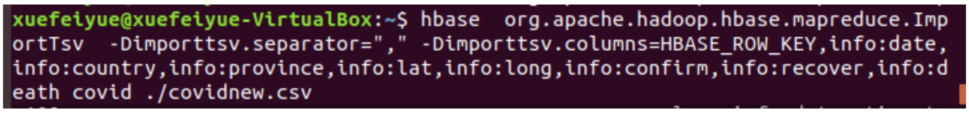
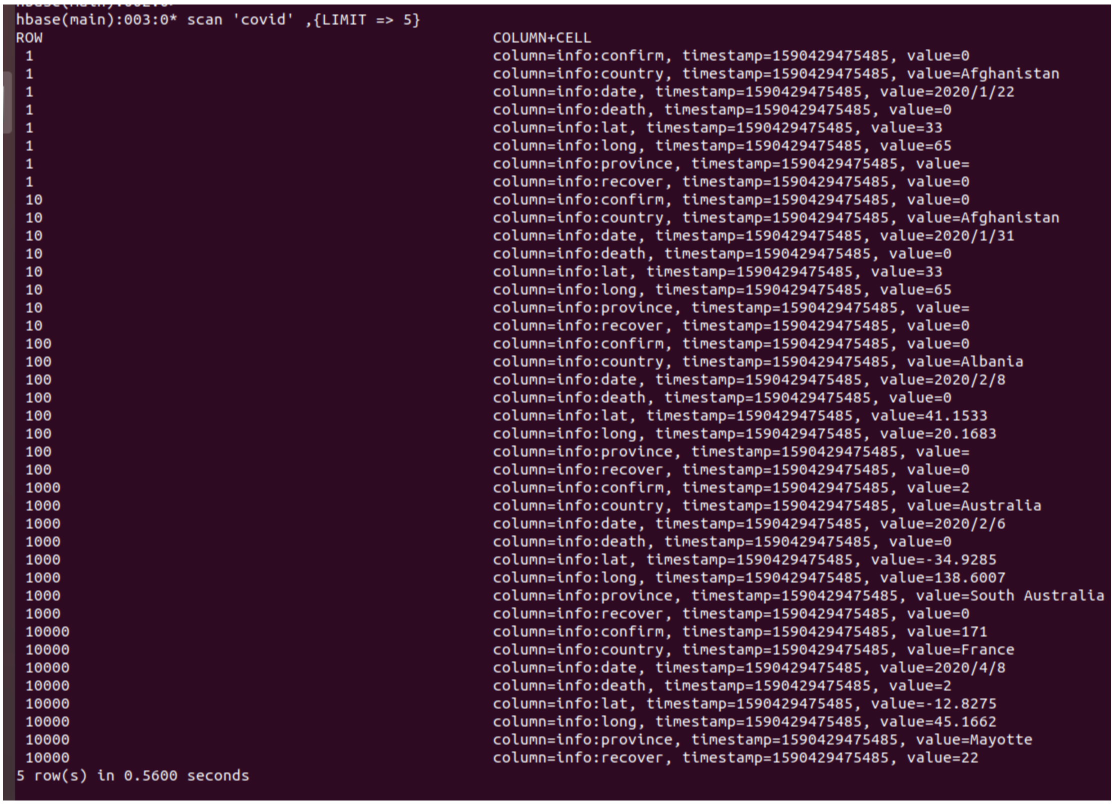
Hbase 实习

薛飞跃 1700017831

一．导入数据到 hbase：

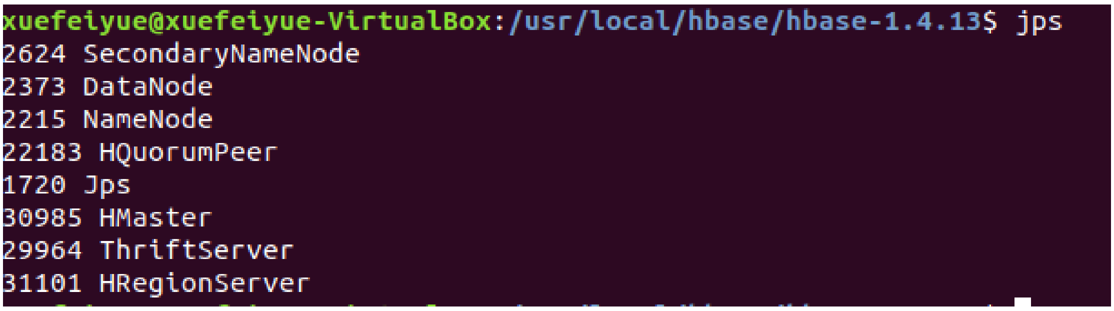


```
xuefeiyue@xuefeiyue-VirtualBox:~$ hbase  org.apache.hadoop.hbase.mapreduce.Imp
ortTsv  -Dimporttsv.separator="," -Dimporttsv.columns=HBASE_ROW_KEY,info:date,
info:country,info:province,info:lat,info:long,info:confirm,info:recover,info:d
eath covid ./covidnew.csv
```



