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. 解决方案
      1. Hadoop（MapReduce 技术）
      2. 流计算（twitter 的 storm 和 yahoo！ 的 S4）