**《大数据导论》课程期末复习资料**

《大数据导论》课程讲稿章节目录：

第1章 大数据概述

（1）大数据的概念

（2）大数据的特征

（3）大数据的数据类型

（4）大数据的技术

（5）大数据的应用

第2章 大数据采集与预处理

（1）大数据采集

（2）大数据预处理概述

（3）数据清洗

（4）数据集成

（5）数据变换

（6）数据规约

第3章 大数据存储

（1）大数据存储概述

（2）数据存储介质

（3）存储系统结构

（4）云存储概述

（5）云存储技术

（6）新型数据存储系统

（7）数据仓库

第4章 大数据计算平台

（1）云计算概述

（2）云计算平台

（3）MapReduce平台

（4）Hadoop平台

（5）Spark平台

第5章 大数据分析与挖掘

（1）大数据分析概述

（2）大数据分析的类型及架构

（3）大数据挖掘

（4）大数据关联分析

（5）大数据分类

（6）大数据聚类

（7）大数据分析工具

第6章 大数据可视化

（1）大数据可视化概述

（2）大数据可视化方法

（3）大数据可视化工具

第7章 社交大数据

（1）社交大数据

（2）国内社交网络大数据的应用

（3）国外社交网络大数据的应用

第8章 交通大数据

（1）交通大数据概述

（2）交通监测应用

（3）预测人类移动行为应用

第9章 医疗大数据

（1）医疗大数据简介

（2）临床决策分析应用

（3）医疗数据系统分析

第10章 大数据的挑战与发展趋势

（1）大数据发展面临的挑战

（2）大数据的发展趋势

一、客观部分：（单项选择、多项选择）

（一）、单项选择

1.以下不是NoSQL数据库的是（）

A.MongoDB

B.HBase

C.Cassandra

D.DB2

★考核知识点:NoSQL与NewSQL主流系统

参考讲稿章节：3.7

附1.1.1（考核知识点解释）：

目前市场上主要的NoSQL数据存储工具有：BigTable、Dynamo 、Hbase、MongoDB、CouchDB、Hypertable

还存在一些其他的开源的NoSQL数据库，Neo4j、Oracle Berkeley DB、Apache Cassandra等

另外，NewSQL数据库。例如：GoogleSpanner、VoltDB、RethinkDB、Clustrix、TokuDB和MemSQL等。

2以下不是目前主流开源分布式计算系统的是（）

A.Azure

B.Hadoop

C.Spark

D.Storm

★考核知识点：主流开源分布式计算系统

参见讲稿章节：4.2

附1.1.2：（考核知识点解释）

由于Google没有开源Google分布式计算模型的技术实现，所以其他互联网公司只能根据Google三篇技术论文中的相关原理，搭建自己的分布式计算系统。

Yahoo的工程师DougCutting和MikeCafarella在2005年合作开发了分布式计算系统Hadoop。后来，Hadoop被贡献给了Apache基金会，成为了Apache基金会的开源项目。Hadoop采用MapReduce分布式计算框架，并根据GFS开发了HDFS分布式文件系统，根据BigTable开发了HBase数据存储系统。尽管和Google内部使用的分布式计算系统原理相同，但是Hadoop在运算速度上依然达不到Google论文中的标准。不过，Hadoop的开源特性使其成为分布式计算系统的事实上的国际标准。Yahoo，Facebook，Amazon以及国内的百度、阿里巴巴等众多互联网公司都以Hadoop为基础搭建自己的分布式计算系统。

Spark也是Apache基金会的开源项目，它由加州大学伯克利分校的实验室开发，是另外一种重要的分布式计算系统。它在Hadoop的基础上进行了一些架构上的改良。

Storm是Twitter主推的分布式计算系统，它由BackType团队开发，是Apache基金会的孵化项目。它在Hadoop的基础上提供了实时运算的特性，可以实时地处理大数据流。

Hadoop，Spark和Storm是目前最重要的三大分布式计算系统，Hadoop常用于离线的、复杂的大数据处理，spark常用于离线的、快速的大数据处理，而storm常用于在线的、实时的大数据处理。

3.Apriori算法是一种（）算法

A.关联规则

B.聚类

C.分类

D.预测

★考核知识点：大数据挖掘算法

参见讲稿章节：5.3-5.7

附1.1.2：（考核知识点解释）

关联分析（Association analysis）是从有噪声的、模糊的、随机的海量数据中，挖掘出隐藏的、事先不知道、但是有潜在关联的信息或知识的过程，或称关联规则学习（Association rule learning）。

Apriori算法是一种最有影响的挖掘布尔关联规则频繁项集的算法，算法有两个关键步骤：一是发现所有的频繁项集；二是生成强关联规则。FP（Frequent Pattern）-growth算法基于Apriori构建，但采用了高级的数据结构减少扫描次数，大大加快了算法速度。

分类（Classification）任务是在给定数据基础上构建分类模型，根据分类模型确定目标对象属于哪个预定义的目标类别。常用的分类算法有：决策树、感知机、K近邻、朴素贝叶斯、贝叶斯网络、逻辑斯谛回归、支持向量机、遗传算法、人工神经网络等。

聚类分析（Cluster analysis）简称聚类（Clustering），是把数据对象划分成子集（类）的过程，每个子集称为一个簇（Cluster），同一个簇中的数据之间存在最大相似性，不同簇之间的数据间存在最大的差异性。K-MEANS（K-均值）算法是一种划分聚类方法，以k 为参数，将n 个对象分为k 个簇，以使簇（类）内具有较高的相似度，而簇间的相似度最低。

（二）、多项选择

1.大数据的特征包括（ ）

A.体量大（Volume）

B.多样性（Variety）

C.速度快（Velocity）

D.价值高（Value）

**★考核知识点:大数据的特征**

**参考讲稿章节：1.2**

附1.2.1（考核知识点解释）：

目前在描述大数据特征时，一般是按照国际数据公司IDC所提的“4V”模型来刻画，即体量大（Volume）、多样性（Variety）、速度快（Velocity）、价值高（Value）。

1). 体量大（Volume）：数据量大是大数据的基本属性。数据规模的大小是用计算机存储容量的单位来计算的，数量的单位从TB级别跃升到PB级别、EB级别，甚至ZB级别。

2). 多样性（Variety）：大数据除了体量大外，另一个最重要的特征就是数据类型的多样化。即数据存在形式包括结构化数据、半结构化数据和非结构化数据。

3) 速度快（Velocity）：大数据环境中速度快有两层含义：一是数据产生速度快； 二是要求数据分析处理速度快。

4) 价值高（Value）：大数据拥有大量有价值信息，通过提炼的信息，能够在更高的层面和视角，将在更大的范围帮助用户提高决策力，洞察未来创造出更大的价值和商机。

2. 按照数据结构分类，数据可分为（ ）

A.结构化数据

B.半结构化数据

C.非结构化数据

D.无结构数据

**★考核知识点:** **按照数据结构分，大数据的数据类型**

**参考讲稿章节：1.3**

附1.2.2（考核知识点解释）：

大数据不仅仅体现在数据量大，也体现在数据类型多。

按照数据结构分，数据可分为结构化数据、半结构化数据和非结构化数据。

在现有大数据的存储中，结构化数据仅有20%，其余80%则在存在于物联网、电子商务、社交网络等 领域的半结构化数据和非结构化数据。据统计，全球结构化数据增长速度约为32%，半结构化数据和非结构化数据的增速高达63%。

