

大数据案例-步骤一:本地数据集上传到数据仓库Hive

一、任务清单

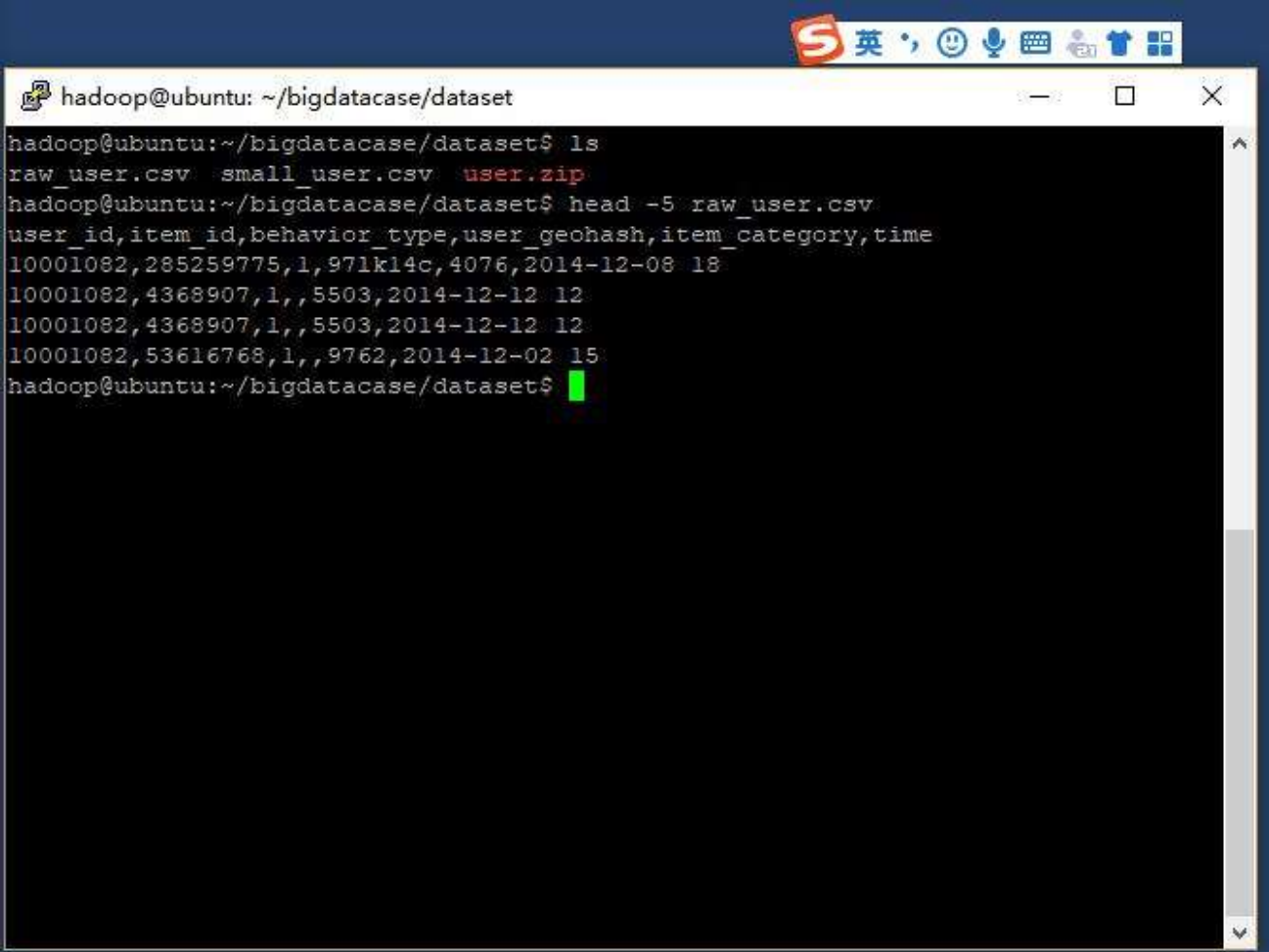
- 安装Linux系统
- 数据集下载与查看
- 数据集预处理
- 把数据集导入分布式文件系统HDFS中
- 在数据仓库Hive上创建数据库

