1. 大数据定义？

海量计算、大量数据管理、数据分析、大量数据存储

1. 什么是大数据？

大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。

1. 大数据概念？

大数据不仅仅是数据的“大量化”，其包含“快速化”、“多样化”、“价值化”等多重价值。

1. 大数据组成？

大数据是由结构化和非结构化数据组成。

1. Jim Gray 提出的四个范式

实验科学、理论科学、计算科学、数据科学。

1. 海量数据来源

“机器制造“和“人工制造“共同贡献了海量数据，集中式的数据中心建设加速数据集中。

1. Google 云计算平台技术架构
   1. 文件存储，Google Distributed File System，GFS
   2. 并行数据处理 MapReduce
   3. 分布式锁 Chubby
   4. 结构化数据表 BigTable
2. Google 的三大法宝

GFS、MapReduce、BigTable

1. 大数据相关技术
   1. 数据采集
      1. ETL
      2. 数据众包（CrowdSouring）
         1. 数据众包是一种新的数据采集方式，由企业通过平台把数据采集任务外包给非特定的大众网络
   2. 数据存储与管理
      1. 结构化、非结构化和半结构化数据
      2. 分布式文件系统
         1. 分布式文件系统（Distributed File System）是指文件系统管理的物理存储资源不一定直接连接在本地节点上，而是通过计算网络与节点相连
         2. Google File System，GFS 是一个可扩展的分布式文件系统。用于大型的、分布式的、对大量数据进行访问的应用。
            1. GFS 将整个系统分为三类角色：Client（客户端）、Master（主服务器）、Chunk Server（数据块服务器）
         3. Hadoop 是一个分布式系统基础架构，它实现了一个分布式文件系统（Hadoop Distributed File System，HDFS）。特点：高容错性。
            1. HDFS 集群由一个 Namenode 和一定数目的 Datanode 组成。
      3. 关系数据库
         1. 局限性
            1. 难以满足**高并发读写**的需求
            2. 难以满足对海量数据高**效率存储和访问**的需求
            3. 难以满足对数据库**高扩展性和高可用性**的需求
      4. 非关系数据库
         1. NoSQL 数据存储不需要固定的表结构，通常也不存在连接操作。在大数据存取上具备关系型数据库无法比拟的性能优势。
         2. 种类
            1. Bigtable - GFS
            2. Amazon DynamoDB
            3. HBASE - HDFS
            4. mongoDB
            5. Redis
            6. Neo4j
      5. 数据仓库
      6. 云计算和云存储
         1. 云计算（cloud computing）是分布式计算技术的一种，其最基本的概念，是透过网络将庞大的计算处理成相互自动分拆成无数个较小的子程序，再交由多部服务器所组成的庞大系统经搜寻、计算分析之后将处理结果回传给用户。
            1. 云计算是一种资源交付和使用模式，指通过网络获得应用所需的资源（硬件、平台、软件）。提供资源的网络被称为“云“。
         2. 云存储是在云计算概念上延申和发展出来的一个新概念，指通过集群应用、网络技术或分布式文件系统等功能，将网络中大量各种不同类型的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储和业务访问功能的一个系统。
      7. 实时流处理
         1. 传统的分布式计算往往是积累大量的数据，再进行拆分和聚合。而实时流处理则是让数据流动起来，数据从内存中流过，截取需要的数据，进行实时分析计算。
         2. 种类
            1. Puma – Facebook
            2. Storm – Twitter
            3. S4（Simple Scalable Streaming System）-- Yahoo!
   3. 数据分析与挖掘
      1. A/B Testing
         1. 核心为：同时实验两个元素或版本（A和B）,确定哪个更好。
      2. 关联规则分析
      3. 分类
      4. 聚类
      5. 遗传算法
      6. 神经网络
      7. 预测模型
      8. 模式识别
      9. 时间序列分析
      10. 回归分析
      11. 系统仿真
      12. 机器学习
      13. 优化
      14. 空间分析
      15. 社会网络分析
      16. 自然语言分析
      17. MapReduce
          1. Google 提出的软件架构，用于大规模数据集的并行运算
      18. R 语言
          1. 是一种自由软件编程语言与操作环境，主要用于统计分析、绘图、数据挖掘。
   4. 计算结果展示
      1. 标签云（Tag Cloud）
         1. 一套相关的标签以及与次对应的权重，权重影响标签的字体大小
      2. 聚类图（Clustergram）
         1. 指用图形方式展示聚类分析结果的技术，可以有助于判断簇组数量不同时的聚类效果。
      3. 空间信息流（Spatial information flow）
         1. 展示信息空间状态的一种可视化技术。
      4. 热图（Heatmap）
         1. 一种数据展示技术，将变量值使用不同的颜色或高亮形式描绘出来。
2. 数据处理技术分布式演进趋势
   1. SMP
      1. 特点：单机、Scale up
      2. 性能存在瓶颈
      3. 扩展性差
   2. SMP + MPP 混合
      1. 特点：集群、Share Everything
      2. 结构化、关系型
      3. Flash Cache + 分布式块存储 + IB
   3. MPP
      1. 特点：集群、Share Nothing
      2. 结构化、关系型
      3. 通用的硬件
   4. Hadoop
      1. 特点、集群、Share Nothing
      2. 开放、全球生态
      3. 结构化、半结构化、非结构化
      4. 高性能、实时
3. Hadoop 核心设计
   1. MapReduce
      1. Map：任务的分解
      2. Reduce：结构的汇总
   2. HDFS
      1. NameNode：文件管理
      2. DataNode：文件存储
      3. Client：文件获取
4. 大数据的“5V”特性
   1. Volume – 数据量巨大
   2. Variety – 数据类型众多
   3. Velocity – 处理速度快
   4. Veracity – 数据来源真实
   5. Value – 潜在价值大（数据价值密度低）
5. 大数据应用的流程
   1. 数据采集
      1. 又称数据获取，从传感器和互联网等渠道自动采集信息的过程
         1. 网络爬虫工具
            1. Python

