期末实训作业

1. **目的**

对课程综合实践，对大数据进行分析处理，掌握理论与实践结合，锻炼学生实际动手的能力

1. **实训环境**

Windows 虚拟机软件，centos8.X，java hadoop spark python3

1. **实训内容**

环境安装，数据处理，可视化输出

1) 统计美国截止每日的累计确诊人数和累计死亡人数。

2) 统计美国每日的新增确诊人数和新增死亡人数。

3) 统计截止5.19日，美国各州的累计确诊人数和死亡人数。

4) 统计截止5.19日，美国确诊人数最多的十个州。

5) 统计截止5.19日，美国死亡人数最多的十个州。

6) 统计截止5.19日，美国确诊人数最少的十个州。

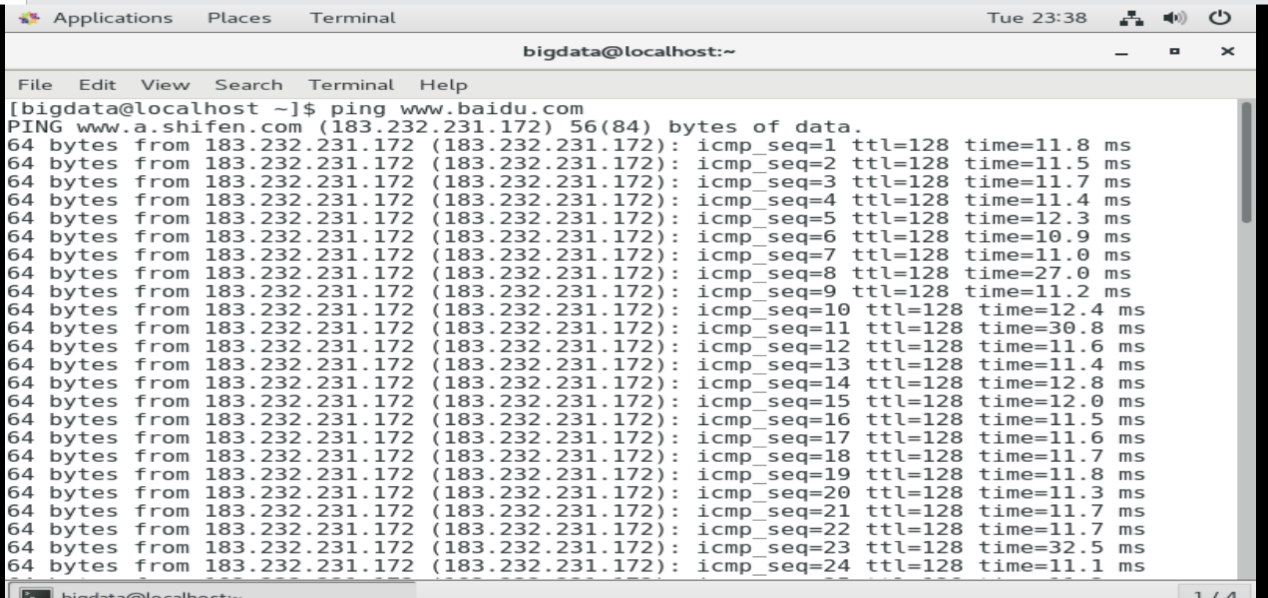
7) 统计截止5.19日，美国死亡人数最少的十个州。

8) 统计截止5.19日，全美和各州的病死率。

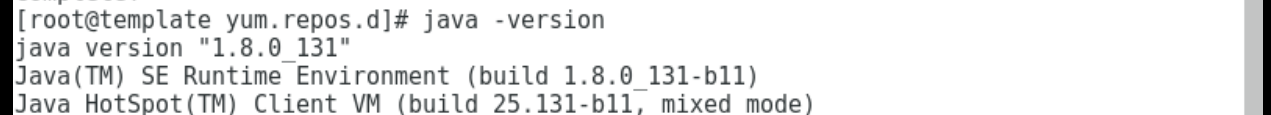
1. **实训步骤**

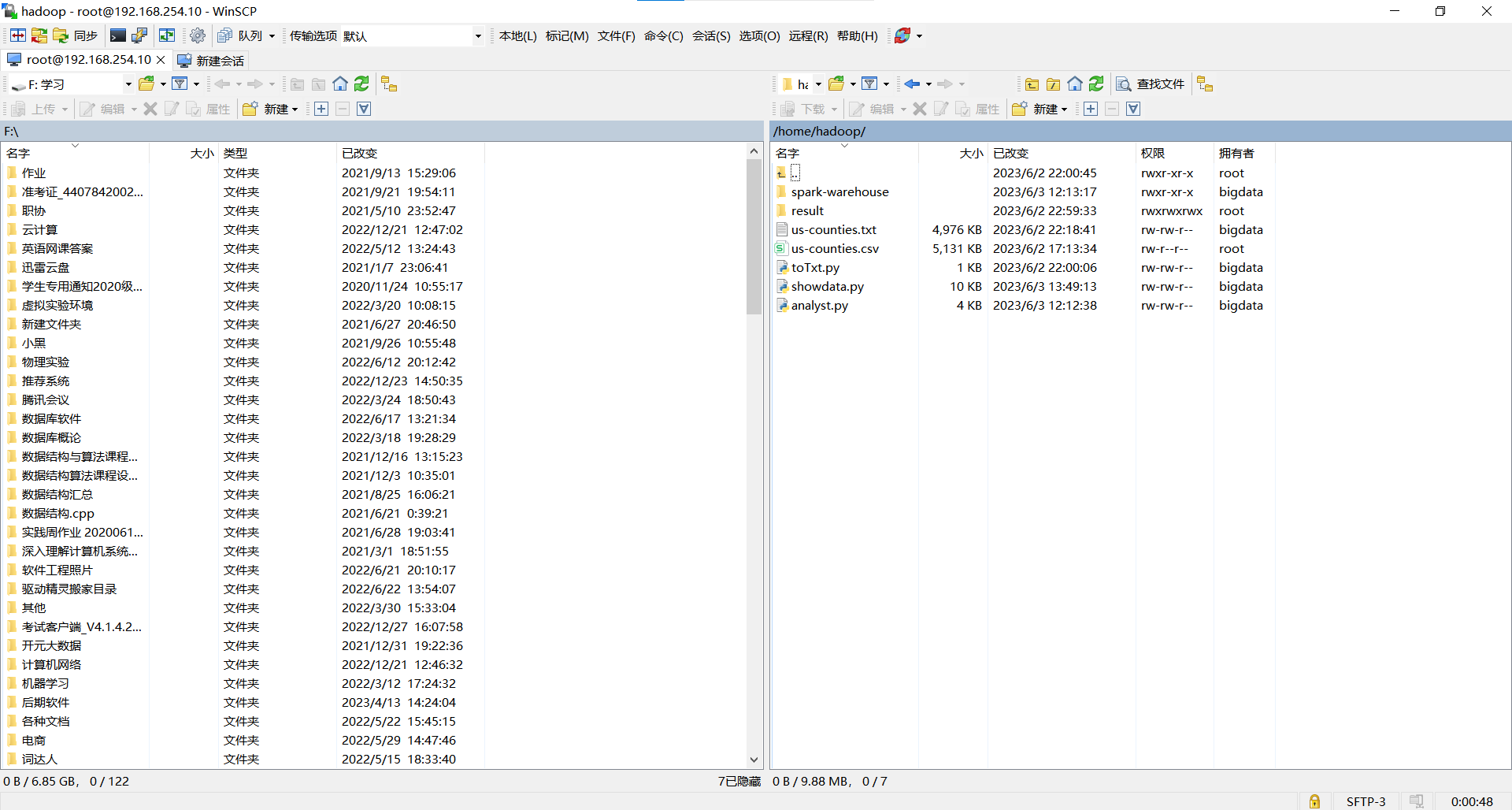
环境搭建：

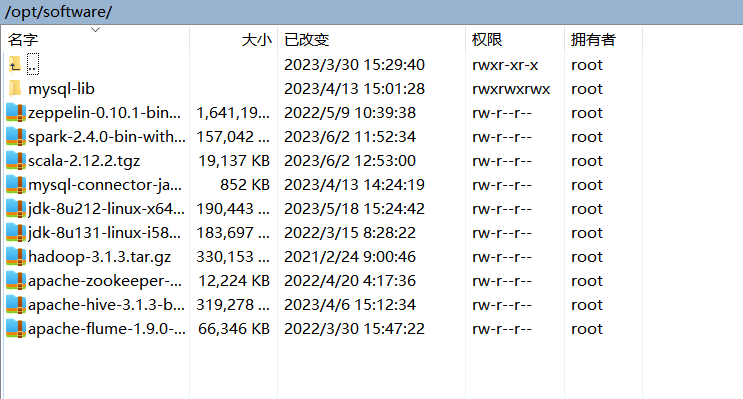
1. 从阿里云官网下载了centOS8的镜像安装到虚拟机上，接着再搭建基础组件，首先可以ping通网络。