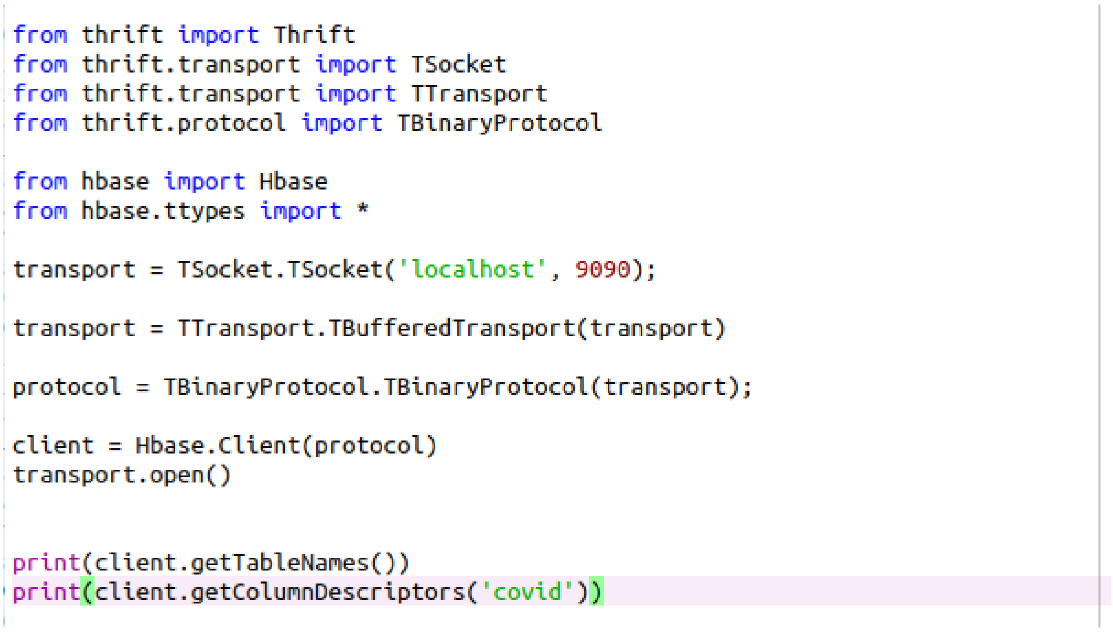
```
hbase(main):003:0* scan 'covid' ,{LIMIT => 5}
ROW                                                     COLUMN+CELL
 1                                                      column=info:confirm, timestamp=1590429475485, value=0
 1                                                      column=info:country, timestamp=1590429475485, value=Afghanistan
 1                                                      column=info:date, timestamp=1590429475485, value=2020/1/22
 1                                                      column=info:death, timestamp=1590429475485, value=0
 1                                                      column=info:lat, timestamp=1590429475485, value=33
 1                                                      column=info:long, timestamp=1590429475485, value=65
 1                                                      column=info:province, timestamp=1590429475485, value=
 1                                                      column=info:recover, timestamp=1590429475485, value=0
 10                                                     column=info:confirm, timestamp=1590429475485, value=0
 10                                                     column=info:country, timestamp=1590429475485, value=Afghanistan
 10                                                     column=info:date, timestamp=1590429475485, value=2020/1/31
 10                                                     column=info:death, timestamp=1590429475485, value=0
 10                                                     column=info:lat, timestamp=1590429475485, value=33
 10                                                     column=info:long, timestamp=1590429475485, value=65
 10                                                     column=info:province, timestamp=1590429475485, value=
 10                                                     column=info:recover, timestamp=1590429475485, value=0
 100                                                    column=info:confirm, timestamp=1590429475485, value=0
 100                                                    column=info:country, timestamp=1590429475485, value=Albania
 100                                                    column=info:date, timestamp=1590429475485, value=2020/2/8
 100                                                    column=info:death, timestamp=1590429475485, value=0
 100                                                    column=info:lat, timestamp=1590429475485, value=41.1533
 100                                                    column=info:long, timestamp=1590429475485, value=20.1683
 100                                                    column=info:province, timestamp=1590429475485, value=
 100                                                    column=info:recover, timestamp=1590429475485, value=0
 1000                                                   column=info:confirm, timestamp=1590429475485, value=2
 1000                                                   column=info:country, timestamp=1590429475485, value=Australia
 1000                                                   column=info:date, timestamp=1590429475485, value=2020/2/6
 1000                                                   column=info:death, timestamp=1590429475485, value=0
 1000                                                   column=info:lat, timestamp=1590429475485, value=-34.9285
 1000                                                   column=info:long, timestamp=1590429475485, value=138.6007
 1000                                                   column=info:province, timestamp=1590429475485, value=South Australia
 1000                                                   column=info:recover, timestamp=1590429475485, value=0
 10000                                                  column=info:confirm, timestamp=1590429475485, value=171
 10000                                                  column=info:country, timestamp=1590429475485, value=France
 10000                                                  column=info:date, timestamp=1590429475485, value=2020/4/8
 10000                                                  column=info:death, timestamp=1590429475485, value=2
 10000                                                  column=info:lat, timestamp=1590429475485, value=-12.8275
 10000                                                  column=info:long, timestamp=1590429475485, value=45.1662
 10000                                                  column=info:province, timestamp=1590429475485, value=Mayotte
 10000                                                  column=info:recover, timestamp=1590429475485, value=22
5 row(s) in 0.5600 seconds
```

