基于底层平台之上，偏应用的部分。

centos7.5（都可）

分3G内存给虚拟机，尽量不要打开ide

Hadoop：

伪分布：一台虚拟机中不同进程充当不同角色

完全分布：多个虚拟机充当不同角色

Hadoop编程Java

网站上Hadoop内容不包含hbase

hadoop

实验一、2可选。核心存储系统hdfs 计算 mapreduce。 3～6简单。7～9Java编程。hbase安装和使用。

hdfs 分布式文件系统，Hadoop的一个组件，存储在哪里🧐？

hdfs中使用filesystem类或URL方式读取文件。hdfs的api使用 即封装了读取文件的方式，生成jar包并导出，用户在hdfs上运行即可读取文件。

Q：datanode跟namenode的关系？

Q：

hive ：底层依赖hadoop，需要装mysql

1-4

spark：基于分布式的内存文件系统。

spark数据来源于hadoop，也可从hive中获取案例。单节点至少3G内存。

sprk编程Java、python也可。

1-5 stream不做。

综合案例用hive或spark完成较简单。

1环境是所有内容。很多软件。

2数据集压缩包在文件夹里。

3可视化。代码段不能用dataframe无法做对比。