



## 大数据课程开发实践期末报告 1

实验题目：\_\_\_\_\_ 淘宝双 11 数据分析 \_\_\_\_\_

姓 名：\_\_\_\_\_ 姚光明 \_\_\_\_\_

学 号：\_\_\_\_\_ 20051133 \_\_\_\_\_

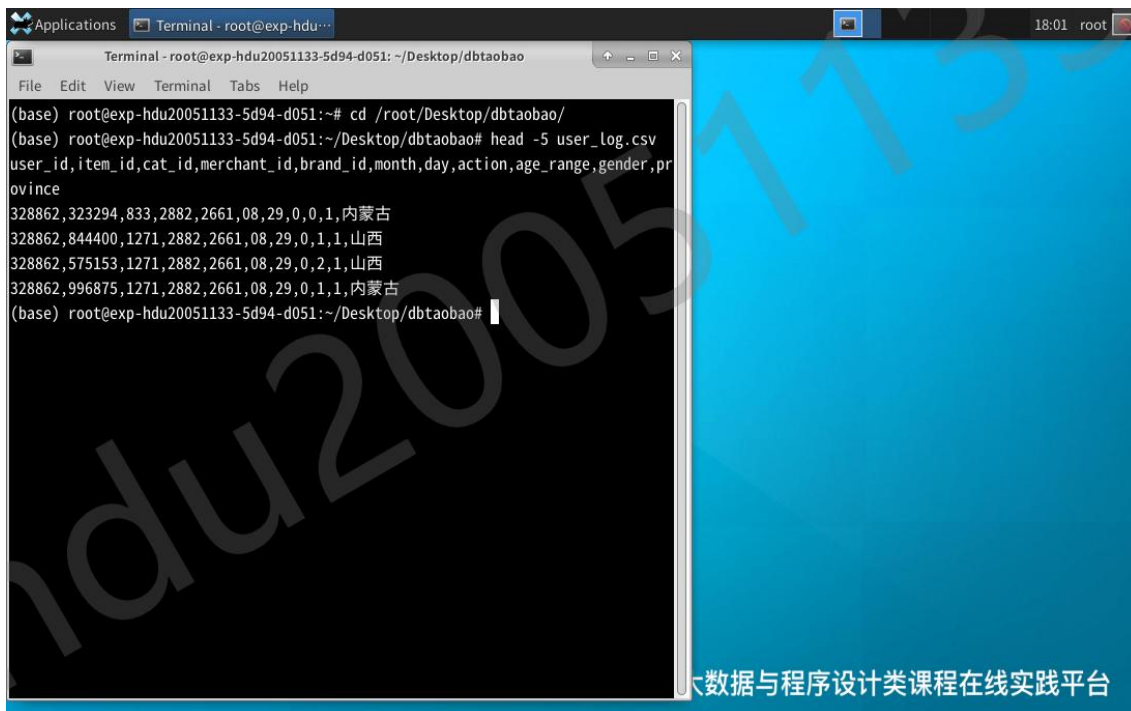
专 业：\_\_\_\_\_ 计算机科学与技术 \_\_\_\_\_

## 目 录

1	HDFS 数据处理 .....	2
2	Hive 查询案例 .....	4
3	Spark 基础编程 .....	10
4	Spark 回头客预测 .....	12

## 1 HDFS 数据处理

### 1.1 数据查看与预处理

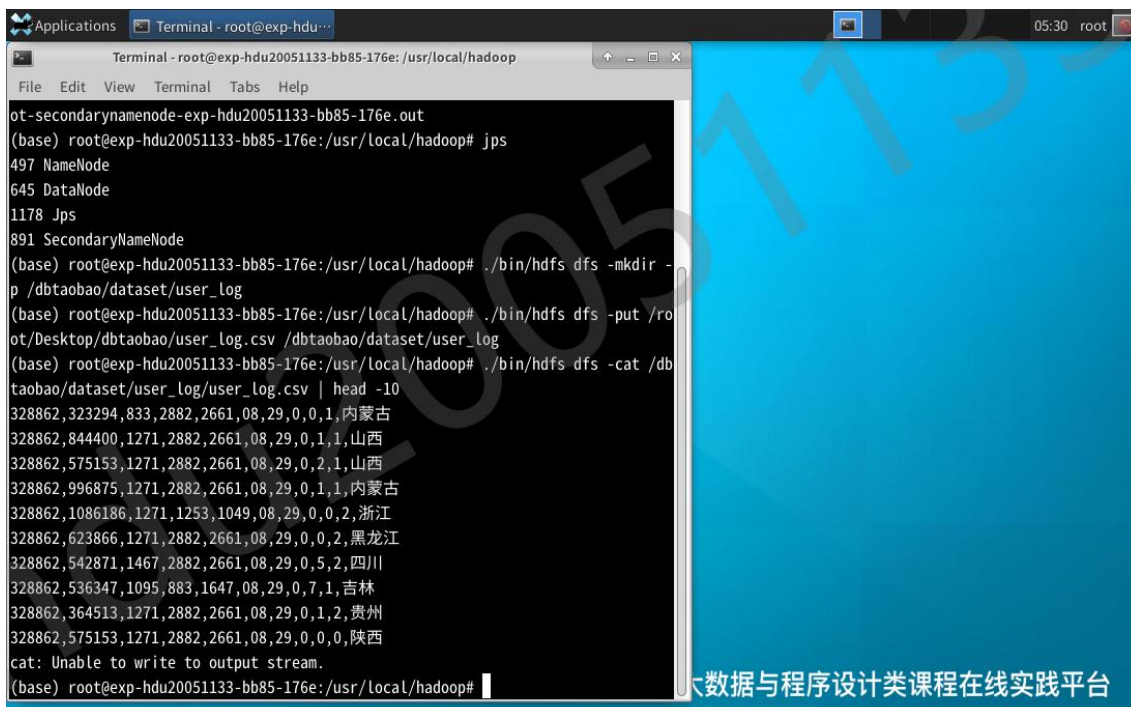


A terminal window titled "Terminal - root@exp-hdu..." shows the following commands and output:

```
(base) root@exp-hdu20051133-5d94-d051:~# cd /root/Desktop/dbtaobao/
(base) root@exp-hdu20051133-5d94-d051:~/Desktop/dbtaobao# head -5 user_log.csv
user_id,item_id,cat_id,merchant_id,brand_id,month,day,action,age_range,gender,province
328862,323294,833,2882,2661,08,29,0,0,1,内蒙古
328862,844400,1271,2882,2661,08,29,0,1,1,山西
328862,575153,1271,2882,2661,08,29,0,2,1,山西
328862,996875,1271,2882,2661,08,29,0,1,1,内蒙古
(base) root@exp-hdu20051133-5d94-d051:~/Desktop/dbtaobao#
```

大数据与程序设计类课程在线实践平台

### 1.2 启动 HDFS

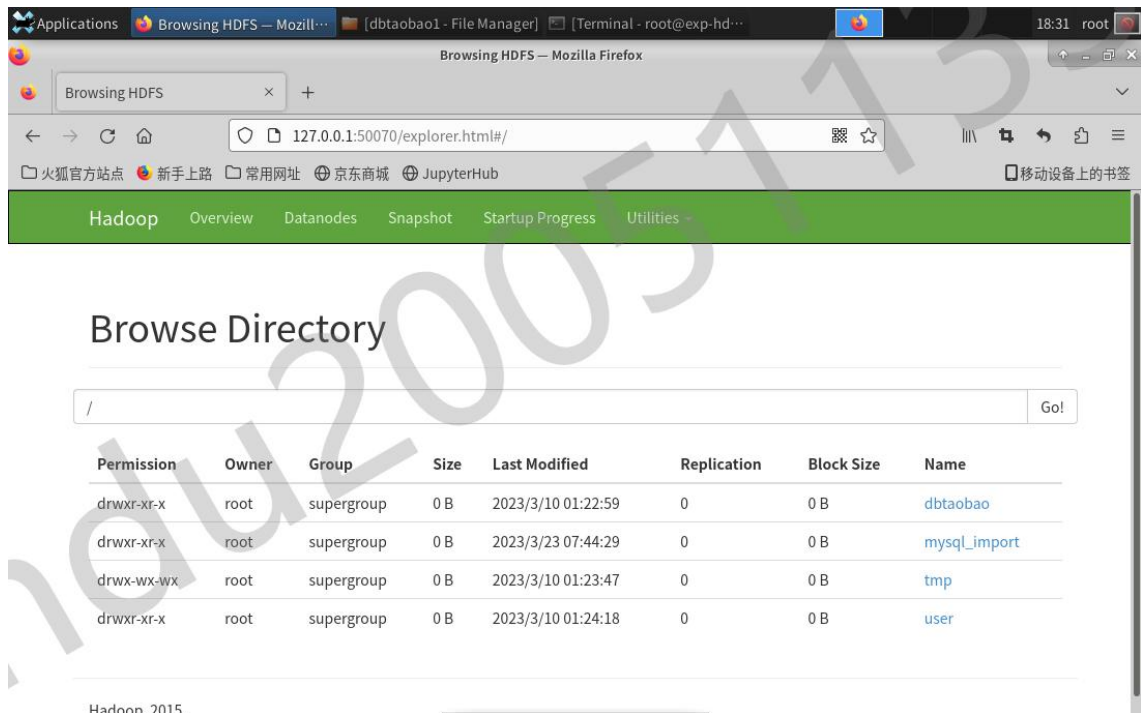


A terminal window titled "Terminal - root@exp-hdu..." shows the following commands and output:

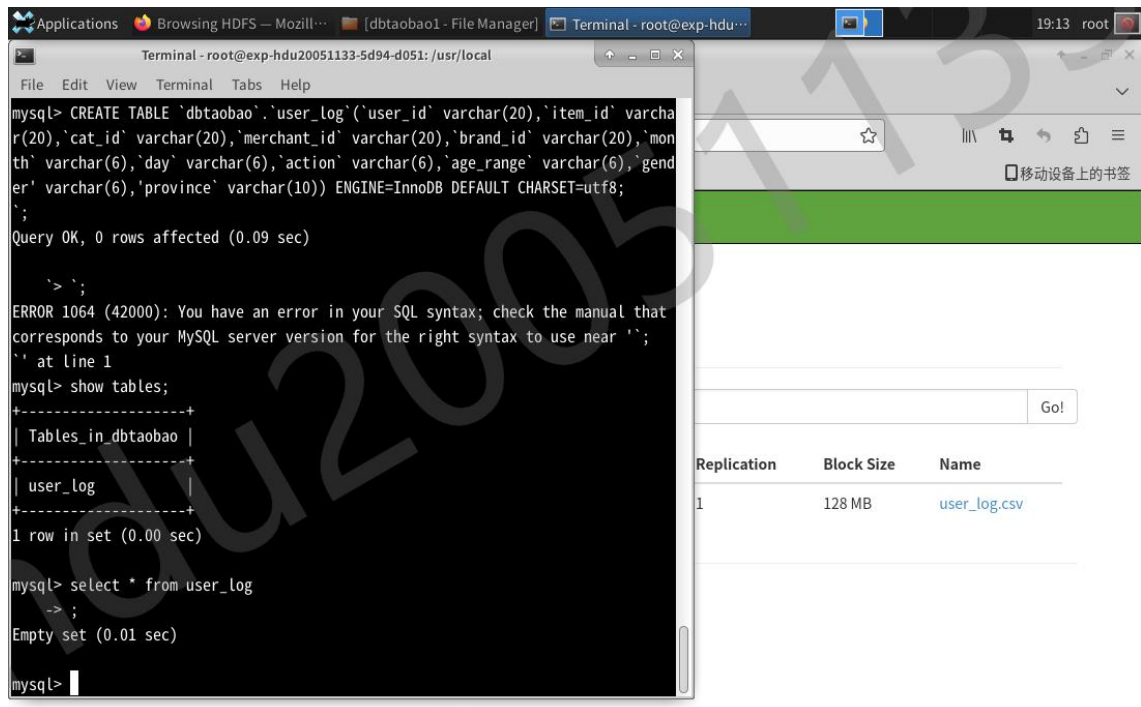
```
ot-secondarynamenode-exp-hdu20051133-bb85-176e.out
(base) root@exp-hdu20051133-bb85-176e:/usr/local/hadoop# jps
497 NameNode
645 DataNode
1178 Jps
891 SecondaryNameNode
(base) root@exp-hdu20051133-bb85-176e:/usr/local/hadoop# ./bin/hdfs dfs -mkdir -p /dbtaobao/dataset/user_log
(base) root@exp-hdu20051133-bb85-176e:/usr/local/hadoop# ./bin/hdfs dfs -put /root/Desktop/dbtaobao/user_log.csv /dbtaobao/dataset/user_log
(base) root@exp-hdu20051133-bb85-176e:/usr/local/hadoop# ./bin/hdfs dfs -cat /dbtaobao/dataset/user_log/user_log.csv | head -10
328862,323294,833,2882,2661,08,29,0,0,1,内蒙古
328862,844400,1271,2882,2661,08,29,0,1,1,山西
328862,575153,1271,2882,2661,08,29,0,2,1,山西
328862,996875,1271,2882,2661,08,29,0,1,1,内蒙古
328862,1086186,1271,1253,1049,08,29,0,0,2,浙江
328862,623866,1271,2882,2661,08,29,0,0,2,黑龙江
328862,542871,1467,2882,2661,08,29,0,5,2,四川
328862,536347,1095,883,1647,08,29,0,7,1,吉林
328862,364513,1271,2882,2661,08,29,0,1,2,贵州
328862,575153,1271,2882,2661,08,29,0,0,0,陕西
cat: Unable to write to output stream.
(base) root@exp-hdu20051133-bb85-176e:/usr/local/hadoop#
```

大数据与程序设计类课程在线实践平台

### 1.3 将数据传入到 HDFS 中



### 1.4 sqoop 导出 HDFS 数据到 Mysql



The screenshot shows a terminal window with the following commands and output:

```

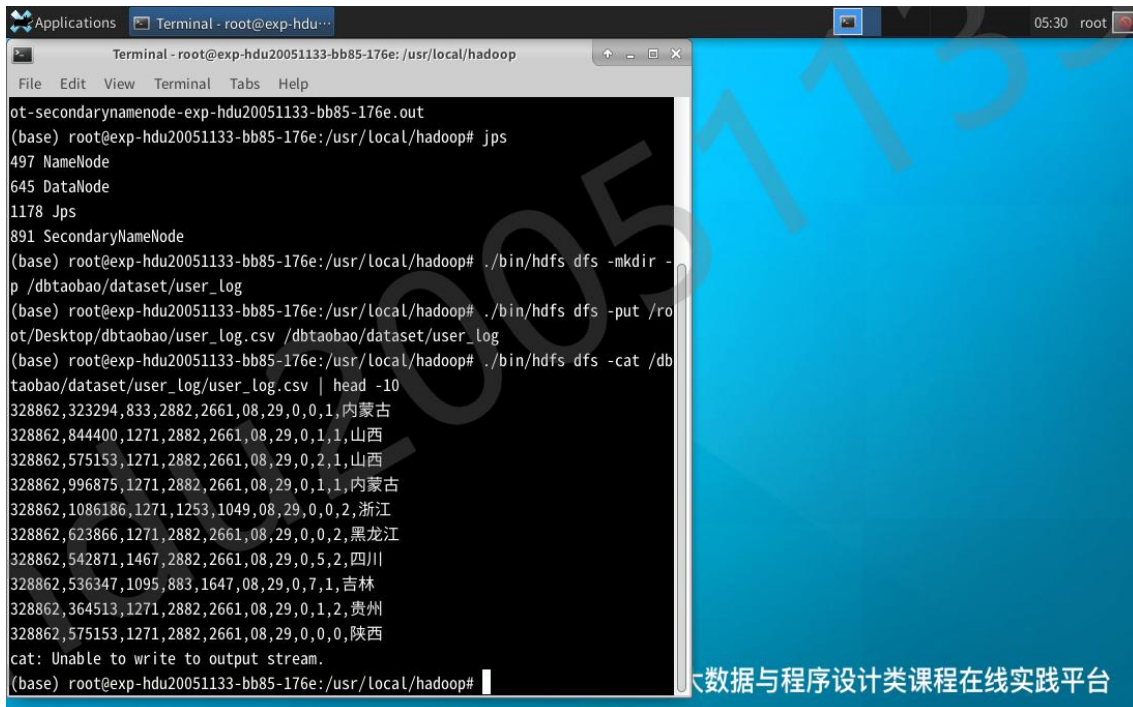
ot-secondarynamenode-exp-hdu20051133-5d94-d051.out
(base) root@exp-hdu20051133-5d94-d051:/usr/local/hadoop# jps
768 DataNode
594 NameNode
1028 SecondaryNameNode
1275 Jps
(base) root@exp-hdu20051133-5d94-d051:/usr/local/hadoop# ./bin/hdfs dfs -mkdir -p /dbtaobao/dataset/user_log
(base) root@exp-hdu20051133-5d94-d051:/usr/local/hadoop# ./bin/hdfs dfs -put /root/Desktop/dbtaobao1/user_log.csv /dbtaobao/dataset/user_log
(base) root@exp-hdu20051133-5d94-d051:/usr/local/hadoop# ./bin/hdfs dfs -cat /dbtaobao/dataset/user_log/user_log.csv | head -10
328862,323294,833,2882,2661,08,29,0,0,1,内蒙古
328862,844400,1271,2882,2661,08,29,0,1,1,山西
328862,575153,1271,2882,2661,08,29,0,2,1,山西
328862,996875,1271,2882,2661,08,29,0,1,1,内蒙古
328862,1086186,1271,1253,1049,08,29,0,0,2,浙江
328862,623866,1271,2882,2661,08,29,0,0,2,黑龙江
328862,542871,1467,2882,2661,08,29,0,5,2,四川
328862,536347,1095,883,1647,08,29,0,7,1,吉林
328862,364513,1271,2882,2661,08,29,0,1,2,贵州
328862,575153,1271,2882,2661,08,29,0,0,0,陕西
cat: Unable to write to output stream.
(base) root@exp-hdu20051133-5d94-d051:/usr/local/hadoop#
  
