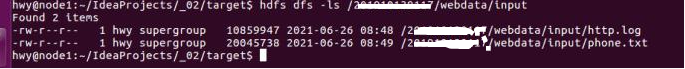
**（1）将数据文件http.log、phone.txt上传到hdfs，保存的目录以个人学号区分，如001为学号：**

Hdfs dfs -put http.log /001/webdata/input

Hdfs dfs -put phone.txt /001/webdata/input



**结果：**

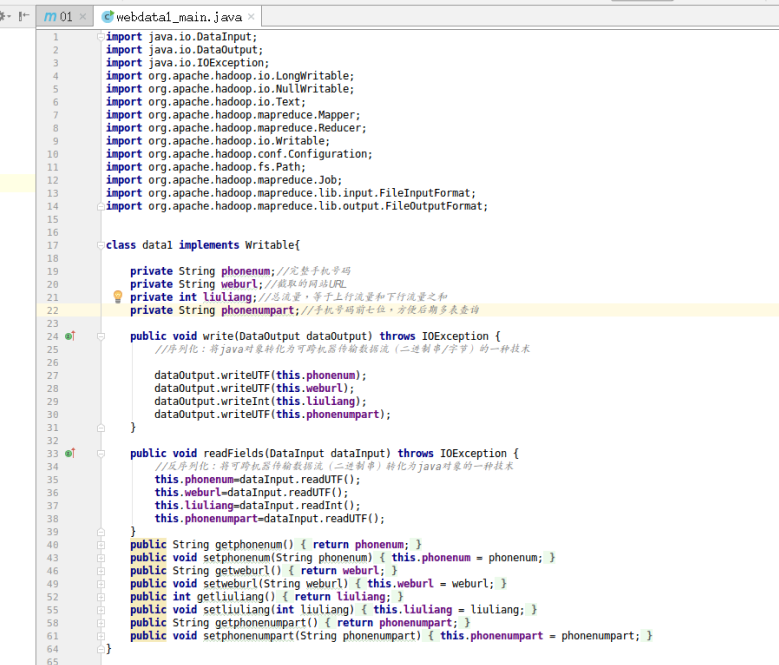


**（2）数据清洗**

**编写MapReduce程序实现数据清洗，将不完整的数据过滤掉，保证输出数据以 ','分割，输出数据保存在两个文件中。**

**因为题目的后续要求等值连接，而http.log文件没有完全与phone.txt相同的字段，所以我构造了一个新字段，手机号码的前七位，用这个字段连接phone.txt的手机号段字段。