（1）结构化数据：结构化数据，通常存储在关系数据库中，并用二维表结构通过逻辑表达实现。

所有关系型数据库（如SQL Server、Oracle、MySQL、DB2等）中的数据全部为结构化数据。

生活中我们常见的结构化数据有企业计划系统（Enterprise Resource Planning，ERP）、医疗的医院信息系统（Hospital Information System，HIS）、校园一卡通核心数据库

（2）半结构化数据就是介于完全结构化数据和完全无结构化的数据之间的数据。

例如邮件、HTML、报表、具有定义模式的XML数据文件等。典型应用场景如邮件系统、档案系统、教学资源库等。半结构化数据的格式一般为纯文本数据，其数据格式较为规范，可以通过某种方式解析得到其中的每一项数据。最常见的半结构化数据是日志数据，采用XML、JSON等格式的数据

（3）非结构化数据是指非纯文本类数据，没有标准格式，无法直接解析出相应的值。

非结构化数据无处不在，常风的包括Web网页.即时消息或者时间数据（如微博、微信、Twitter等数据）、富文本文档（Rich Text Format , RTF）、富媒体文件（Rich Media）、实时多媒体数据（如各种视频，音频、图像文件）

3. 根据产生主体的不同，大数据可以分为（）

A.产量企业应用产生的数据

B.大量个人用户产生的数据

C.由巨量机器产生的数据

D.科研数据

**★考核知识点:** **根据产生主体分，大数据的数据类型**

**参考讲稿章节：1.3**

附1.2.3（考核知识点解释）：

数据可根据产生主体的不同分为三类：

（1）由少量企业应用而产生的数据。关系型数据库中的数据、数据仓库中的数据。

（2）大量个人用户产生的数据。社交媒体，如微博、博客、QQ、微信、Facebook、Twitter等产生的大量文字 、图片、视频、音频数据）、企业应用的相关评论数据、电子商务在线交易、供应商交易的日志数据。

（3）由巨量机器产生的数据。应用服务器日志（Web站点、游戏）、传感器数据（天气、水、智能电网)、图像和视频监控、RFID、二维码或者条形码扫描的数据。

4. 根据作用方式不同，大数据可以分为（）

A.交互数据

B.社交数据

C.交易数据

D.个人数据

**★考核知识点:** **根据作用方式的不同，大数据的数据类型分类**

**参考讲稿章节：1.3**

附1.2.4（考核知识点解释）：

数据还可根据作用方式的不同分为两类：

（1）交互数据：指相互作用的社交网络产生的数据，包括人为生成的社交媒体交互和机器设备交互生成的新型数据。

（2）交易数据：交易数据是指来自于电子商务和企业应用的数据。包括EPR（网络公关系统）、B2B（企业对企业）、B2C（企业对个人）、C2C（个人对个人）、O2O（线上线下）、团购等系统产生的数据。

这些数据存储在关系型数据库和数据仓库中，可以执行联机分析处理（OLAP)和联机事务处理（OLTP)。随着大数据的发展，此类数据的规模和复杂性一直在提高。

交互和交易这两类数据的有效融合是大数据发展的必然趋势，大数据应用要有效集成这两类数据，并在此基础上，实现对这些数据的处理和分析。

5. Google分布式计算模型不包括（）

A. GFS

B. BigTable

C. MapReduce

D.RDD

**★考核知识点：Google的分布式计算模型**

**参见讲稿章节：4.2、4.3**

**附1.2.5：（考核知识点解释）**

2003年到2004年间，Google发表了MapReduce、GFS（Google File System）和BigTable三篇技术论文 ，提出了一套全新的分布式计算理论。 MapReduce 是分布式计算框架。GFS是分布式文件系统。BigTable是基于Google File System的数据存储系统。这三大组件组成Google的分布式计算模型。

在Google云计算平台的技术架构中，除了少量负责特定管理功能的节点（如GFSmaster、分布式锁Chubby和Scheduler等），所有的节点都是同构的，即同时运行GFS chunkserver、BigTable Server和MapReduce Job等核心功能模块，与之相对应的则是数据存储、数据管理和编程模型三项关键技术。

6. 根据数据分析深度，可将数据分析分为（ ）

A. 关联性分析

B. 预测性分析

C. 规则性分析

D. 描述性分析

★考核知识点：根据数据分析深度，数据分析的类型

参见讲稿章节： 5.2

附1.2.6：（考核知识点解释）

根据数据分析深度，可将数据分析分为3个层次：描述性分析(Descriptive Analysis)，预测性分析（Predictive Analysis）和规则性分析（Prescriptive Analysis）。

1描述性分析基于历史数据来描述发生的事件。

例如，利用回归分析从数据集中发现简单的趋势，并借助可视化技术来更好地表示数据特征。

2预测性分析用于预测未来事件发生的概率和演化趋势。

例如，预测性模型使用对数回归和线性回归等统计技术发现数据趋势并预测未来的输出结果。

3规则性分析用于解决决策制定和提高分析效率。

例如，利用仿真来分析复杂系统以了解系统行为并发现问题，并通过优化技术在给定约束条件下给出最优解决方案。

7. 根据数据分析的实时性，可将数据分析分为（ ）

A. 实时数据分析

B. 预测性分析

C. 规则性分析

D. 离线数据分析

★考核知识点：按照数据分析的实时性，数据分析的类型

参见讲稿章节： 5.2

附1.2.7：（考核知识点解释）

按照数据分析的实时性，一般将数据分析分为实时数据分析和离线数据分析。

实时数据分析也称在线数据分析，能够实时处理用户的请求。

离线数据分析通过数据采集工具将日志数据导入专用分析平台进行分析，非实时处理数据。

二、主观部分：

（一）、名词解释

1. 流处理

★考核知识点: 数据处理

参考讲稿章节：1.5

附2.1.1（考核知识点解释）：

数据处理有两种范式，批处理和流处理。

批处理：“静止数据”转变为“正使用数据”，先存储后处理（Store-then-Process），先把信息存下来，稍后一次性地处理掉；对于批量数据，多采用批处理，批处理擅长全时智能，但速度慢，需要批处理加速。

流处理：“动态数据”转变为“正使用数据”，直接处理（Straight-through Process），任务来一件做一件，信息来一点处理一点，有的直接过滤掉，有的存起来。对于流数据，多采用流处理，获得实时智能，速度快。

