1.第一次信息化浪潮主要解决什么问题
信息（传输，处理，爆炸，转换）
2.下列哪个不属于 hadoop 的特性
成本高，高可靠，高容错，运行在 Linux 上
3.哪项不属于 hdfs 采用抽象的块概念带来的好处
简化系统设计，支持大规模文件存储，强大的跨平台兼容性，适合数据备份
4.关于 Hbase 和 bigtable 的底层技术对应关系错误的是
GFS 与 HDFS 相对应
GFS 与 Zookeeper 相对应
MapReduce 与 hadoop MapReduce 相对应
Chubby 与 Zookeeper 相对应
5.关于 NoSql 和关系数据库的简单比较，哪个错误？
理论基础，横向扩展，数据库模式，面相复杂查询的索引
6.关于云数据库的描述错误的是
部署和虚拟化在云计算环境，在云计算的大背景下发展的共享基础架构的方法，价格不菲维护费用昂贵，高扩展性高可用性采用多租形式支持资源有效分发
7.MapReduce 工作流程，哪个正确
数据交换都通过框架自身实现，不同的 map 任务之间可通信，不同 reduce 任务之间可以信息交换，可以显式地从一台机器向另一台发送消息
8.哪个不属于 hadoop1.0 的问题
单一名称节点单点失效，单一命名空间无法资源隔离，资源管理效率低，很难上手
9.哪个不可能是 Hive 的执行引擎
MapReduce，Tez，Storm，Spark
10.关于 Scala 特性错误的是
语法复杂但能优雅 API 计算，强大并发性支持函数式编程更好支持分布式系统，兼容 java 运行速度快且能融合到 hadoop 生态圈中，它是 Spark 的主要编程语言

多选题
1.信息科技为大数据时代提供哪些技术支撑
存储设备容量不断增加，网络带宽不断增加，CPU 处理能力大幅提升，数据量不断增大
2.hadoop 的特性包括哪些
高可扩展性，支持多种编程语言，成本低，运行在 linux 平台上
3.hdfs 的应用局限性，主要包括哪几个方面？
较差跨平台兼容性，无法高效存储大量小文件，不支持多用户写入及任意修改文件，不适合低延迟数据访问
4.Hbase 访问接口类型包括哪些
Native java API，HbaseShell，ThriftGateway，REST Gateway
5.关系数据库已经无法满足 web2.0 的需求，主要表现在
无法满足海量数据的管理需求，无法满足数据高并发的需求，无法满足高可扩展性和高可用性需求，使用难度高
6.云数据库具有以下哪些特性
动态可扩展，高可用性，免维护，安全
7.MapReduce 体系结构主要由以下哪几个部分构成
Client，JobTracker，TaskTracker，Task

8.Hadoop 的优化与发展主要体现在
自身核心组件 MapReduce 的架构设计改进，自身核心组件 HDFS 的架构设计改进，生态系统其他组件的不断丰富，生态系统减少不必要的组件、整合系统
9.数据仓库 Hive 的执行引擎可以是
Tez，MapReduce，Pig，Spark
10.Spark 具有以下哪几个主要特点
运行速度快，容易使用，通用性，运营模式单一

问答题
1.阐述 HDFS 中名称结点和数据节点的作用
2.阐述 NoSql 数据库有哪些特点
3.描述 MapReduce 算法的执行过程
4.阐述 HDFS 联邦的设计是为了 HDFS1,0 中存在的哪些问题
5.画出 Spark 的运行架构图

