**四川工商职业技术学院2020-2021学年第1学期《数据采集系统》期末考试A卷**

**专业班级:**  **学号** **姓名**

**命题教师：** 任 亮 **使用班级：** **教研室主任(签字)：** **系主任(签字)**：

装 订 线

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 题号 | 一 | 二 | 三 | 四 | 五 | 六 | 七 | 八 | 九 | 十 | 总分 |
| 得分 |  |  |  |  |  |  |  |  |  |  |  |
| 评分人 |  |  |  |  |  |  |  |  |  |  |  |

|  |
| --- |
| **一. 单项选择题（每题一分，共16分）**  1．下面不属于CSS子选择器的是（ ）  A.标签选择器 B.Class（类）选择器 C.id选择器 D.对象选择器  2．网页抓取数据时，会根据内容协商技术进行处理，一下不属于内容协商技术的是（ ）。  A.服务器驱动协商 B.客户端驱动协商 C.妥协协商 D.透明协商  3．在使用爬虫时，常用获取数据的方式不包括 ( )  A.警察局盗用数据 B.企业产生的数据 C.数据平台购买的数据 D.政府、机构公开的数据  4. 在爬虫抓取网页的过程中，使用HTTP的GET方法请求返回200表示（ ）。  A.无法找到请求页面 B.请求成功 C.拒绝访问 D.请求未完成  5．从待爬取的URL队列中依次读出URL、并通过( )解析URL、把链接地址转换为网站服务器所对应的IP地址。  A.MAC B.DNS  C.URL D.HTTP  6．以下哪一项不属于消息中间件（ ）  A.Apache Kafka B.RabbitMQ  C.RocketMQ D.share memory  7．HDFS含义 ( )  A. 高级分布式系统 B. Hadoop定向文件系统  C.高级分布式Shell D. Hadoop分布式文件系统  8．Hadoop中的HDFS架构源于( )  A. Google分布式文件系统 B. Yahoo分布式文件系统  C. Facebook分布式文件系统 D. Sina分布式文件系统  9．HDFS数据块的默认大小是多少（ ）。  A. 16MB B. 32MB  C. 64MB D. 128MB  10.HDFS中文件不能( ).  A. 读取 B. 删除  C. 执行 D. 存档  11．大数据的数据量现在已经达到了哪个级别（ ）。  A. GB B. TB  C. PB D. ZB  12． 下列哪个选项不是HDFS架构的组成部分（ ）。  A. NameNode B. DataNode  C. SecondaryNameNode D. GFS  13．与Hadoop 1.x 相比， Hadoop 2.x采用全新的架构，最明显的变化是增加了哪个组件 （ ）  A. MapReduce B. Pig  C. HBase D. Yarn  14．建立在Hadoop文件系统之上的分布式的列式数据库是（ ）。  A. HBase B. Hive  C. Mahout D. Yarn  15． 下面哪个选项不属于Google的三架马车（ ）  A. HDFS B. MapReduce  C. BigTable D. GFS  16． 在HDFS中，对于129MB文件，将创建多少个块（ ）  A. 1 B. 2  C. 3 D. 4  **二、多项选择题（每题3分，共30分）**  1．Apache Kafka通过将以下哪些内容分布在不同的节点（机器上），构成分布式系统( )。  A.生产者(producer) B.代理（broker） C.消费者（consumer） D.用户（user） E.命名空间（namespace）  2．Kafka使用主题来组织数据，每个主题被分为若干个分区，每个分区有多个副本，副本包含以下哪几种类型（ ）。  A.leader副本 B.follower副本 C.topic副本 D.partition副本 E.offset副本  3. Kafka的关键特性包括( )  A.持久化 B.复制和扩展性  C.高性能 D.分区再均衡  E.提交和偏移量  4. 构建一个真实数据通道需要考虑的问题包括( )  A.及时性和可靠性 B.高吞吐量和动态吞吐量  C.数据格式和转换 D.故障处理能力  E.安全性与灵活性  5．ETL设计分为哪3部分（ ）。  A.数据提取 B.数据清洗  C.数据转换 D.数据加载  E.数据测试  6．HDFS的稳健性表现在以下哪几方面（ ）。  A．数据磁盘故障、心跳和重新复制 B．群集重新平衡 C．数据的完整性 D．元数据磁盘故障 E．快照  7. Hadoop支持的文件压缩格式，常见的是( )  A．DEFLATE B．gzip C．bzip2 D．LZo E．rar  8.构建一个网页爬虫的实践过程包括( )  A. 规范化URL B. 广度优先  C.URL/站点黑名单 D. 内容指纹  E. CSS颜色内容识别  9. HDFS提供以下权限处理模式包括( )  A. 匿名 B. 只读权限（r）  C. 超级权限（s） D. 写入权限（w）  E. 可执行权限（x）  10．通过客户端爬虫建立HTTP请求，该格HTTP请求式包含（）  A. 请求行 B. 请求头部  C. 状态行 D. 空行  E. 请求数据  **三、判断题（每题1分，共10分）**  1. 为了提升数据写入的效率，Kafka采用批量写入方式。( )  2. Kafka中生产者创建消息，并把消息发布到一个或多个特定的主题上。( )  3. 大数据存储与查询在计算广告系统中主要用于广告日志存储、用户数据库和历史广告数据库等，所用技术为HDFS和HBase。( ) 4. HTTP的GET方法最初是用来向服务器输入数据的。( )  5. 404 Not Found: 该状态码表明服务器上无法找到请求的资源。( )  6. 消费者读取数据，只能订阅一个主题，并按照消息生成的顺序读取它们。（ ）  7. Apache Kafka是一个分布式的流式处理平台，具有高性能、持久化、多副本备份、横向扩展能力。（ ）  8. 消费者通过向被指派为消费群组协调器的broker发送心跳来维持它们和消费群组的从属关系以及它们对分区的所有权关系。（ ）  9. 在广告系统中，保存广告的点击信息，可以通过生产者写入Kafka中。（ ）  10． 一个Kafka服务称为一个代理（broker），它不能同时为生产者和消费者提供服务。（ ）  **四、填空题（每空1分，共20分）**  1. 网页一般由三个部分组成包括 , , 。  2. Kafka中每一个分区（partition）是一个 ,写入的数据会 排列，其中每一个元素都会按照顺序被标记上id，称为 。  3. 大数据采集和处理在计算机广告系统中主要用于 和 所用的技术为 和 。  4. 计算广告是在 下找到 和 之间的最佳匹配。  5． HDFS的设计初衷是为了支持 和 的流式读写操作。  6. HDFS中的文件和目录都有所属 , 及 。  7. 复制功能是Kafka架构的核心，Kafka分区中的副本包含 和  两种类型。  五、简答题（每题6分，共24分）  1. 大数据的基本特征可以用哪4个V总结？简述他们各具有什么含义。  2. 简述网络爬虫的工作流程。  3.Kafka具备的三个特点是什么?有哪些关键特性？  4.大数据呈现出一种全新的思维角度，请结合学习内容和现实生活给与阐释。 |