2.磁盘阵列

★考核知识点: 磁盘阵列

参考讲稿章节：3.2

附2.1.2（考核知识点解释）：

磁盘阵列（Redundant Arrays of Independent Disks，RAID），全称为“冗余的独立磁盘阵列”。冗余是为了补救措施、保证可靠性而采取的一种方法，独立是指磁盘阵列不在主机内而是自成一个系统。磁盘阵列是由很多价格较便宜的磁盘，组合成一个容量巨大的磁盘组，利用个别磁盘提供数据所产生加成效果提升整个磁盘系统效能。RAID可以让很多磁盘驱动器同时传输数据，在逻辑上又是一个磁盘驱动器，故使用此技术可以达到单个磁盘几倍、几十倍甚至上百倍的速率。在很多RAID模式中都有较为完备的相互校验/恢复功能，大大提高了系统容错度和稳定性。

3.云存储

★考核知识点:云存储的定义

参考讲稿章节：3.4

附2.1.3（考核知识点解释）：

云存储是在云计算（cloud computing）概念上延伸和发展出来的一个新的概念，是一种新兴的网络存储技术。它是云计算的重要组成部分，也是云计算的重要应用之一。云存储是指通过集群应用、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。

4.NoSQL

★考核知识点:NoSQL

参考讲稿章节：3.7

附2.1.4（考核知识点解释）：

NoSQL(Not Only SQL)泛指非关系型、分布式和不提供ACID的数据库设计模式，它不是单纯地反对关系型数据库，而是强调键值（Key-Value）存储数据库和文档数据库的优点。

5.数据仓库

★考核知识点:数据仓库的定义

参考讲稿章节：3.8

附2.1.5（考核知识点解释）：

William H. Inmon在1992年出版Building the Data Warehouse一书，第一次给出了数据仓库的清晰定义和操作性极强的指导意见，真正拉开了数据仓库得到大规模应用的序幕。在该书中，将数据仓库定义为：“一个面向主题的（subject oriented）、集成的（integrate），相对稳定的（non-volatile）、反映历史变化（time variant）的数据集合，用于支持管理决策。

6.云计算

★考核知识点：云计算的定义

参见讲稿章节：4.1

附2.1.6：（考核知识点解释）

云计算（Cloud Computing） 是一种分布在大规模数据中心、能动态的提供各种服务器资源以满足科研、电子商务等领域需求的计算平台。同时，云计算是分布式计算、并行计算和网络计算的发展 ，是虚拟化、效用计算、 IaaS（基础设施即服务）、PaaS（平台即服务）、SaaS（软件即服务）等概念混合演进并跃升的结果。

简单的说，云计算是基于互联网相关服务的增加、使用和交付模式，通过互联网来提供一般为虚拟化的动态易扩展资源。

狭义云计算指IT基础设施的交付和使用模式；广义云计算指服务的交付和使用模式。两种云计算均通过网络以按需、易扩展的方式获得所需服务。这种服务可以是IT和软件、互联网相关，也可是其他服务。

云计算的核心思想，是将大量用网络连接的计算资源统一管理和调度，构成一个计算资源池，向用户按需服务。

提供资源的网络被称为“云”。“云”中的资源在使用者看来是可以无限扩展的，并且可以随时获取、按需使用、随时扩展、按使用付费。

7.RDD

★考核知识点：RDD弹性分布式数据集

参见讲稿章节： 4.5

附2.1.7：（考核知识点解释）

RDD弹性分布式数据集，简单来说，是一种自定义的可并行数据容器，可以存放任意类型的数据。弹性是指有容错的机制，若一个RDD分片去失，Spark可以根据粗粒度的日志数据更新记录的信息（Spark中称为“血统”）重构它：分布式指的是能对其进行并行的操作。除了这两点，它还能通过persist或者cache函数被缓存在内存里或磁盘中，共享给其他计算机，可以避免Hadoop那样存取带来的开销。

8.大数据分析

★考核知识点：大数据分析的概念

参见讲稿章节： 5.1

附2.1.8：（考核知识点解释）

大数据分析是大数据理念与方法的核心，是指对海量增长快速、内容真实、类型多样的数据进行分析，从中找出可以帮助决策的隐藏模式、未知的相关关系以及其他有用信息的过程。

9.数据挖掘

★考核知识点：数据挖掘的概念

参见讲稿章节： 5.3

附2.1.9：（考核知识点解释）

数据挖掘（Data Mining，DM）简单来说就是在大量的数据中提取或挖掘信息，通过仔细分析来揭示数据之间有意义的联系、趋势和模式。

10.关联分析

★考核知识点：关联分析的概念

参见讲稿章节： 5.4

附2.1.10：（考核知识点解释）

关联分析（Association analysis）是从有噪声的、模糊的、随机的海量数据中，挖掘出隐藏的、事先不知道、但是有潜在关联的信息或知识的过程，或称关联规则学习（Association rule learning）。

11.分类分析

★考核知识点：分类的概念

参见讲稿章节： 5.6

附2.1.11：（考核知识点解释）

分类（Classification）任务是在给定数据基础上构建分类模型，根据分类模型确定目标对象属于哪个预定义的目标类别。

构建分类模型：通过分析已知训练样本类别的数据集属性，通过训练建立相应分类模型，是监督学习(supervised learning）过程，数据集被称为训练数据集。

使用模型分类：评估模型的分类预测准确率，使用测试数据集进行评估；当准确率可以接受时，用分类模型对未知数据进行分类。

12.聚类分析

★考核知识点：聚类分析的概念

参见讲稿章节： 5.7

附2.1.12：（考核知识点解释）

聚类分析（Cluster analysis）简称聚类（Clustering），是把数据对象划分成子集（类）的过程，每个子集称为一个簇（Cluster），同一个簇中的数据之间存在最大相似性，不同簇之间的数据间存在最大的差异性。

（二）、简答

1.人类社会的数据产生方式经历了哪些阶段？简述各阶段的特点。

**★考核知识点:数据产生方式变革、大数据的数据来源**

**参见讲稿章节：**1.1

附2.2.1（考核知识点解释）：

人类历史上从未有哪个时代和今天一样产生如此海量的数据，人类社会的数据产生方式大致经历了3个阶段：运营式系统、用户原创内容阶段、感知式系统阶段。

（1）运营式系统：

数据库的出现使得数据管理的复杂度大大降低，实际中数据库大都为运营系统所采用，作为运营系统的数据管理子系统，如超市的销售记录系统、银行的交易记录系统、医院病人的医疗记录等。人类社会数据量第一次大的飞跃正是建立在运营式系统广泛使用数据库开始，这些数据规范、有秩序、强调数据的一致性，且这些数据的产生方式是被动的。

（2）用户原创内容阶段：

互联网的诞生促使人类社会数据量出现第二次大的飞跃，但真正的数据爆发产生于Web2.0时代，其重要标志就是用户原创内容。以博客、微博为代表的新型社交网络的出现和快速发展，使得用户产生数据的意愿更加强烈；新型移动设备出现，易携带、全天候接入网络的移动设备使得人员在网上发现自己意见的途径更为便捷