```

The web browser shows the HDFS file listing for `user_log.csv`:

Replication	Block Size	Name
1	128 MB	<a href="#">user_log.csv</a>

## 2 Hive 查询案例

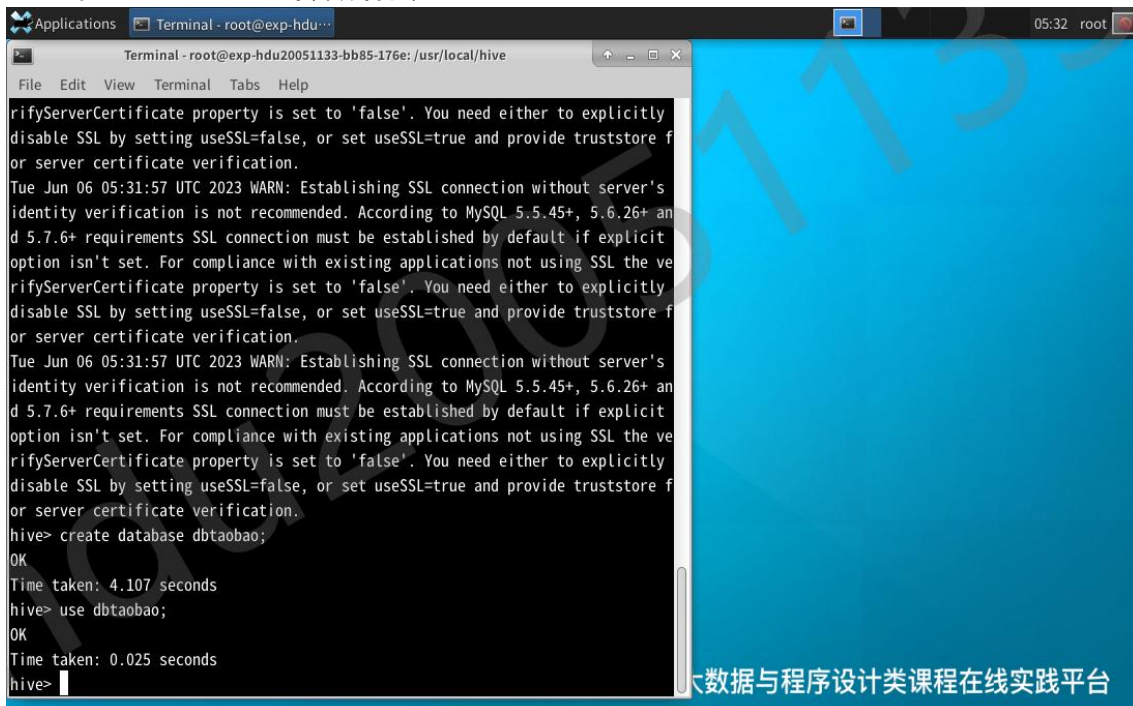
### 2.1 将数据传入 HDFS 中

A terminal window titled 'Terminal - root@exp-hdu...' shows a series of HDFS commands and their outputs. The user is in the directory /usr/local/hadoop. They run 'jps' and see processes for NameNode, DataNode, and SecondaryNameNode. Then they create a directory /dbtaobao/dataset/user\_log, upload a local CSV file to it, and use 'cat' to display the first 10 lines of the file. The data represents user logs with fields like user ID, age, gender, and province.

```
Terminal - root@exp-hdu20051133-bb85-176e: /usr/local/hadoop
File Edit View Terminal Tabs Help

ot-secondarynamenode-exp-hdu20051133-bb85-176e.out
(base) root@exp-hdu20051133-bb85-176e: /usr/local/hadoop# jps
497 NameNode
645 DataNode
1178 Jps
891 SecondaryNameNode
(base) root@exp-hdu20051133-bb85-176e: /usr/local/hadoop# ./bin/hdfs dfs -mkdir -p /dbtaobao/dataset/user_log
(base) root@exp-hdu20051133-bb85-176e: /usr/local/hadoop# ./bin/hdfs dfs -put /root/Desktop/dbtaobao/user_log.csv /dbtaobao/dataset/user_log
(base) root@exp-hdu20051133-bb85-176e: /usr/local/hadoop# ./bin/hdfs dfs -cat /dbtaobao/dataset/user_log/user_log.csv | head -10
328862,323294,833,2882,2661,08,29,0,0,1,内蒙古
328862,844400,1271,2882,2661,08,29,0,1,1,山西
328862,575153,1271,2882,2661,08,29,0,2,1,山西
328862,996875,1271,2882,2661,08,29,0,1,1,内蒙古
328862,1086186,1271,1253,1049,08,29,0,0,2,浙江
328862,623866,1271,2882,2661,08,29,0,0,2,黑龙江
328862,542871,1467,2882,2661,08,29,0,5,2,四川
328862,536347,1095,883,1647,08,29,0,7,1,吉林
328862,364513,1271,2882,2661,08,29,0,1,2,贵州
328862,575153,1271,2882,2661,08,29,0,0,0,陕西
cat: Unable to write to output stream.
(base) root@exp-hdu20051133-bb85-176e: /usr/local/hadoop#
```

### 2.2 把 HDFS 上的数据传入 hive

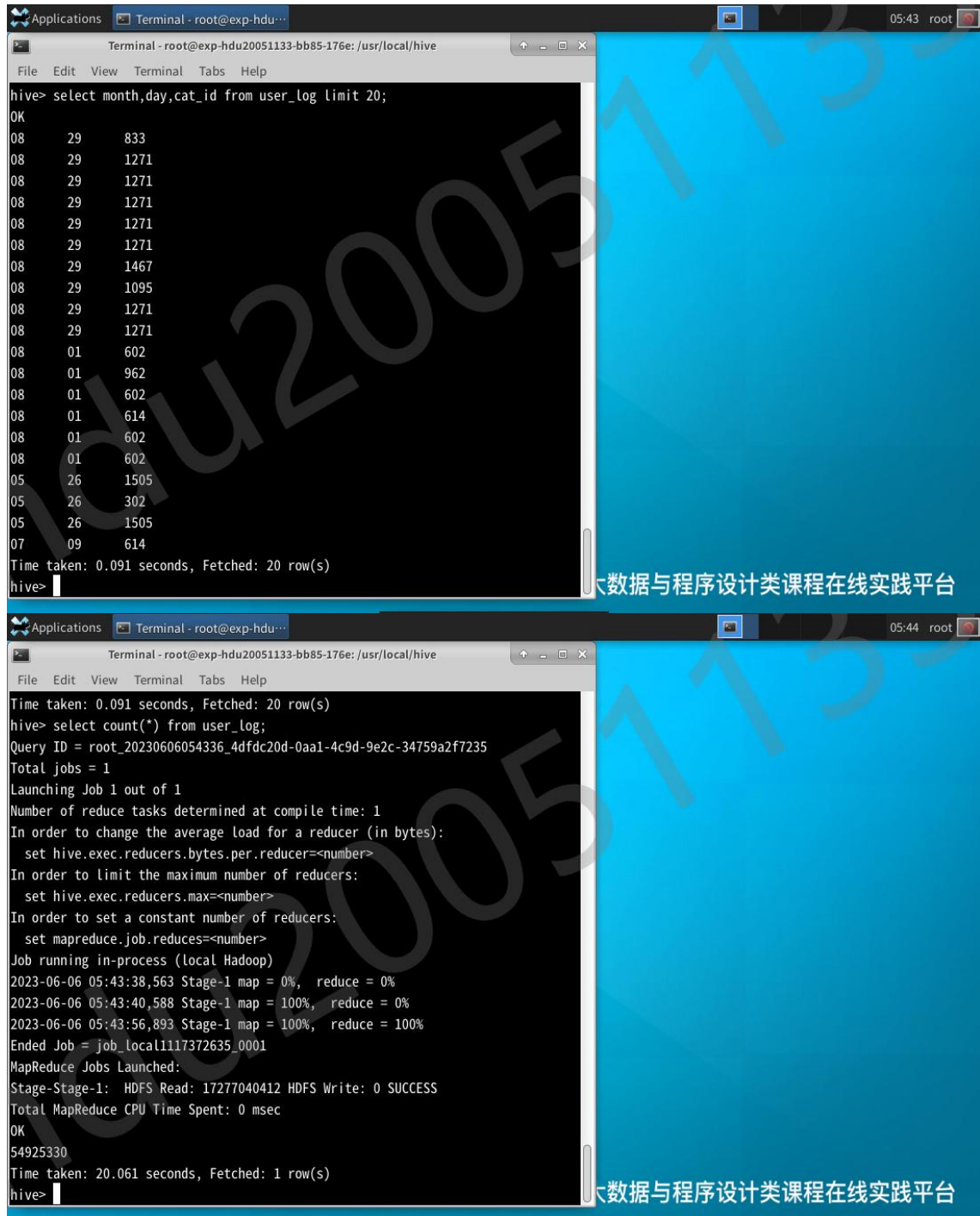
A terminal window titled 'Terminal - root@exp-hdu...' shows Hive commands. It starts with a warning about SSL connection. The user creates a database named 'dbtaobao', switches to it, and the terminal shows the time taken for each operation.

```
Terminal - root@exp-hdu20051133-bb85-176e: /usr/local/hive
File Edit View Terminal Tabs Help

rifyServerCertificate property is set to 'false'. You need either to explicitly
disable SSL by setting useSSL=false, or set useSSL=true and provide truststore f
or server certificate verification.
Tue Jun 06 05:31:57 UTC 2023 WARN: Establishing SSL connection without server's
identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ an
d 5.7.6+ requirements SSL connection must be established by default if explicit
option isn't set. For compliance with existing applications not using SSL the ve
rifyServerCertificate property is set to 'false'. You need either to explicitly
disable SSL by setting useSSL=false, or set useSSL=true and provide truststore f
or server certificate verification.
Tue Jun 06 05:31:57 UTC 2023 WARN: Establishing SSL connection without server's
identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ an
d 5.7.6+ requirements SSL connection must be established by default if explicit
option isn't set. For compliance with existing applications not using SSL the ve
rifyServerCertificate property is set to 'false'. You need either to explicitly
disable SSL by setting useSSL=false, or set useSSL=true and provide truststore f
or server certificate verification.
hive> create database dbtaobao;
OK
Time taken: 4.107 seconds
hive> use dbtaobao;
OK
Time taken: 0.025 seconds
hive>
```

