项目背景：运营商日志分析

需求分析：记录每位用户的移动轨迹

架构设计：日志分析系统架构

算法模型：Map-Reduce并行算法

程序开发：

–用Maven构建Hadoop项目

–MapReduce程序实现

# 1 项目背景

通信某站产生的数据如下：



**应用前景：**

**（1）通过用户移动轨迹识别客户身份：**

•在没有客户身份数据的情况下

•认定上课在学校附近出现, 而且上课时间通话很少的用户作为疑似学生

通过用户移动轨迹识别客户身份

•分析目标地点的用户量和用户类型, 为户外广告选址和促销提供决策支持

**（2）分析轨迹流量**

•分析土地利用情况，为政府的土地分配和使用作决策支撑

**（3）决策支撑**

分析土地利用情况，为政府的土地分配和使用作决策支撑

## 1.1 数据处理的挑战



## 1.2 数据解决思路

（1）架设Hadoop分布式计算平台, 搭建数据存储体系

（2）对数据进行初步的分析和统计

（3）只保留用户的活动位置数据

（4）如有必要导入结构化数据库进行分析

## 1.3 数据来源

（1）500万客户

每天接收20G的数据

（2）数据分两种类型

•上网数据

•位置数据

（3）只存储每个用户在不同时段停留最长的三个位置

•记录 （用户ID, 时间段, 地点）

•区分凌晨, 上午通勤时间, 工作时间, 下午回家时间和晚上

•可以根据需求更改时间段的划分

# 2 需求分析

## 2.1 案例介绍

只存储每个用户在不同时段停留最长的三个位置

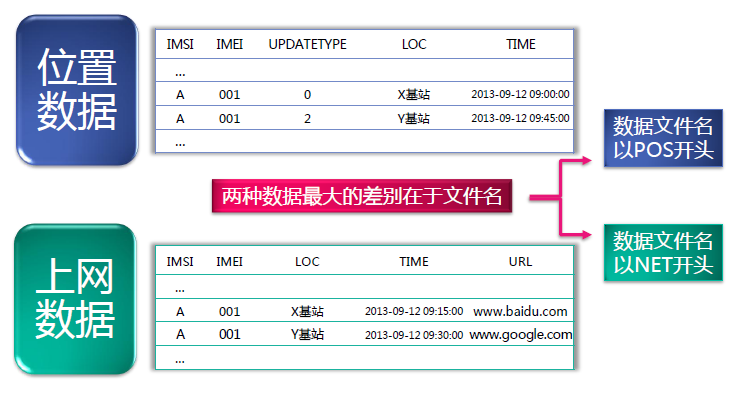
•记录 （用户ID, 时间段, 地点）

•区分凌晨, 上午通勤时间, 工作时间, 下午回家时间和晚上

•可以根据需求更改时间段的划分

## 2.2 数据格式

日志的来源有两种数据格式，分别以不同的字符串开头，需要对这两种格式同时进行分析，具体如下：



# 3 架构设计

## 3.1 日志的收集

通过网络或者别的途径将日志拷贝到集群中。

## 3.2 计算架构

MapReduce集群进行日志的分析，产生相应的分析结果，分析结果可用R等方式以图形的方式展示出来。

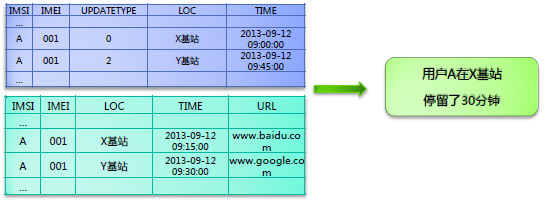
# 4 程序设计

## 4.1 计算规则

（1）认为用户在任何时间的停留位置都取决于之前一次位置更新的基站位置（两个文件中相同用户的记录，以时间排序）

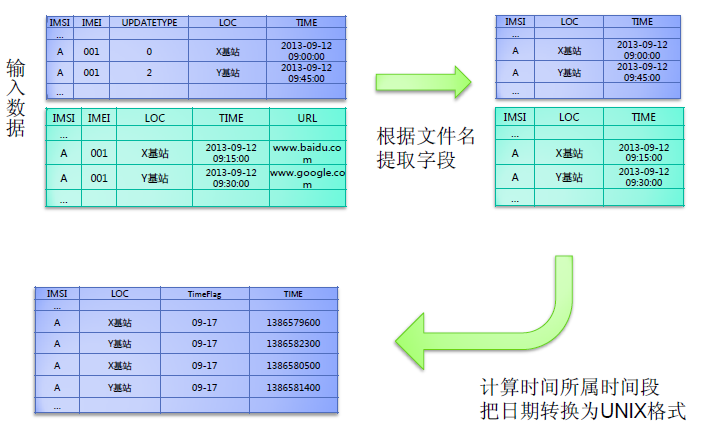
