# Spark Core API实验

**实验目的：**

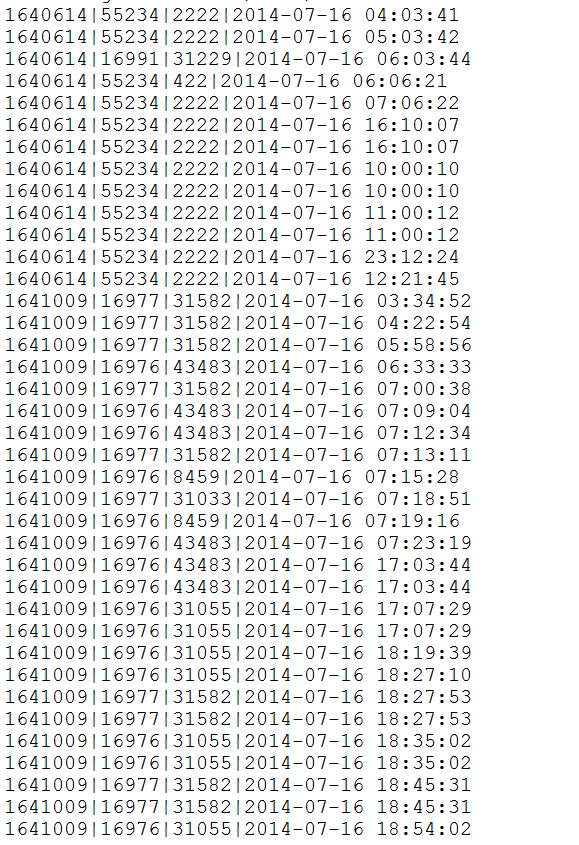
1. 掌握Spark内核API的数据处理中的数据流转换。
2. 实践reduce函数和fold函数的差异。
3. 实践aggregate函数，combineByKey函数。
4. 综合实践（优化等）。

**实验内容：**

1. 回顾上次的四个实验，解答难点。
2. 通过实验大公司合并小公司的案例，学生实践编程感觉reduce和fold的差异。
3. 通过简单实例动手实践aggregate和combineByKey。
4. 位置数据的数据处理。

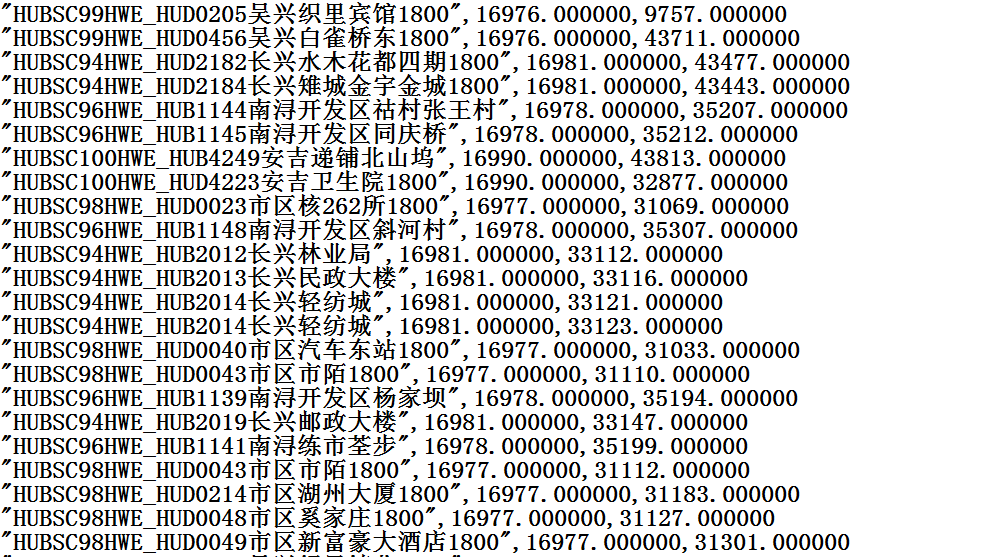
**实验数据说明：**

1. 略
2. 略
3. 略
4. 本实验数据包含两个部分的文本文件（文件相对较大，目的在于给予大家调优的测试），user-location-time和businessHall-location。user-location-time数据片段如下（文件2G多，可在命令行使用more命令等查看）：



本数据采用了字符“|”作为字段分隔符。每行为一个完整记录。数据的第一列表示了用户的ID标号如（1640614）。数据第二列和第三列分别表示了用户的区域编码和小区编码如（16977 | 31055），用户的所在位置范围又区域编码和小区编码确定。数据第四列描述用户在该区域地点出现的时刻入（2014-07-16 18:54:02）。

businessHall-location数据片段如下：



数据有“，”分隔。第一个列表示位置的名字编号。第二，三列分别表示了用户的区域编码和小区编码。和user-location-time的位置字段对应。

**数据处理的目标：**

1. **找出用户都在哪些些地点出现过,用户位置索引；**

**数据格式形如：此处无严格格式限制**

1640614 [55234|2222 , 16991|31229 ,55234|422]

表示1640614用户分别在55234|2222 ,16991|31229 ,55234|422这三个地点出现过。

1. **用户在各个地点出现过并且在各个地点分别停留的时间**

**数据格式形如：此处无严格格式限制**

**1640614 {**55234|2222:500 , 16991|31229:700 ,55234|422:800**}或者其他能够表达这个意思的数据格式。**

**时间单位是秒。计算方式：**

1. **对于间隔时间大于T，T参数由用户传入，当相邻时间间隔大于T，则该间断的时间视为无效（或者为0）。否则算作有效时间。例如**

**141451 1245|545 t1**

**141451 1245|545 t2**

**141451 1245|545 t3**

**141451 1245|545 t4**

**141451 1245|545 t5**

**且按照时间先后顺序排序的。假如t2-t1 <T, t3-t2<T , t4-t3 >T,t5-t4 <T. 则有效时间总和为**

**Time = (t2-t1) +(t3-t2) +(0)+(t5-t4).最后结果如同：**

**141451 1245|545 Time**

1. **识别出各个营业厅所有的用户群如：**

**HUBSC99HWE\_HUD0205吴兴织里宾馆1800 [user1，user2..]**

1. **识别出所有用户在营业厅停留过的时长。形如：**

**User1 HUBSC99HWE\_HUD0205吴兴织里宾馆1800 Time1**

**User1 HUBSC96HWE\_HUB1145南浔开发区同庆桥 Time2**

**User2 HUBSC100HWE\_HUB4249安吉递铺北山坞 Time3**

**User3 HUBSC94HWE\_HUB2012长兴林业局 Time4**