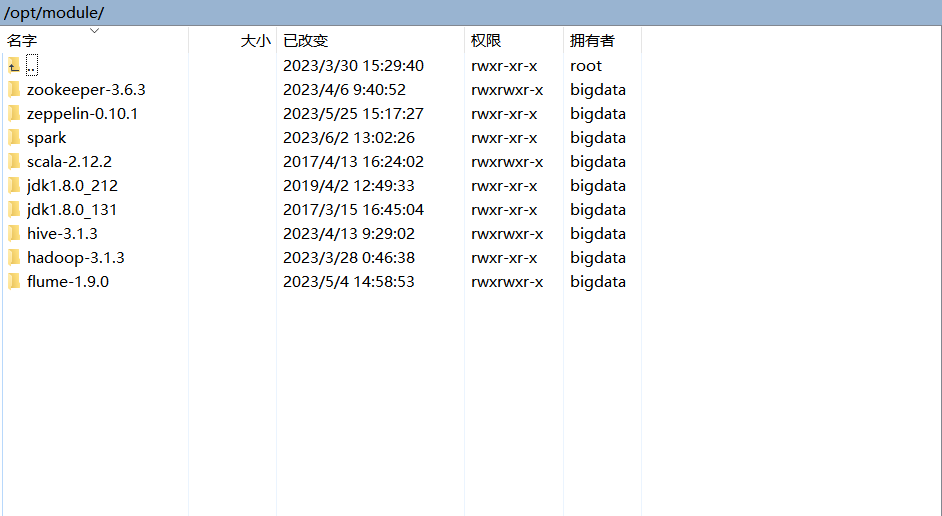
1. 在ping通网络的基础上安装JDK-1.8.0以上的版本。



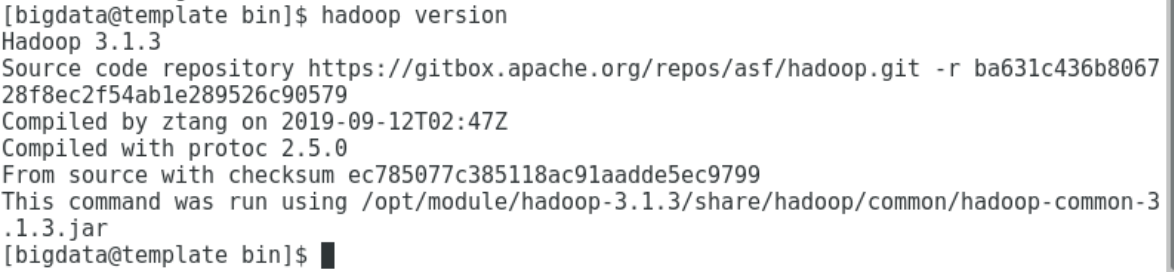
1. 用虚拟机连接wincp，传输需要的文件压缩包。

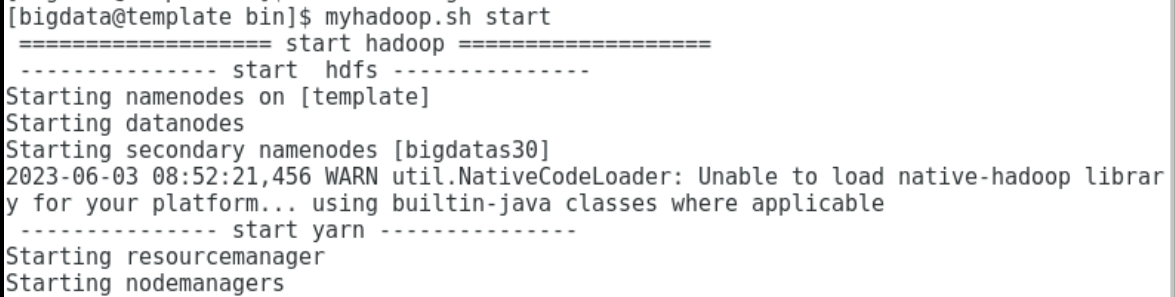


1. 将压缩包解压到software

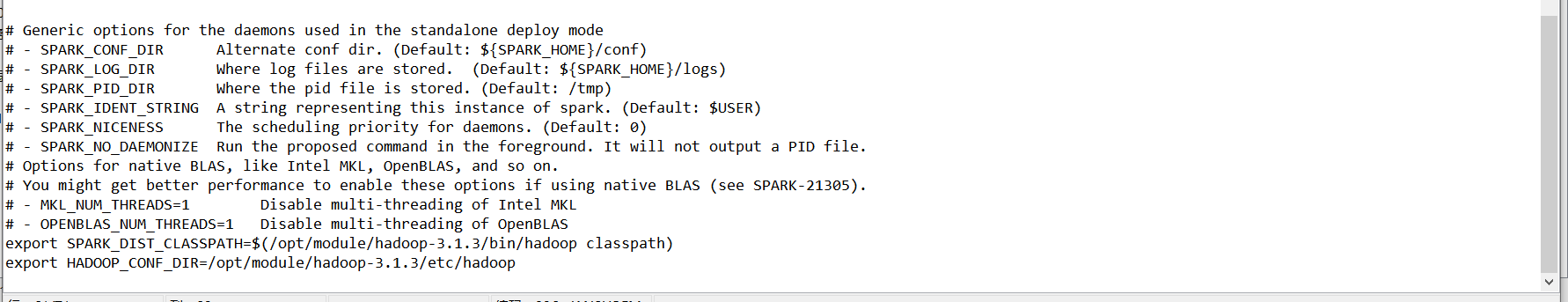


1. 安装好JDK后开始配置hadoop-3.1.3的环境,启动集群。

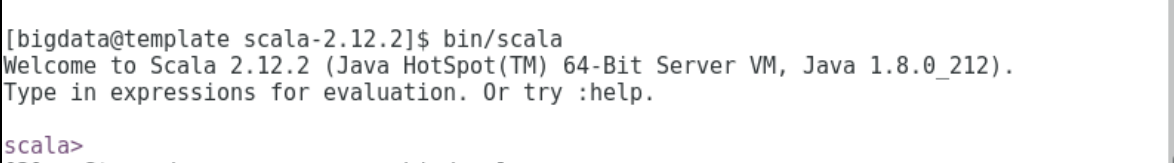




