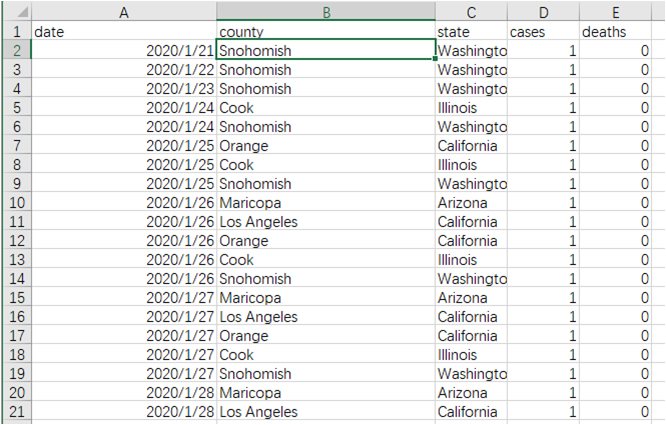
# 2020年美国新冠肺炎疫情数据分析

**一、题目简介**

本题目是根据所学的大数据基础上的相关知识对提供的2020年美国新冠肺炎疫情数据作为数据集进行分析，并对分析结果进行可视化。包括价格分布、各区域租域、租赁类型、地理、时间分布等。

1. **数据集说明**

本次作业使用的数据集来自数据网站Kaggle的美国新冠肺炎疫情数据集，该数据集以数据表us-counties.csv组织，其中包含了美国发现首例新冠肺炎确诊病例至今（2020-05-19）的相关数据。数据包含以下字段：  
**字段名称 字段含义 例子**  
date 日期 2020/1/21；2020/1/22；etc  
county 区县（州的下一级单位） Snohomish；  
state 州 Washington  
cases 截止该日期该区县的累计确诊人数 1,2,3…  
deaths 截止该日期该区县的累计确诊人数 1,2,3…

