等行为记录日志
  - 被曝光时间：对应曝光日志的曝光时间(浏览)
  - 被曝光场景：对应曝光日志的曝光场景(浏览)
  - 用户唯一标识
  - 行为时间
  - 行为类型
  - 商品ID
  - 商品类别ID
  - 停留时长(浏览)
  - 评分(评论)
  - 搜索词(搜索)

埋点日志意义

用户行为偏好分析

- 利用点击流日志分析个体/群体用户的行为特征，预测出用户行为的偏好

统计指标分析

- 点击率：顾名思义被点击的概率，计算公式通常是：点击次数/曝光次数。如某商品共曝光或展示了100次，曝光后总共被点击了10次，那么点击率则是10%
- 跳出率：用户访问一个页面后，之后没有再也没有其他操作，称为跳出，计算公式常用的是：访问一次就退出的访问量/总的访问量
  - 整体(整个板块/应用)跳出率
  - 单页面的跳出率。如某页面共计有100个用户访问，但其中有10个用户访问当前页面后就再也没有其他访问了，那么当前页面的跳出率是10%

- 转化率：电商中的转化率计算：商品订单成交量/商品访问量。如某商品累计访问量是100个，最终提交订单的只有5个，那么该商品转化率就是5%

注意：跳出率和转化率中的访问量指的是独立用户访问量。独立用户访问量：首先独立用户访问量并不等价于来访问的总用户个数。比如某用户A在1月1日访问了商品1，1月2日又访问了商品1，那么这里商品1的访问量应算作2次，而不是1次。

In [1]:
```
import logging#log：记录

def get_logger(logger_name, path, level):

    # 创建logger
    logger = logging.getLogger(logger_name)
    # level:  OFF、FATAL、ERROR、WARN、INFO、DEBUG、ALL或者自己定义的级别
    logger.setLevel(level)

    # 创建formatter
    # %(asctime)s: 打印日志的时间
    # %(message)s: 打印日志信息
    fmt = '%(asctime)s: %(message)s'
    datefmt = '%Y/%m/%d %H:%M:%S'
    formatter = logging.Formatter(fmt,datefmt)

    # 创建handler
    # FileHandler：writes formatted logging records to disk files
    handler = logging.FileHandler(path)
    handler.setLevel(level)

    # 添加handler和formatter 到 logger
    handler.setFormatter(formatter)
    logger.addHandler(handler)

    return logger
```

In [14]:
```
import time
import logging

exposure_logger = get_logger('exposure','/root/workspace/3.rs_project/project2/meiduo
                             logging.DEBUG)
# 曝光日志
exposure_timesteamp = time.time()
exposure_loc = 'detail'
uid = 1
sku_id = 1
cate_id = 1

```

In [15]:
```
exposure_logger.info("exposure_timesteamp<%d> exposure_loc<%s> uid<%d> sku_id<%d> cate
                     %(exposure_timesteamp, exposure_loc, uid, sku_id, cate_id))
```

In [16]:
```
# 每运行上条指令一次，就会在log中记录一条信息
# 显示最后五条信息
!cat /root/workspace/3.rs_project/project2/meiduoSourceCode/logs/exposure.log | tail
```

```
2020/12/23 21:16:43:exposure_timesteamp<1608729319> exposure_loc<detail> uid<1> sku_id<
1> cate_id<1>
2020/12/23 21:21:21:exposure_timesteamp<1608729679> exposure_loc<detail> uid<1> sku_id<
1> cate_id<1>
2020/12/23 21:21:21: exposure_timesteamp<1608729679> exposure_loc<detail> uid<1> sku_id
<1> cate_id<1>
2020/12/24 11:09:20: exposure_timesteamp<1608779359> exposure_loc<detail> uid<1> sku_id
<1> cate_id<1>
2020/12/24 11:09:20: exposure_timesteamp<1608779359> exposure_loc<detail> uid<1> sku_id
<1> cate_id<1>
```