### 2.3 查询数据





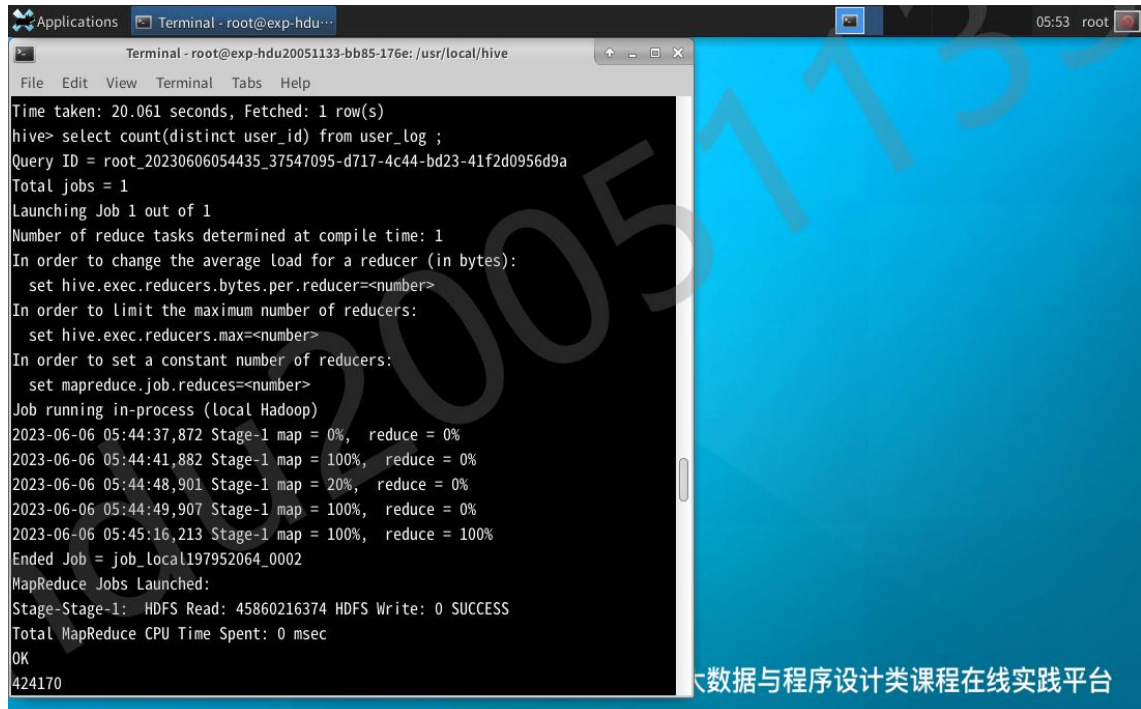
The image displays two screenshots of a Hive terminal window. The top screenshot shows the execution of a SQL query to select the first 20 rows from the 'user\_log' table, displaying columns for month, day, and cat\_id. The bottom screenshot shows the execution of a COUNT query on the same table, followed by a series of status messages and progress updates from the Hive engine.

```
Applications Terminal - root@exp-hdu... 05:43 root
Terminal - root@exp-hdu20051133-bb85-176e: /usr/local/hive
File Edit View Terminal Tabs Help
hive> select month,day,cat_id from user_log limit 20;
OK
08 29 833
08 29 1271
08 29 1271
08 29 1271
08 29 1271
08 29 1271
08 29 1467
08 29 1095
08 29 1271
08 29 1271
08 01 602
08 01 962
08 01 602
08 01 614
08 01 602
08 01 602
05 26 1505
05 26 302
05 26 1505
07 09 614
Time taken: 0.091 seconds, Fetched: 20 row(s)
hive>
```

大数据与程序设计类课程在线实践平台

```
Applications Terminal - root@exp-hdu... 05:44 root
Terminal - root@exp-hdu20051133-bb85-176e: /usr/local/hive
File Edit View Terminal Tabs Help
Time taken: 0.091 seconds, Fetched: 20 row(s)
hive> select count(*) from user_log;
Query ID = root_20230606054336_4dfdc20d-0aa1-4c9d-9e2c-34759a2f7235
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (Local Hadoop)
2023-06-06 05:43:38,563 Stage-1 map = 0%, reduce = 0%
2023-06-06 05:43:40,588 Stage-1 map = 100%, reduce = 0%
2023-06-06 05:43:56,893 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local1117372635_0001
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 17277040412 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
54925330
Time taken: 20.061 seconds, Fetched: 1 row(s)
hive>
```

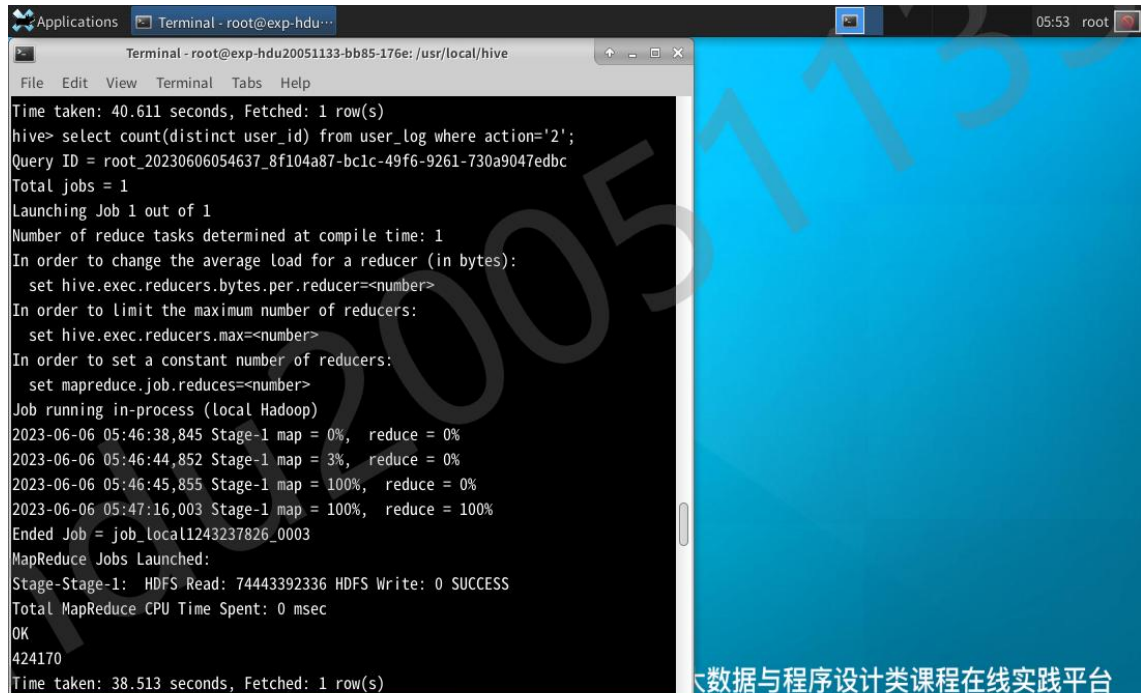
大数据与程序设计类课程在线实践平台



```
Applications Terminal - root@exp-hdu... 05:53 root
Terminal - root@exp-hdu20051133-bb85-176e: /usr/local/hive
File Edit View Terminal Tabs Help
Time taken: 20.061 seconds, Fetched: 1 row(s)
hive> select count(distinct user_id) from user_log ;
Query ID = root_20230606054435_37547095-d717-4c44-bd23-41f2d0956d9a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (Local Hadoop)
2023-06-06 05:44:37,872 Stage-1 map = 0%, reduce = 0%
2023-06-06 05:44:41,882 Stage-1 map = 100%, reduce = 0%
2023-06-06 05:44:48,901 Stage-1 map = 20%, reduce = 0%
2023-06-06 05:44:49,907 Stage-1 map = 100%, reduce = 0%
2023-06-06 05:45:16,213 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local197952064_0002
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 45860216374 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
424170
```