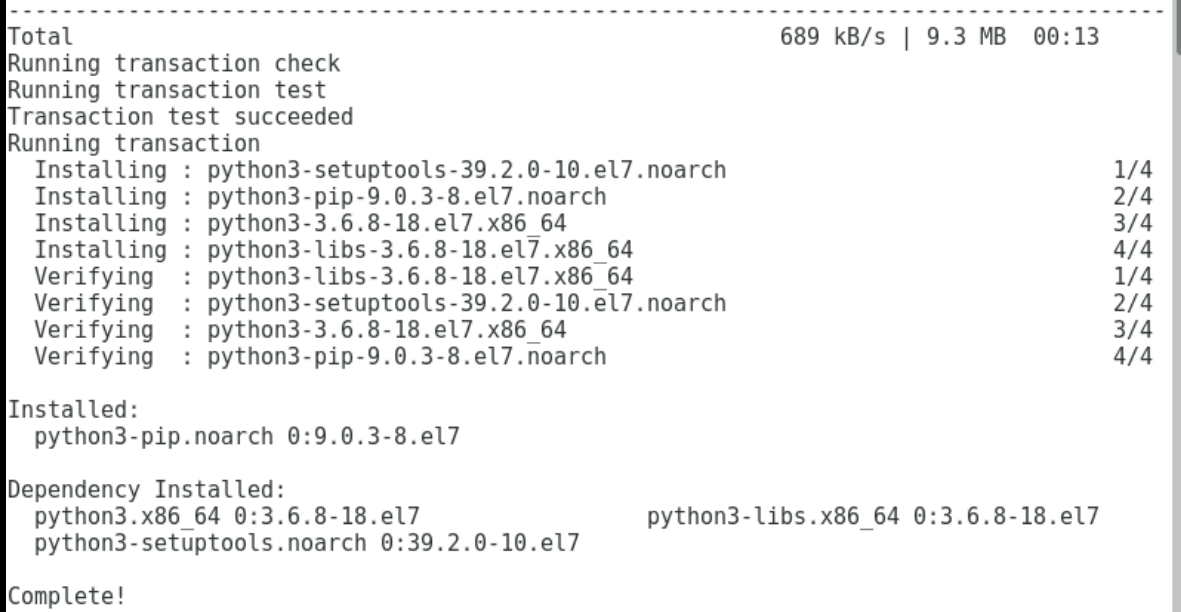
6、安装好spark并且配置环境变量，scala也安装好。





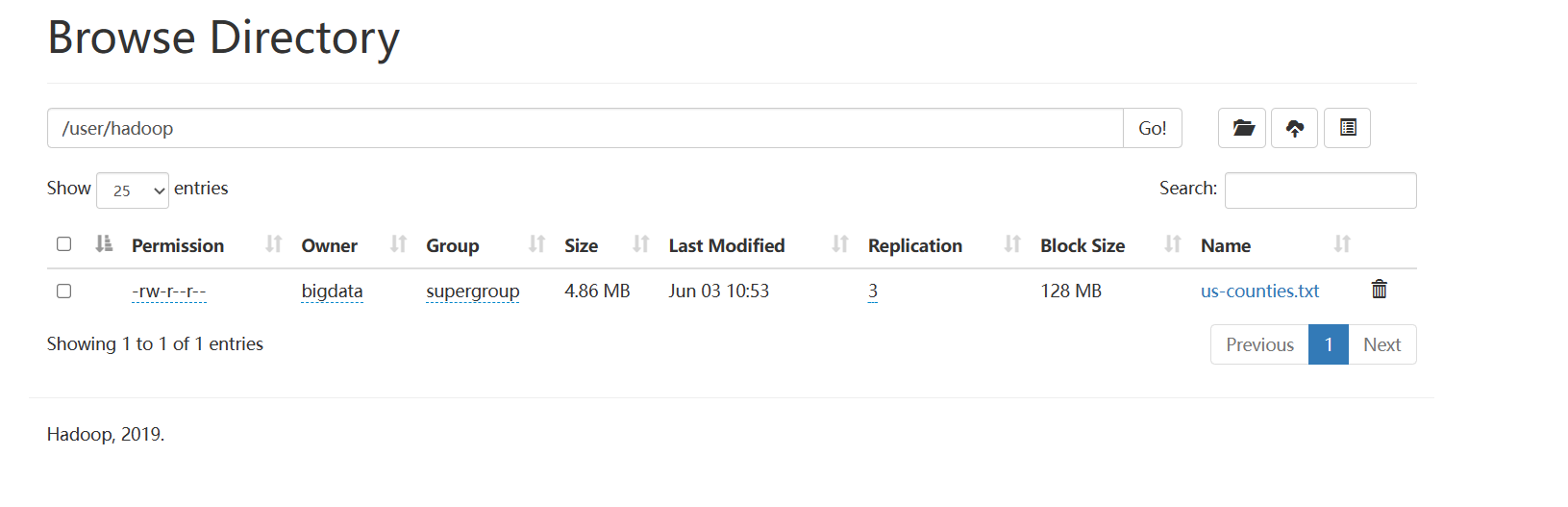


7、通过yum下载python



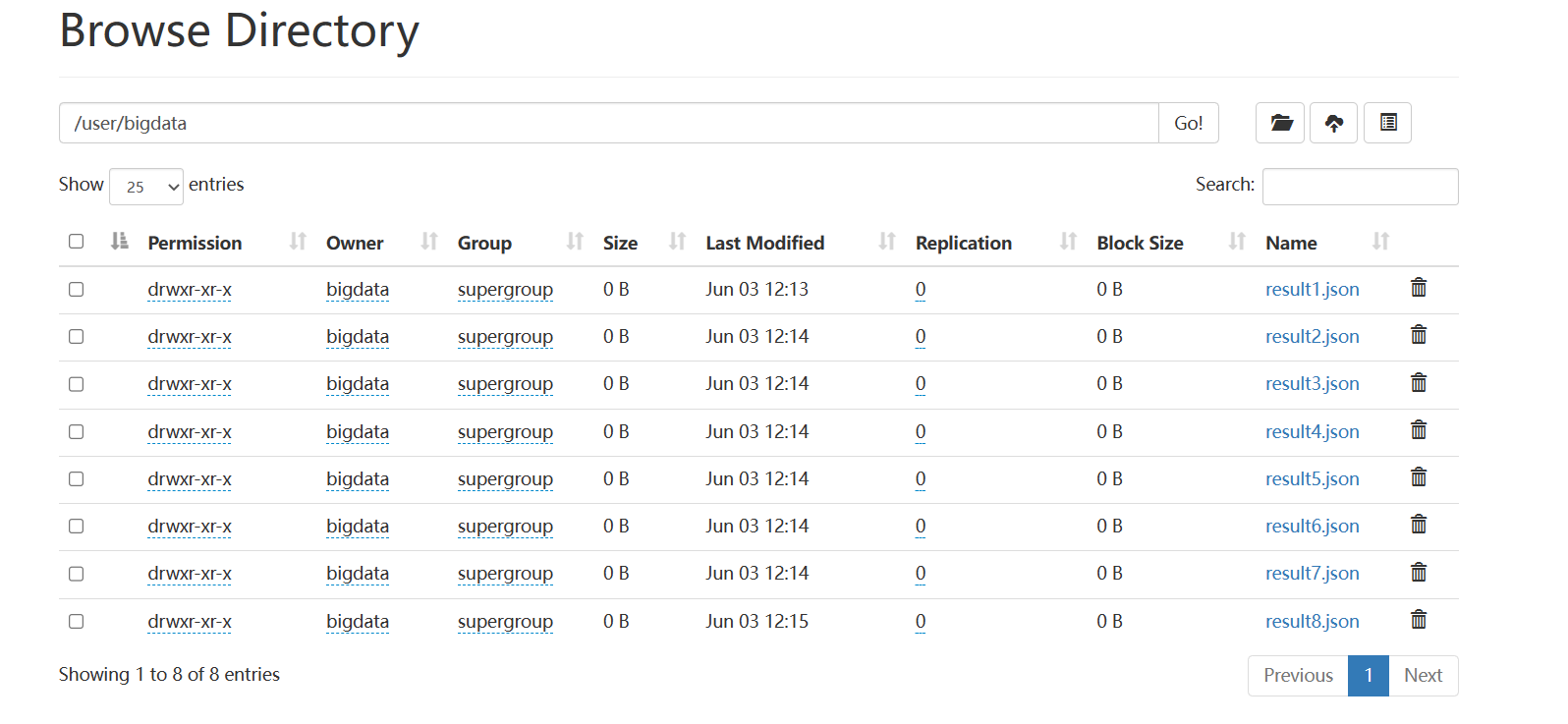
1. 通过python文件将csv文件转化为txt文件，并且上传到HDFS文件系统





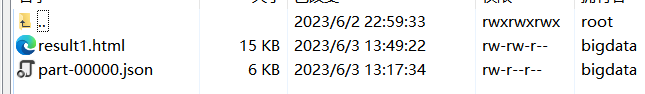
1. 创建一个数据分析的py文件，在本地系统建好一个result文件，py代码运行后的文件会上传到HDFS文件系统。