In [17]:
```
import time
import logging
click_trace_logger = get_logger('click_trace','/root/workspace/3.rs_project/project2/r
                                logging.DEBUG)
```

In [18]:
```
# 点击流日志
exposure_timesteamp = exposure_timesteamp
exposure_loc = exposure_loc
timesteamp = time.time()
behavior = 'pv' # pv fav cart buy
uid = 1
sku_id = 1
cate_id = 1
stay_time = 60
# 假设某点击流日志记录格式如下：
click_trace_logger.info("exposure_timesteamp<%d> exposure_loc<%s> timesteamp<%d> behav
                        %(exposure_timesteamp, exposure_loc, timesteamp, behavior, ui

```

In [19]:
```
!cat /root/workspace/3.rs_project/project2/meiduoSourceCode/logs/click_trace.log
```

```
2018/11/30 03:20:24: exposure_timesteamp<1543519181> exposure_loc<detail> timesteamp<
1543519224> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 11:54:45: exposure_timesteamp<1543519181> exposure_loc<detail> timesteamp<
1543550085> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 11:54:45: exposure_timesteamp<1543519181> exposure_loc<detail> timesteamp<
1543550085> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 13:17:16: exposure_timesteamp<1543519181> exposure_loc<detail> timesteamp<
1543555036> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 13:17:16: exposure_timesteamp<1543519181> exposure_loc<detail> timesteamp<
1543555036> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 13:17:16: exposure_timesteamp<1543519181> exposure_loc<detail> timesteamp<
1543555036> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 13:18:16: exposure_timesteamp<1543555093> exposure_loc<detail> timesteamp<
1543555096> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 13:18:16: exposure_timesteamp<1543555093> exposure_loc<detail> timesteamp<
1543555096> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 13:18:16: exposure_timesteamp<1543555093> exposure_loc<detail> timesteamp<
1543555096> behavior<pv> uid<1> sku_id<1> cate_id<1> stay_time<60>
2018/11/30 13:18:16: exposure_timesteamp<1543555093> exposure_loc<detail> timesteamp<
```

flume采集日志

In [20]:
```
!hadoop fs -ls /meiduo_mall/logs
```

```
Found 2 items
drwxr-xr-x   - root supergroup          0 2020-12-24 08:55 /meiduo_mall/logs/click-trac
e
drwxr-xr-x   - root supergroup          0 2020-12-23 22:31 /meiduo_mall/logs/click_trac
e
```

In [21]:
```
import re
s = '2018/12/01 02:35:13: exposure_timesteamp<1543601846> exposure_loc<detail> timeste

match = re.search("\
exposure_timesteamp<(?P<exposure_timesteamp>.*?)> \
exposure_loc<(?P<exposure_loc>.*?)> \
timesteamp<(?P<timesteamp>.*?)> \
behavior<(?P<behavior>.*?)> \
uid<(?P<uid>.*?)> \
sku_id<(?P<sku_id>.*?)> \
cate_id<(?P<cate_id>.*?)> \
stay_time<(?P<stay_time>.*?)>", s)

result = []
if match:
    result.append(("exposure_timesteamp", match.group("exposure_timesteamp")))
    result.append(("exposure_loc", match.group("exposure_loc")))
    result.append(("timesteamp", match.group("timesteamp")))
    result.append(("behavior", match.group("behavior")))
    result.append(("uid", match.group("uid")))
    result.append(("sku_id", match.group("sku_id")))
    result.append(("cate_id", match.group("cate_id")))
    result.append(("stay_time", match.group("stay_time")))
result
```

Out[21]:
```
[('exposure_timesteamp', '1543601846'),
 ('exposure_loc', 'detail'),
 ('timesteamp', '1543602913'),
 ('behavior', 'pv'),
 ('uid', '1'),
 ('sku_id', '1'),
 ('cate_id', '1'),
 ('stay_time', '60')]
```

In [ ]:
```

```