其他具体数据处理清洗逻辑如程序图中注释所示。**

**第一个程序如下所示，用于处理http.log：（共四个类，Mapper，Redecer，主类，还有一个用作传递多维度信息的data1类）**

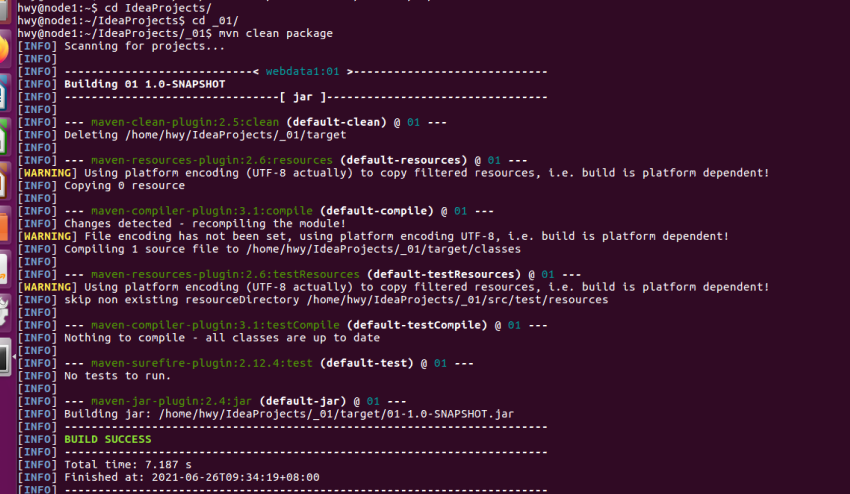






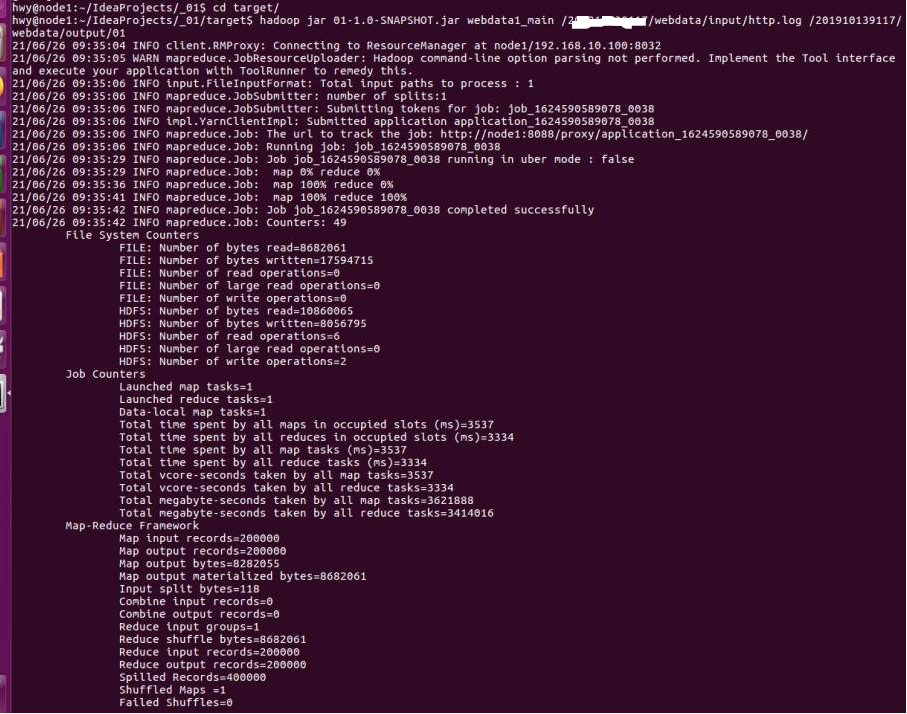
**程序打包：**

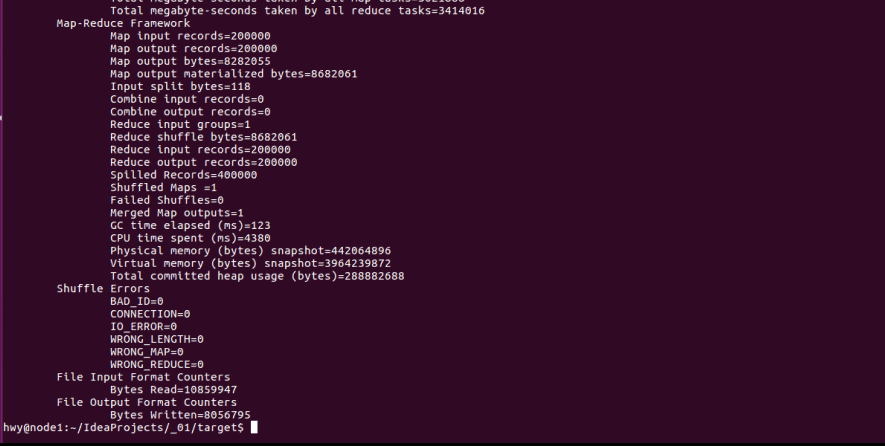
mvn clean package



**程序运行：**

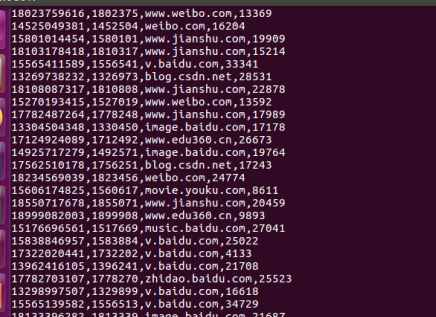
hadoop jar 01-1.0-SNAPSHOT.jar webdata1\_main /001/webdata/input/http.log /001/webdata/output/01