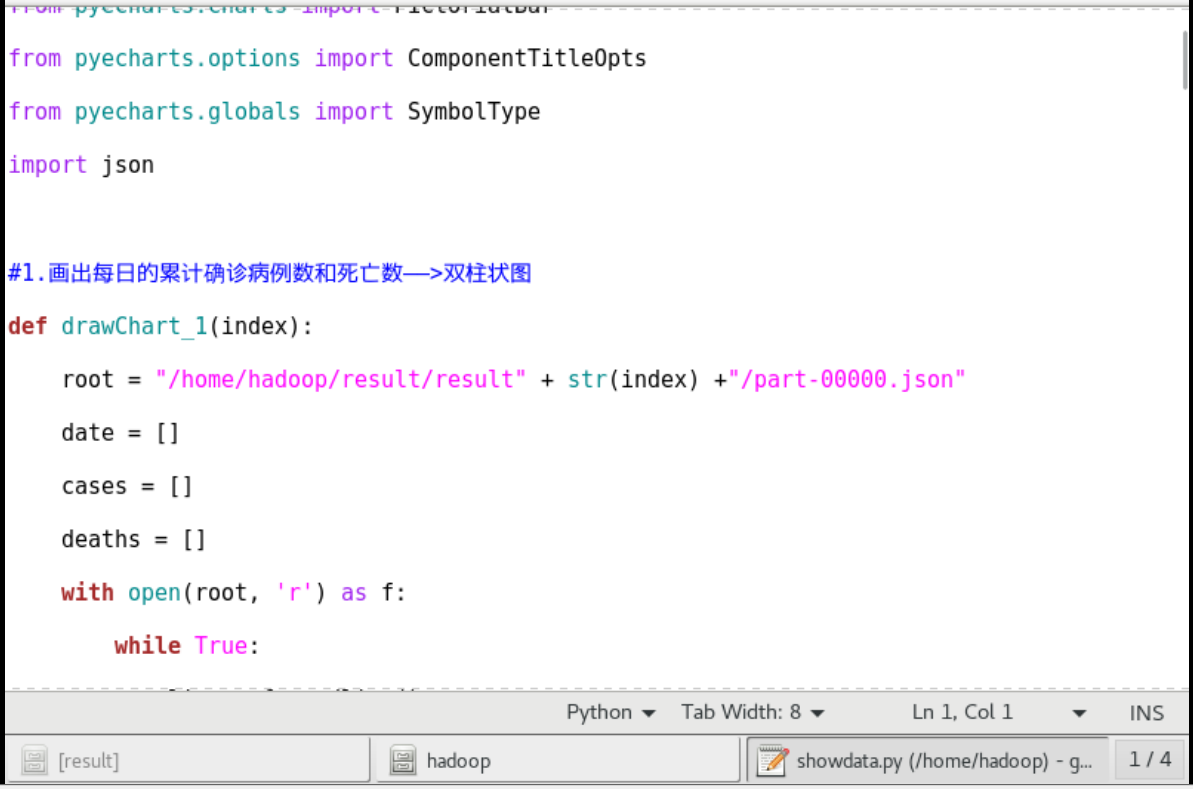


1. 将HDFS生成的八个文件传到本地文件系统

例：



1. 编写一个showdata文件，进行可视化分析，得出八个html结果文件。





数据处理可视化输出：

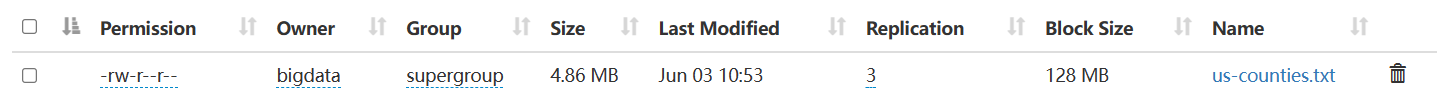
1. 使用Python将us-counties.csv转换为.txt格式文件us-counties.txt。



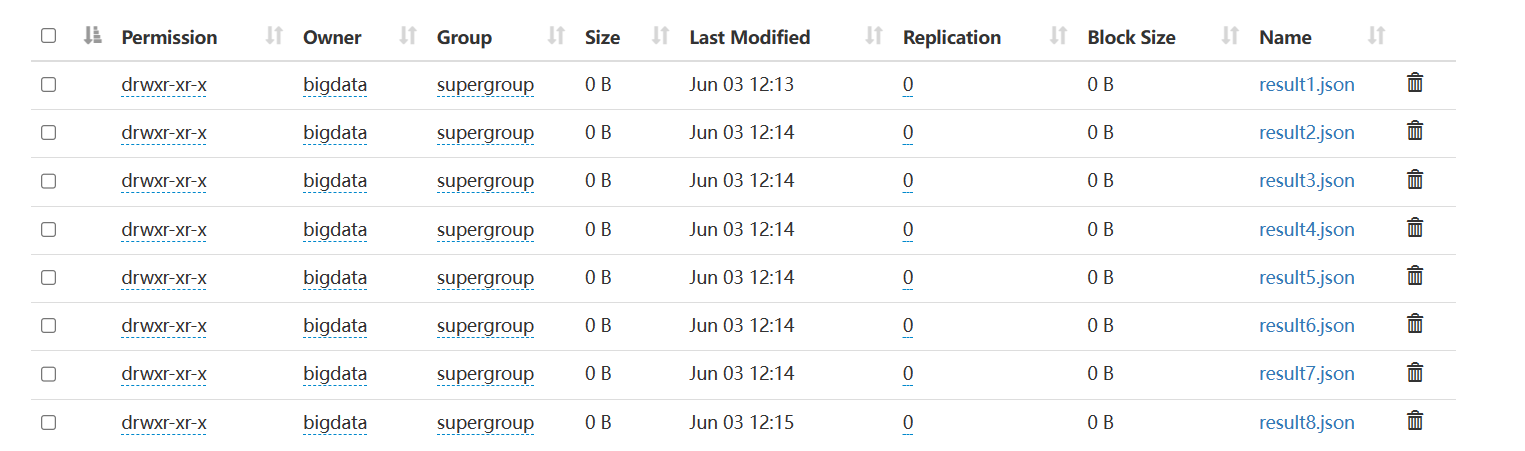
1. 将转好的文件通过本地系统上传到HDFS文件系统中。

$./bin/hdfs dfs -mkdir -p /user/hadoop

$./bin/hdfs dfs -put /home/hadoop/us-counties.txt /user/hadoop



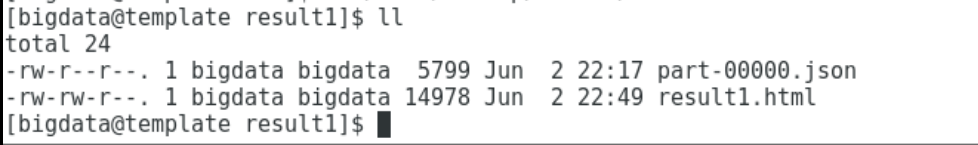
1. 用一个数据分析的analyst.py文件，在HDFS上生成八个result文件。



1. 将HDFS上结果文件转储到本地文件系统中。

用八次如下的命令将result.json文件上传到本地系统中。

$./bin/hdfs dfs -get /user/hadoop/result1.json/\*.json /home/hadoop/result/result1

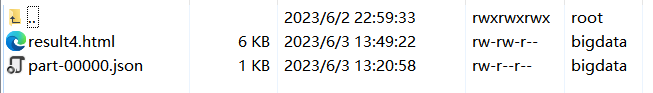


1. 使用python第三方库pyecharts作为可视化工具,用pip3 install --user pyecharts命令下载，使得可视化结果是.html格式的

