2023年上海市高等学校信息技术水平考试试卷

四级 大数据与云计算——大数据平台（模拟卷）

（本试卷考试时间 150 分钟）

一、单选题（ 本大题 30 道小题 ，每小题 1 分，共 30 分），从下面题目给出的A、B、C、D四个可供选择的答案中选择一个正确答案。

1.在Linux操作系统中，___命令可以用来改变一个文件的权限。
A.chmod
B.chdir
C.change
D.file

2.Linux上查看网络连接、路由表、接口统计使用的命令是__。
A.netstat
B.top
C.ps
D.iostat

3.NAT工作在OSI模型中的__。
A.网络层
B.链路层
C.传输层
D.应用层

4.HTTPS协议采用的默认TCP端口号是__。
A.443
B.80
C.22
D.3306

5.__不属于服务器安全测试的项目。
A.单元测试
B.泄露测试
C.接地电阻测试
D.耐压测试

6.__是公有云计算基础架构的基石。
A.分布式
B.虚拟化
C.并行计算
D.高可用

7.大大降低用户在硬件上的开销，按需要租用相应的计算能力和存储能力，这是__的特

点。
A.基础设施即服务（IaaS）
B.数据即服务（DaaS）
C.平台即服务（PaaS）
D.软件即服务（SaaS）

8.云计算按照提供的服务类型进行分类，包括 IaaS、PaaS、__。
A.SaaS
B.Docker
C.Xen
D.KVM

9.在 SQL 语句中，用于表示任意字符的是__。
A.%
B.*
C.LIKE
D._

10.在 Oracle 中，可用于提取日期时间类型特定部分（如年、月、日、时、分、秒）的函数是__。
A.TRUNC
B.DATEPART
C.EXTRACT
D.TO_CHAR

11.有关系 S（SNO，SNAME，SEX），C（CNO，CNAME），SC（SNO，CNO，GRADE）；其中 SNO 是学生号，SNAME 是学生姓名，SEX 是性别，CNO 是课程号，CNAME 是课程名称。要查询选修“数据库”课的全体男生姓名的 SQL 语句是 SELECT SNAME FROM S，C，SC WHERE 子句。这里的 WHERE 子句的内容是__。
A.S.SNO = SC.SNO and C.CNO = SC.CNO and SEX='男' and CNAME='数据库'
B.S.SNO = SC.SNO and C.CNO = SC.CNO and SEX is '男'and CNAME is '数据库'
C.SEX '男' and CNAME ' 数据库'
D.S.SEX='男' and CNAME=' 数据库'

12.str = "Hello,Python";
suffix = "Python";
print (str.endswith(suffix,2));
上述代码的输出结果是__。
A.True
B.False
C.语法错误
D.P

13.在 HDFS Shell 执行语法中，下列关于 Hadoop fs 和 HDFS dfs 说法正确的是__。

A.Hadoop fs 只能操作 HDFS 文件系统
B.HDFS dfs 可以操作任何文件系统
C.HDFS dfs 可以操作 Linux 本地文件
D.Hadoop fs 可以操作任何文件系统

14.编写 MapReduce 程序时，下列错误的是__。
A.不同的 Map 任务之间可以进行通信
B.MapReduce 采用非共享式架构，容错性好
C.MapReduce 采用" 分而治之"策略
D.Hadoop MapReduce 是 MapReduce 的开源实现，后者比前者使用门槛低很多

15.关于 Hadoop 说法错误的是__。
A.Hadoop 框架是用 Java 实现的，MapReduce 应用程序则一定要用 Java 来写
B.Map 函数将输入的元素转换成<key,value>形式的键值对
C.不同的 Map 任务之间不能互相通信
D.MapReduce 框架采用了 Master/Slave 架构，包括一个 Master 和若干个 Slave

16.关于 Kafka 中,描述错误的是__。
A.分区和副本之间没有主从之分
B.分区是一个提交日志
C.一个分区可以有若干副本
D.消息以追加的形式写入分区

17.Flume 的最小可部署单位是__。
A.Event
B.Agent
C.Sink
D.Source

18.使用-compress 参数从 sqoop 导入生成文件,该文件的默认扩展名是__。
A..orc
B..gz
C..tar
D..textfile

19.关于分布式系统特征描述错误的是__。
A.分布式系统节点间的网络通信带来的延时低于单机操作
B.同一个分布式系统中的多个节点，可能会并发地操作一些共享的资源
C.分布式系统在各个节点之间通过网络进行通信；但是，由于网络本身的不可靠，每次网络通信都会伴存在网络不可用的风险
D.当网络分区现象出现时，分布式系统会出现局部小集群；在极端情况下，这些局部小集群会独立完成原本需要整个分布式系统才能完成的功能