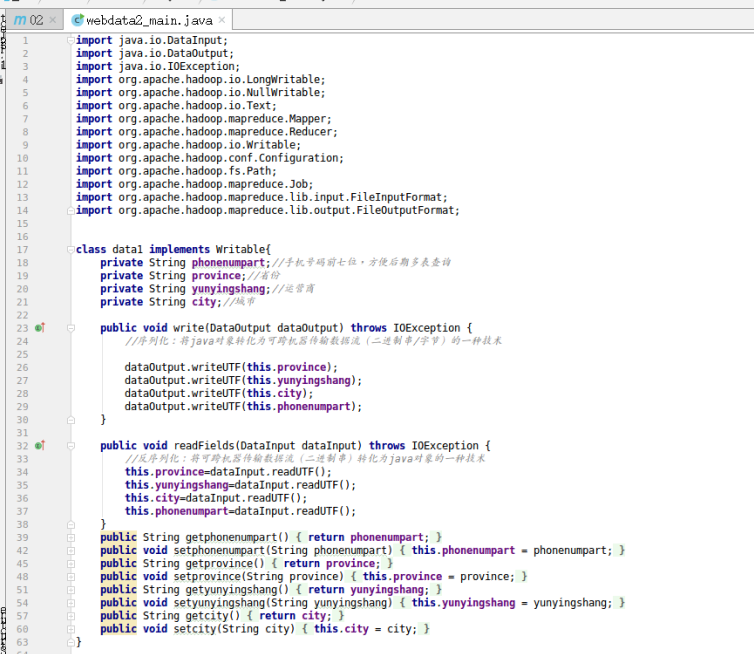
**输出结果查看（输出数据已按 ','分割）：**

hdfs dfs -cat /001/webdata/output/01/part-r-00000

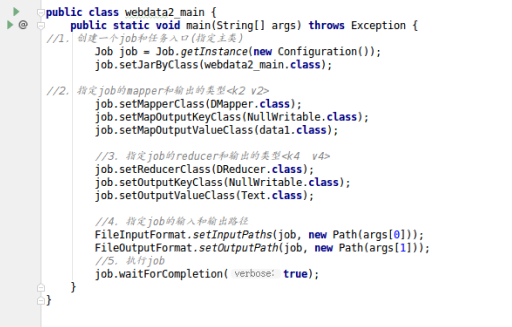


**第二个程序如下，用于处理Phone.txt，数据处理清洗逻辑如程序图中注释所示。**

**（共四个类，Mapper，Redecer，主类，还有一个用作传递多维度信息的data1类）:**

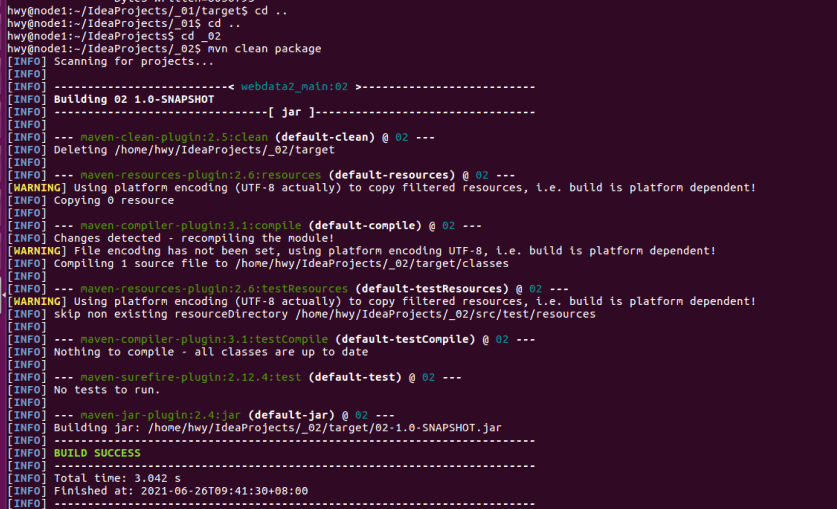






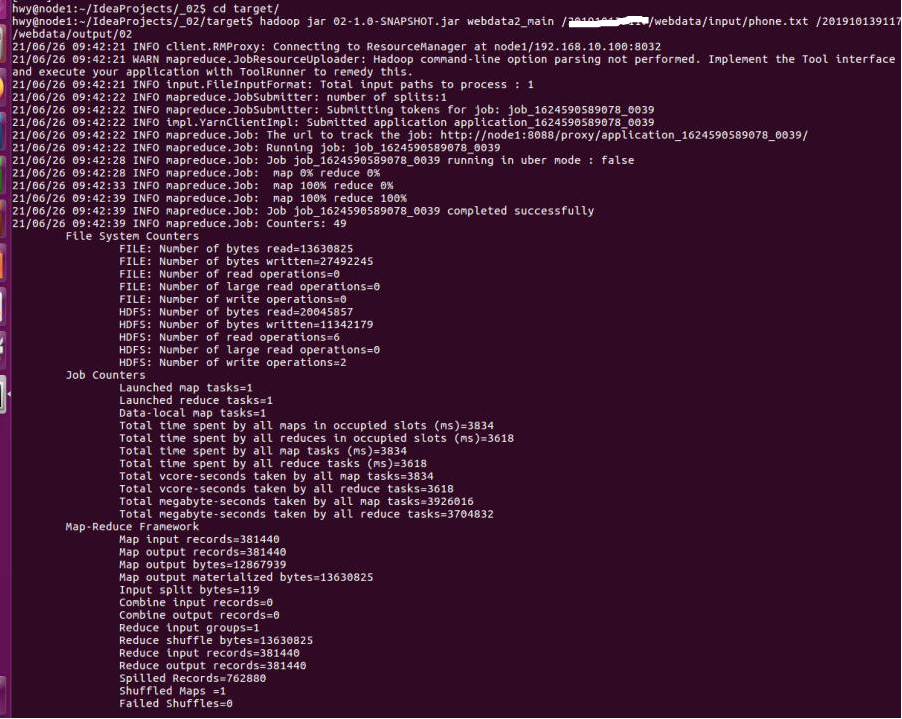