二． 安装 thrift 并运行，jps 中出现 HMaster，HQuorumPeer，HRegionSever 说明 HBase 启动， ThriftServer 说明 thrift 启动成功

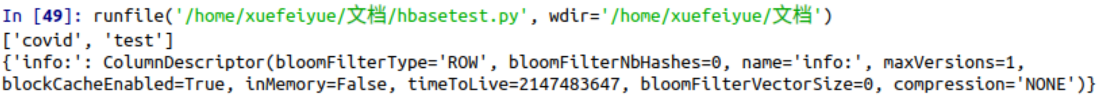


```
xuefeiyue@xuefeiyue-VirtualBox:/usr/local/hbase/hbase-1.4.13$ jps
2624 SecondaryNameNode
2373 DataNode
2215 NameNode
22183 HQuorumPeer
1720 Jps
30985 HMaster
29964 ThriftServer
31101 HRegionServer
```

三．python 安装 thrift 库和 hbase-thrift 库，可以在 python 调用 hbase 接口

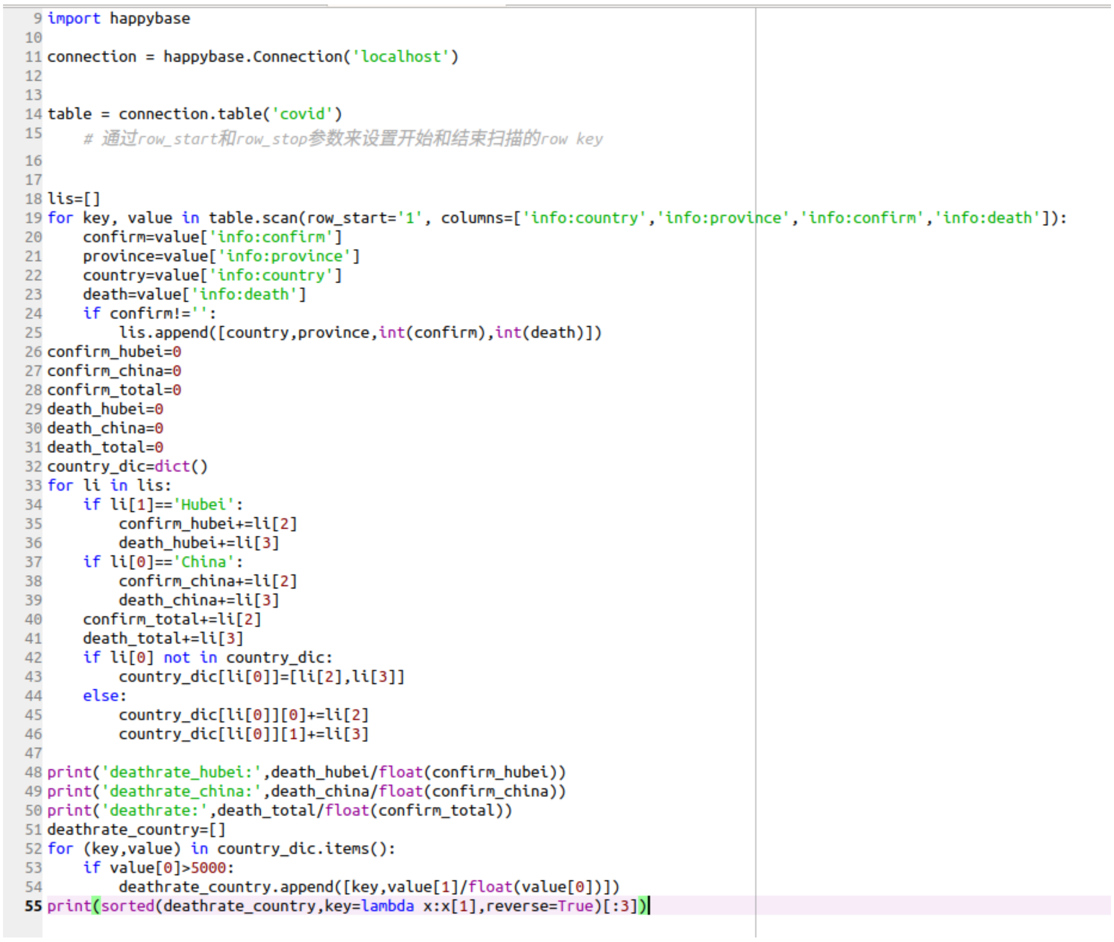


```
from thrift import Thrift
from thrift.transport import TSocket
from thrift.transport import TTransport
from thrift.protocol import TBinaryProtocol

from hbase import Hbase
from hbase.ttypes import *

transport = TSocket.TSocket('localhost', 9090);

transport = TTransport.TBufferedTransport(transport)

protocol = TBinaryProtocol.TBinaryProtocol(transport);

client = Hbase.Client(protocol)
transport.open()


print(client.getTableNames())
print(client.getColumnDescriptors('covid'))
```