大数据与程序设计类课程在线实践平台


## 2.4 查询



```
Applications Terminal - root@exp-hdu... 05:53 root
Terminal - root@exp-hdu20051133-bb85-176e: /usr/local/hive
File Edit View Terminal Tabs Help
Time taken: 40.611 seconds, Fetched: 1 row(s)
hive> select count(distinct user_id) from user_log where action='2';
Query ID = root_20230606054637_8f104a87-bc1c-49f6-9261-730a9047edbc
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (Local Hadoop)
2023-06-06 05:46:38,845 Stage-1 map = 0%, reduce = 0%
2023-06-06 05:46:44,852 Stage-1 map = 3%, reduce = 0%
2023-06-06 05:46:45,855 Stage-1 map = 100%, reduce = 0%
2023-06-06 05:47:16,003 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local1243237826_0003
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 74443392336 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
424170
Time taken: 38.513 seconds, Fetched: 1 row(s)
```

大数据与程序设计类课程在线实践平台





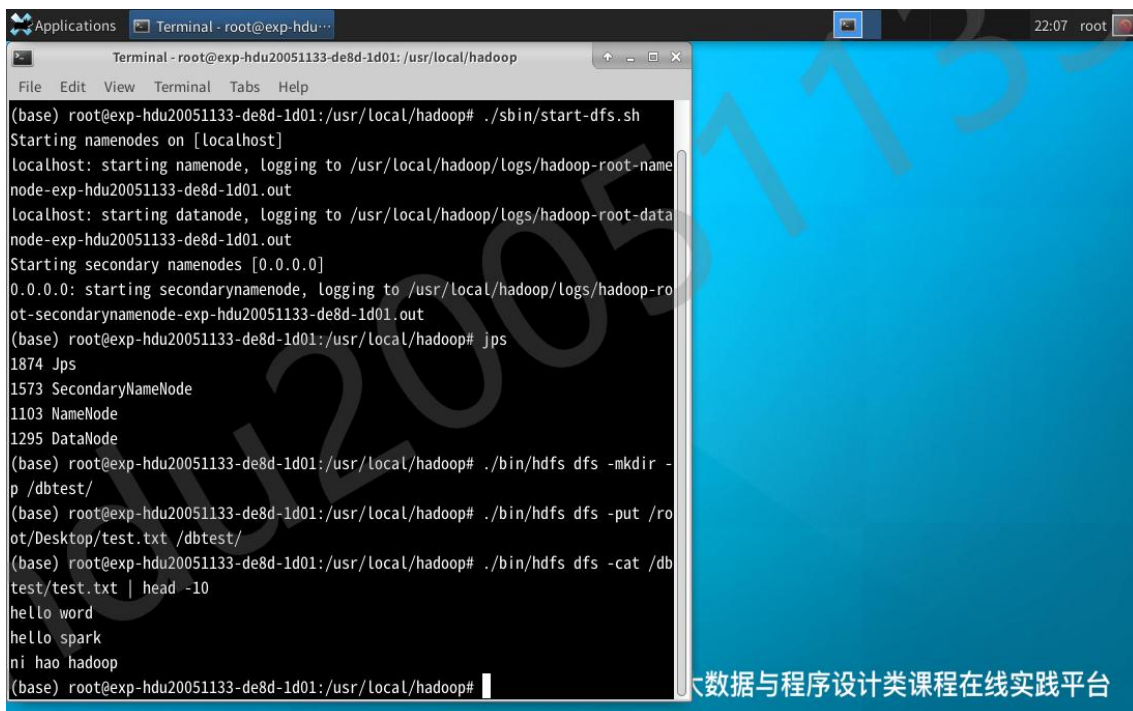
```
Applications Terminal - root@exp-hdu... 05:53 root
Terminal - root@exp-hdu20051133-bb85-176e: /usr/local/hive
File Edit View Terminal Tabs Help
hive> select user_id from user_log where action='2' group by user_id having count(action='2')>5 limit 10;
Query ID = root_20230606054900_4369a5cc-d602-4b9d-9be1-58f4209ec5d3
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 11
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (Local Hadoop)
2023-06-06 05:49:02,082 Stage-1 map = 0%, reduce = 0%
2023-06-06 05:49:05,087 Stage-1 map = 100%, reduce = 0%
2023-06-06 05:49:32,191 Stage-1 map = 100%, reduce = 27%
2023-06-06 05:49:33,198 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local1950233245_0004
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 206965512858 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
22
44
大数据与程序设计类课程在线实践平台

Applications Terminal - root@exp-hdu... 05:53 root
Terminal - root@exp-hdu20051133-bb85-176e: /usr/local/hive
File Edit View Terminal Tabs Help
Time taken: 31.533 seconds, Fetched: 1 row(s)
hive> select count(*) from user_log where gender='1';
Query ID = root_20230606055151_0e1bb834-67ca-4a87-b4d8-ec8dc2667e4c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (Local Hadoop)
2023-06-06 05:51:52,874 Stage-1 map = 0%, reduce = 0%
2023-06-06 05:51:55,878 Stage-1 map = 100%, reduce = 0%
2023-06-06 05:52:22,922 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local1855070104_0006
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 160192920222 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
18311977
Time taken: 31.322 seconds, Fetched: 1 row(s)
hive>
```



## 3 Spark 基础编程

### 3.1 环境配置

A screenshot of a terminal window titled "Terminal - root@exp-hdu...". The terminal shows the execution of Hadoop commands to start the distributed filesystem (DFS) and perform basic file operations. The background of the terminal window is a blue gradient with a large, faint watermark "2025". The terminal output includes the following commands and their results:

```
(base) root@exp-hdu20051133-de8d-1d01: /usr/local/hadoop# ./sbin/start-dfs.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-name
node-exp-hdu20051133-de8d-1d01.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-data
node-exp-hdu20051133-de8d-1d01.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-ro
ot-secondarynamenode-exp-hdu20051133-de8d-1d01.out
(base) root@exp-hdu20051133-de8d-1d01: /usr/local/hadoop# jps
1874 Jps
1573 SecondaryNameNode
1103 NameNode
1295 DataNode
(base) root@exp-hdu20051133-de8d-1d01: /usr/local/hadoop# ./bin/hdfs dfs -mkdir -
p /dbtest/
(base) root@exp-hdu20051133-de8d-1d01: /usr/local/hadoop# ./bin/hdfs dfs -put /ro
ot/Desktop/test.txt /dbtest/
(base) root@exp-hdu20051133-de8d-1d01: /usr/local/hadoop# ./bin/hdfs dfs -cat /db
test/test.txt | head -10
hello word
hello spark
ni hao hadoop
(base) root@exp-hdu20051133-de8d-1d01: /usr/local/hadoop#
```