二、实验结果

- 安装Linux系统
 - 本实验采用ubuntu 16.04 server版
 - IP:192.169.227.10

```
hadoop@ubuntu: ~  
docker0  Link encap:Ethernet  HWaddr 02:42:13:45:b7:ce  
          inet addr:172.17.0.1  Bcast:0.0.0.0  Mask:255.255.0.0  
          UP BROADCAST MULTICAST  MTU:1500  Metric:1  
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0  
          TX packets:0 errors:0 dropped:0 overruns:0 carrier:0  
          collisions:0 txqueuelen:0  
          RX bytes:0 (0.0 B)  TX bytes:0 (0.0 B)  
  
enp0s3  Link encap:Ethernet  HWaddr 08:00:27:b5:81:bf  
          inet addr:10.0.2.15  Bcast:10.0.2.255  Mask:255.255.255.0  
          inet6 addr: fe80::a00:27ff:feb5:81bf/64 Scope:Link  
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1  
          RX packets:236 errors:0 dropped:0 overruns:0 frame:0  
          TX packets:98 errors:0 dropped:0 overruns:0 carrier:0  
          collisions:0 txqueuelen:1000  
          RX bytes:279351 (279.3 KB)  TX bytes:9624 (9.6 KB)  
  
enp0s8  Link encap:Ethernet  HWaddr 08:00:27:a9:13:76  
          inet addr:192.168.227.10  Bcast:192.168.227.255  Mask:255.255.255.0  
          inet6 addr: fe80::a00:27ff:fea9:1376/64 Scope:Link  
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1  
          RX packets:39 errors:0 dropped:0 overruns:0 frame:0  
          TX packets:75 errors:0 dropped:0 overruns:0 carrier:0  
          collisions:0 txqueuelen:1000  
          RX bytes:5238 (5.2 KB)  TX bytes:11257 (11.2 KB)  
  
enp0s9  Link encap:Ethernet  HWaddr 08:00:27:00:1c:2a  
          inet6 addr: fe80::alfa:2e07:7d7e:9810/64 Scope:Link  
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1  
          RX packets:0 errors:0 dropped:0 overruns:0 frame:0  
          TX packets:48 errors:0 dropped:0 overruns:0 carrier:0  
          collisions:0 txqueuelen:1000  
          RX bytes:0 (0.0 B)  TX bytes:7798 (7.7 KB)  
  
lo  Link encap:Local Loopback  
     inet addr:127.0.0.1  Mask:255.0.0.0  
     inet6 addr: ::1/128 Scope:Host  
     UP LOOPBACK RUNNING  MTU:65536  Metric:1  
     RX packets:9 errors:0 dropped:0 overruns:0 frame:0  
     TX packets:9 errors:0 dropped:0 overruns:0 carrier:0  
     collisions:0 txqueuelen:1  
     RX bytes:467 (467.0 B)  TX bytes:467 (467.0 B)  
  
hadoop@ubuntu:~$
```

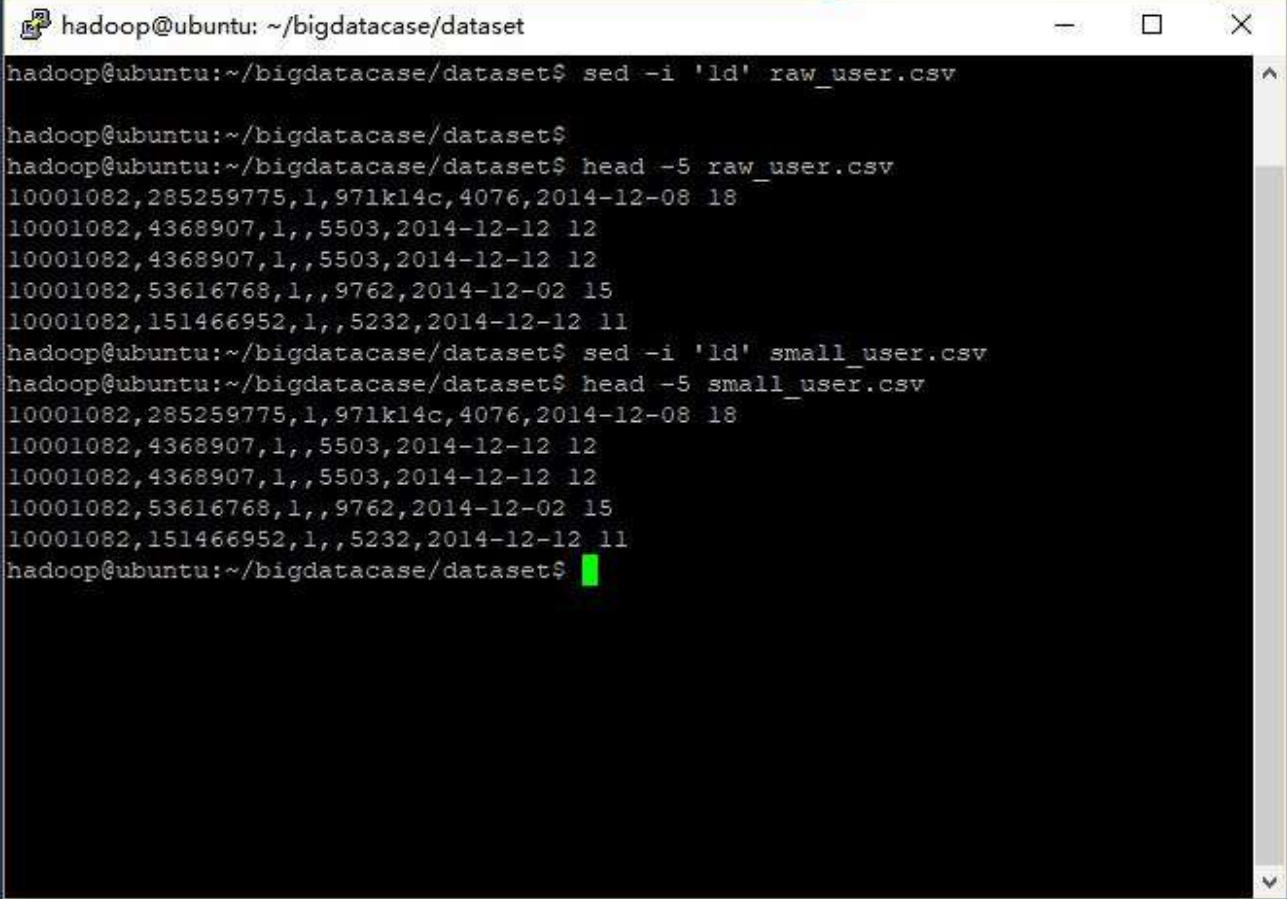
- 数据集下载与查看
 - 查看raw_user.csv前五条数据: `head -5 raw_user.csv`