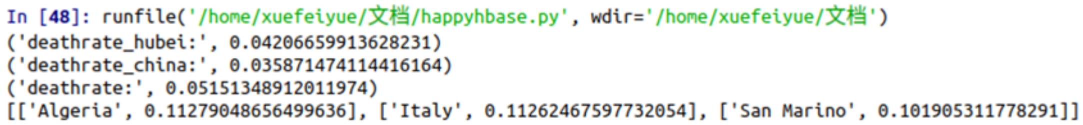


```
In [49]: runfile('/home/xuefeiyue/文档/hbasetest.py', wdir='/home/xuefeiyue/文档')
['covid', 'test']
{'info:': ColumnDescriptor(bloomFilterType='ROW', bloomFilterNbHashes=0, name='info:', maxVersions=1,
blockCacheEnabled=True, inMemory=False, timeToLive=2147483647, bloomFilterVectorSize=0, compression='NONE')}
```

四 . python 安装 happybase 库，更方便的调用接口来执行查询



```
import happybase

connection = happybase.Connection('localhost')


table = connection.table('covid')
    # 通过row_start和row_stop参数来设置开始和结束扫描的row key


lis=[]
for key, value in table.scan(row_start='1', columns=['info:country','info:province','info:confirm','info:death']):
    confirm=value['info:confirm']
    province=value['info:province']
    country=value['info:country']
    death=value['info:death']
    if confirm!='':
        lis.append([country,province,int(confirm),int(death)])
confirm_hubei=0
confirm_china=0
confirm_total=0
death_hubei=0
death_china=0
death_total=0
country_dic=dict()
for li in lis:
    if li[1]=='Hubei':
        confirm_hubei+=li[2]
        death_hubei+=li[3]
    if li[0]=='China':
        confirm_china+=li[2]
        death_china+=li[3]
    confirm_total+=li[2]
    death_total+=li[3]
    if li[0] not in country_dic:
        country_dic[li[0]]=[li[2],li[3]]
    else:
        country_dic[li[0]][0]+=li[2]
        country_dic[li[0]][1]+=li[3]

print('deathrate_hubei:',death_hubei/float(confirm_hubei))
print('deathrate_china:',death_china/float(confirm_china))
print('deathrate:',death_total/float(confirm_total))
deathrate_country=[]
for (key,value) in country_dic.items():
    if value[0]>5000:
        deathrate_country.append([key,value[1]/float(value[0])])
print(sorted(deathrate_country,key=lambda x:x[1],reverse=True)[:3])
```



```
In [48]: runfile('/home/xuefeiyue/文档/happyhbase.py', wdir='/home/xuefeiyue/文档')
('deathrate_hubei:', 0.04206659913628231)
('deathrate_china:', 0.035871474114416164)
('deathrate:', 0.05151348912011974)
[['Algeria', 0.11279048656499636], ['Italy', 0.11262467597732054], ['San Marino', 0.101905311778291]]
```

五．实验心得

1. hbase 中的数据本身几乎不蕴含格式，统一为字符串，取到数据后， 需要把数据转化为相应的格式。

2. Hbase 的访问接口很多，Native Java API、HBase Shell、Thrift 各有各的优势，Java API 最常用，也能执行最复杂的查询，Shell 最简单，但只能提供简单的查询，Thrift 支持其他语言访问 Hbase 数据库。