第一章

数据：对客观事件进行记录并可以鉴别的符号。

信息：比较宏观，由数据的有序排列组合而成

数据类型：

1、数字

2、文字

3、图像

4、声音

1、文本：不能参与算术运算的任何字符，字符型数据。ASCII、MIME、TXT

2、图片：BMP、JPG 属于点阵图。flash 制作的 SWF 等和 ps 做的 PSD 等属于矢量图。

3、音频：数字化的声音数据就是音频数据。CD、WAV、MP3、MID、WMA、RM

4、视频：视频数据是指连续的图像序列。MPEG-4、AVI、RM、MOV、ASF、WMV、DivX

数据组织形式：

1、文件：文件由文件系统负责管理

2、数据库：1968 年 IBM 公司推出第一个大型商用数据库管理系统 IMS，数据库经历了，层次数据库、网状数据库、关系数据库

NoSQL 数据库。关系数据库仍是目前数据库的主流。Web2.0 兴起，非结构化数据迅速增加，目前人类社会中有 90%

是非结构化数据。NoSQL 数据库更好的支持非结构化数据管理

数据的使用：

1、数据清洗：把数据变成可用的状态。工具：“古老”的 UNIX 工具 AWK、XML 解析器和机器学习库。脚本语言，如 Perl 和 Python 在此过程

中发挥重要作用。

2、数据管理：数据经历清洗后，被存放到数据库系统中进行管理。

3、数据分析：分析数据需要借助于数据挖掘和机器学习算法，同时需要相关的大数据处理技术。

数据的价值：

数据的价值在于可以为人们找出答案。数据的价值不会因为不断使用而消减，反而会因为不断重组而变得更大。

数据爆炸

第三次信息化浪潮：

第一次 1980 前后 标志 个人计算机 解决 信息处理 代表企业 Intel AMD IBM 等

第二次 1995 前后 标志 互联网 解决 信息传输 代表企业 雅虎、谷歌

第三次 2010 前后 标志 物联网、云计算 大数据 解决 信息爆炸 代表企业 亚马逊 谷歌 IBM

信息科技为大数据时代提供技术支撑 ：

信息科技进步是大数据时代的物质基础：解决信息储存 、信息处理、信息传输 3 个核心问题

1、存储设备容量不断提升

2、CPU 处理性能大幅增加

3、网络带宽不断增加

数据产生方式的变革促成大数据时代的来临

1、运行式系统阶段

只有当实际的企业业务发生时，才会产生新的数据并存入数据库

2、用户原创内容阶段

weibo qq

3、感知式阶段

物联网导致数据量第三次跃升

大数据发展三个阶段，萌芽期、成熟期大规模应用期

1、数据挖掘理论和数据库技术的逐步成熟，一些工具和技术被应用

2、Web2.0 大量数据产生

3、大数据应用于各行各业

各国大数据发展战略：考就摆烂，（英国应对脱欧经济挑战）

大数据的概念：

1、数据量大：每年 50%增长，两年产生之前总和

2、数据类型繁多

3、处理速度快

4、价值密度低

大数据对科学研究的影响

1、实验科学

2、理论科学

3、计算科学

4、数据密集型科学

大数据对社会发展的影响

1、大数据决策成为一种新的决策方式

2、大数据成为提升国家治理能力的新方法

3、大数据应用促进信息技术与各行业的深度融合

4、大数据开发推动新技术和新应用不断涌现。

大数据对就业市场影响

数据科学家，零售、金融、互联网企业。

中国用户还主要局限在结构化数据分析方面，未来空间很大

也有人认为，未来采用自动化处理，对人才需求降低

大数据对人才培养的影响：

需要学科杂，数据多，高校不具备。多位实际工作环境中成长起来

高校应引进企业，走出实践

大数据应用：略 靠想象

大数据产业

1、IT 基础设施层

2、数据源层

3、数据管理层

4、数据分析层

5、数据平台层

6、数据应用层

习题：

1、阐述数据的基本类型

数字，文本，图像，音频，视频

2、阐述数据可用经历步骤

数据清洗、数据管理、数据分析

3、阐述三次浪潮时间 标志 解决问题

看前面好好背

4、阐述信息技术对大数据时代的支撑

存储、处理、传递（详细见前）

5、数据产生方式的三个阶段

运行式系统阶段，用户原创内容阶段，感知式阶段

6、阐述大数据发展的 3 个重要阶段

萌芽期、成熟期、大规模应用期。（大概一基础理论出现，二数据大规模出现，三应用普遍）

7、大数据 4v 特性

（ 大大大小）数据量大，数据种类多，数据处理速度快，数据价值密度低。

8、大数据对科学研究有什么影响

（继实验科学范式，理论科学范式，计算科学范式后）给人们带来了数据密集型科学范式，为人类提供了认识复杂系统的新思维和新手段。

9、举例说明大数据应用

（随便说几个啥什么商家分析客户习惯，智能 xxx）

10、阐述高校大数据专业知识体系

（太多了建议摆烂，一部分：）从大数据分析来说，数据采集与预处理，数据储存与管理，数据处理与分析，数据可视化

第二章

云计算的概念

从商业模式：通过网络、以服务的方式为千家万户提供非常廉价的 IT 资源和技术。

AWS 亚马逊最早推出的云计算网络服务

以挖井取水为例子

1、初期成本高、周期长

2、后期需要自己维护

3、供水量有限

自来水

1、初期 0 成本、使用成本低

2、后期免维护费，使用成本低

3、在供水方面“予取予求”

传统 IT 资源

1、初期成本高、周期长

2、后期需要自己维护、成本高

3、IT 资源供应有限

云计算优点

1、初期 0 成本，瞬时可以获得

2、后期免维护，使用成本低

3、在 IT 资源供应方面“予取予求”