A terminal window titled 'hadoop@ubuntu: ~/bigdatacase/dataset' with standard window controls. The terminal shows the following commands and output:

```
hadoop@ubuntu:~/bigdatacase/dataset$ ls
raw_user.csv  small_user.csv  user.zip
hadoop@ubuntu:~/bigdatacase/dataset$ head -5 raw_user.csv
user_id,item_id,behavior_type,user_geohash,item_category,time
10001082,285259775,1,971k14c,4076,2014-12-08 18
10001082,4368907,1,,5503,2014-12-12 12
10001082,4368907,1,,5503,2014-12-12 12
10001082,53616768,1,,9762,2014-12-02 15
hadoop@ubuntu:~/bigdatacase/dataset$
```

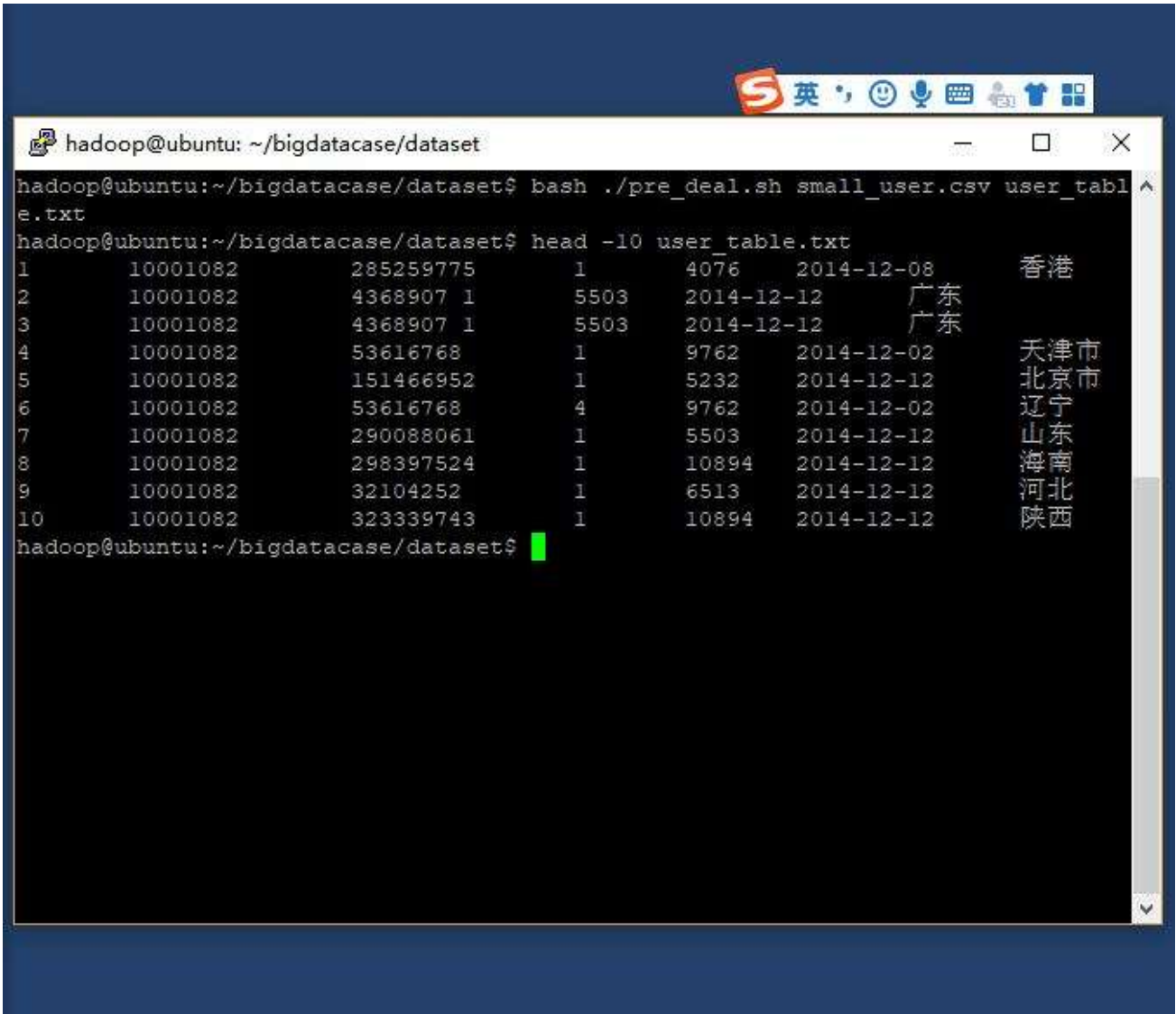
- 数据集预处理
 - 删除文件第一行记录，即字段名称



A terminal window titled 'hadoop@ubuntu: ~/bigdatacase/dataset' with standard Ubuntu window controls. The terminal shows the following commands and output:

```
hadoop@ubuntu:~/bigdatacase/dataset$ sed -i '1d' raw_user.csv
hadoop@ubuntu:~/bigdatacase/dataset$
hadoop@ubuntu:~/bigdatacase/dataset$ head -5 raw_user.csv
10001082,285259775,1,971k14c,4076,2014-12-08 18