命令：pip3 install --user pyecharts

1. 运行showdata.py得到八个html文件。





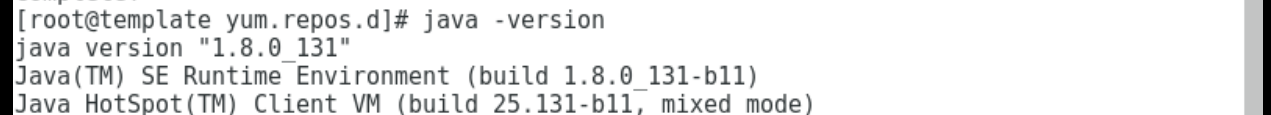
1. **实训结论**

环境搭建：

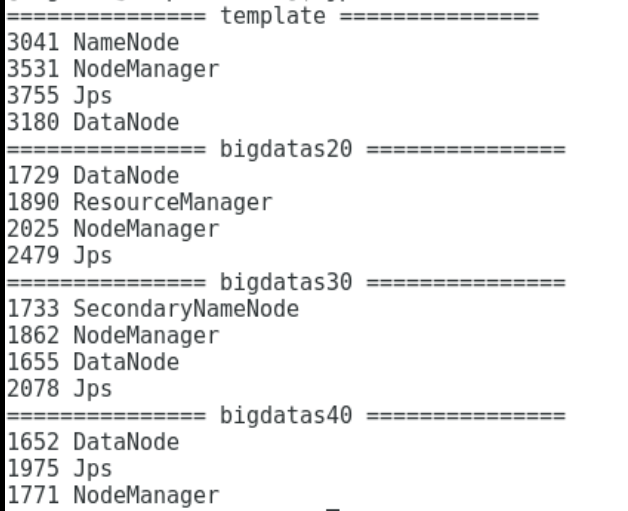
1. 系统正常运行结论，截图

系统正常运行，环境配置配好。

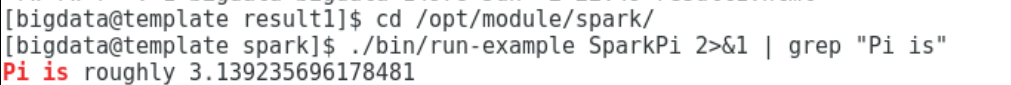
2、JDK配置好：



3、Hadoop正常运行：



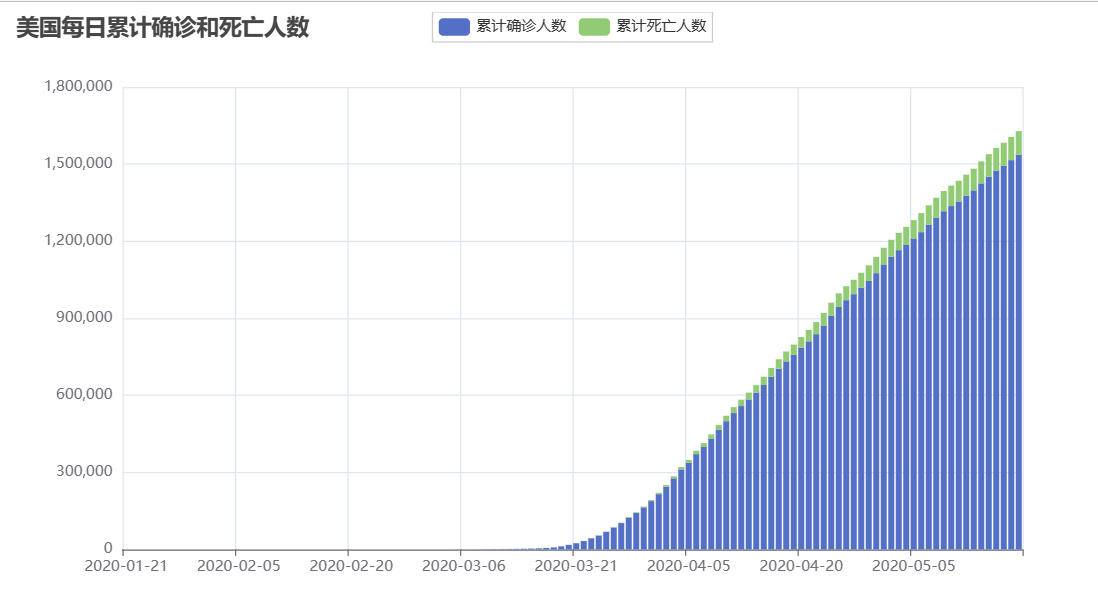
4、spark配置没有问题，启动服务器。

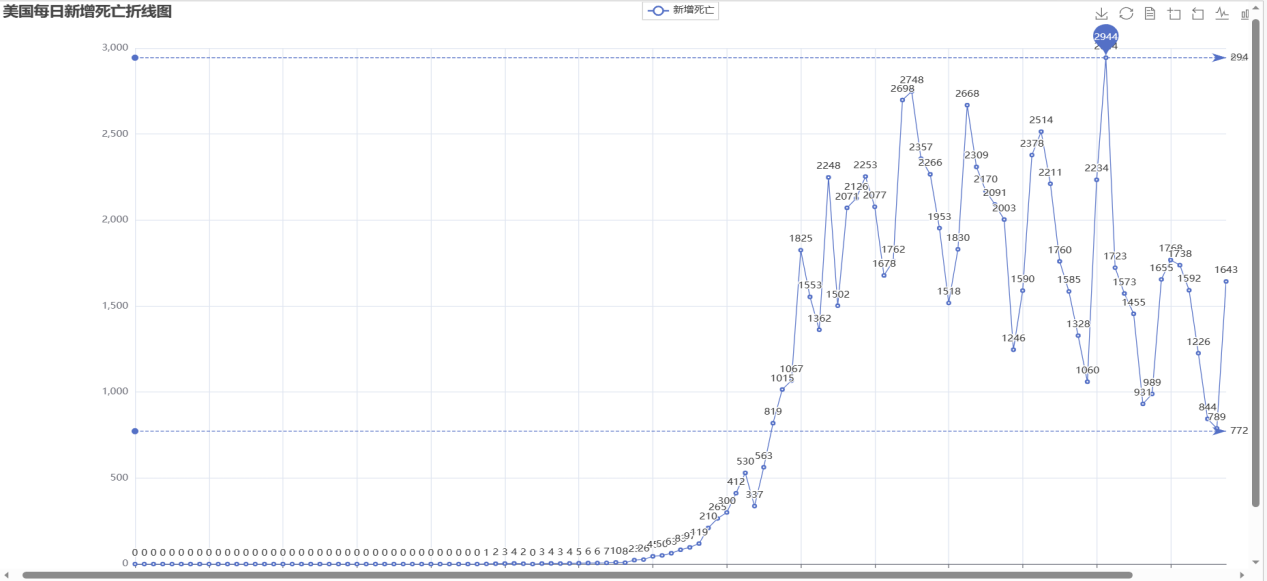




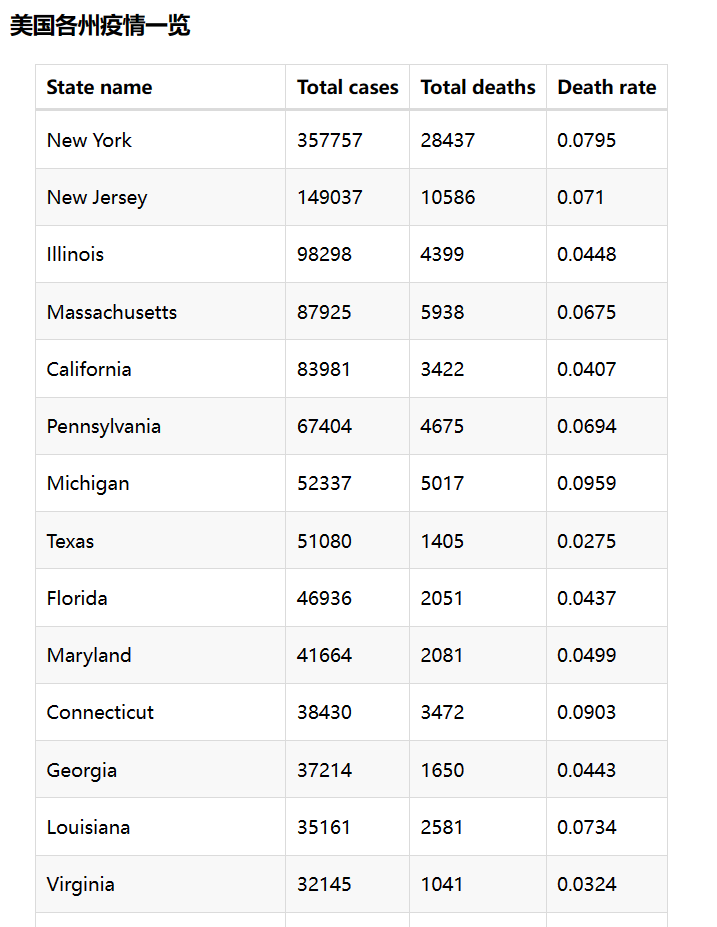
数据处理可视化输出：

任务输出，截图

Result1、

Result2、

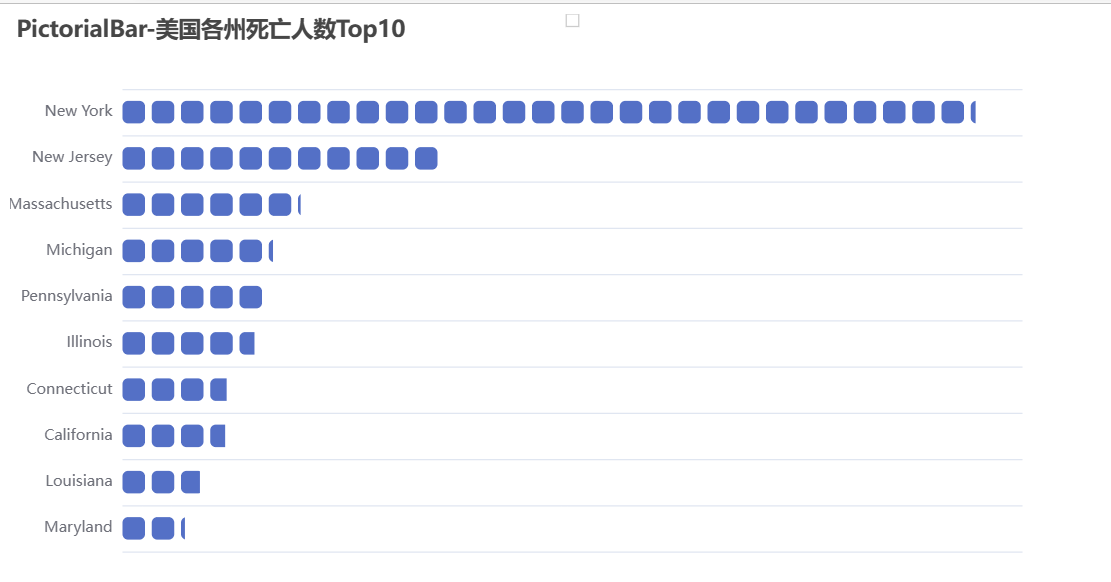
Result3

****

Result4



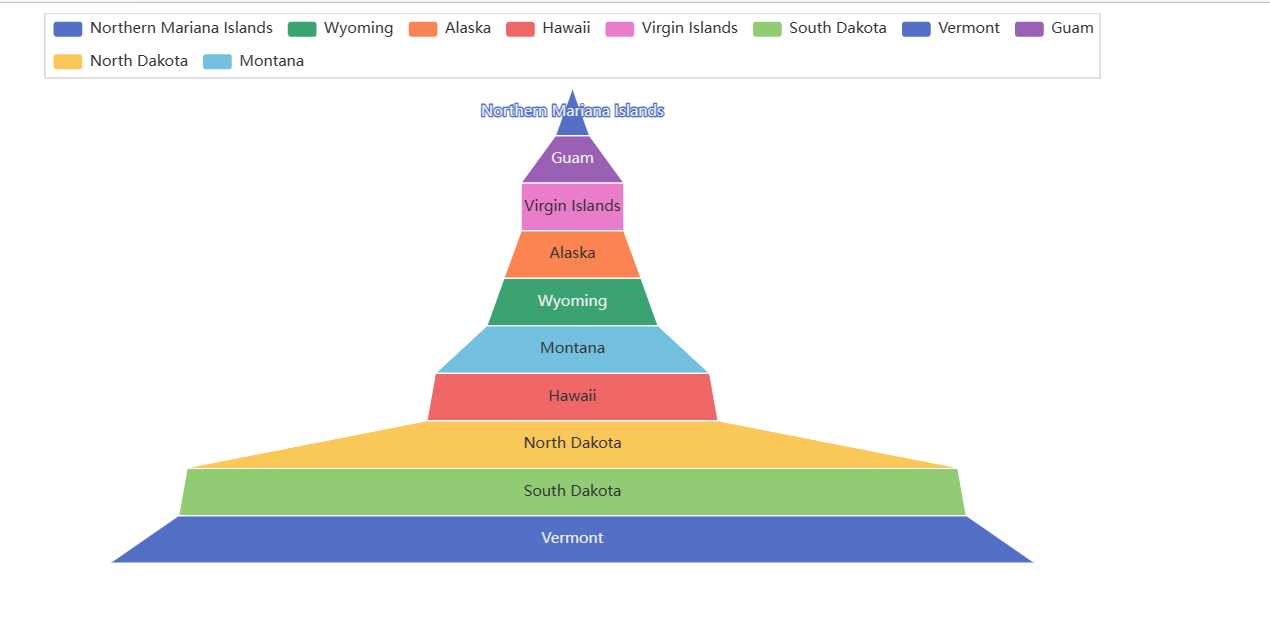
Result5



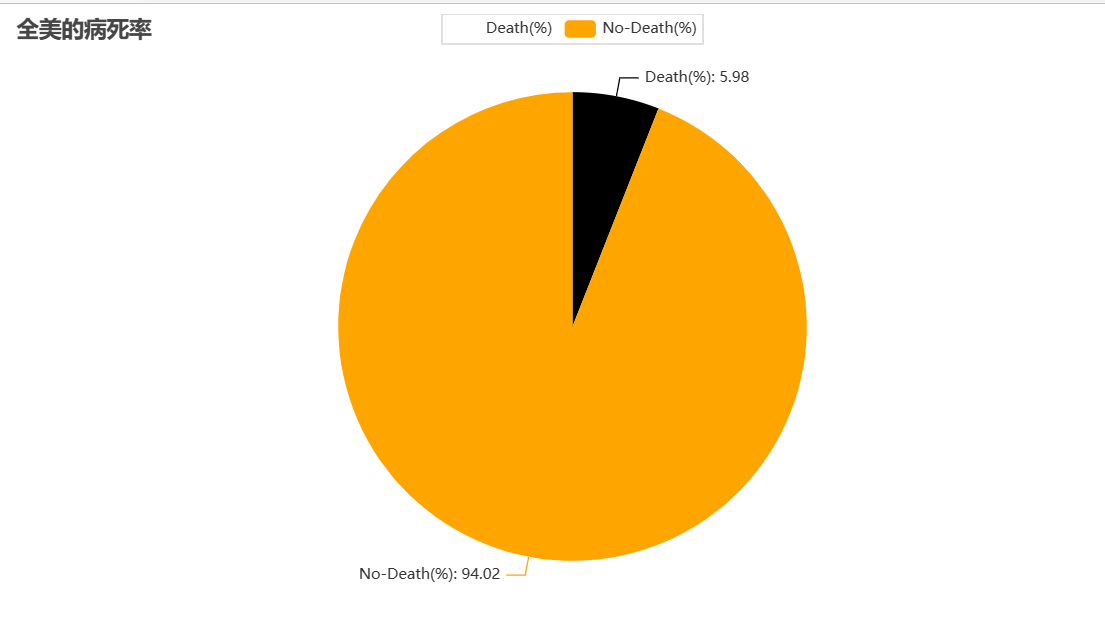
Result6



Result7



