# 数据分层方案设计

## 设计目的

1. 分离采集数据存储库与产品应用库

## 技术选型与术语描述

1. 选用python语言开发
2. M1: 现有正在线上使用的app\_data库
3. M2: 未来需要分离提供给应用组使用的数据库，原则上M1与M2基本一致，但会存在少量数据差异，比如删除标记字段等信息。

## M2全量数据库前置准备

1. 找运维确定M2数据库地址与名称
2. 全量dump M1 app\_data 数据库到文件，然后从文件导入数据到M2
3. 确认M2中的索引与M1中的索引完全一致。
4. 在M2全量数据库创建导出初期需要记录时间点，后期需要补数据程序根据时间差补上M1数据库在M2创建初期变更的数据信息

## 同步功能

1. **同步修改的数据**

\_utime字段发生变化的数据，这里不区分是新增的数据还是修改部分字段的数据

1. **同步删除的数据**
   1. 整个document 都在库中被消除的数据，这种数据在删除之前需要把整条记录都存在删除记录表中，具体的格式如下:

{

# \_id 可以随意， 当前这个字段不关注

\_id: xxxx

# \_record\_id 是数据索引字段，也是必须字段， 且是唯一索引

\_record\_id: xxxxxxxx

# \_utime是指当前删除数据的时间，必须字段，且时间必须准确

\_utime: 2017-08-12:00:00

}

* 1. 每一个业务表都对应一个记录删除表，记录删除表命名方式如下: delete\_{业务表名}
  2. 在同一个同步周期中，不允许同时存在对同一条document进行删除和修改操作，这种行为必须在执行层面进行人为控制。

1. **同步周期**

10分钟为一个同步周期，当前时间点向前推移3分钟，同步3分钟之前的10分钟之内的数据。z

1. **逻辑删除字段**

在M1中如document被整个删除掉了，则需要新增logic\_delete 字段， 该字段为int类型字段，类型如下:

0: 数据处于可用状态

1: 数据已经被删除

1. **单向同步约束 （重要）**

假设 M1、M2 同时拥有字段a, 当前 M1 与 M2中a字段都为1，在某一时刻M2的字段a被修改为100，而M1的字段a还是为1，则在某一个同步周期中M1中的字段a会覆盖M2中字段a，a的值会复原为1

通过以上案例，我们可以确定M2不适合写操作，如需要进行必要的写操作需要在M1上进行。

## 同步任务记录

1. 每一个同步任务都需要记录在任务记录表中，在任务执行之前生成任务记录，任务执行完后需要记录任务完成状态，如任务执行失败则需要在下一个更新周期中进行重试，针对失败任务需要进行记录，如失败达到一定次数则需要邮件提醒

## 同步流程