Urllib：HTTP 请求库，用于请求、下载网页

Requests：基于 Urllib，但更加快捷方便

Scrapy：爬虫框架，用于从网页中提取数据

* + - * 1. Java

Webmagic：垂直爬虫框架，简化爬虫开发流程，专注与逻辑功能的开发

* + - * 1. C#

DotnetSpider：跨平台、高性能、轻量级爬虫软件

* + - * 1. 分布式采集工具

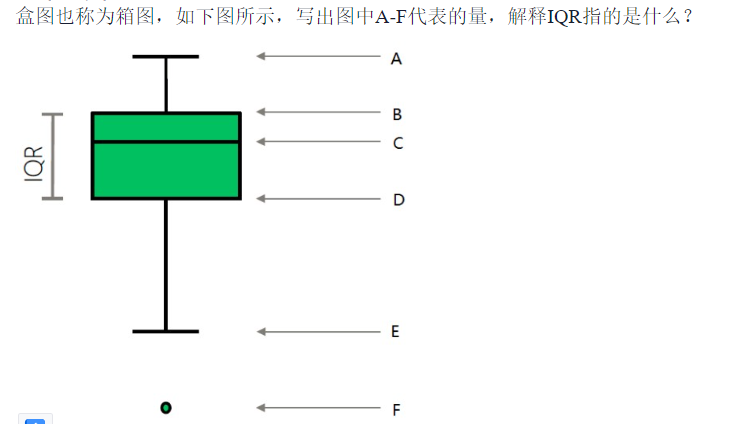
Chukwa（Hadoop）：用于监控大型分布式系统的数据收集系统

Flume（Cloudera）：分布式海量日志采集、聚合和传输系统

Kafka：由 Apache 开源，高吞吐量的分布式发布订阅消息系统

* 1. 数据存储
     1. 数据以目中格式记录在计算机内部或外部存储介质上
     2. 关系型数据库（SQL）
        1. MySQL
        2. Oracle
        3. SQL Server
     3. 非关系型数据库（NoSQL）
        1. Hbase
        2. MongoDB
        3. Redis
  2. 数据清洗
     1. 为了便于后续的处理和分析，对数据进行质量诊断、数据整合、数据转换、缺失值处理和异常值处理等操作；
     2. 数据清洗的方法
        1. 缺失值处理：对存在缺失的数据进行插补
        2. 异常值处理：对数据集中存在的不合理值进行处理
        3. 数据转换：将数据从一种表现形式转换为另一种表现形式
     3. 数据清洗的工具
        1. MapReduce（Hadoop）：基于集群的高性能并发计算框架
        2. Pandas（Python）：解决数据分析任务的 Python 库
        3. OpenRefine：能够对数据进行可视化操作
  3. 数据分析
     1. 利用机器学习等技术从数据中构建模型，从而挖掘出有价值的信息
        1. 机器学习的流程
           1. 明确目标
           2. 准备数据
           3. 训练模型
           4. 测试模型
     2. 数据分析的方法
        1. 机器学习
           1. 降维
           2. 聚类
           3. 分类
           4. 回归
           5. 深度学习
     3. 数据分析的工具
        1. Scikit-learn（Python）：机器学习库，包含众多机器学习算法
        2. Torch：Facebook 力推的深度学习框架，有较好的灵活性和速度
        3. Spark：Apache 开源的为大规模数据处理而设计的快速通用的计算引擎
  4. 数据可视化
     1. 将数据分析的过程和结果用图标等形式进行展示。
     2. 数据可视化工具
        1. Matplotlib（Python）：一个 2D 绘图库，可以绘制许多高质量的图形
        2. Seaborn（Python）：基于 Matplotlib 的高级绘图库
        3. Tableau：一个强大的数据可视化工具
        4. Echarts：基于 JS 的数据可视化图表库