Result8

****

**六、总结**

通过实训，学到那些东西？对自己好处，困难解决，后期怎么学习，等等。

通过大数据实时计算的实训，我学到了以下内容：

1、实时计算框架：从大数据实时计算中，我学习了一些流行的实时计算框架，如hadoop、spark、scala等。这些框架提供了快速、高效的数据处理能力，并可以轻松地进行大规模分布式计算。并且学会了运用linux中的python做数据分析。

2、数据存储和管理：我学习了分布式文件系统（如Hadoop HDFS）来存储和管理大数据。学会了HDFS的文件管理系统，使得自己对大数据框架又认识了不少专业知识。

3、数据处理和分析：我学会了运用spark和scala做数据分析。在数据分析的过程中也认识到了疫情给我们人类带来了不可磨灭的影响并一直持续到现在。对社会造成了负面影响。这次的实时计算也使得我对于数据的分析技术提高了不少，处于进步状态。也学会了用虚拟机做实时计算分析。

4、我还学习到了在搭建过程中需要花费耐心去查找搭建资料，这也致使了我学会了一套框架搭建流程。

通过这些学习，我获得以下好处：

1、提高工作效率：掌握实时计算技术，可以帮助您更快速地处理海量数据，提高工作效率，做出有效的判断。

2、市场竞争力：随着大数据技术的普及，掌握实时计算技术将使我在市场上更具竞争力，拥有实时计算可以为自己的职业道路添上一笔浓厚的色彩，并且可以为自己以后深究大数据分析打下良好基础。