云计算的服务模式和类型

1、基础设施即服务：IaaS

2、平台即服务：PaaS

3、软件即服务：SaaS

云计算包括

1、公有云：面向所有用户提供服务。如 AWS

2、私有云：只为特定用户提供服务。

3、混合云：综合前两个特点。因为一些企业出于安全性吧数据放在私有云，一方面希望获得公有云资源。

云计算数据中心

1、数据中心包括一套复杂的设施，包括刀片服务器、宽带网络、环境控制设备、监控设备以及安全装置等

2、数据中心是云计算的重要载体，数据中心里 CPU、内存、磁盘、带宽等 IT 资源汇集成一个庞大的 IT 资源池。

3、福建有两大重点数据中心。贵州被公认为中国南方最适合建设数据中心的地方，中国移动、联通、电信将南方的数据中心建在贵州。

云计算的应用

政务云，教育云，中小企业云，医疗云。

云计算产业

硬件与设备制造、基础设施运营、软件与解决方案供应商、基础设施即服务、平台即服务、软件即服务

物联网的概念

物联网是物物相连的互联网，是互联网的的延伸，它利用局部网络或互联网等技术把传感器、控制器、机器、人员或物等通过新的方式连在一起，形成人与物、物与物相连，实现信息化和远程管理控制。