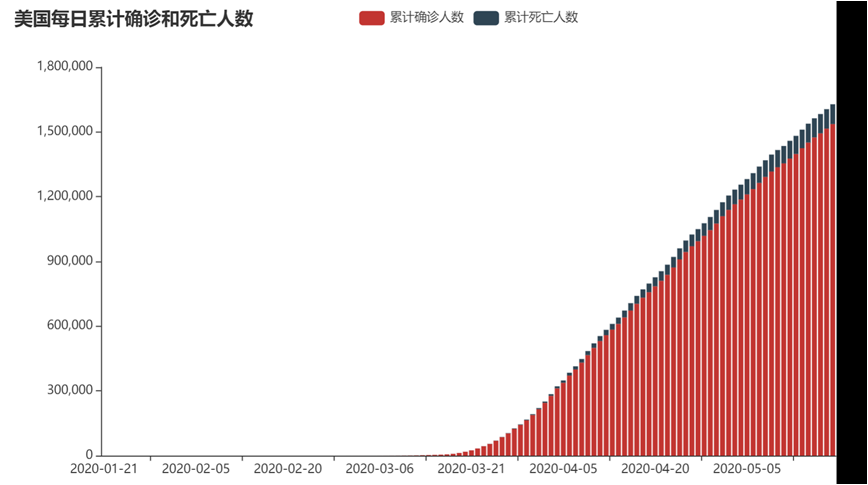


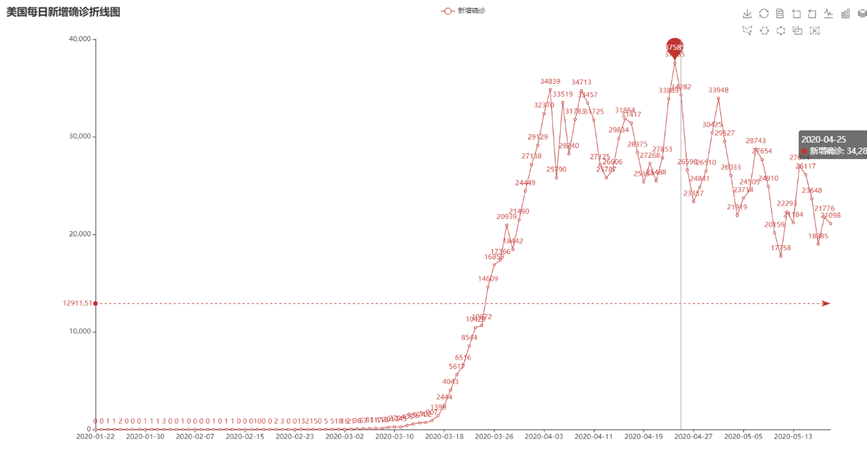
**三、步骤概述**

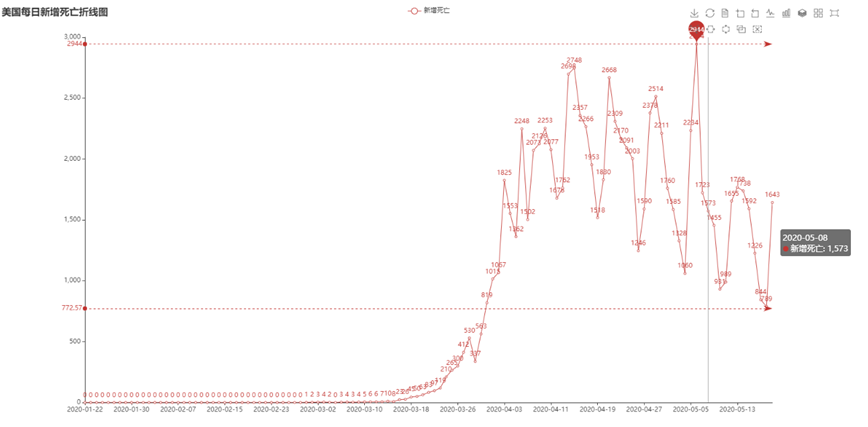
完成课题，大致包含以下5个步骤：  
（1）建立环境。（Hadop、HBase、Hive、MySQL、Sqoop)  
（2）创建数据源读取数据,并将数据存储进MySQL。  
（3）对数据集指定转换操作，保存到HBase。

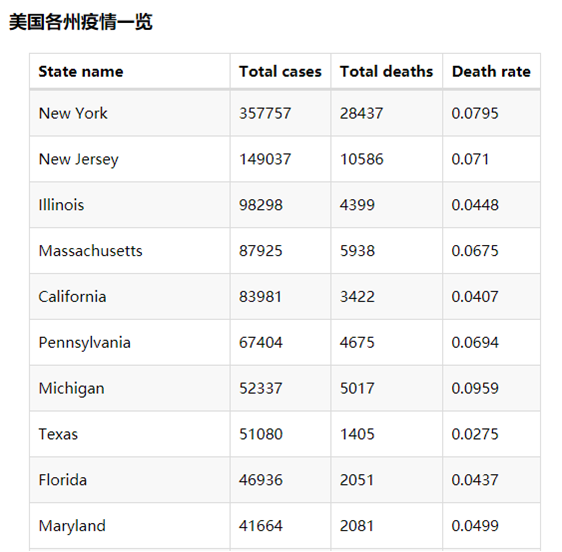
（4）利用Hive实数据进行分析，并将结果保存到MySQL。  
（5）对结果进行可视化。