数据结构复杂，无秩序，不强调数据的一致性或只强调弱一致性，这些数据的产生方式 是主动的。

（3）感知式系统：

人类社会数据量第三次大的飞跃最终导致了大数据的产生，这次飞跃的根本原因在于感知式系统的广泛使用。微小带着处理功能的传感器设备广泛布置于社会的各个角落，通过这些设备对整个社会的运转进行监控，这些设备会源源不断地产生新数据，这些数据的产生方式是自动的，数据呈现多源异构、分布广泛、动态演化等。

简单来说，数据产生经历了被动、主动和自动三个阶段，这些被动、主动和自动的数据共同构成了大数据的数据来源。

2. 大数据处理的关键技术都有哪些？并做简要描述。

**★考核知识点:大数据处理的关键技术（处理流程）**

**参考讲稿章节：1.4**

附2.2.2（考核知识点解释）：

大数据处理的关键技术主要包括：数据采集和预处理、数据存储、数据计算处理、数据分析和挖掘、数据可视化展示等。

1).数据采集，又称数据获取，是大数据生命周期的第一个环节，通过RFID射频识别技术、传感器、交互型社交网络以及移动互联网等方式获得的各种类型的结构化、半结构化及非结构化的海量数据。

2).数据预处理是数据分析和挖掘的基础，是将接收数据进行抽取、清洗、转换、归约等并最终加载到数据存储的过程。

3).数据存储，需要将采集到的数据进行存储管理，建立相应的数据库。

4).数据计算处理。单台计算机必然无法完成海量的数据处理工作，需要分布式架构的计算平台。

5).数据分析与挖掘，是基于商业目的，有目的的进行收集、整理、加工和分析数据，提炼有价值信息的一个过程。

6).大数据可视化技术，可以提供更为清晰直观的数据表现形式，将错综复杂的数据和数据之间的关系，通过图片、映射关系或表格，以简单、友好、易用的图形化、智能化的形式呈现给用户，供其分析使用。

3. 简述网络大数据的一般采集过程。

★考核知识点：大数据采集

参见讲稿章节： 2.2

附2.2.3：（考核知识点解释）

大数据采集主要包括：系统日志采集、网络数据采集、数据库采集和其他数据采集四种。

网络数据采集常用的是通过网络爬虫或网站公开API等方式从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。

网络大数据的一般采集过程：先在URL队列中写入一个或多个目标链接作为爬虫爬取信息的起点；爬虫从URL队列中读取链接，并访问该网站；从该网站爬取内容；从网页内容中抽取出目标数据和所有URL链接；从数据库中读取已经抓取过内容的网页地址；过滤URL，将当前队列中的URL和已经抓取过的URL进行比较；如果该网页地址没有被抓取过，则将该地址（Spider URL ）写入数据库，并访问该网站；如果该地址已经被抓取过，则放弃对这个地址的抓取操作；获取该地址的网页内容，并抽取出所需属性的内容值；将抽取的网页内容写入数据库，并将抓取到的新链接加入URL队列。

4. 解释为什么要进行数据预处理。

★考核知识点: 影响数据质量因素

参考讲稿章节：2.3

附2.2.4（考核知识点解释）：

高质量的数据是能够满足应用需求的数据。数据质量涉及很多因素，包括准确性、完整性、一致性、时效性、可信性和可解释性。

1)不完整数据的出现可能有多种原因：重要的信息并非总是可以得到、用户输入时的遗漏、用户理解错误导致相关数据没有记录、设备故障导致的输入缺失、记录中不一致数据的删除、记录历史或被修改的数据被忽略、缺失的数据，特别是某些属性缺失值的元组。

2)不正确数据的出现原因有：收集数据的设备出现故障、人为或计算机内部错误在数据输入时出现、数据传输中的错误也可能出现、出于个人隐私考虑，用户故意向强制输入字段输入不正确的信息。

3)不一致数据，如命名约定或所用的数据代码不一致、输入字段（如日期）的格式不一致等。

4)时效性：数据更新不及时对数据质量产生负面影响。

5)可信性反映了有多少数据是用户信赖的。

6)可解释性反映了数据是否容易被理解。

以上因素影响数据质量，低质量的数据将导致低质量的挖掘效果，因此需要对数据进行预处理。

5. 试给出几种对数据缺失值的处理方法。

**★考核知识点:数据清洗**

**参考讲稿章节：2.4**

附2.2.5（考核知识点解释）：

对缺失值的处理一般是想方设法的把他补充上或者干脆弃之不用，一般的处理方法有以下几种：

（1）忽略元组。

通常当在缺少类标号时，通过这样的方法来填补缺失值。除非元组有多个属性缺少值，否则该方法的有效性不高，而且大量有价值的数据有可能被忽略。

（2）人工填写缺失值。

由于用户自己最了解关于自己的数据，因此，这个方法产生数据偏离的问题最小，但该方法十分费时，尤其是当数据集很大、存在很多缺失值时，靠人工填写的方法不具备实际的可操作性。

（3）使用一个全局常量填充缺失值。

该方法是将缺失的属性值用同一个常数进行替换，如"Unkown”。这种方法虽然简单，但可用性较差。由于此方法大量采用同一属性值，又可能会误导挖掘程序得出有偏差甚至错误的结论，因此，也要谨慎使用。

（4）使用属性的中心度量（如均值或中位数）填充缺失值：

均值和中位数从不同角度反映了数据的某些统计特征，例如，对于对称分布的数据而言，缺失的数据与均值的偏差期望是最小的，因此用均值补充缺失值可以在最大限度上控制人工添加的值对数据整体特征的影响。

（5）使用与给定元组属同一类的所有样本的属性均值或中位数：

例如，如果将顾客按信用风险分类，并假设顾客收入的数据分布是对称的，则将具有相同信用风险顾客的平均收入替代数据库列表中收入income列的缺失值；如果顾客收入的数据分布是倾斜的，则中位数是更好的选择。

（6）使用最可能的值填充缺失值。

可以用回归、使用贝叶斯形式化的基于推理的工具或决策树归纳确定。例如，利用数据集中其他客户顾客的属性，可以构造一棵决策树来预测家庭月总收入的缺失值。

6. 大数据预处理技术都有哪些？并做简要描述。

**★考核知识点:大数据预处理技术**

**参考讲稿章节：2.3-2.7**

附2.2.6（考核知识点解释）：

主流数据预处理技术：数据清洗、数据集成、数据变换、数据规约

1.数据清洗。数据清洗过程主要包括数据的缺省值处理、噪声数据处理、数据不一致处理。

2.数据集成。数据集成过程是将多个数据源中的数据合并存放到一个一致的数据存储（如数据仓库）中。其中数据源可以包含多个数据库、数据立方体或一般文件。

数据集成需要考虑诸多问题，如数据集成中对象匹配问题、冗余问题和数据值的冲突检测与处理问题。

3.数据变换。数据变换是把原始数据转化为适合于数据挖掘的数据形式。

数据转换主要包括光滑、聚集、数据泛化、数据规范化和新属性构造。

4.数据规约。数据归约得到数据集的简化表示，它小得多，但能够产生同样的（或几乎同样的）分析结果。数据归约是指在尽可能保持数据原貌的前提下，最大限度地精简数据量保持数据的原始状态。