1. 大数据行业应用
   1. 根据用户的喜好向其推荐商品
   2. 利用用户画像进行精准营销
   3. CTR预估
      1. CTR（Click-Through-Rate）即点击通过率，指网络广告的点击到达率
   4. 改进 CTR 预估算法应用于精准定向广告场景
   5. 大数据在金融领域的应用
      1. 智能投顾
      2. 智能客服
      3. 智能获客
      4. 征信与风控
   6. 大数据在健康医疗中的应用
      1. 流感预测
      2. 疾病诊断
      3. 健康管理
   7. 大数据在交通领域的应用
      1. 交通人口及出行监测
      2. 道路拥堵状况预测
      3. 出行路径规划
   8. 大数据在教育行业的应用
      1. 学习评估系统
      2. 智能招生
      3. 校园生活管理
      4. 学生心理健康管理
   9. 大数据在环境保护的应用
      1. 森林监视
      2. 尾气监测
      3. 空气质量预测
2. 大数据挖掘的四项策略
   1. 尽量设想挖掘的场景
   2. 尽量多方面收集数据
   3. 尽量将数据整合
   4. 悉心观察数据特征
3. 数据挖掘技术
   1. 技术分类
      1. 描述（Description）：了解数据中潜在的规律
      2. 预言（Prediction）：用历史预测未来
   2. 数据挖掘技术
      1. 分类（预测）
      2. 聚集
      3. 关联分析
      4. 序列模式
      5. 异常检测
4. 度量中心趋势
   1. 算数平均值
   2. 加权算术平均
   3. 截断均值
      1. 去除高、低极端值得到的均值
   4. 中位数
      1. 有序集的中间值或者中间两的平均
   5. 众数（Mode，也叫 模）
      1. 集合中出现频率最高的值
      2. 对于适度倾斜（非对称的）单峰频率曲线可以使用以下公式
         1. 平均值 – 众数 = 3 X（平均值 – 中位值）
5. 度量数据的离散度
   1. 极差（range）：数据集的最大值和最小值之差
   2. 百分位数（percentile）：第 k 个百分位数是具有如下性质的值 x：k% 的数据项位于或低于 x
      1. 中位数就是第 50 个百分位数
   3. 四分位数
      1. Q1（25 percentile）、Q3（75 percentile）
   4. 中间四分位数极差（IQR）
      1. IQR = Q3 – Q1
   5. 孤立点
      1. 通常我们认为：挑出落在至少高于第三个四分位数或低于第一个四分位数 1.5 X IQR 处的值。
   6. 五数概括
      1. min、Q1、Median（中位数）、Q3、max
   7. 盒图：数据分布的一种直观表示
   8. 方差和标准差
6. 传感器检测得到的原始数据属于 零次数据
7. 大数据时代的数据分析与传统的技术手段最大的区别是 基于云计算
8. 曼哈顿距离位样本各个差值的绝对值之和；欧式距离位样本数据映射在多维坐标系中点的直线距离。
9. ETL（Extract-Transform-Load）：数据抽取、转换、加载
10. Hadoop1.x 可以支持的计算框架有 Map-Reduce
11. 对于数据标准化错误的描述：转换为标准的字符体系
12. 数据的标准化：是将原始各指标数据按比例缩放，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。数据标准化最典型的处理就是归一化处理，将数据统一映射到[0,1]区间上。
