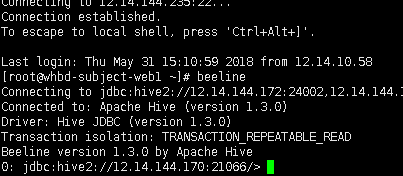
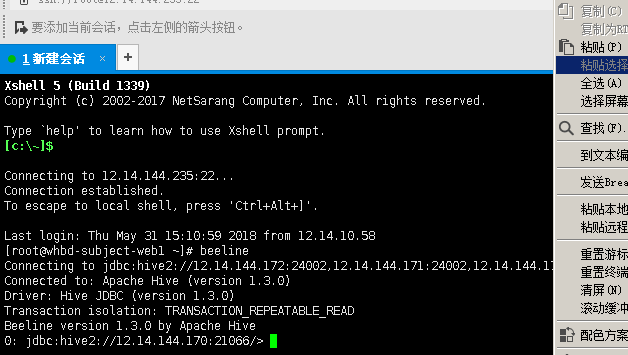
Hive数据库抽取特定人群电围流程

打开d:\python\_codes\hive\form\_hive\_sql.py,根据需要更改特定人群列表，生成相应的sql语句，如create table new\_tabel\_name as select sfzh,msisdn from tb\_sub\_realname\_new where sfzh IN ('420104196212081217','420103196112251614','420103197003010015','420123197112056210','412326197211024257','420102198109110637','640324199211050831','640324199408010454','640324199007092436','642127196603213210','640324198308100474','420984198707050417',………);

将生成的语句复制，然后打开Xshell远程连接到12.14.144.235 用户名root，密码thinker，然后输入beeline进入hive平台。



右键粘贴刚复制的sql语句,回车执行

0: jdbc:hive2://12.14.144.170:21066/> select \* from qzls\_phone;

+---------------------+--------------------+--+

| qzls\_phone.sfzh | qzls\_phone.msisdn |

+---------------------+--------------------+--+

| 420984199004281731 | 18318162490 |

| 420117198505097934 | 13871130263 |

| 420122197210155814 | 02791288430 |

| 420122197210155814 | 18086010777 |

| 420122197210155814 | 17771781881 |

| 420122197210155814 | 15337137699 |

查询可见已经生成了身份证号与手机号码的对应表

4.由于电围表tb\_evt中每一天的电围数据量可以达到1亿条，因此如对电围表全量查询会报GC overhead limit exceeded错，意思是耗尽了HIVE集群的内存，因此需根据表的分区来筛选数据，并根据筛选后数据建新表。首先show partitions tb\_evt;看表的分区情况，| cp=2018053014/ld=2018053019 |

| cp=2018053015/ld=2018053019 |

| cp=2018053015/ld=2018053020 |

| cp=2018053015/ld=2018053021 |

| cp=2018053016/ld=2018053107 |

| cp=2018053016/ld=2018053109 |

| cp=2018053017/ld=2018053110 |

以上可以看出cp就是按日期时间分区2018/05/30/15代表5.30号15点的纪录，所以新表的sql语句可以这么写:

Create table tj\_dw\_7 as select usernum,relatenum,lai,ci,longitude,latitude,cp from tb\_evt where cp>=2018052000 and cp<=2018052723; 其中电围表中的这些列比较有用。而日期可以按需增加

5. 建完select \* from tj\_dw\_7 limit 10;结果为

+------------------+--------------------+--------------+-------------+--------------------+-------------------+-------------+--+

| tj\_dw\_7.usernum | tj\_dw\_7.relatenum | tj\_dw\_7.lai | tj\_dw\_7.ci | tj\_dw\_7.longitude | tj\_dw\_7.latitude | tj\_dw\_7.cp |

+------------------+--------------------+--------------+-------------+--------------------+-------------------+-------------+--+

| 15871415356 | | 57617 | 857 | 114.34683 | 30.54955 | 2018052515 |

| 15926282594 | 18164250093 | 28735 | 5583 | 114.39156 | 30.621174 | 2018052515 |

