乐清中学 2024 级高二信息技术

晚读、练（四）必修 1 第四章

参考答案

2.数据缺失、数据重复、数据异常、逻辑错误、格式不一致等。

3.分治思想

4.静态数据　　批处理　　Hadoop
流数据　　流计算
图数据　　图计算

5.分布式文件系统、分布式数据库

7.分词、特征提取
①词典、jieba
②统计
③规则

8.词

9.词频

10.智能交通、电子商务

11.[11,22,33],index=['羊','牛','猪']

12.index、values

13.BD

14.index、columns、values、T

15.

df["地区"]　　df.地区
df[["省份","羊数"]]
①df[0:5]　　df[:5] ②df.head()　　df.head(5)
df[df["省份"]=="浙江"]
df.at[8,"羊数"]　　df["羊数"][8]
df[df["羊数"]>100000]

16.BD

17.df.groupby("地区",as_index=False).count()
df.groupby("地区",as_index=True).sum()

18.df.groupby("地区",as_index=False)["省份"].count()
df.groupby("地区",as_index=True)["羊数"].sum()

19.df.sort_values("羊数",ascending=False).head()
df1.省份、df1.羊数
df.groupby("地区",as_index=False).sum()
g1.sort_values('羊数',ascending=False)