13. 数据分析思维模式与传统思维模式不同，主要表现在
    1. 前者主要从“数据“入手，最终改变”业务“；
    2. 后者从“业务“或”决策“等要素入手，最终改变”业务“
14. Precision（准确率）指的是：TP/(TP + FP)
15. 关于 ROC 曲线，曲线越陡效果越理想。
16. 决策树算法在构建数时，根节点的选择策略正确的是：信息增益最大的特征
17. 从数据维度来看
    1. 零次数据：原始数据
    2. 一次数据：干净数据
    3. 二次数据：增值数据
    4. 三次数据：洞见数据
18. 云计算、大数据和物联网代表了IT领域最新的技术发展趋势，三者相辅相成，既有联系又有区别
19. 0-1（Min - Max）标准化转换函数是：(1)(x - min)(max - min)
20. 简述机器学习的基本过程，比如监督、非监督、半监督学习，过拟合指的是什么？出现的原因是什么？
    1. 机器学习的基本过程是用来训练数据（包含输入数据和预期输出的分类或者数值）训练一个模型（Model），利用这个模型，就可以对新的实例数据（Instances）进行分类和计算一个预测值。
    2. 监督学习是机器学习的一种类别，训练数据由输入特征（features）和预期的输出构成，输出可以是一个连续的值（称为回归分析），或者是一个分类的类别标签（称为分类）
    3. 无监督学习与监督学习的区别是，它没有训练样本，直接对数据进行建模
    4. 半监督学习，是监督学习与无监督学习相结合的一种学习方法
    5. 过拟合是指模型参数过多，以至于所选模型对已知数据预测得很好，但对未知数据预测很差。如果一味追求提高对训练数据的预测能力，所选模型的复杂度往往比真模型更高。
21. 简述 HDFS主要特点，并简要说明适合做什么，不适合做什么
    1. HDFS主要特点：
       1. 存储大文件；
       2. 将大文件分割成很多小块存储
       3. 滚式数据读取
       4. 本身是分布式的，具备良好的可扩展性
       5. 通过放开 POSIX 要求，极大地改善数据读写性能
    2. HDFS 不适合于：
       1. 存储大量小文件（< 1MB）
       2. 实时数据读取
       3. 需经常修改数据的场景
22. Hadoop 集群在生产环境下部署的最小规模由哪些节点组成？为什么？HDFS 在写入文件时的具体过程和机架感知策略的用途
    1. 2 个 Master 和 3 个 Slave 节点
    2. 第一个副本：放置在上传文件的 DN；如果是集群外体骄傲，则随机挑选一台磁盘不太满，CPU 不太忙的节点
    3. 第二个副本：放置在与第一个副本不同的机架的节点上
    4. 第三个副本：与第二个副本相同集群的节点
    5. 机架感知：用于容错和负载均衡，提高可用性和优化性能
23. 举例说明什么是结构化数据、半结构化数据和非结构化数据
24. 举例说明大数据时代数据分析思维模式与传统思维模式有什么不同？
    1. 数据科学（Data Science）是指以数据为中心的科学 ，数据分析思维模式与传统思维模式不同，
    2. 前者主要从“数据”入手，最终改变 “业务”；
    3. 后者从“业务”或“决策”等要素入手，最终改变“数据”。如电影电视剧制作等，和传统方式极大不同。
25. 盒图也成为箱图，如下图所示，写出图中 A-F 代表的量，解释 IQR 指的是什么？



盒图中A-F分别表示最大非异常值、上四分位数、Q2中位数、Q1下四分位数、最小非异常值、异常值。

IQR四分位距（interquartile range, IQR），又称四分差，Q3-Q1。（1分）

1. 数据标准化 MIN-Max 标准化和 Z-Score 标准化



1. 计算并填写下表混淆矩阵中的空缺数值