（物联网就是通过一系列机器把物品联入互联网）

四个层次

1、感知层：获取

2、网络层：传递

3、处理层：处理

4、应用层：应用

关键技术

1、识别和感知技术

二维码 RFID 传感器

2、网络与通信技术

短距离无线通信 WIFI 蓝牙 和长距离通信 2/3/4G 移动通信网络

3、数据挖掘与融合技术

物联网的应用

智能 xx

（交通、医疗、农业、家居）

物联网产业

核心感应器件供应商

感知层末端设备供应商

网络运营商

软件与行业解决方案提供商

系统集成商

运营及服务提供商

大数据、云计算、互联网的关系

联系：

云计算为大数据提供了技术基础、大数据为云计算提供了用武之地

物联网是大数据的重要来源、大数据为物联网数据分析提供了技术支撑

云计算为物联网提供了海量数据存储能力、物联网为云计算技术提供了广阔的应用空间。

区别：

大数据侧重于海量数据存储、处理与分析；

云计算旨在整合和优化各种 IT 资源并通过网络以服务的方式，廉价地提供给用户

物联网实现物物相连

人工智能的概念

研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门技术学科。

人工智能的关键技术

1、机器学习：算法、经验、性能

2、知识图谱：知识域可视化、知识域映射地图

3、自然语言处理：

4、人机交互

5、计算机视觉：可分为成像学、图像理解、三维视觉、动态视觉和视频编解码

6、生物特征识别

7、VR/AR

人工智能的应用

1、智能制造

2、智能家居

3、智能金融

4、智能交通

5、智能安防

6、智能医疗

7、智能物流

8、智能零售

人工智能产业

1、基础设施和建设：智能芯片、智能传感器、分布式计算框架

2、智能信息和数据：数据提供商、数据采集存储分析综合性产商

3、智能技术服务：提供算法模型、提供解决方案、提供人工智能在线服务

4、智能产品；智能机器人、智能终端、自然语言处理（翻译）、计算机视觉、生物特征识别（人脸识别）。。。

大数据与人工智能关系

联系：

1、人工智能需要数据来建立其智能，特别是机器学习

2、大数据技术为人工智能提供了强大的存储能力和计算能力。

区别：

1、人工智能是一种计算形式，会根据结果反馈；大数据是会寻找结果

2、大数据通过数据对比推演出更优方案；人工智能的开发是为了辅助或代替我们。

比特币以及区块链

1、去中心化

比特币需要解决问题一：防篡改（哈希函数特性 略，见密码学）

如何交易：比特币地址对应银行卡号，私钥对应密码（还会应用密码的不可抵赖性）

问题二：

去中心化记账：太长略

区块链定义

用块链式数据结构来验证与存储数据的方式

三要素：

交易：一次操作会导致账本状态的一次改变

区块：一个区块记录了一段时间内发生的交易和状态结果，是对当前账本状态的一次共识

链：由一个个区块按照发生顺序串联而成，是整个状态变化的日志记录

区块链应用

1、金融领域

2、物流领域

3、物联网领域

4、版权保护

5、教育行业

6、数字政务

7、公益和慈善

8、实体资产

9、社交

大数据与区块链的区别

1、数据量：区块链技术是分布式数据存储，数据量小处理更加细致；大数据是海量数据，处理更粗糙。

2、结构化和非结构化：区块链是典型的结构化数据；大数据需要处理更多的非结构化数据。

3、独立和整合：区块链信息相对独立以保证安全性；大数据的重点是信息的整合分析。

4、直接和间接：区块链是一个分布式账本，本质上是一个数据库；大数据是对数据处理，适合一种间接的数据。

5、CAP 理论：C 一致性 A 可用性 P 分区容忍性；三者不可同时满足，区块链 CP，大数据 AP

6、基础网络：区块链 P2P，大数据计算机集群

7、价值来源：区块链数据是资产；大数据数据是信息，价值需要提炼

8、计算模式：区块链多人重复做一事；大数据一事分摊多人做。

联系：

区块链的可信任性、安全性呵呵不可篡改性，让更多的数据被释放出来

1、区块链使大数据极大的降低信用成本

2、区块链是构建大数据时代的信任基石

3、区块链时候促进大数据价值流通的管道

习题

1、阐述云计算概念

为所有人提供便利可用的低成本 IT 资源。

（通过网络、以服务的方式为千家万户提供非常廉价的 IT 资源和技术。）

2、云计算有哪几种服务模式和那几种类型

IaaS PaaS SaaS 类型：混合云、私有云、公有云。

3、阐述什么是数据中心和数据中心在云计算中的作用

数据中心是与云计算服务最终数据存储的位置。

（数据中心包括一套复杂的设施，包括刀片服务器、宽带网络、环境控制设备、监控设备以及安全装置等）

数据中心是云计算的重要载体，数据中心里 CPU、内存、磁盘、带宽等 IT 资源汇集成一个庞大的 IT 资源池。

4、云计算经典应用

政务云，教育云，中小企业云，医疗云。

5、阐述物联网概念和物联网各个层次的功能

物联网是物物相连的互联网，是互联网的的延伸，它利用局部网络或互联网等技术把传感器、控制器、机器、人员或物等通过新的方式连在一起，形成人与物、物与物相连，实现信息化和远程管理控制。