有许多数据归约策略，包括数据聚集（例如建立数据立方体）、属性子集选择（例如通过相关分析去掉不相关的属性）、维度归约（例如使用诸如最小长度编码或小波等编码方案）和数据数值归约（例如使用聚类或参数模型等较小的表示“替换"数据）、数据离散化。

7. 简述数据集成过程应考虑哪些问题及如何解决。

**★考核知识点:数据集成**

**参考讲稿章节：2.5**

附2.2.7（考核知识点解释）：

数据集成需要考虑诸多问题，如数据集成中对象匹配问题、冗余问题和数据值的冲突检测与处理问题。

（1）对象匹配：模式集成和对象匹配涉及到实体识别问题。例如，如何才能确定一个数据库中的customer\_id和另一个数据库中的cust\_number指的是相同属性？在集成期间，当一个数据库的属性与另一个数据库的属性匹配时，必须特别注意数据的结构。这旨在确保源系统中的函数依赖和参照约束与目标系统中的匹配。

（2）冗余：一个属性如果能由另一个或一组属性导出，则这个属性可能是冗余的。有些冗余可以被相关分析检测到。对于数值属性，可以使用相关系数（Correlation Coefficient）和协方差（Covariance）来评估一个属性的值如何随另一个属性变化。

（3）数据值的冲突检测与处理：对于来自同一个世界的某一实体，在不同的数据库中可能有不同的属性值。例如：某一表示长度的属性在不同数据库中分别用“厘米”和“分米”表示。检测到这类数据值冲突后，可以根据需要修改某一数据库的属性值以使来自不同的数据库中但为同一实体的属性值统一起来。

8. 简述大数据面临存储的问题与挑战。

**★考核知识点:大数据存储面临的问题与挑战**

**参考讲稿章节：3.1**

附2.2.8（考核知识点解释）：

随着结构化数据和非结构化数据数量的不断增长，以及分析数据来源的多样化，之前的存储系统设计已经无法满足大数据应用的需求。

对于大数据的存储，存在以下问题和挑战：

1.容量问题：大数据通常可达到pb级的数据规模，因此大数据存储系统需要达到相应等级的扩展能力。

2.延迟问题：大数据应用还存在实时性的问题，很多大数据应用环境，如涉及网上交易或者金融类相关的应用，都需要较高的每秒进行读写操作的次数

3.安全问题：某些特殊行业的应用，例如金融数据、医疗信息以及政府情报等又都自己的安全标准和保密性需求

4.成本问题 对于使用大数据环境的企业，成本控制是关键问题

5.数据的积累，需要基于大数据的应用要求较长的数据保存时间，为了实现长期的数据保存，需要存储厂商开发出能持续进行数据一致性检测、备份和容灾等保证长期高可用性的技术

6.灵活性 大数据存储系统的基础设置规模庞大，保证存储系统的灵活性和扩展性是一大挑战。

为了应对大数据对存储系统的挑战，数据存储领域的工作者通过不懈努力提升数据存储系统的能力，主要提升有3个方面：提升系统的存储容量、提升系统的吞吐量、系统的容错性等。

9. 简述传统存储系统架构分类

**★考核知识点:存储系统架构**

**参考讲稿章节：3.3**

附2.2.9（考核知识点解释）：

经过多年发展，存储系统架构由早期的DAS（Direct-AttachedStorage，直连式存储）发展到NAS（Network-AttachedStorage，网络附加存储）和SAN（StorageAreaNetwork，存储区域网络），现在已经进入到云存储阶段。

1.直连式存储（Direct Attached Storage，DAS）是最早出现的最直接的扩展数据存储模式 ，即与普通的PC架构一样，存储设备与主机系统直接相连，挂接在服务器内部总线上。

2. 网络附加存储（Network Attached Storage，NAS ）是一种采用直接与网络介质相连的特殊设备实现数据存储的模式。

3. 存储区域网络（Storage Area Network，SAN）指存储设备相互连接并与服务器群相连而成网络，创造了存储的网络化。通过专用高速网将一个或多个网络存储设备和服务器连接起来的专用存储系统，数据处理服务器上的操作系统可以像访问本地盘数据一样对这些存储设备进行高速访问。

10.云存储技术具有哪些特点？并加以解释

**★考核知识点:云存储的特点**

**参考讲稿章节：3.4**

附2.2.10（考核知识点解释）：

云存储技术具有以下特点

（1)可靠性。云存储通过增加冗余度提高存储的可靠性。但是增加可靠性受到可靠性原理、成本及性能等方面的制约，因此在在保证可靠性的同时，提高系统的整体运行效率是当前一个亟待解决问题。

（2）可用性。企业需要全天候地为世界不同地区的用户提供服务支持，因此可用性至关重要。对于云存储平台，冗余的架构部分可以减少停机风险。同时，多路径、控制器、不同的光纤网、RAID技术、端到端的架构控制/监控和成熟的变更管理过程等方案均可提高云存储可用性。

（3）安全性。云存储服务间传输以及保存的数据都有被截取或篡改的隐患，因此当服务通过云交付时，数据分片混淆存储和数据加密传输成为了实现用户数据私密性和保证安全性的重要手段。

（4）动态伸缩性。指的是读/写性能和存储容易的扩展与缩减。一个设计优良的云存储系统可以在系统运行过程中简单地通过添加或移除节点来自由扩展和缩减，这些操作对用户来说是透明的。

（5）低成本。云存储可以降低企业级存储成本，包括购置存储的成本、驱动存储的成本、修复存储的成本及管理存储的成本。

11.云存储架构分哪些层次，各层实现了什么功能？

**★考核知识点:云存储架构**

**参考讲稿章节：3.4**

附2.2.11（考核知识点解释）：

云存储是一个由网络设备、存储设备、服务器、应用软件、公用访问接口、接入网和客户端程序等组成的复杂系统。以存储设备为核心，通过应用软件来对外提供数据存储和业务访问服务。

云存储的架构由上而下可以分为访问层、应用接口层、基础管理层和存储层。

1）存储层：云存储的最基础部分。云存储中的存储设备通常分布在不同地域且数量非常庞大，通过互联网、广域网或FC光纤通道网络把各个存储设备连接在一起。统一存储设备管理系统在存储设备的上一层，它能够完成多链路冗余管理，存储设备的逻辑虚拟化管理以及硬件设备的状态监控与故障维护。

2）基础管理层：云存储最核心最难以实现的部分，基础管理层的主要功能是使云存储中多个存储设备之间可以协同工作，以便对外提供同一种服务，能够提供更大、更好、更强的数据访问性能，它所采用的技术主要有集群系统、分布式文件系统和网格计算等。为了保证云存储中的数据不会被未授权的用户所访问，它还提供了CDN内容分发系统以及数据加密技术。同时，为了确保云存储中的数据不丢失以及云存储自身的安全和稳定，它还采取了各种数据备份、数据容灾技术和措施。

3）应用接口层：云存储中灵活性最好的部分，根据实际业务类型的不同，不同的云存储运营单位开发的应用服务接口及提供的应用服务也不一样。例如在线音乐播放应用平台、网络硬盘应用平台、IPTV和视频点播应用平台、远程教学应用平台等。