10001082,4368907,1,,5503,2014-12-12 12
10001082,4368907,1,,5503,2014-12-12 12
10001082,53616768,1,,9762,2014-12-02 15
10001082,151466952,1,,5232,2014-12-12 11
hadoop@ubuntu:~/bigdatacase/dataset$ sed -i '1d' small_user.csv
hadoop@ubuntu:~/bigdatacase/dataset$ head -5 small_user.csv
10001082,285259775,1,971k14c,4076,2014-12-08 18
10001082,4368907,1,,5503,2014-12-12 12
10001082,4368907,1,,5503,2014-12-12 12
10001082,53616768,1,,9762,2014-12-02 15
10001082,151466952,1,,5232,2014-12-12 11
hadoop@ubuntu:~/bigdatacase/dataset$
```

- 对字段进行预处理
 - 为每行记录增加一个id字段（让记录具有唯一性）、增加一个省份字段（用来后续进行可视化分析），并且丢弃user_geohash字段（后面分析不需要这个字段）。



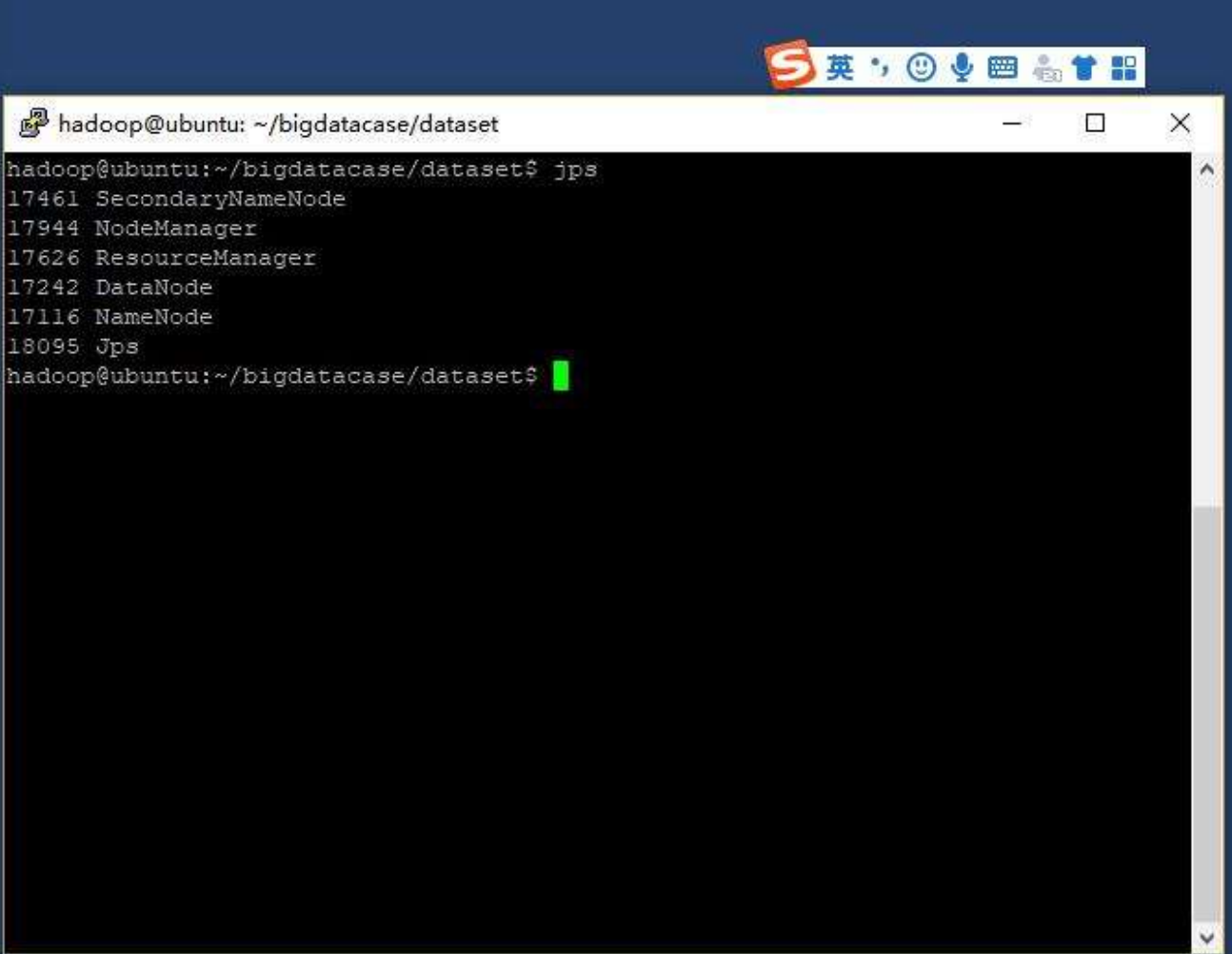
A terminal window titled 'hadoop@ubuntu: ~/bigdatacase/dataset' with standard window controls. The terminal shows the execution of a script and the head of a file.

```
hadoop@ubuntu:~/bigdatacase/dataset$ bash ./pre_deal.sh small_user.csv user_table.txt
hadoop@ubuntu:~/bigdatacase/dataset$ head -10 user_table.txt
```