（物联网就是通过一系列机器把物品联入互联网）

四个层次

感知层：获取

网络层：传递

处理层：处理

应用层：应用

6、请阐述物联网关键技术

识别与感知技术 网络与通信技术 数据挖掘与融合及技术

7、大数据 云计算 物联网的相互关系

联系：

云计算为大数据提供了技术基础、大数据为云计算提供了用武之地

物联网是大数据的重要来源、大数据为物联网数据分析提供了技术支撑

云计算为物联网提供了海量数据存储能力、物联网为云计算技术提供了广阔的应用空间。

区别：

大数据侧重于海量数据存储、处理与分析；

云计算旨在整合和优化各种 IT 资源并通过网络以服务的方式，廉价地提供给用户

物联网实现物物相连

8、阐述人工智能的概念

研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门技术学科。

9、人工智能关键技术

机器学习 知识图谱 自然语言处理 人机交互 生物特征识别 计算技术视觉 VR/AR

10、人工智能和大数据的关系

联系：

1、人工智能需要数据来建立其智能，特别是机器学习

2、大数据技术为人工智能提供了强大的存储能力和计算能力。

区别：

1、人工智能是一种计算形式，会根据结果反馈；大数据是会寻找结果

2、大数据通过数据对比推演出更优方案；人工智能的开发是为了辅助或代替我们。

11、区块链概念以及区块链和比特币的关系

用块链式数据结构来验证与存储数据的方式，是比特币的基础技术

12、区块链如何解决防篡改

运用到了哈希函数的单向性和弱碰撞性，难以在合理的时间内计算出合理的值篡改

13、区块链和大数据关系

「

1、数据量：区块链技术是分布式数据存储，数据量小处理更加细致；大数据是海量数据，处理更粗糙。

2、结构化和非结构化：区块链是典型的结构化数据；大数据需要处理更多的非结构化数据。

3、独立和整合：区块链信息相对独立以保证安全性；大数据的重点是信息的整合分析。

4、直接和间接：区块链是一个分布式账本，本质上是一个数据库；大数据是对数据处理，适合一种间接的数据。

5、CAP 理论：C 一致性 A 可用性 P 分区容忍性；三者不可同时满足，区块链 CP，大数据 AP

6、基础网络：区块链 P2P，大数据计算机集群

7、价值来源：区块链数据是资产；大数据数据是信息，价值需要提炼

8、计算模式：区块链多人重复做一事；大数据一事分摊多人做。

联系：

区块链的可信任性、安全性呵呵不可篡改性，让更多的数据被释放出来

1、区块链使大数据极大的降低信用成本

2、区块链是构建大数据时代的信任基石

3、区块链时候促进大数据价值流通的管道

」

第三章习题

1、传统数据安全的威胁主要包括哪些

计算机病毒 黑客攻击 数据信息存储介质的损坏

2、大数据安全与传统的数据安全不同

大数据成为网络攻击的显著目标

大数据家的隐私泄漏风险

大数据技术被应用到攻击手段中

大数据成为高级可持续攻击的载体

3、大数据安全事例

棱镜门

维基解密

Facebook 数据滥用事件

手机 app 过度采集个人信息

12306 数据泄露

免费 wifi 窃取用户信息

收集个人信息的“探针盒子‘

4、机械思维的核心思想

世界变化的规律是确定的

因为有确定性，所以规律不仅可以被认识，而且可以用简单的公式或者语言来描述清楚

规律应该放之四海而皆准，可以应用到任何场合都死活正确的

5、大数据时代为什么需要新思维

不确定性在生活中无处不在，现在的数据量相比于过往打了很多，量的积累带来质变

6、大数据时代人的思维转变在哪些方面

全样而非抽样

效率而非精确

相关而非因果

以数据为中心

我为人人人人为我

7、举个例子 ：啤酒与尿布相关而非因果

8、大数据伦理概念

由大数据技术产生和使用引发的社会问题。

9、举个例子 ：大数据杀熟

10、大数据伦理表现在：

隐私泄漏问题

数据安全问题

数字鸿沟问题：大数据技术优先在发达地方普及，进一步拉大差距

数据独裁问题：大数据时代数据量爆炸增长过，迫使人们必须完全依赖数据才能做出决策

数据垄断问题：大企业因为拥有更多数据而产生垄断的行为

数据的真实可靠性

人的主体地位问题：在一些皆数据的情况下，人的主体地位正在逐渐消失

11--14 见上

15--19:

政府数据孤岛：政府职能部之间难以实现对数据的共享

企业数据孤岛：企业内各部门数据无法共享

政府原因：有些政府任务数据资源占有就是财富，有些政府因为数据标准不一样，不能与外界共通

企业原因：以功能为标准的部门划分，不同版本类型的信息化管理系统

消除的意义：对政府可以提高自用利用率，有助于推动政府转型；对企业提高企业生产率，企业间由更好的发展能力

20、政府开放数据理论基础：数据资产理论 数据权理论 开放政府理论

21、政府信息公开与数据开放联系与区别

信息是数据加工后得到的。信息公开满足公众知情权，数据开放是信息公开的自然延伸。

22、政府数据开放的意义：

有利于促进开放透明的政府形成

有利于创新创业和经济增长

有利于社会治理创新

23、交易平台包括的类型：

综合数据服务平台

第三方数据交易平台

24、交易平台数据来源

政府公开数据

企业内部数据

数据供应方数据

网页爬虫数据

25、交易平台的产品类型

API

数据包

云服务

解决方案

数据定制服务

数据产品

26、举例简述运营模式

（太长摆烂，记点脑补）

一种兼具中介和数据处理加工功能

以种植具备中介功能

27、具有代表意义的大数据交易平台

贵阳大数据交易所

上海数据交易中心

华东江苏大数据交易中心

浙江大数据交易中心

习题未提及：

政府层面的挑战：不愿共享开放、不敢共享开放、不会共享开放、数据中心共享开放作用不强

企业层面的挑战：系统孤岛挑战、组织架构挑战、数据合作挑战

数据共享案例：菜鸟物流等

数据交易形式：

1、大数据交易公司：向买方出售数据和为用户出售个人数据（= =有啥区别）

2、数据交易所

3、API 模式

4、其他：中国知网、北大法宝等，通过收取费用向用户提供文章等

第五章习题

1、传统采集和大数据采集区别：

传统采集： 数据源，来源单一，数据量相对较少；数据类型，结构单一；数据存储，关系数据库和并行数据库

大数据采集：数据源，来源广泛，数据量巨大；数据类型，丰富，包括结构化、半结构化、和非结构化数据；数据存储，分布式数据库， 分布式文件系统

2、数据采集三大特点：

全面性

多维性

高效性

3、数据采集的数据源

传感器数据

互联网数据

日志文件

企业业务系统数据

4、数据采集方法

系统日志采集

分布式消息订阅分发

5、什么是网络爬虫

网络爬虫是自动抓取网页的程序

6、网络爬虫组成：

控制节点 爬虫节点 资源库

7、网络爬虫类型：

通用网络爬虫

聚焦网络爬虫

增量式网络爬虫

深层网络爬虫

8、Scrapy 爬虫的体系架构

Scrapy 引擎（engine）

爬虫（Spiders）

下载器（Downloader）

调度器（Scheduler）

项目管道（Item Pipeline）

下载器中间件（Downloader Middlewares）

爬虫中间件（Spiders Middlewares）

调度器中间件（Scheduler Middlewares）

9、数据清洗主要内容：

缺失值处理：估算 整列删除 变量删除 成对删除

异常值处理

数据类型转化

重复值处理

10、数据清理注意事项

缺失值、异常值、数据类型转化、重复值的顺序处理

处理数据按照业务要求

数据清洗前了解数据表的结构和要处理的值

数据量大小影响清洗操作（补或删）

看起来可用的数据导入也要清洗

11、数据转换策略：

平滑处理

聚集处理

数据泛化处理

规范化处理

属性构造处理

12、数据脱敏原则

保持原有数据特征

保持数据间一致性

保持业务规则的关联性

多次脱敏数据之间的数据一致性

13、数据脱敏方法：

数据替换

无效化

随机化

偏移和取证

掩码屏蔽

灵活编码

习题未提及：

核心是 Agent 一个 Agent 就是一个 java 虚拟机（Java Virtual Machine JVM）Agent 是完成的数据采集工具

三个核心组件数据源（Source）数据通道（Channel）数据槽（Sink）

分布式消息订阅分发中 Kafka 是一个高吞吐量的分布式发布/订阅消息系统，实时在线处理的低延迟，批量离线处理的高吞吐量

四个组件：

话题（Topic）产生特定类型的信息流

生产者（Producer）能够发布消息

服务代理（Broker）保存已发布的消息的服务器，被称为代理或 Kafka 集群

消费者（Consumer）可以订阅一个或多个话题，并从服务代理拉数据，从而"消费"这些已经发布的消息

ETL：常用于数据仓库的数据采集和与处理环节

从原系统中抽取数据，转换后加载到目标数据存储中

它既可以用于数据采集环节也可以用于数据预处理环节

Scrapy 工作流：取 URL 抓去 下载 解析 反馈

反爬机制：

成因：一企业不愿自己数据被免费获得。二低级爬虫访问请求多影响企业业务

双刃剑：一保护企业数据。二误伤用户

第六章习题

1、传统的数据储存与管理技术

文件系统、关系数据库、数据仓库、并行数据库

2、关系数据库特性

（产品：Oracle、SQL sever MySQL、DB2）

储存方式：关系数据库采用表格方式，数据以行列方式进行存储，读取和查询十分方便

存储结构：结构化的方式存储数据。

存储规范：数据按照最小关系表的形式进行存储

扩展方式：数据存在表中，多表提取数据时操作受限

查询方式：结构化查询语言查询，灵活强大

事务性：关系数据库支持 ACID（原子性 一致性 隔离性 持久性）

连接方式：不同的关系数据库产品都遵守一个统一的数据库连接接口标准。

3、数据仓库特性

面向主题

集成

相对稳定：不可更新主要是查询

反映历史变化

4、hadoop 特性

高可靠性

高效性

高扩展性

高容错性

成本低

基于 java，可运行在 linux

支持多种编程语言 如 c++

5、hadoop 生态系统及其功能

HDFS：分布式文件管理系统处理超大数据、流式处理

Hbase：一个提供高可靠性、高性能、可伸缩、实时读写、分布式的列式数据库

MapReduce：将复杂的、运行于大规模集群上的并行计算高度的抽象到 Map 和 Reduce 函数上

Hive：对 hadoop 文件中的数据集进行过数据整理、特殊查询和分析存储

Pig：是一种数据流语言和运行环境

Mahout：提供一些可扩展的机器学习领域经典算法的实现

ZooKeeper：是高效可高的协同工作系统，提供分布式所之类的服务。用 java 编写，可用 java c 接入

Flume：是一个高可用的、高可靠的、分布式的海量日记采集、聚集和传输工具。有收集数据然后将数据简单处理写到接收方的能力

Sqoop：SQL-to-Hadoop 的缩写，主要用来在 Hadoop 和关系数据库中交换数据

Ambari：基于 web 的工具，支持 Apache HHAdoop 集群的安装、部署、配置和管理

6、HDFS 设计实现要点：

兼容廉价的硬件设备

流数据读写

大数据集

简单的文件模型

强大的跨平台兼容性

7、名称节点和数据节点的具体功能

名称节点作为中心服务器，负责管理文件系统的命名空间及客户端对文件的访问