4）访问层：用户获得云存储系统的授权后，就可以通过标准的公用应用接口进行登录并享受云存储服务。云存储提供的访问类型和访问手段会根据云存储运营单位的不同而有所不同。

12. 存储虚拟化技术有哪几个实施层次，分别叙述这几个层次的特点。

**★考核知识点:存储虚拟化实施层次**

**参考讲稿章节：3.5**

附2.2.12（考核知识点解释）：

根据不同的虚拟化实现位置，虚拟化还可以分为基于主机虚拟化、基于存储设备虚拟化和基于存储网络虚拟化。

1）基于主机虚拟化。基于主机的虚拟化存储的核心技术是，通过增加一个运行在操作系统下的逻辑卷管理软件将磁盘上的物理块号映射成逻辑卷号，从而把多个物理磁盘阵列映射成一个统一的虚拟逻辑块，来进行存储虚拟化的控制和管理。

2）基于存储设备虚拟化。基于存储设备虚拟化技术依赖于提供相关功能的存储设备的阵列控制器模块，常见于高端存储设备，其主要应用针对异构的SAN存储构架。

3）基于存储网络虚拟化。基于存储网络虚拟化技术的核心是在存储区域网中增加虚拟化引擎实现存储资源的集中管理，其具体实施一般是通过具有虚拟化支持能力的路由器或交换机实现的。

13. 云存储关键技术

**★考核知识点:云存储技术**

**参考讲稿章节：3.5-3.6**

附2.2.13（考核知识点解释）：

云存储技术是多种技术的集合体，这些技术涉及硬件、软件和网络等计算机技术的各个方面，具有高可用性、高可靠性、高安全性和低成本等特征。

1.存储虚拟化。存储虚拟化可以将系统中不同厂商、不同型号、不同通信技术、不同类型的存储设备映射为一个统一的存储资源池，屏蔽了存储实体之间的物理位置及异构特征，从而对这些存储资源进行统一分配管理。

在虚拟化存储环境中，服务器及应用系统面对的都是物理设备的逻辑映像，且不会随物理设备的改变而变化，实现了资源对系统管理员的透明性，在降低构建存储系统成本的同时使管理和维护资源变得容易。

云存储的虚拟化将存储资源虚拟化为全局命名空间，并通过多租户技术给使用者提供存储资源，在此过程中，数据可以在存储资源池中跨节点、跨数据中心流动。

根据不同的虚拟化实现位置，虚拟化还可以分为基于主机虚拟化、基于存储设备虚拟化和基于存储网络虚拟化。

2.分布式存储。

（1）分布式块存储。块存储就是服务器直接通过读写存储空间中的一个或一段地址来存取数据。

（2）分布式对象存储。对象存储是为海量数据提供Key-Value这种通过键值查找数据文件的存储模式；对象存储引入对象元数据来描述对象特征，对象元数据具有丰富的语义；引入容器概念作为存储对象的集合。

对象存储系统底层基于分布式存储系统来实现数据的存取，其存储方式对外部应用透明。这样的存储系统架构具有高可扩展性，支持数据的并发读写，一般不支持数据的随机写操作。

（3）分布式文件系统。文件存储系统可提供通用的文件访问接口，如POSIX、NFS、CIFS、FTP等，实现文件与目录操作、文件访问、文件访问控制等功能。

目前的分布式文件系统存储的实现有软硬件一体和软硬件分离两种方式。主要通过NAS虚拟化，或者基于x86硬件集群和分布式文件系统集成在一起，以实现海量非结构化数据处理能力。

3.数据缩减

云存储中的数据缩减技术，满足了海量信息爆炸式增长趋势，一定程度上节约企业存储成本，提高效率。

（1）自动精简配置。自动精简配置技术是利用虚拟化方法减少物理存储空间的分配，最大限度提升存储空间利用率。

（2）自动存储分层。自动存储分层(AutomatedStorageTier'AST)技术主要用来帮助数据中心最大限度地降低成本和复杂性。

（3）重复数据删除。“重复删除”技术（De-duplication）作为一种数据缩减技术可对存储容量进行优化。它通过删除数据集中重复的数据，只保留其中一份，从而消除冗余数据。

4.负载均衡。庞大的数据量必然会用来支持海量的请求，云存储一个典型特点就是实现这些请求在系统内部的负载均衡。在传统的负载均衡中，处于网络边缘的设备将来自不同地址的请求均匀地、最优化地发送到各个承载设备上。而在云存储中，除了在网络边缘实现DNS动态均匀解析的负载均衡设备，还有在系统内部的负载均衡机制，即在节点资源之间的负载均衡。

14. NoSQL普遍采用的技术

★考核知识点:NoSQL普遍采用的技术

参考讲稿章节：3.7

附2.2.14（考核知识点解释）：

NoSQL系统普遍采用的一些技术有以下几种。

（1）简单数据模型。不同于分布式数据库，大多数NoSQL系统采用更加简单的数据模型.

（2）元数据和应用数据的分离。

NoSQL数据管理系统需要维护两种数据：元数据和应用数据。

元数据是用于系统管理的，如数据分区到集群中节点和副本的映射数据。

应用数据就是用户存储在系统中的商业数据。

（3）弱一致性。NoSQL系统通过复制应用数据来达到一致性。

这种设计使得更新数据时副本同步的开销很大，为了减少这种同步开销，弱一致性模型如最终一致性和时间轴一致性得到广泛应用。

15.请分析相比于关系型数据库，NoSQL数据存储管理系统的主要优势与劣势。

★考核知识点：新兴数据存储系统

参见讲稿章节： 3.7

附2.2.15：（考核知识点解释）

相对于关系型数据库，NoSQL数据存储管理系统的主要优势有以下几方面。

（1）避免不必要的复杂性。关系型数据库提供各种各样的特性和强一致性，但是许多特性只能在某些特定的应用中使用，大部分功能很少被使用。NoSQL系统则提供较少的功能来提高性能。

（2）高吞吐量。一些NoSQL数据系统的吞吐量比传统关系数据管理系统要高很多，如Google使用MapReduce每天可处理20PB存储在Bigtable中的数据。

（3）高水平扩展能力和低端硬件集群。NoSQL数据系统能够很好地进行水平扩展，与关系型数据库集群方法不同，这种扩展不需要很大的代价。而基于低端硬件的设计理念为采用NoSQL数据系统的用户节省了很多硬件上的开销。

（4）避免了昂贵的对象-关系映射。许多NoSQL系统能够存储数据对象，这就避免了数据库中关系模型和程序中对象模型相互转化的代价。

虽然NoSQL数据库提供了高扩展性和灵活性，但是它也有自己的缺点，主要有以下几方面．

（1)数据模型和查询语言未经数学验证。SQL这种基于关系代数和关系演算的查询结构有着坚实的数学保证，即使一个结构化的查询本身很复杂，但是它能够获取满足条件的所有数据．由于NoSQL系统都没有使用SQL,而使用SQL的一些模型还未有完善的数学基础。这也是NoSQL系统较为混乱的主要原因之一