大数据与程序设计类课程在线实践平台

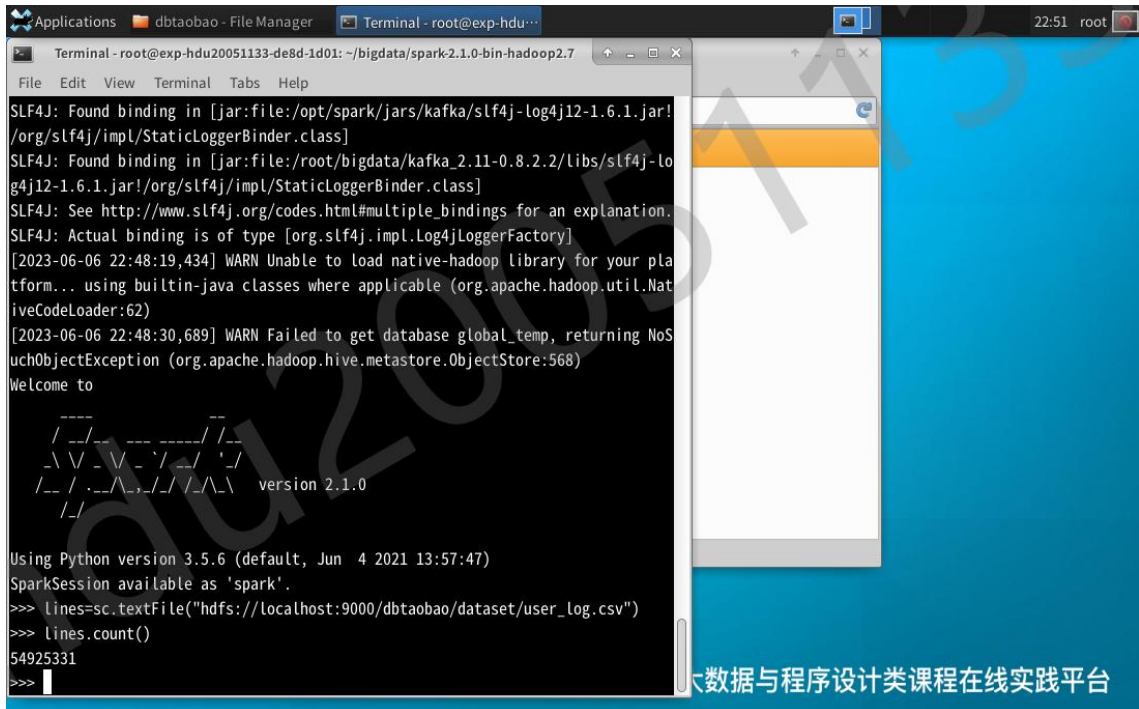
## 3.2 RDD 操作

The screenshot shows a terminal window titled "Terminal - root@exp-hdu...". The prompt indicates the user is at "/bigdata/spark-2.1.0-bin-hadoop2.7". The output displays the ASCII art logo for Apache Spark version 2.1.0, followed by information about Python version 3.5.6 being used as the default interpreter. Then, a series of Scala commands are executed:  

```
>>> Lines=sc.textFile("hdfs://localhost:9000/dbtest/test.txt")  
>>> Lines.count()  
3  
>>> Lines.take(1)  
['hello word']  
>>> LinesWithHello=Lines.filter(lambda line:"hello" in line)  
>>> LinesWithHello.count()  
2  
>>> Linesmap =Lines.map(lambda line: len(line.split(' ')))  
>>> Linesmap.foreach(print)  
2  
2  
3  
>>> Linesmap.reduce(lambda a,b:a+b)  
7  
>>> Linesmap =lines.flatMap(lambda x: x.split(' '))  
>>> Linesmap.foreach(print)
```

The background features a large blue watermark reading "大数据与程序设计类课程在线实践平台".

```
Applications Terminal - root@exp-hdu... 22:17 root
Terminal - root@exp-hdu20051133-de8d-1d01: ~/bigdata/spark-2.1.0-bin-hadoop2.7
File Edit View Terminal Tabs Help
2
3
>>> linesmap.reduce(lambda a,b:a+b)
7
>>> linesmap = lines.flatMap(lambda x: x.split(' '))
>>> linesmap.foreach(print)
hello
word
hello
spark
ni
hao
hadoop
>>> linesmap = lines.map(lambda x:(x,1))
>>> linesmap.foreach(print)
('hello word', 1)
('hello spark', 1)
('ni hao hadoop', 1)
>>> linesmap = linesmap.reduceByKey(lambda a,b :a+b)
>>> linesmap.foreach(print)
('ni hao hadoop', 1)
('hello spark', 1)
('hello word', 1)
>>>
```



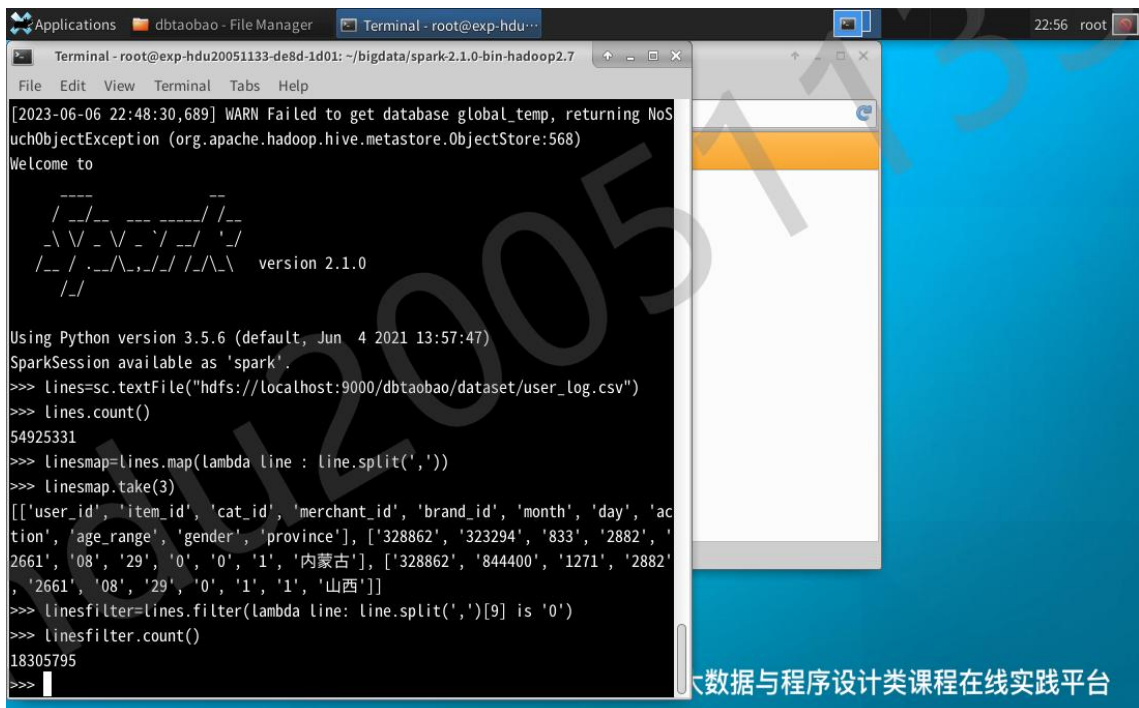
```
SLF4J: Found binding in [jar:file:/opt/spark/jars/kafka/slf4j-log4j12-1.6.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/root/bigdata/kafka_2.11-0.8.2.2/libs/slf4j-log4j12-1.6.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
[2023-06-06 22:48:19,434] WARN Unable to load native-hadoop library for your platform... using builtin-java classes where applicable (org.apache.hadoop.util.NativeCodeLoader:62)
[2023-06-06 22:48:30,689] WARN Failed to get database global_temp, returning NoSuchObjectException (org.apache.hadoop.hive.metastore.ObjectStore:568)
Welcome to

  ____
 /  __ \  _ __ ___
 \  __ \| '_ \ / __|
  | |__) | |_) | | |_\__|
  |____/|___|_|_|_|\___|
                        version 2.1.0

Using Python version 3.5.6 (default, Jun  4 2021 13:57:47)
SparkSession available as 'spark'.
>>> lines=sc.textFile("hdfs://localhost:9000/dbtaobao/dataset/user_log.csv")
>>> lines.count()
54925331
>>>
```