数据节点负责处理文件系统客户端的读写请求

8、各数据库的适用场合和优缺点

键值数据库，会使用哈希表通过 key 来定位 value，扩展性好，灵活性好，对大量写操作时候性能高；无法存储结构化信息，条件查询效率低。代表产品 Redia

列族数据库，查找速度快、可扩展性强、容易进行分布式扩展、复杂性低；缺点是功能较少、大多不支持强事物一致性。代表产品 Hbase

文档数据库，

文档数据库，性能好、灵活性高、复杂性低、数据结构灵活；缺点是缺乏统一的查询语法。代表产品 MongoDB

图数据库，灵活性高、支持复杂的图运算、可用于构建复杂的关系图谱；缺点是复杂性高，只能支持一定的数据规模

9、云数据库概念

云数据库是部署在云计算环境中的虚拟化数据库。 10、云数据库特性：

动态可扩展

高可用性

较低的使用代价

易用性

高性能

免维护

安全

11、云数据库和其他数据库之间的关系：

云数据库后端以 MySQL 为主 NoSQL 为辅。提供的服务关系数据库和非关系数据库都有

12、云数据库厂商代表产品

Google Google Cloud SQL

百度 百度云数据库

腾讯 腾讯云数据库

13、Hadoop 中国 Hbase 和其他的关系

利用 Hadoop MapReduce 来处理 Hbase 中海量数据，实现高性能计算

利用 ZooKeeper 作为协同服务，实现稳定服务和失败恢复；

使用 HDFS 作为高可靠的低层数据存储系统

Sqoop 为 Hbase 提供了高效便捷的关系数据管理系统（RSBMA）数据导入功能

Pig 与 Hive 为 HBase 提供了高层语言支持

14、见 p197

15、

行健：每个 HBase 表由若干行组成，每个行由行健来标识

列族：他是基本的访问控制单元

时间戳：每个单元格都保存着同一份数据的多个版本，这些版本采用时间戳进行索引

16、HBase 系统架构以及功能

客户端：包含访问 Hbase 借口，缓存已经访问过 Region 位置

ZooKeeper 服务器：监控工作状态并通知 Masiter

Master 主服务器：管理各部分工作

Region 服务器：包含位于某个值域的所有数据，负载均衡和数据分发的基本单位

17、Spanner 服务器的组织方式

zonemaster 把数据分配给 Spanserver、Spanserver 吧数据提供给客户端，客户端使用每个 zone 上的 Locatinproxy 来定位可以为自己提供数据的 spanserver。universemaster 是一个控制台，显示各 zone 信息，placementdriver 还会周期性的与 spanserver 交互，发现需要转移数据。

习题未涉及

数据仓库通常包括：数据源，数据存储和管理，OLAP 服务器，前段工具和应用

并行数据库：无共享的体系结构中进行数据操作的数据库系统，大多采用了关系数据库模型并且支持 SQL 语句查询；缺点似乎没有较好的弹性，对中小企业有利；系统的容错性差 。

Web2.0 到来时，出现大量非结构化数据让久的数据库模型使用不灵活

HFDS 不足：不适合访问低延迟数据；无法高效存储大量小文件

Spanner 特性：数据副本的配置的配置上可以在很细的粒度上动态控制；提供读和写操作的外部一致性

第七章习题

1、数据分析概念以及与数据处理的关系

数据分析可以分为广义的和狭义的，广义的数据分析包括狭义的数据分析和数据挖掘，指用适当的分析方法对数据进行分析，提取有用的信息和形成结论。

关系：两者是难以分割的，用户进行大量数据分析时也会进行大量数据处理

2、机器学习概念和数据挖掘关系

机器学习研究计算机怎么模拟人的行为进行学习，数据挖掘很多技术来自于机器学习。

3、数据挖掘和机器学习常见算法

分类

聚类

回归分析

关联规则

协同过滤

4、协同过滤的算法种类：

UserCF 算法

ItemCF 算法

ModelCF 算法

5、大数据处理分析技术的种类以及代表产品

批处理计算：Spark

流计算：Spark Streaming

图计算：Spark GraphX

查询分析计算：Hive

6、流计算概念以及处理流程

流计算平台实施获取来自不同数据源的海量数据，经过实时分析处理，获得有价值的信息

数据实时采集；数据实时计算；实时查询服务

7、通用的图计算软件：

第一种基于遍历算法的、实时的图数据库，如 Neo4j 等

第二种基于 BSP 模型实现并行图处理系统，如 Pregel

8、阐述 MapReduce 的工作流程

一个大的 MapReduce 作业，首先会被拆分成多个 Map 任务再多台机器上并行执行，当 map 结束后，会生成许多<key,value>形式的中间结果，分发到多个 reduce 任务、在多台机器上并行执行。reduce 任务会对中间结果进行汇总计算得到最后结果，输出到分布式文件系统。

9、阐述 mapreduce 不足

表达能力有限

磁盘 IO 开销大

延迟高

10、数据仓库 Hive 和传统数据库对比分析：

？？？？ 

11、Hive 的体系架构

用户接口模块：实现外部对 Hive 的访问

驱动模块：编译、优化、执行

元数据存储模块：保存表模式和其他系统数据

12、Spark 相对于 MapReduce 的优点

（spark 特点运行速度快；容易使用；通用性强；运行模式多样）

更加灵活；提供内存运算；基于 DAG 的任务调度执行机制

13、Spark 与 Hadoop 关系

Spark 只能解决计算问题，无法解决存储问题，仅取代了 hadoop 中的计算框架 mapreduce

14、Spark 的体系架构包括哪些组件

Spark Core：被简称为 Spark 包含最基础核心的功能，如内存计算，任务调度等，主要面向批数据处理

Spark SQL：用于结构化数据处理组件

Spark Streaming：一种流计算框架，支持高吞吐量、可容错处理的实时流数据处理

Structured Streaming：基于 Spark SQL 引擎构建的，可扩展且容错的流处理引擎

MLlib：提供了常用机器学习的实现

GraphX：用于图计算的 PI

15、Spark 的部署方式有哪几种

Standalone 模式

Spark on Mesos

Spark on YARN

Spark on Kubernetes

16、为什么推出 Spark SQL

他可以对内部恶化外部的数据源执行各种关系操作

其次，可以支持和大量的数据源和数据分析算法

17、Spark Streaming 的基本原理

将实时输入的数据流以时间片（秒级）为单位进行拆分，然后经 Spark 引擎以类似批处理的方式处理每个时间片数据

18、Structured Streaming 有哪几种处理模型

微批处理；持续处理（至少一次的特性）

19、Structured Streaming 和 Spark SQL、Spark Streaming 进行对比分析

Structured Streaming 与 Spark Streaming 处理数据流，区别前者用的数据抽象时候 DataFrame 后者用的是 DStream