Line	ID	Card	Age	Gender	Date	Location
1	10001082	285259775	1		2014-12-08	香港
2	10001082	4368907	1	5503	2014-12-12	广东
3	10001082	4368907	1	5503	2014-12-12	广东
4	10001082	53616768	1		2014-12-02	天津市
5	10001082	151466952	1		2014-12-12	北京市
6	10001082	53616768	4		2014-12-02	辽宁
7	10001082	290088061	1		2014-12-12	山东
8	10001082	298397524	1		2014-12-12	海南
9	10001082	32104252	1		2014-12-12	河北
10	10001082	323339743	1		2014-12-12	陕西

```
hadoop@ubuntu:~/bigdatacase/dataset$
```

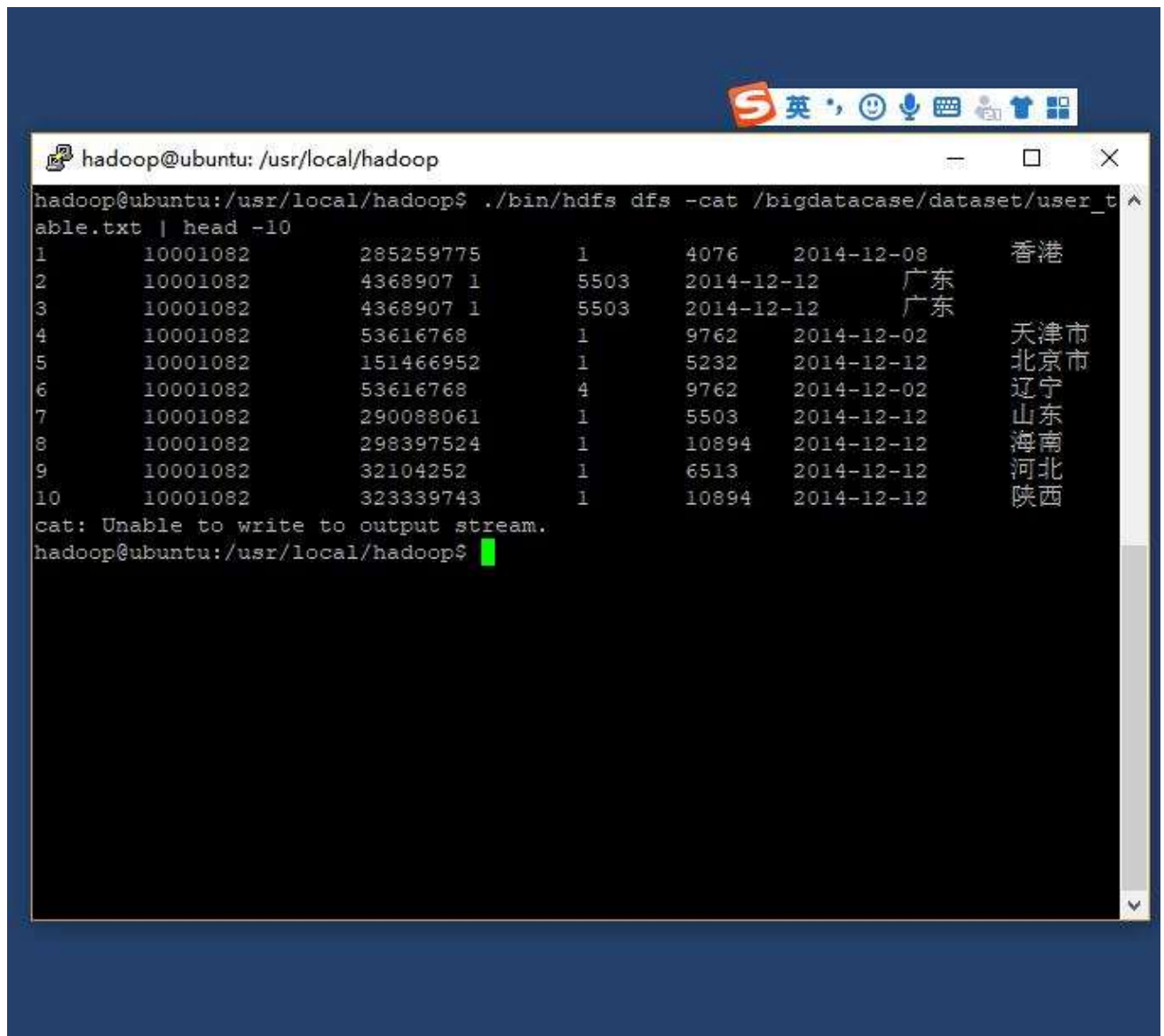
- 导入数据库
 - 启动hadoop: `./sbin/start-all.sh`