大数据与程序设计类课程在线实践平台

### 3.3 数据操作



```
[2023-06-06 22:48:30,689] WARN Failed to get database global_temp, returning NoSuchObjectException (org.apache.hadoop.hive.metastore.ObjectStore:568)
Welcome to

  ____
 /  __ \  _ __ ___
 \  __ \| '_ \ / __|
  | |__) | |_) | | |_\__|
  |____/|___|_|_|_|\___|
                        version 2.1.0

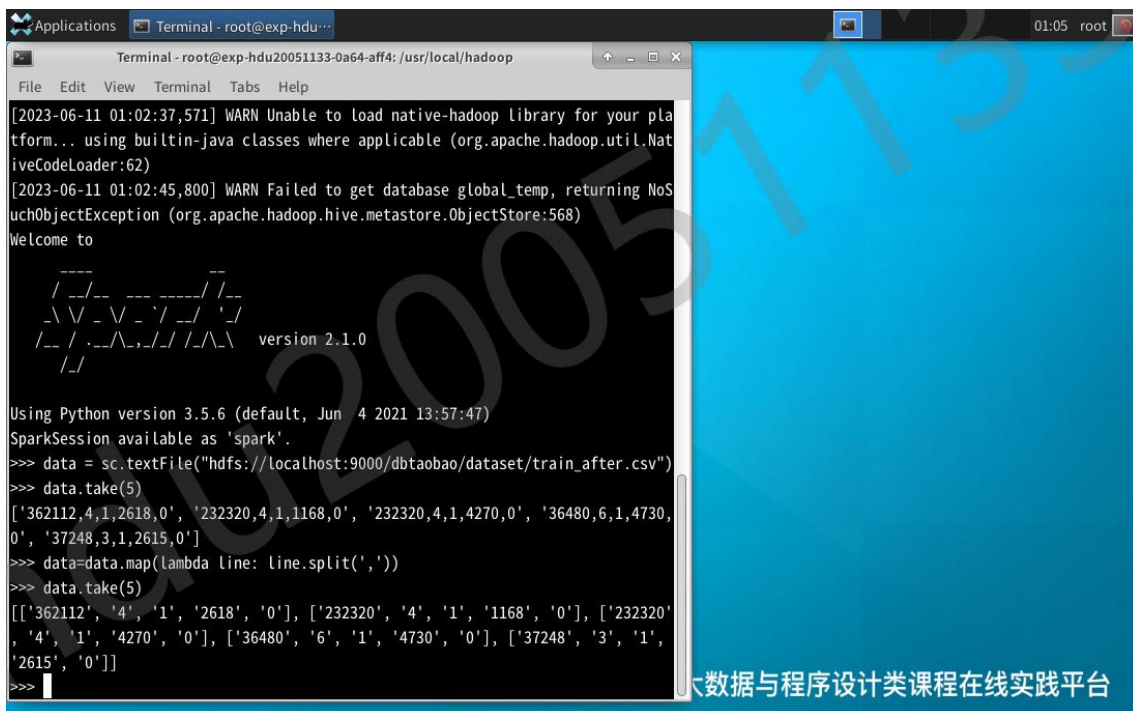
Using Python version 3.5.6 (default, Jun  4 2021 13:57:47)
SparkSession available as 'spark'.
>>> lines=sc.textFile("hdfs://localhost:9000/dbtaobao/dataset/user_log.csv")
>>> lines.count()
54925331
>>> linesmap=lines.map(lambda line : line.split(','))
>>> linesmap.take(3)
[('user_id', 'item_id', 'cat_id', 'merchant_id', 'brand_id', 'month', 'day', 'action', 'age_range', 'gender', 'province'), ('328862', '323294', '833', '2882', '2661', '08', '29', '0', '0', '1', '内蒙古'), ('328862', '844400', '1271', '2882', '2661', '08', '29', '0', '1', '1', '山西')]
>>> linesfilter=lines.filter(lambda line: line.split(',')[9] is '0')
>>> linesfilter.count()
18305795
>>>
```

大数据与程序设计类课程在线实践平台



## 4 Spark 回头客预测

### 4.1 预处理



A terminal window titled 'Terminal - root@exp-hdu...' showing the execution of Spark code. The code loads a CSV file from HDFS, takes the first 5 rows, and maps them to a list of lists. The output shows the first 5 rows of the dataset, each containing 10 numerical values. The terminal also displays warning messages about loading native Hadoop libraries and a Hive metastore exception.

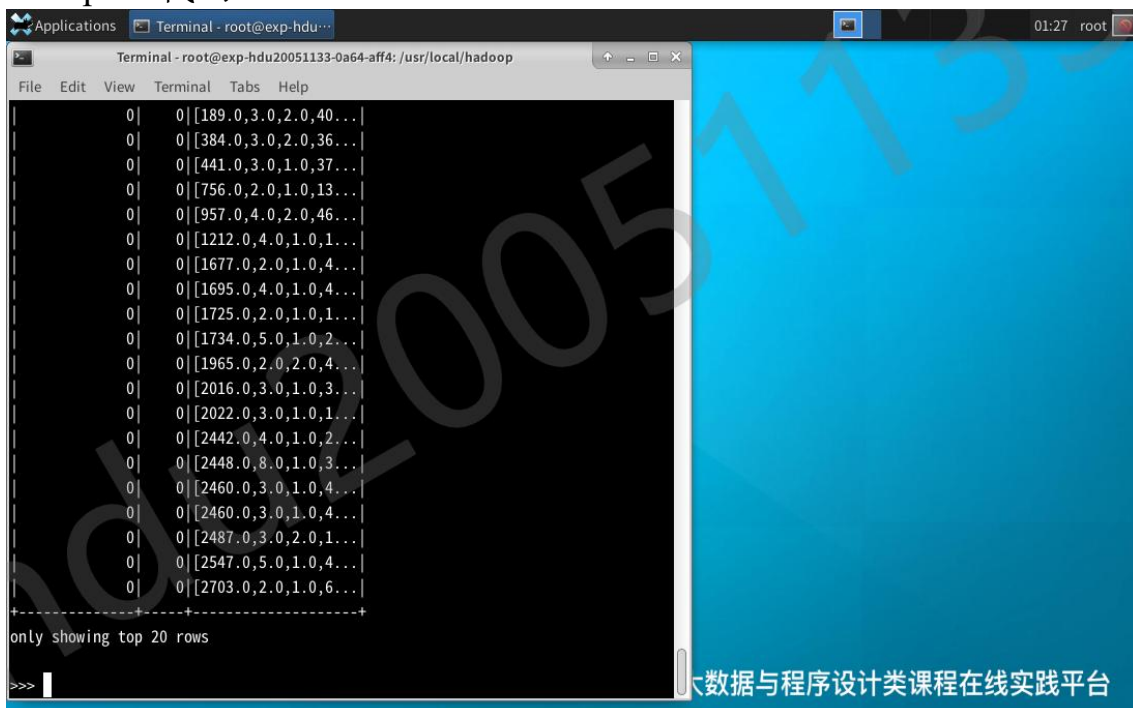
```
Terminal - root@exp-hdu20051133-0a64-aff4: /usr/local/hadoop
[2023-06-11 01:02:37,571] WARN Unable to load native-hadoop library for your platform... using builtin-java classes where applicable (org.apache.hadoop.util.NativeCodeLoader:62)
[2023-06-11 01:02:45,800] WARN Failed to get database global_temp, returning NoSuchObjectException (org.apache.hadoop.hive.metastore.ObjectStore:568)
Welcome to

  ____
 /  _ \  _ __| | | |
/_  \ \ / /  _ \| | | |
 \  __/ / _ \| |_| |
  \___/_/ \_\_|_|_|_|

version 2.1.0

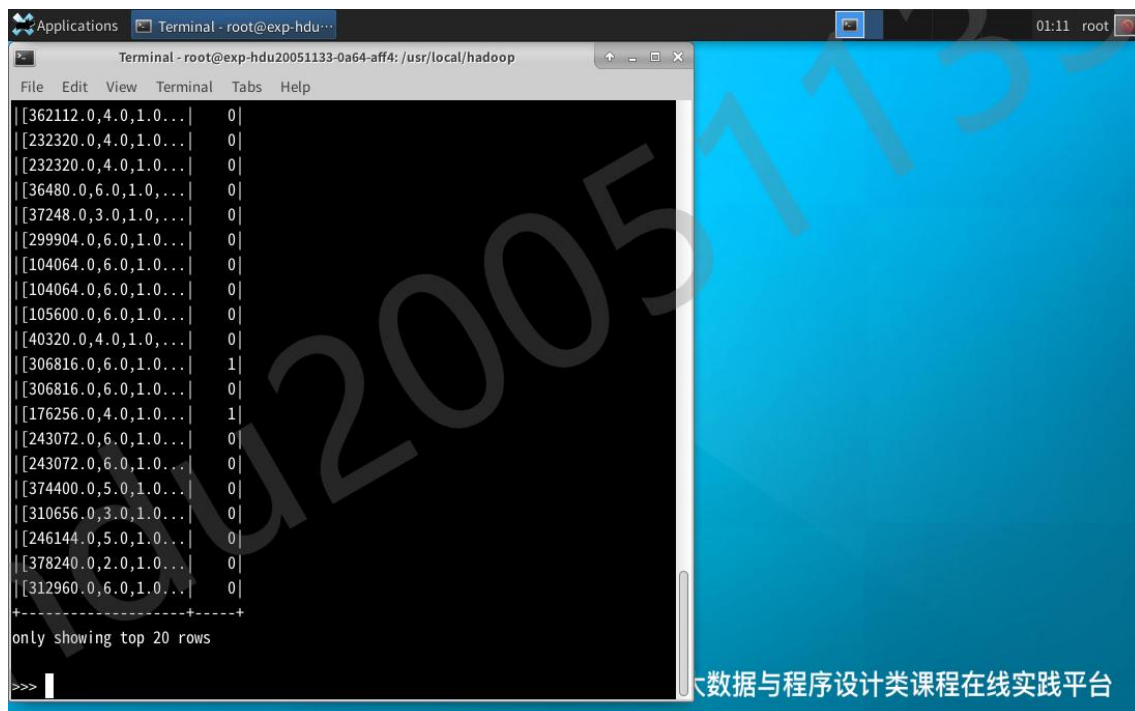
Using Python version 3.5.6 (default, Jun 4 2021 13:57:47)
SparkSession available as 'spark'.
>>> data = sc.textFile("hdfs://localhost:9000/dbtaobao/dataset/train_after.csv")
>>> data.take(5)
['362112,4,1,2618,0', '232320,4,1,1168,0', '232320,4,1,4270,0', '36480,6,1,4730,0', '37248,3,1,2615,0']
>>> data=data.map(lambda line: line.split(','))
>>> data.take(5)
[['362112', '4', '1', '2618', '0'], ['232320', '4', '1', '1168', '0'], ['232320', '4', '1', '4270', '0'], ['36480', '6', '1', '4730', '0'], ['37248', '3', '1', '2615', '0']]
>>>
```