前者可以使用 Spark SQL 的 DataFrame 来处理数据流，虽然 Spark SQL 也采用 DataFrame，但是他只能处理静态数据，Structured Streaming 可以处理结构化数据；

Structured Streaming 将另外两个的特点结合了起来

20、Spark MLlib 的功能以及它提供了哪些工具

提供了主要的机器学习算法

算法工具；特征化工具；流水线工具；持久性工具；实用工具

21、TensorFlowOnSpark 的 Spark 应用程序包括哪几个基本过程

预留；启动；训练/推理；关闭

22、画出 Storm 的集群架构并加以简要说明

Nimbus《=》多个 ZooKeeper《=》更多个 Supervisor

master-worker 的节点方式，master 节点运行名为 nimbus，分发任务给 worker 节点，运行名为 supervisor，zookeeper 在之间协调工作

P234

23、Storm 的工作流程

客户端提交 topology 到 storm 集群中

nimbus 分配给 supervisor 的任务写入 zookeeper

supervisor 从 zookeeper 中获取所分配的任务，并启动 worker 进程

worker 进程执行具体的任务。

24、Spark Streaming 和 Storm 简要对比

前者无法实现毫秒级响应，后者可以

25、为什么流计算场景比较适合 Flink

流处理架构需要具备低延迟、高吞吐、高性能的特性（storm 做不到高吞吐；spark streamin 牺牲了低延迟；structure streaming 牺牲了一致性）

26、Flink 的体系架构包含哪些组件

jobmanager；taskmanager

27、Beam 的设计目标

统一、规范分布式数据处理的需求

生成的分布式数据处理任务能够在各个引擎上执行

28、查询分析系统 Dremel 有哪些特点

大规模、稳定

是 mapreduce 交互式查询能力不足的补充

数据模型是嵌套的

数据是列式存储的

结合了 web 搜索和并行 DBMS 的技术

习题未涉及

数据挖掘是指从大量数据中通过算法搜索隐藏于其中的信息的过程

数据分析和数据挖掘关系：太多了懒得写

分类典型方法：决策树；朴素贝叶斯；支持向量机；人工神经网络

聚类：太多摆烂

回归分析：太多摆烂

流计算特性：高性能；海量式；实时性；分布式；易用性；可靠性；

数据采集系统三部分：Agent；Collector；Store

BSP 每个超步需要三个组件：局部计算；通信；栅栏同步

数据仓库 impala：。。。。。

资源管理框架 YARN 的好处：计算资源按需伸缩；不同负载应用混搭，集群利用率高；共享底层存储，避免数据跨集群迁移。

Spark 运行架构 ：集群管理器（Cluster Manager）；工作节点（Worker Node）；控制节点（Driver）；执行进程（Executor）

Spark 的数据抽象 RDD：过程：读入数据；转化操作；行动操作处理，输出到外部数据源

Storm 特点：整合性；简易的 API；可扩展性；容错性；可靠的消息处理；支持各种编程语言；快速部署；免费开源；

flink 四个特性：太长不写 p236

第八章习题

1、试述数据可视化概念

将大型数据集中的数据以图形、图像的形式表示，并利用数据分析和开工具发现其中未知信息的处理过程

2、可视化的重要应用

观测、跟踪数据；分析数据；辅助理解数据；增强数据吸引力

3、常见的图表类型及其应用场景

柱状图 二维 指定一个分析轴比较其中一维

折线图 二维 按照时间序列分析数据的变化趋势，适用于较大的数据集

饼图 二维 指定一个分析轴进行所占比例的比较，只适用于反映部分与整体的关系

散点图 二维或三维 有两个维度需要比较

气泡图 三维或四维 其中只有两维内恩狗精确辨识

雷达图 四维以上 数据点不超过 6 个

漏斗图 适用于业务流程规范、周期长、环节多的流程分析

树图 把要实现的目的与需要采取的手段或措施系统地展开

热力图 以特殊高亮的形式显示访客热衷的区域

关系图 3d 空间中表征各节点关系

词云 通过形成“关键词云层”或“关键词渲染”对网络文本中出现频率较高的关键词给予视觉上的突出

桑基图 特定类型的流程图，宽度对应流量大小

日历图 以日历为基本维度的、对单元格加以修饰的图表

4、可视化工具有哪些类型，各自的代表性产品有哪些：

入门级工具：excel

信息图表工具：Google Chart API；ECharts；D3；Tableau；大数据魔镜

地图工具：Google Fusion Tables ；Modest Maps；Leaflet

时间线工具：Timetoast；Xtimeline

高级分析工具：R；Python；Weka；Gephi

5、可视化案例

全球黑客活动

互联网地图

可视化互联网地图 app