A terminal window titled 'hadoop@ubuntu: ~/bigdatacase/dataset' with standard window controls. The terminal output shows the command 'jps' and its results: '17461 SecondaryNameNode', '17944 NodeManager', '17626 ResourceManager', '17242 DataNode', '17116 NameNode', and '18095 Jps'. The prompt 'hadoop@ubuntu:~/bigdatacase/dataset\$' is followed by a green cursor.

```
hadoop@ubuntu: ~/bigdatacase/dataset
hadoop@ubuntu:~/bigdatacase/dataset$ jps
17461 SecondaryNameNode
17944 NodeManager
17626 ResourceManager
17242 DataNode
17116 NameNode
18095 Jps
hadoop@ubuntu:~/bigdatacase/dataset$
```

- 把usertable.txt上传到HDFS中, 查看一下HDFS中的usertable.txt的前10条记录

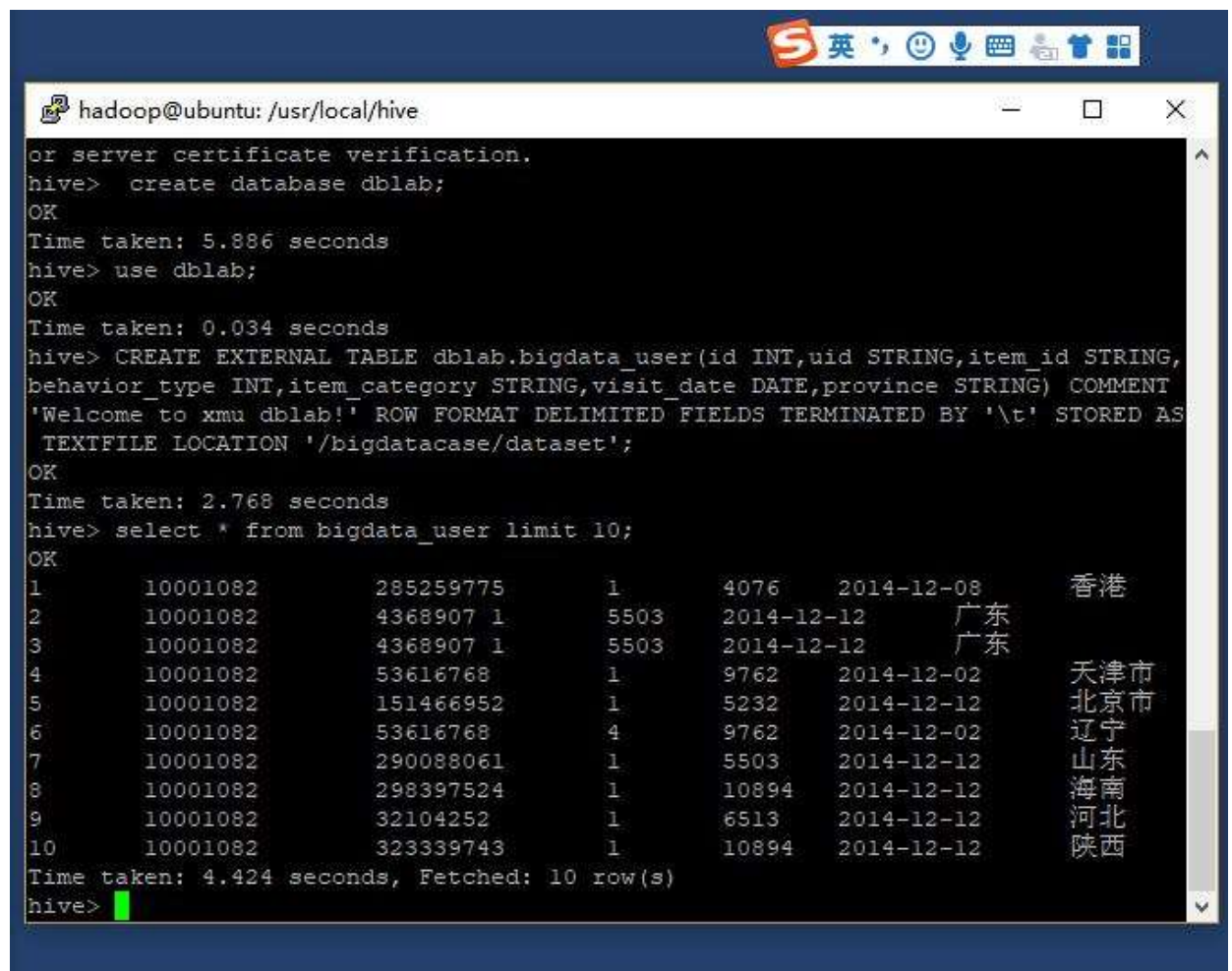


A terminal window titled 'hadoop@ubuntu: /usr/local/hadoop' displays the execution of the command `./bin/hdfs dfs -cat /bigdatacase/dataset/user_table.txt | head -10`. The output shows 10 lines of data with columns for line number, user ID, a numeric value, a count, a date, and a location. The locations are: 香港, 广东, 广东, 天津市, 北京市, 辽宁, 山东, 海南, 河北, and 陕西. An error message 'cat: Unable to write to output stream.' is visible at the bottom of the output.

Line	User ID	Value	Count	Date	Location
1	10001082	285259775	1	2014-12-08	香港
2	10001082	4368907	1	2014-12-12	广东
3	10001082	4368907	1	2014-12-12	广东
4	10001082	53616768	1	2014-12-02	天津市
5	10001082	151466952	1	2014-12-12	北京市
6	10001082	53616768	4	2014-12-02	辽宁
7	10001082	290088061	1	2014-12-12	山东
8	10001082	298397524	1	2014-12-12	海南
9	10001082	32104252	1	2014-12-12	河北
10	10001082	323339743	1	2014-12-12	陕西