（2）时间间隔超过超过60分钟的判定为关机

如下图中A用户表示为在X基站停留了30分钟：

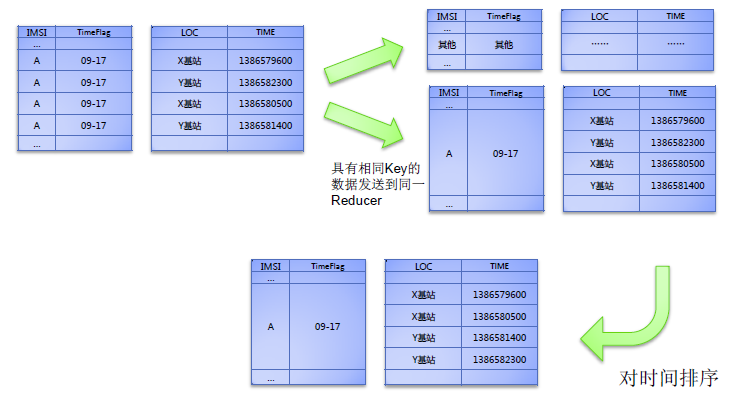
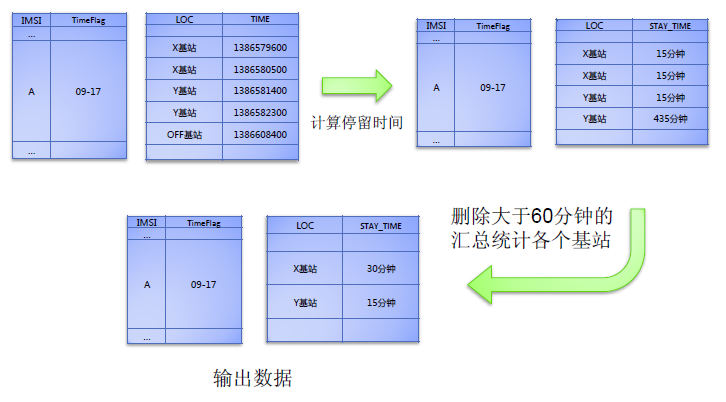


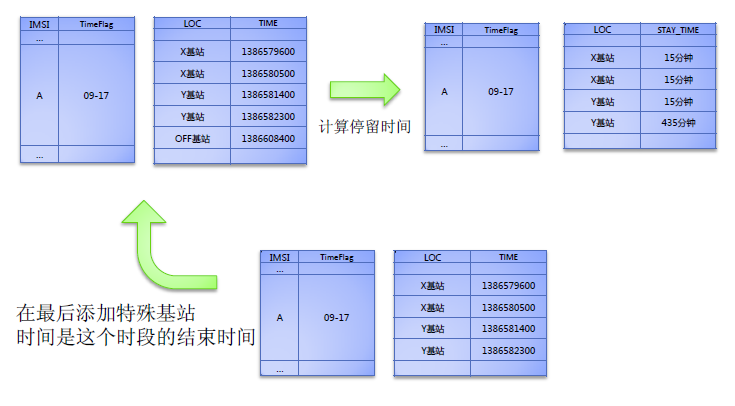
## 4.2 算法模型

### （1）Map的过程

 

### （2）Reduce过程





## 4.3 程序开发环境

（1）开发环境：

–Win7 64bit

–Java 1.7.0\_17

–Maven3

–Eclipse Juno Service Release 2

（2）Hadoop集群环境：

–Linux: Ubuntu 14.04.4 LTS 64bit Server

–Java: 1.7.0\_17

–Hadoop: hadoop-2.7.2，单节点，IP:192.168.1.210



## 4.4 开发流程

### 4.4.1 将要分析的日志上传到HDFS系统中

参考下面的命令操作

~ hadoop fs -mkdir /test/in

~ hadoop fs -copyFromLocal /home/access.log.10 /test/in

注：本实例中的日志文件在项目目录中的：\template-kpi\testData\mobile，里面提供了一些简单的测试日志

### 4.4.2 代码流程

（1）对”日志行”的解析

（2）Map函数实现

（3）Reduce函数实现

（4）启动程序实现

### 4.4.3 源码路径

git@github.com:liulong3712/hadoopcode.git中的template-kpi目录com.llw.mobileoperators.mr包中

### 4.4.5 程序运行方法

示例：hadoop jar ~/BaseStationDataPreprocess.jar BaseStationDataPreprocess \test\in test\out 2013-09-12 00-05-12-24

**BaseStationDataPreprocess.jar**:代码打出的jar包

**BaseStationDataPreprocess**:程序主类名

**\test\in**:HDFS中日志存放路径

**test\out**:HDFS中结果输出路径

**2013-09-12**:要统计的日期

**00-05-12-24**:要分的三个时间段