**《大数据导论》课程期末复习资料**

1.

**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。**

Spark也是Apache基金会的开源项目，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。

2.

关联分析海量数据挖掘出隐藏的关联的信息或知识的过程。

Apriori算法：一是发现所有的频繁项集；二是生成强关联规则。

FP-growth算法基于Apriori构建，高级的数据结构，大大加快了算法速度。

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。

K-MEANS（K-均值）算法是一种划分聚类方法。

3.

**★大数据的特征：**数据规模**体量大（Volume）、**数据类型**多样性（Variety）、产生处理速度快（Velocity）、价值高（Value）。**

按照**数据结构**分，分为结构化数据、半结构化数据和非结构化数据。

（1）结构化数据：存在关系数据库中、二维表。

常见：企业计划系统、医院信息系统、校园一卡通

（2）半结构化数据：介于二者，一般为纯文本数据，数据格式较为规范，可以通过某种方式解析得到数据。

常见：邮件、报表、部分XML文件等。典型应用场景如邮件系统、档案系统、教学资源库等。

（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。

常见：即时消息或者时间数据（如微博、微信、Twitter等数据）、web网页、实时多媒体数据（如各种视频，音频、图像文件）

根据**产生主体**分，大数据的数据类型：

（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、QQ、微信；企业应用的数据（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。

根据**作用方式**的不同，大数据的数据类型分类

（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。

4.

数据分析分为3个层次：

1描述性分析基于历史数据来描述发生的事件。回归分析

2预测性分析用于预测未来事件发生的概率和演化趋势。对数回归和线性回归

3规则性分析用于解决决策制定和提高分析效率。利用仿真来分析复杂系统

**5.**

批处理：**“静止数据”转变为“正使用数据”**，先存储后处理，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，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。

6.

磁盘阵列（RAID），全称为“冗余的独立磁盘阵列”。由很多价格较便宜的磁盘，组合成一个容量巨大的磁盘组，效能高、有较为完备的相互校验/恢复功能。

7.

云计算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。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。

**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。**

**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。

8.

**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。**

10.

RDD弹性分布式数据集，弹性是指有容错的机制，可以重构它：分布式指的是能对其进行并行的操作。

11.数据挖掘、关联分析

大量模糊随机、提取、挖掘、隐藏的潜在的之间意义、联系、趋势和模式。

12.聚类分析

划分成子类的过程、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。

**13.人类社会的数据产生方式经历了哪些阶段？**

（1）运营式系统：开始广泛使用数据库、运营系统管理（超市销售记录系统、银行交易记录系统、医院医疗记录等）强调数据的一致性、秩序，产生方式被动。

（2）用户原创内容阶段：互联网、博客、微博快速发展、移动设备、数据结构复杂无序，产生方式主动。

（3）感知式系统：感知式系统广泛使用、传感器设备广泛布置、监控社会信息，自动产生，数据呈现多源异构、分布广泛、动态演化等。

**14.大数据处理的关键技术都有哪些？**

1)数据采集： RFID、传感器、移动互联网等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：抽取、清洗、转换、归约。

3)数据存储：管理、建立相应的数据库。

4).数据计算处理。分布式架构的计算平台。

5).数据分析与挖掘，有目的、收集、整理、加工和分析、提炼有价值信息。

6).大数据可视化技术，图片或表格智能化的形式呈现给用户。

**15. 简述网络大数据的一般采集过程。**

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。

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。非结构化抽取存储结构化。

采集过程：URL爬虫起点-读取链接访问-爬取-抽取目标数据和所有URL链接-和已经抓取过的URL进行比较-写入新url-访问

16. 解释为什么要进行数据预处理。

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。包括**准确性、完整性、一致性、时效性、可信性和可解释性。**

17. 试给出几种对数据缺失值的处理方法。

（1）忽略。（2）人工填写。（3）常量填充。（4）使用总体均值或中位数填充（5）使用同类的所有样本的属性均值或中位数：（6）推理工具填充。

18. 大数据预处理技术都有哪些？并做简要描述。

1.数据清洗。缺省值处理、噪声、一致。

2.数据集成。多个数据源合并

3.数据变换。把原始数据转化为适合于数据挖掘的数据形式

4.数据规约。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。

19. 简述数据集成过程应考虑哪些问题及如何解决。

（1）对象匹配

（2）冗余

（3）数据值的冲突检测与处理

20. 简述大数据面临存储的问题与挑战。

容量、延迟、安全、成本、长期的数据保存、灵活扩展

21. 简述传统存储系统架构分类

1.直连式存储（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（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（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。

22.云存储架构分哪些层次，各层实现了什么功能？

1）存储层

2）基础管理层：CDN内容分发系统以及数据加密技术

3）应用接口层

4）访问层

23. 云存储关键技术

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。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。

3.数据缩减

（1）自动精简配置（2）自动存储分层（3）重复数据删除

4.负载均衡。

**24. NoSQL普遍采用的技术**

（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。

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。

**优势**

（1）高吞吐量。

（2）高水平扩展能力和低端硬件集群。

**缺点**

（1) 没有统一的查询模型。数据模型和查询语言未经数学验证。

（2）不支持ACID特性。

（3）功能简单。

**25.数据仓库有哪些特点？并做简要描述。**

（1）**面向主题组织的。**每个主题对应一个宏观分析领域。

对分析对象的数据作个完整、一致的描述，有效刻画涉及数据联系。适合较高层次的数据分析

（2）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。

**26. 简述数据仓库与数据库的区别。**

（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）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。

**27.云计算有哪些特点？并做简要描述。**

（1）超大规模。

（2）高可靠性。地理上相互隔离

（3）虚拟化。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”

（4）高扩展性。大量中间件软件和设备提供针对该平台的通用接口

（5）按需服务。

（6）极其廉价。自动化集中式管理

**28. 简述云计算的核心服务模型。**

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。

**29. 试对MapReduce编程模型原理进行描述。**

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。

**30. 请写出MapReduce程序的执行过程。**

（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。

**31. 简述HDFS集群的构成及各部分的功能。**

一个HDFS集群由一个名字节点和若干个数据节点构成。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。

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。

**32.简述Spark对比 Hadoop的优势**

（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。

33. 简述Apriori算法的核心思想。

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。

34.请分析大数据未来的发展趋势。

价值化、安全与隐私、分析与可视化、商品化、共享、深度学习、数据科学的兴起