## 区域人口分析详细设计方案

（需求细化 V1）

高胜杰

## 数据导入模块

#### 模拟数据生成

我们需要使用基站数据和营帐系统的数据：基站的数据格式为文本；营帐系统的数据为关系性数据。

基站的数据格式为：手机号\t\经度\t纬度\t时间\

营帐的表段：手机号 用户名 套餐类型 消费额度 性别 。

#### 数据导入

基站数据存放在ＦＴＰ服务器上在，我们需要先把数据存放到集群的ＨＤＦＳ上面去．

营帐数据存放在关系型数据库中，我们需要使用ｓｑｏｏｐ把他们存放到HBase中去。

这样就完成了数据导入工作。

## 数据关联模块

#### 原生数据处理

我们需要对基站的数据进行处理分析。确定该用户的居住地点，工作地点的。就是得到这些地点的经度和纬度。

目前基本的分析方法如下：

1：该用户长期早上0-6出现在某个区域，可认为该用户为该区域的常住人口。取得该区域内的一个点为该用户的确切的经纬度。

2：该用户长期早上10-16出现在某个区域，可认为该用户为该区域的工作人口。取得该区域内的一个点为该用户的确切的经纬度。（如果和该用户常住人口位置相近就删除该记录）

3：如果该用户长期变换区域，那么该用户设置为流动人口。不做分析。

#### 数据关联

把处理好的基站数据和营帐数据通过用户号码进行关联存储到Hbase中去。数据关联这块使用hive技术，hive需要和hbase实现读写功能，这块技术需要仔细琢磨，网上有教程但是需要验证。

## 用户信息分析模块

该块是对用户的信息进行分析，分析步骤如下：

1：该用户的消费能力，通过该用户的电话费用来划分，还有用户消费场所

2：该用户的消费习惯，通过商场基站和餐厅等商业区的基站数据来分析用户的行为习惯。

初期我们就做这两块。

这块使用Hive来对基站数据进行分析得到，结果存储到hbase中去。

## 区域人口分析模块

该模快分析有两块：

1：分析一个区域内的常住人口和流动人口。

2：分析一个区域在哪个范围内的人口密度最大。

这块需要mahout来实现。

## 数据查询展现模块

这块目前需求还不是特别的明确，我的个人想法如下：

1：管理员通过输入经纬度坐标和半径，等到一个区域内的人口分布：常住人口和工作人口。（通过图表显示）

2：该区域人的消费能力详解。（通过图表显示）

3：分析该区域的人口密集点，该坐标为商家提供咨询服务(导入关系型数据库)

4：该区域人口的年龄分布，性别分布。