cat: Unable to write to output stream.
hadoop@ubuntu: /usr/local/hadoop\$

- 在Hive上创建数据库，创建外部表，查询数据
 - `select * from bigdata_user limit 10;`

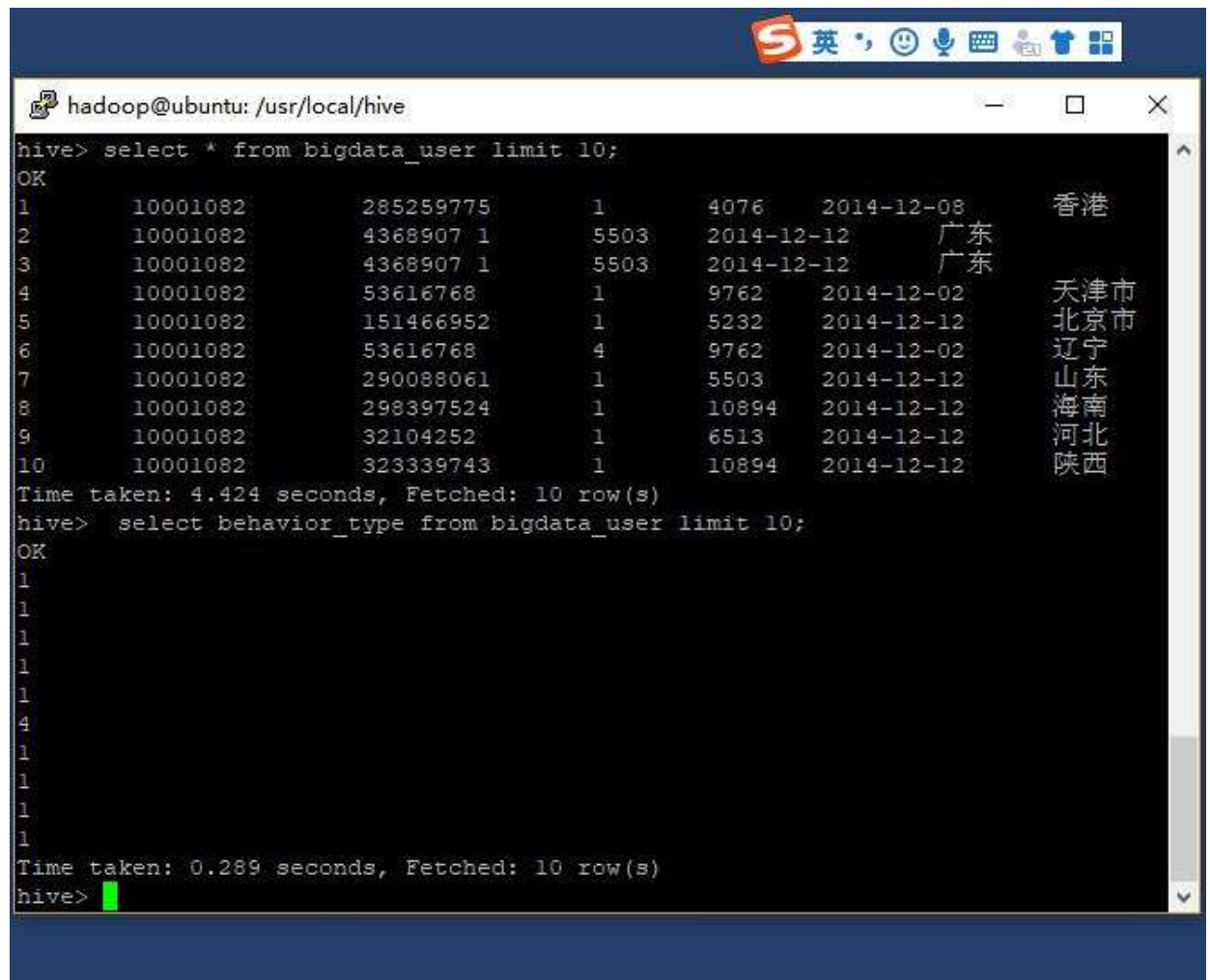


```

hadoop@ubuntu: /usr/local/hive
or server certificate verification.
hive> create database dblab;
OK
Time taken: 5.886 seconds
hive> use dblab;
OK
Time taken: 0.034 seconds
hive> CREATE EXTERNAL TABLE dblab.bigdata_user(id INT,uid STRING,item_id STRING,
behavior_type INT,item_category STRING,visit_date DATE,province STRING) COMMENT
'Welcome to xmu dblab!' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS
TEXTFILE LOCATION '/bigdatacase/dataset';
OK
Time taken: 2.768 seconds
hive> select * from bigdata_user limit 10;
OK
1      10001082      285259775      1      4076      2014-12-08      香港
2      10001082      4368907 1      5503      2014-12-12      广东
3      10001082      4368907 1      5503      2014-12-12      广东
4      10001082      53616768      1      9762      2014-12-02      天津市
5      10001082      151466952      1      5232      2014-12-12      北京市
6      10001082      53616768      4      9762      2014-12-02      辽宁
7      10001082      290088061      1      5503      2014-12-12      山东
8      10001082      298397524      1      10894      2014-12-12      海南
9      10001082      32104252      1      6513      2014-12-12      河北
10     10001082      323339743      1      10894      2014-12-12      陕西
Time taken: 4.424 seconds, Fetched: 10 row(s)