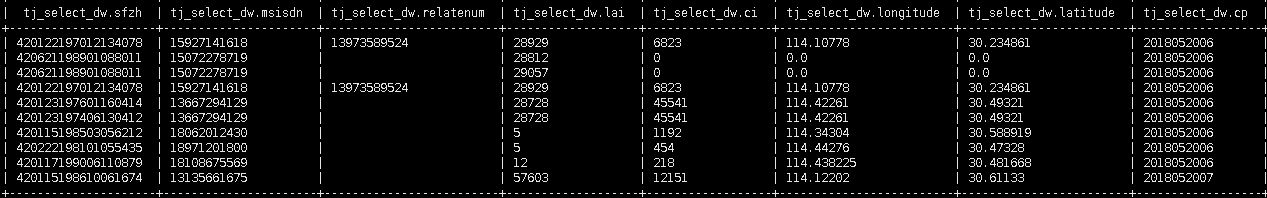
| 17771704029 | #777 | 13 | 125 | 111.14066 | 30.691711 | 2018052515 |

| 13661173750 | | 29019 | 23681 | 111.64445 | 30.31361 | 2018052515 |

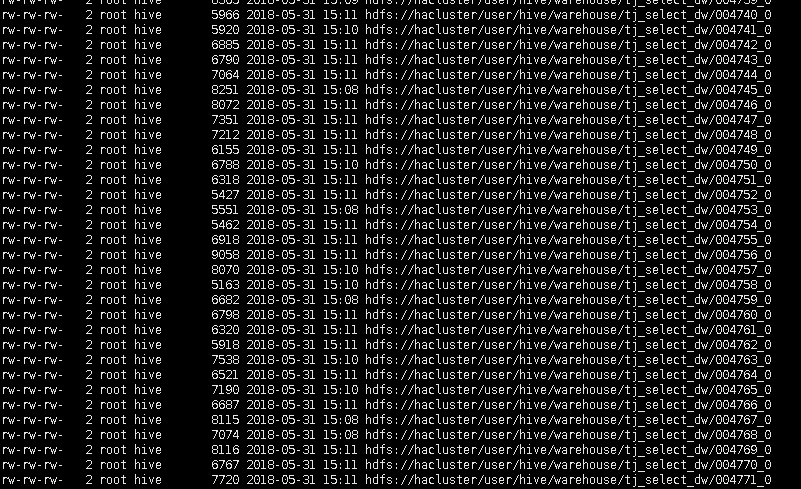
6. 因为电围表中不包含身份证号，而我们的模型根据身份证号来建模，因此需要联表查询，sql语句为 create table tj\_select\_dw as select a.sfzh,a.msisdn,b.relatenum,b.lai,b.ci,b.longitude,b.latitude,b.cp from qzls\_phone a join tj\_dw\_7 b on a.msisdn=b.usernum;

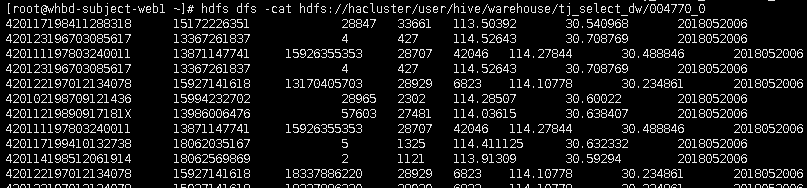
但是这样在hdfs中储存的文件将会是压缩的格式org.apache.hadoop.hive.ql.io.RCFileOutputFormat，出来的文件无法直接读，因此需要加入stored as textfile将其改为org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat这种输出格式。同时设置输出的文件以’\t’分割各列

因此sql语句改为：create table tj\_select\_dw row format delimited fields terminated by ‘\t’ stored as textfile as select a.sfzh,a.msisdn,b.relatenum,b.lai,b.ci,b.longitude,b.latitude,b.cp from qzls\_phone a join tj\_dw\_7 b on a.msisdn=b.usernum; 插入后的表如下所示：



6.下一步需从hdfs文件系统中找到这张表的位置，show create table tj\_select dw;后在LOCATION中能找到该表在hdfs中的位置hdfs://hacluster/user/hive/warehouse/tj\_select\_dw，随后退出HIVE。退出的命令为!q，回到服务器命令行界面。

7. 随后在命令行中输入hdfs dfs –ls hdfs://hacluster/user/hive/warehouse/tj\_select\_dw查看该表存储情况

可以看到有一大堆文件，随后可用hdfs dfs –cat hdfs://hacluster/user/hive/warehouse/tj\_select\_dw/004770\_0

8.随后使用-getmerge 命令将文件夹下所有内容合并为一个本地文件 命令为 hdfs dfs –getmerge hdfs://hacluster/user/hive/warehouse/tj\_select\_dw /home/tongji\_datas/dw.txt

[root@whbd-subject-web1 ~]# cd /home/tongji\_datas

[root@whbd-subject-web1 tongji\_datas]# ls

data.txt dw.txt

[root@whbd-subject-web1 tongji\_datas]#

9. 再使用 sz dw.txt命令将文本文件保存至电脑中。