**程序打包：**

mvn clean package



**程序运行：**

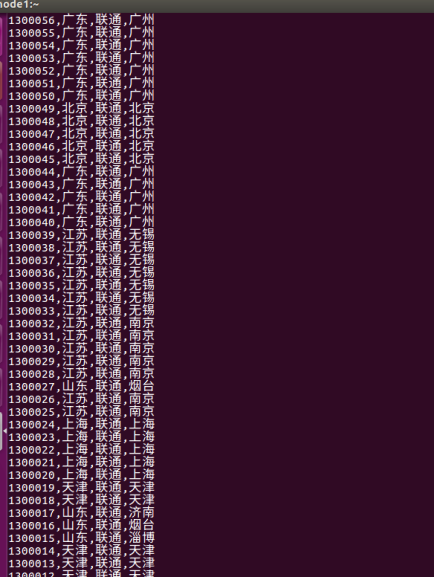
hadoop jar 02-1.0-SNAPSHOT.jar webdata2\_main /001/webdata/input/phone.txt /001/webdata/output/02





**输出结果查看（输出数据已按 ','分割）：**

hdfs dfs -cat /001/webdata/output/02/part-r-00000



1. **创建hive表(要求表名以姓名缩写首字母+学号)**

**首先创建hive库：**

Create database if not exists hwy001;