### 4.2 Spark 代码



A terminal window titled 'Terminal - root@exp-hdu...' showing the execution of Spark code. The code displays the first 20 rows of the dataset, each containing 10 numerical values. The terminal also displays warning messages about loading native Hadoop libraries and a Hive metastore exception.

```
Terminal - root@exp-hdu20051133-0a64-aff4: /usr/local/hadoop
0| 0|189.0,3.0,2.0,40...
0| 0|384.0,3.0,2.0,36...
0| 0|441.0,3.0,1.0,37...
0| 0|756.0,2.0,1.0,13...
0| 0|957.0,4.0,2.0,46...
0| 0|1212.0,4.0,1.0,1...
0| 0|1677.0,2.0,1.0,4...
0| 0|1695.0,4.0,1.0,4...
0| 0|1725.0,2.0,1.0,1...
0| 0|1734.0,5.0,1.0,2...
0| 0|1965.0,2.0,2.0,4...
0| 0|2016.0,3.0,1.0,3...
0| 0|2022.0,3.0,1.0,1...
0| 0|2442.0,4.0,1.0,2...
0| 0|2448.0,8.0,1.0,3...
0| 0|2460.0,3.0,1.0,4...
0| 0|2460.0,3.0,1.0,4...
0| 0|2487.0,3.0,2.0,1...
0| 0|2547.0,5.0,1.0,4...
0| 0|2703.0,2.0,1.0,6...
+-----+
only showing top 20 rows
>>>
```



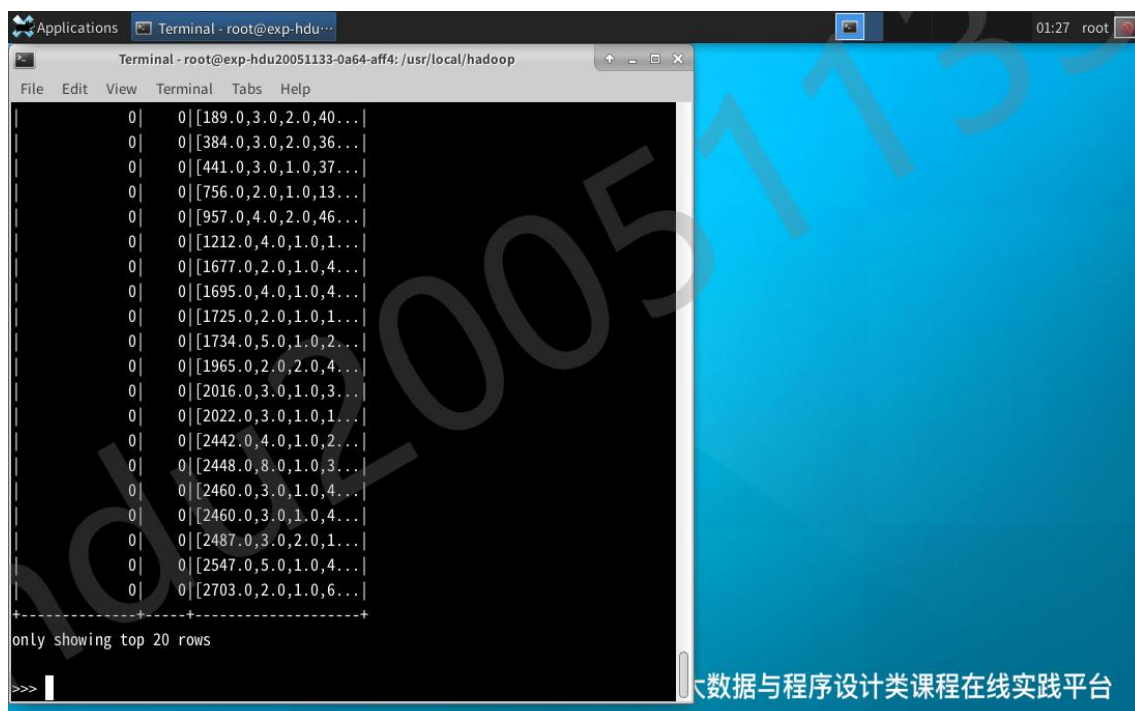
```

Terminal - root@exp-hdu20051133-0a64-aff4: /usr/local/hadoop
File Edit View Terminal Tabs Help

[362112.0,4.0,1.0...] | 0|
[232320.0,4.0,1.0...] | 0|
[232320.0,4.0,1.0...] | 0|
[36480.0,6.0,1.0...] | 0|
[37248.0,3.0,1.0...] | 0|
[299904.0,6.0,1.0...] | 0|
[104064.0,6.0,1.0...] | 0|
[104064.0,6.0,1.0...] | 0|
[105600.0,6.0,1.0...] | 0|
[40320.0,4.0,1.0...] | 0|
[306816.0,6.0,1.0...] | 1|
[306816.0,6.0,1.0...] | 0|
[176256.0,4.0,1.0...] | 1|
[243072.0,6.0,1.0...] | 0|
[243072.0,6.0,1.0...] | 0|
[374400.0,5.0,1.0...] | 0|
[310656.0,3.0,1.0...] | 0|
[246144.0,5.0,1.0...] | 0|
[378240.0,2.0,1.0...] | 0|
[312960.0,6.0,1.0...] | 0|
+-----+
only showing top 20 rows
>>>

```

大数据与程序设计类课程在线实践平台



```

Terminal - root@exp-hdu20051133-0a64-aff4: /usr/local/hadoop
File Edit View Terminal Tabs Help

0| 0| [189.0,3.0,2.0,40...]
0| 0| [384.0,3.0,2.0,36...]
0| 0| [441.0,3.0,1.0,37...]
0| 0| [756.0,2.0,1.0,13...]
0| 0| [957.0,4.0,2.0,46...]
0| 0| [1212.0,4.0,1.0,1...]
0| 0| [1677.0,2.0,1.0,4...]
0| 0| [1695.0,4.0,1.0,4...]
0| 0| [1725.0,2.0,1.0,1...]
0| 0| [1734.0,5.0,1.0,2...]
0| 0| [1965.0,2.0,2.0,4...]
0| 0| [2016.0,3.0,1.0,3...]
0| 0| [2022.0,3.0,1.0,1...]
0| 0| [2442.0,4.0,1.0,2...]
0| 0| [2448.0,8.0,1.0,3...]
0| 0| [2460.0,3.0,1.0,4...]
0| 0| [2460.0,3.0,1.0,4...]
0| 0| [2487.0,3.0,2.0,1...]
0| 0| [2547.0,5.0,1.0,4...]
0| 0| [2703.0,2.0,1.0,6...]
+-----+
only showing top 20 rows
>>>

```

大数据与程序设计类课程在线实践平台

The screenshot shows a terminal window with a menu bar (File, Edit, View, Terminal, Tabs, Help) and a title bar (Terminal - root@exp-hdu20051133-0a64-aff4: /usr/local/hadoop). The terminal displays a list of 20 IP addresses, each followed by a vertical bar and a list of numbers in brackets. Below this list, it says "only showing top 20 rows". Then, a Python script is executed, which imports the MulticlassClassificationEvaluator from pyspark.ml.evaluation, sets up the evaluator with specific columns and metric name, and prints the accuracy. The output shows an accuracy of 0.9432530517980865.

```

0| 0|[1677.0,2.0,1.0,4...|
0| 0|[1695.0,4.0,1.0,4...|
0| 0|[1725.0,2.0,1.0,1...|
0| 0|[1734.0,5.0,1.0,2...|
0| 0|[1965.0,2.0,2.0,4...|
0| 0|[2016.0,3.0,1.0,3...|
0| 0|[2022.0,3.0,1.0,1...|
0| 0|[2442.0,4.0,1.0,2...|
0| 0|[2448.0,8.0,1.0,3...|
0| 0|[2460.0,3.0,1.0,4...|
0| 0|[2460.0,3.0,1.0,4...|
0| 0|[2487.0,3.0,2.0,1...|
0| 0|[2547.0,5.0,1.0,4...|
0| 0|[2703.0,2.0,1.0,6...|
+-----+
only showing top 20 rows

>>> from pyspark.ml.evaluation import MulticlassClassificationEvaluator
>>> evaluatorClassifier=MulticlassClassificationEvaluator().setLabelCol("indexed
Label").setPredictionCol("prediction").setMetricName("accuracy")
>>> accuracy=evaluatorClassifier.evaluate(predictionsClassifier)
>>> print(accuracy)
0.9432530517980865
>>>

```

大数据与程序设计类课程在线实践平台