（2）不支持ACID特性。这为NoSQL带来优势的同时也是其缺点，毕竟事务在很多场合下还是需要的，ACID特性使系统在中断的情况下也能够保证在线事务能够准确执行。

（3）功能简单。大多数NoSQL系统提供的功能都比较简单，这就增加了应用层的负担，例如，如果在应用层实现ACID特性，那么编与代码的程序员一定非常痛苦。

（4）没有统一的查询模型。NOSQL系统一般提供不同查询模型，这一定限度上增加了开发者的负担。

16.数据仓库有哪些特点？并做简要描述。

★考核知识点:数据仓库的特点

参考讲稿章节：3.8

附2.2.16（考核知识点解释）：

（1）数据仓库中的数据是面向主题组织的。

数据仓库是按照面向主题的方式进行数据组织的，也就是在较高层次上对分析对象的数据作个完整、一致的描述，能有效地刻画出分析对象所涉及的各项数据及数据间的联系。这种数据组织方式更能适合较高层次的数据分析，便于发现数据中蕴含的模式和规律。

主题通常是在一个较高层次上将数据归类的标准，每个主题对应一个宏观分析领域。

（2）数据仓库的数据是集成的。

数据仓库中每一主题对应的源数据在原有的各分散数据库中可能是重复出现的、不一致的，数据仓库中的数据不能从原有的数据库系统中直接得到。

事务处理系统中的操作型数据在进入数据仓库之前，必须经过统一和综合，演变为分析型数据。

（3）数据仓库的数据是稳定的。

数据仓库中存放的是供分析决策用的历史数据，而不是联机事务处理的当前数据，涉及的数据操作主要是数据查询，一般不进行数据的增、删、改操作，业务系统中的数据经集成进入数据仓库之后极少或根本不再更新。

如果对数据仓库中的数据进行了修改，就失去了统计分析正确性的基础一一数据的真实性。

（4）数据仓库的数据是随时间不断变化的。

数据仓库中的数据不是永远不变的。数据仓库数据是随时间变化的，数据仓库系统需要不断获取联机事务处理系统不同时刻的数据，经集成后追加到数据仓库中，因此数据仓库中数据的码（键）都包含时间项，以表明数据的历史时期，并可在时间维度上对数据进行分析。

此外，数据仓库中的数据也有时间期限，在新数据不断进入的同时，过时的数据也要从数据仓库中排除出去。

17. 简述数据仓库与数据库的区别。

★考核知识点:数据仓库

参考讲稿章节：3.8

附2.2.17：（考核知识点解释）

数据仓库是在数据库的基础上发展起来的，数据仓库把数据从各个信息源中提取出来后，

依照数据仓库使用的公共数据模型，进行相应变换后与仓库中现有数据集成在一起。在数据

仓库中，数据可以被直接访问，查询和分析处理速度很快。数据仓库的特点决定了它与传统

的数据库系统之间必然存在很大的差异。二者之间的区别主要体现在以下几个方面。

（1）数据库中存储的都是当前使用的值，而数据仓库中的数据都是一些历史的、存档的、归纳的、计算的数据。

（2）数据库的数据主要是面向业务操作程序的，可以重复处理，主要是用来进行事务处理的。而数据仓库却是面向主题，主要是用来分析应用的。

（3）数据库的数据结构是高度结构化的，比较复杂，适用于操作计算。而数据仓库的数据却比较简单，适用于分析处理。

（4）数据库中的数据使用频率是很高的。数据仓库中的数据则不是很高。

（5）通常对数据库中事务的访问，只需要访问少量的记录数据。而对数据仓库中事务的

访问就可能需要访问大量的记录。

（6）对数据的响应时间一般要求比较高，通常是以秒为单位。而对数据仓库的响应时间要求则较低，通常比较长。

18.云计算有哪些特点？并做简要描述。

★考核知识点：云计算的特点

参见讲稿章节：4.1

附2.2.18：（考核知识点解释）

云计算的特点如下：

（1）超大规模。“云”具有相当大的规模，Google云计算己经拥有上百万台服务器；Amazon、IBM、Microsoft、Yahoo等公司的“云”均拥有几十万台服务器；一般企业私有云则可拥有数百上千台服务器。“云”能赋予用户前所未有的计算能力。

（2）高可靠性。分布式数据中心可将云端的用户信息备份到地理上相互隔离的数据库主机中，甚至连用户自己也无法判断信息的确切备份地点。该特点不仅提供了数据恢复的依据，也使得网络病毒和网络黑客的攻击因为失去目的性而变成徒劳，大大提高系统的安全性和容灾能力。

（3）虚拟化。云计算支持用户在任意位置、使用各种终端获取应用服务。所请求的资源来自“云”，而非固定的有形的实体。应用在“云”中某处运行，但用户无需了解，也不用担心应用运行的具体位置。

（4）高扩展性。目前主流的云计算平台均根据SPI架构，构建在各层集成功能各异的软硬件设备和中间件软件。大量中间件软件和设备提供针对该平台的通用接口，允许用户添加本层的扩展设备。部分云与云之间提供对应接口，允许用户在不同云之间进行数据迁移。类似功能更大程度上满足了用户需求，集成了计算资源，是未来云计算的发展方向之一

（5）按需服务。“云”是一个庞大的资源池，可以像自来水、电、煤气那样计费，并按需购买。

（6）极其廉价。“云”的特殊容错措施可以采用极其廉价的节点来构成云。“云”的自动化集中式管理，使大量企业无需负担日益高昂的数据中心管理成本，“云”的通用性使资源的利用率较之传统系统大幅提升，因此用户可以充分享受“云”的低成本优势。

19. 简述云计算的体系架构分层及每层的含义。

★考核知识点：云计算体系架构

参见讲稿章节：4.1

附2.2.19：（考核知识点解释）

云计算可以按需提供弹性资源，它的表现形式是一系列服务的集合。

结合当前云计算的应用与研究，其体系架构可分为核心服务、服务管理、用户访问接口三层。

1）核心服务层将硬件基础设施、软件运行环境、应用程序抽象成服务，这些服务具有可靠性强、可用性高、规模可伸缩等特点，满足多样化的应用需求。

2）服务管理层为核心服务提供支持，进一步确保核心服务的可靠性、可用性与安全性。

3）用户访问接口层实现端到云的访问。

20. 简述云计算的核心服务模型。

★考核知识点：云计算服务模型

参见讲稿章节：4.1

附2.2.20：（考核知识点解释）

IaaS、PaaS、SaaS 是云计算的三种服务模型。

基础设施即服务（IaaS）：消费者通过Internet可以从完善的计算机基础设施获得服务。

平台即服务（PaaS）：PaaS实际上是指将软件研发的平台作为一种服务，以SaaS的模式提交给用户。

软件即服务（SaaS）：软件即服务。它是一种通过Internet提供软件的模式，用户无需购买软件，而是向提供商租用基于Web的软件，来管理企业经营活动。

21. 试对MapReduce编程模型原理进行描述。

★考核知识点：MapReduce编程模型原理

参见讲稿章节： 4.3

附2.2.21：（考核知识点解释）

MapReduce编程模型结合用户实现的Map和Reduce函数，可完成大规模地并行化计算。MapReduce编程模型的原理是：用户自定义的Map函数处理一个输入的基于key-value pair的集合，输出中间基于key-value pair的集合，MapReduce库把中间所有具有相同key值的value值集合在一起后传递给Reduce函数，用户自定义的Reduce函数合并所有具有相同key值的value值，形成一个较小value值的集合。