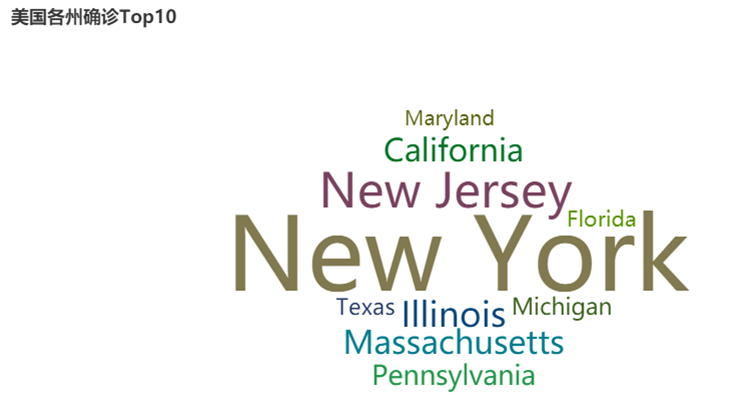
**四、分析结果样例**

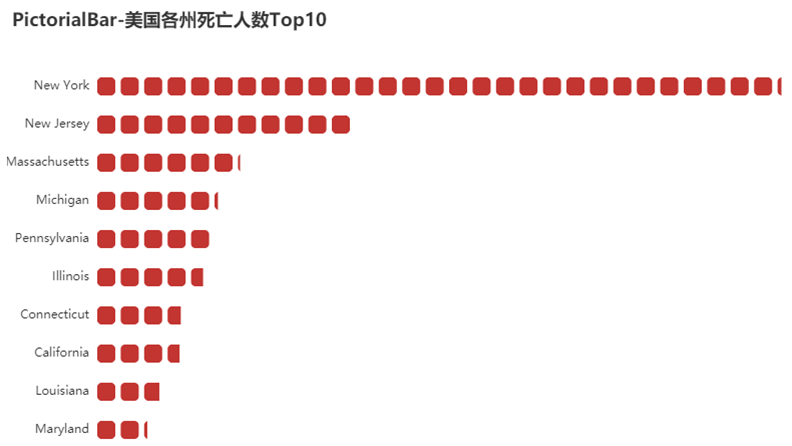
（1）美国每日的累计确诊病例数和死亡数——>双柱状图  


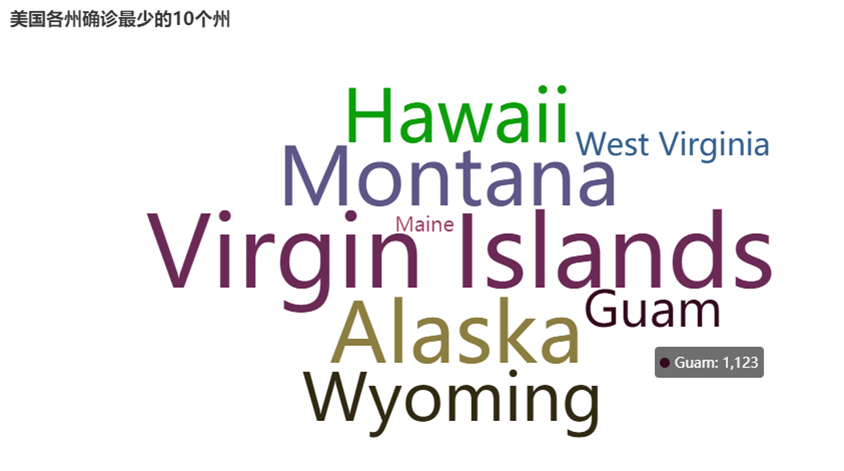
（2）美国每日的新增确诊病例数——>折线图  


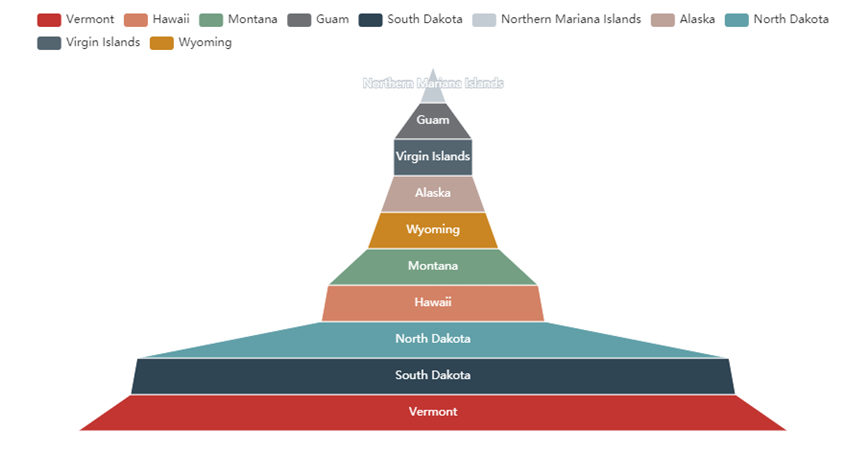
（3）美国每日的新增死亡病例数——>折线图  


（4）截止5.19，美国各州累计确诊、死亡人数和病死率--->表格  


（5）截止5.19，美国累计确诊人数前10的州--->词云图  


（6）截止5.19，美国累计死亡人数前10的州--->象柱状图  


（7）截止5.19，美国累计确诊人数最少的10个州--->词云图  


（8）截止5.19，美国累计死亡人数最少的10个州--->漏斗图  


（9）截止5.19，美国的病死率--->饼状图  