**使用自己的数据库：**

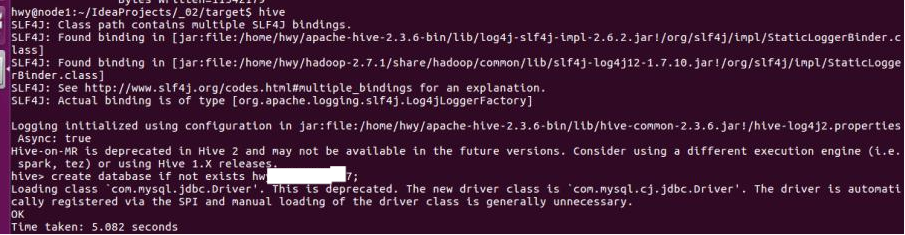
use hwy001;

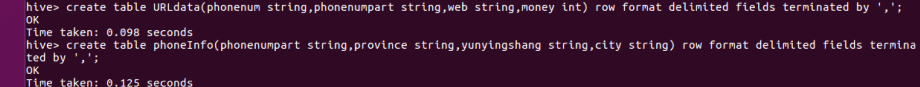
**创建第一个表URLdata，用于保存http.log经程序处理后的结果数据：**

create table URLdata(phonenum string,phonenumpart string,web string,money int) row format delimited fields terminated by ',';

**创建第二个表phoneInfo，用于保存phone.txt经程序处理后的结果数据：**

create table phoneInfo(phonenumpart string,province string,yunyingshang string,city string) row format delimited fields terminated by ',';



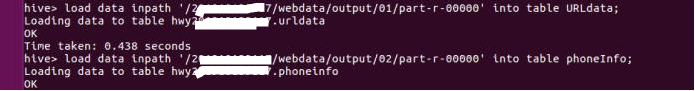


1. **将清洗后的数据导入Hive**

**向两个表中导入数据：**

load data inpath '/001/webdata/output/01/part-r-00000' into table URLdata;

load data inpath '/001/webdata/output/02/part-r-00000' into table phoneInfo;



1. **使用SQL查询满足条件的数据（等连接的多表查询）**

**1、根据给的用户上网日志记录数据，计算出总流量最高的网站Top3(网站例如：v.baidu.com，weibo.com)**

**命令如下：**

select web,sum(money) as mm from URLdata group by web order by mm desc limit 3;



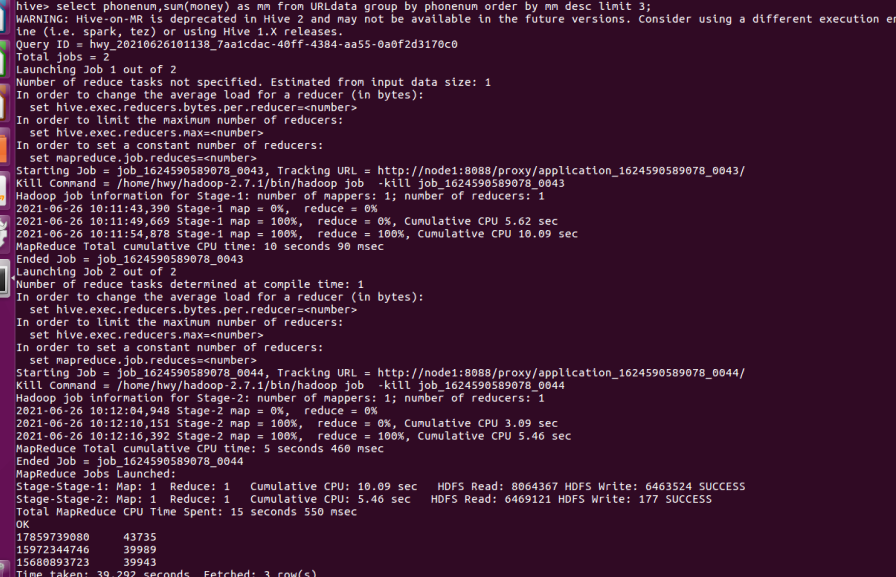
**结果：**



**2、根据给的用户上网日志记录数据，计算出总流量最高的手机号Top3**

**命令如下：**

select phonenum,sum(money) as mm from URLdata group by phonenum order by mm desc limit 3;