3、学习了大数据计算让我明白了很多事情是靠数据来说明的，并不是侃侃而谈，有了计算出来的数据，说的事情才具有说服力，这也是大数据实时计算的必要性。在这门课中我也学到了大数据分析框架的搭建。

在学习过程中，您可能会遇到一些困难，如：

1. 环境配置：实时计算框架需要在分布式环境下运行，因此环境配置可能比较复杂。并且对与JDK、python、hadoop、spark、scala等都需要进行环境的调配，因此常常会出现一系列的错误。
2. 还有就是上传到hdfs文件系统的命令执行会一直报错，导致进度停顿，卡在了上传阶段，这也是较为困扰的一步。
3. 代码测试：由于实时计算是在流上进行的，因此代码测试也花费了一定的时间运行，检查是否可以正常执行。

针对这些困难，我不断通过官方文档、社区交流等途径解决。自己在各大网络平台，如csdn等平台寻找环境配置方法，上传文件方法等，经过了不断的调试的不懈的努力，最终才找到问题的根源，解决问题。

后期我会通过参加相关的培训课程、阅读相关书籍和论文等方式不断深入学习和提高自己的能力。并且加强实践能力，因为实践是检验真理的唯一标准，只有不断学习新知识，并把新知识用于实践中，才可以不断成长。多思考错误的地方，总结经验，收获解决的方法。

这次的大数据实时计算大作业真的让我受益匪浅，自己通过做实验获得了许多不曾有的知识，并且学会了许多解决问题的方法，得到了学习新技能的技巧，也获得了成功的喜悦，从做大作业开始，不断发现问题，不断解决问题，让我的知识得到了巩固和加强，也学习到了新知识。谢谢老师！