22. 请写出MapReduce程序的执行过程。

★考核知识点：MapReduce执行过程

参见讲稿章节： 4.3

附2.2.22：（考核知识点解释）

MapReduce执行过程主要包括以下几方面。

（1）将输入的海量数据切片分给不同的机器处理。

（2）执行Map任务的Worker将输入数据解析成key-value pair，用户定义的Map函数把

输入的key-value pair转成中间形式的key-value pair。

（3）按照key值对中间形式的key-value进行排序、聚合。

（4）把不同的key值和相应的value集分配给不同的机器，完成Reduce运算。

（5）输出Reduce结果。

23. 简述HDFS集群的构成及各部分的功能。

★考核知识点：HDFS集群的构成

参见讲稿章节： 4.4

附2.2.23：（考核知识点解释）

一个HDFS集群由一个名字节点(Name Node)和若干个数据节点(Data Node)构成。此外，还有与这两个角色之间作为沟通桥梁的客户端(Client)。

名字节点是一个中心服务器，负责管理文件系统的命名空间(Namespace)以及客户端对文件的访问。

集群中的数据节点，一般是一个节点上有一个，负责管理它所在节点上的存储。HDFS对外公开文件系统的名字空间，用户能够以文件的形式在上面存储数据。

从内部看，一个文件其实被分成一个或多个数据块（Block），这些块存储在一组数据节点上。名字节点执行文件系统的名字空间操作，例如，打开、关闭、重命名文件或目录，它也负责确定数据块到具体数据节点的映射。数据节点负责处理文件系统客户端的读写请求，在名字节点的统一调度下进行数据块的创建、删除和复制。

Client代表用户通过与NameNode和DataNode来交互访问整个文件系统。因为Client提供了一些文件系统接口，所以在编写程序的时候，不用知道DataNode和NameNode的内部详细情况，就可以编写程序进行操作，实现所需的功能。

24.简述Spark对比 Hadoop的优势

★考核知识点：Spark平台优势

参见讲稿章节： 4.5

附2.2.24：（考核知识点解释）

Spark作为现今最流行的分布式云平台技术，对比Hadoop云平台技术来说，可以总结出以下优势。

（1）内存管理中间结果。MapReduce作为Hadoop的核心编程模型，将处理后的中间结果输出并存储到磁盘上，依赖HDFS文件系统存储每一个输出的结果。spark运用内存缓存输出的中间结果，便于提高中间结果再度使用的读取效率。

（2）优化数据格式。Spark使用弹性分布式数据集(RDD)，这是一种分布式内存存储结构，支持读写任意内存位置，运行时可以根据数据存放位置进行任务的调度，提高任务调度焦虑，支持数据批量转换和创建相应的RDD。

（3）优化执行策略。Spark支持基于哈希函数的分布式聚合，不需要针对Shuffle进行全量任务的排序，调度时使用DAG（有向无环图），能够在一定程度上减少MapReduce在任务排序上花费的大量时间，成为一个优化的创新点。

（4）提高任务调度速率。Spark启动任务采用事件驱动模式，尽量复用线程，减少线程启动和切换的时间开销。Hadoop是以处理庞大数据为目的设计的，在处理略为小规模的数据会出现任务调度上时间开销的增加。

(5）通用性强。spark支持多语言（Scala，Java，Python）编程，支持多种数据形式（流式计算、机器学习、图计算）的计算处理，通用性强且一定程度上方便研究人员对平台代码的复用和重写。

25. 简述大数据分析流程。

★考核知识点：大数据分析流程

参见讲稿章节： 5.1

附2.2.25：（考核知识点解释）

1）发现和问题定义。学习业务领域知识,将业务问题转化为分析挑战。

2）识别和设计数据需求。管理者根据决策和过程控制需求，提出对数据需求。

3）收集和预处理数据。有目的收集数据，数据预处理。

4）分析数据及建立模型。将收集的数据通过加工、整理和分析将其转化为信息。

5）沟通结果及过程改进。

26. 简述Apriori算法的核心思想。

★考核知识点：大数据关联分析

参见讲稿章节： 5.7

附2.2.26：（考核知识点解释）

Apriori算法是一种最有影响的挖掘布尔关联规则频繁项集的算法，算法有两个关键步骤：一是发现所有的频繁项集；二是生成强关联规则。

Apriori算法的核心思想如下：对于给定的一个数据库和最小支持度阈值，首先对其进行扫描，找出所有的频繁1-项集，该集合记作L1；然后得用L1找频繁2-项集的集合L2，L2找L3，如此下去，直到不能再找到任何频繁k-项集。最后在所有的频繁集中提取出强规则，即产生用户感兴趣的关联规则。

27.简述数据可视化的流程和步骤。

★考核知识点：大数据可视化概述

参见讲稿章节： 6.1

附2.2.27：（考核知识点解释）

数据可视化是对数据的综合运用，其操作包括数据获取、数据处理、可视化模式和可视化应用4个步骤。

1）数据获取

数据获取的形式多种多样，大致可以分为主动式和被动式两种。主动式获取是以明确的数据需求为目的，利用相关技术手段主动采集相关数据，如卫星影像、测绘工程等；被动式获取是以数据平台为基础，由数据平台的活动者提供数据来源，如电子商务网站、网络论坛等。

2）数据处理

数据处理是指对原始的数据进行分析、预处理和计算等步骤。数据处理的目标是保证数据的准确性、可用性等。

3）可视化模式

可视化模式是数据的一种特殊展现形式，常见的可视化模式有标签云、序列分析、网络结构、电子地图等。可视化模式的选取决定了可视化方案的雏形。

4）可视化应用

可视化应用主要根据用户的主观需求展开，最主要的应用方式是用来观察和展示，通过观察和人脑分析进行推理和认知，辅助人们发现新知识或者得到新结论。可视化界面也可以帮助人们进行人与数据的交互，辅助人们完成对数据的迭代计算，通过若干步，数据的计算实验，生产系列化的可视化成果。

28.请分析大数据未来的发展趋势。

★考核知识点：大数据的发展趋势

参见讲稿章节： 10.2

附2.2.28：（考核知识点解释）

1. 大数据从概念化走向价值化

2. 大数据安全与隐私越来越重要

3. 大数据分析与可视化成为热点

4. 数据的商品化和数据共享的联盟化

5. 深度学习与大数据性能成为支撑性的技术

6. 数据科学的兴起

7. 大数据产业成为一种战略性产业

8. 大数据生态环境逐步完善

9. 大数据处理架构的多样化模式并存