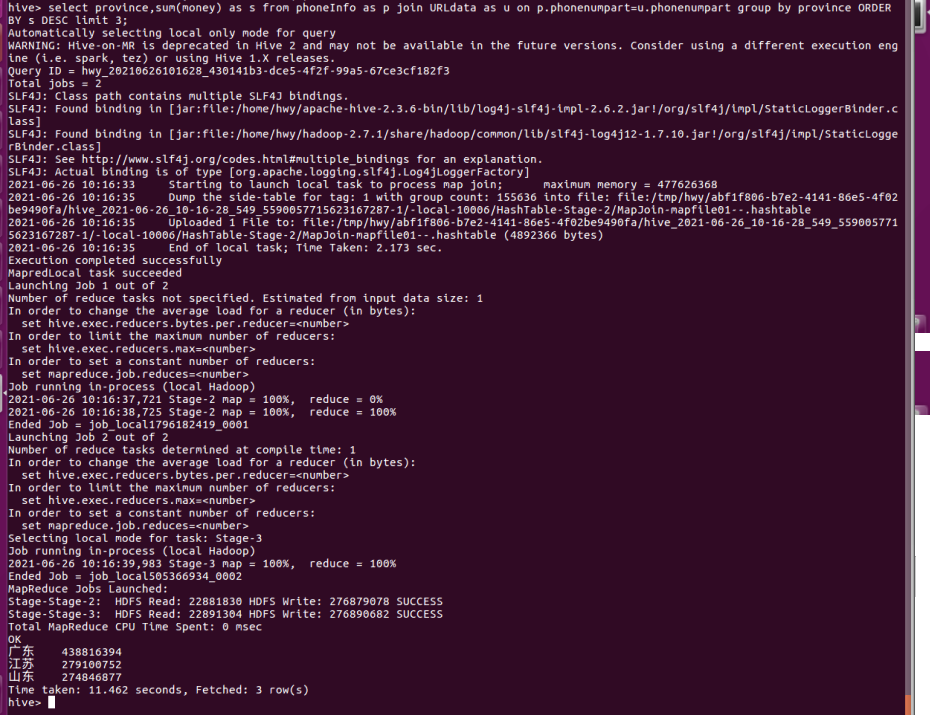
**结果：**



**3、根据给的手机号段归属地规则，计算出总流量最高的省份Top3**

**命令如下：**

select province,sum(money) as s from phoneInfo as p join URLdata as u on p.phonenumpart=u.phonenumpart group by province ORDER BY s DESC limit 3;