hive>

```

- select behavior_type from bigdatauser limit 10;



The image shows a terminal window titled 'hadoop@ubuntu: /usr/local/hive'. It contains two Hive queries and their results. The first query selects all columns from 'bigdata_user' with a limit of 10, returning 10 rows of user data including IDs, phone numbers, genders, ages, birth dates, and locations. The second query selects 'behavior_type' from 'bigdata_user' with a limit of 10, returning 10 rows of behavior types.

```
hadoop@ubuntu: /usr/local/hive
hive> select * from bigdata_user limit 10;
OK
1      10001082      285259775      1      4076      2014-12-08      香港
2      10001082      4368907 1      5503      2014-12-12      广东
3      10001082      4368907 1      5503      2014-12-12      广东
4      10001082      53616768      1      9762      2014-12-02      天津市
5      10001082      151466952      1      5232      2014-12-12      北京市
6      10001082      53616768      4      9762      2014-12-02      辽宁
7      10001082      290088061      1      5503      2014-12-12      山东
8      10001082      298397524      1      10894      2014-12-12      海南
9      10001082      32104252      1      6513      2014-12-12      河北
10     10001082      323339743      1      10894      2014-12-12      陕西
Time taken: 4.424 seconds, Fetched: 10 row(s)
hive> select behavior_type from bigdata_user limit 10;
OK
1
1
1
1
1
1
4
1
1
1
1
Time taken: 0.289 seconds, Fetched: 10 row(s)
hive>
```