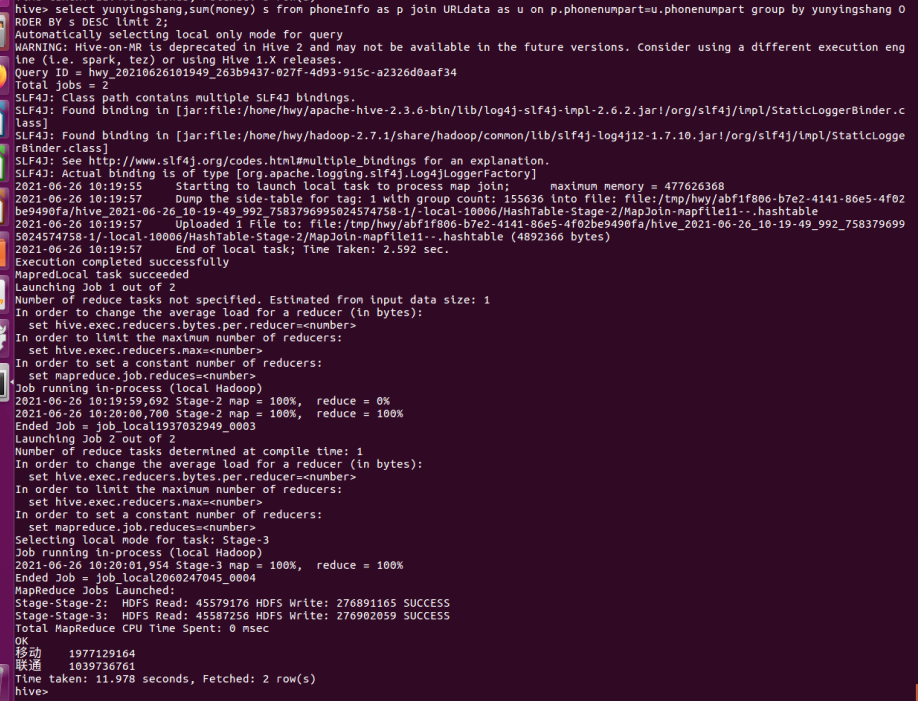
**结果：**



**4、根据给的手机号段运营商规则，计算出总流量最高的运营商Top2**

**命令如下：**

select yunyingshang,sum(money) s from phoneInfo as p join URLdata as u on p.phonenumpart=u.phonenumpart group by yunyingshang ORDER BY s DESC limit 2;



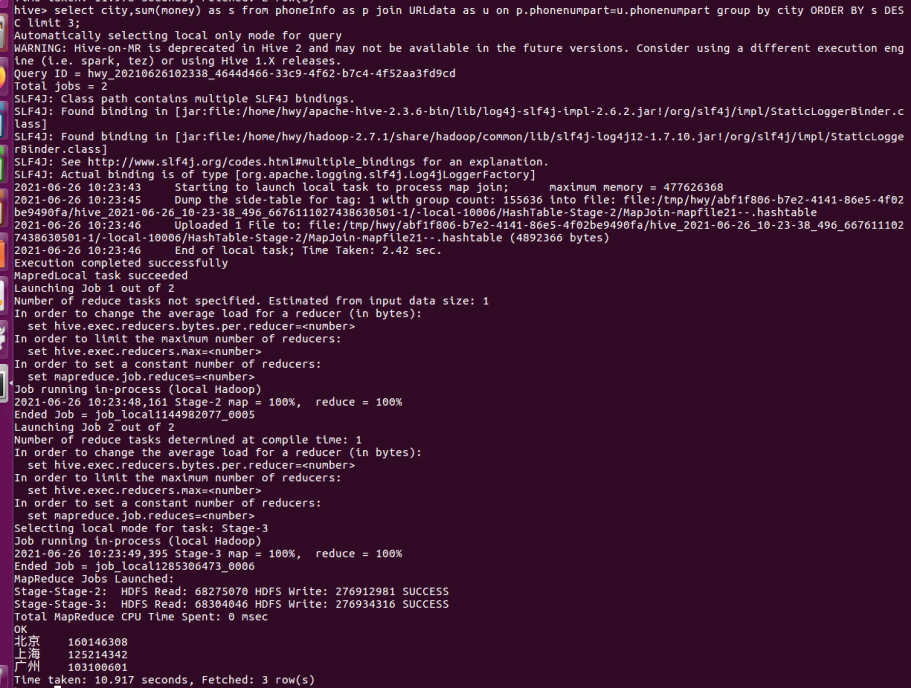
**结果：**



**5、根据给的手机号段归属地规则，计算出总流量最高的城市Top3**

**命令如下：**

select city,sum(money) as s from phoneInfo as p join URLdata as u on p.phonenumpart=u.phonenumpart group by city ORDER BY s DESC limit 3;



**结果：**



1. **删除hive数据库（以姓名缩写首字母+学号创建的数据库）**

**因为数据库里面有数据表，所以要用cascade来强制删除hive库（删除库时，库里面的表也一并删除），命令如下：**

drop database hwy001 cascade;

**结果：**

F:\暂时不用\hadoop大作业\图片\7.png7