20.HDFS 体系结构的设计原则是__。

A.数据与元数据分离
B.动态扩容
C.分治思想
D.表决机制

21.15000 转的 SATA 盘的顺序读取带宽可以达到 100MB/s 以上，磁盘寻道时间大约为 10ms，请问磁盘顺序读取 1MB 数据的时间大约为__。
A.20ms
B.10ms
C.60ms
D.50ms

22.__是 Kafka 强依赖外部组件。
A.Zookeeper
B.HDFS
C.Yarn
D.HBase

23.在 Hive 数据库中,select ceil(2.34) from table 的结果是__。
A.3
B.2
C.2.3
D.2.4

24.已知数组 trans_cnt[1,2,3,4]，__是求数组的元素数量。
A.size(trans_cnt)
B.coalesce(trans_cnt)
C.length(trans_cnt)
D.type(trans_cnt)

25.Spark 重启后，Application 的记录信息需要保留，则应该开启__。
A.EventLog 和 History Server
B.Spark Shell
C.History Server
D.EventLog

26.SparkStreaming 支持用户自定义数据源，但暂不支持__语言。
A.Python
B.Java
C.Golang
D.C 语言

27.Kafka 中 segment 的默认存储数据量大小为__。
A.1GB

B. 2GB
C. 3GB
D. 4GB

28. 关于大数据、云计算和物联网的区别，描述错误的是__。
A. 云计算旨在从海量数据中发现价值，服务于生产和生活
B. 云计算本质上是整合和优化各种 IT 资源并通过网络以服务的方式，廉价地提供给用户
C. 大数据侧重于对海量数据的存储、处理与分析，从海量数据中发现价值，服务于生产和生活
D. 物联网的发展目标是实现物物相连，应用创新是物联网发展的核心

29. HBase 分区数为 1 时，表文件达到__时就会触发分区操作。
A. 256MB
B. 512MB
C. 1GB
D. 2GB

30. 下列不属于大数据产业的产业链环节的是__。
A. 数据循环层
B. 数据源层
C. 数据分析层
D. 数据应用层

二、多选题（ 本大题 5 道小题， 每小题 2 分，共 10 分），从下面题目给出的 A、B、C、D 四个可供选择的答案中选择所有正确答案。

1. 下列属于热插拔设备的是__。
A. 硬盘
B. 电源
C. PCI 卡
D. 机箱风扇

2. 云管理平台的作用包括__。
A. 将各种接口、工具和流程进行组合并以服务的形式提供
B. 以软件和硬件相结合的方式提供服务
C. 各种自动化的工作流程
D. 提供云资源的监控、运维和计费等功能

3. __是 Flink 组件。
A. 作业管理器
B. 任务管理器
C. 资源管理器
D. 分发器

4. HBase 特性描述正确的是__。

A.高可靠性
B.面向列存储
C.可伸缩
D.稀疏存储

5.__是 HDFS2.x 版本相对 HDFS1.x 增加的地方
A.NameNode HA
B.DataNode 水平扩容
C.NameNode 联邦
D.Secondary NameNode

三、是非题（ 本大题 5 道小题 ，每小题 2 分，共 10 分）。
1.一般而言，传统软件开发是业务（功能）驱动的，这就使得数据具有一定的开放性。

2.非 BulkLoad 数据导入 HBase 时，数据会先写入 MemStore。

3.Zookeeper 是一个开放源代码的分布式协调服务。

4.HBase Shell 是 HBase 的数据访问接口。

5.集群管理不是 Zookeeper 的功能。

四、填空题（ 本大题 5 道小题 ，每空 2 分，共 10 分）。
1.大数据的计算模式包括____计算、流处理计算、图计算和查询分析计算。

2.HDFS 的 NameNode 负责管理文件系统的命名空间，将所有的文件和文件夹的元数据保存在一个文件系统树中，这些信息也会在硬盘上保存成__和命名空间镜像。

3.Yarn 是一个__的资源管理系统，用以提高分布式的集群环境下的资源利用率，这些资源包括内存、IO、网络、磁盘等。

4.Hive 中 metastore 默认存储在__数据库中。

5.MapReduce 编程模型中最后执行的组件是__。

五、操作题
以下第（一）题（简答题）、第（二）题（综合题）题目请在文件"C:\KS\大数据与云计算-答题纸.docx"中作答！

(一)简答题
1.请简述MapReduce中Shuffle组件的特点。

请在答题纸作答！此处答题一律无效！

2.请简述Spark的出现解决了哪些问题。

请在答题纸作答！此处答题一律无效！

3.请简述HBase中Rowkey的设计原则。

请在答题纸作答！此处答题一律无效！

4.请简要描述Yarn的设计思想。

请在答题纸作答！此处答题一律无效！

5.有两个文本文件，文件中的数据按行存放，示例如下图所示；请编写MapReduce程序，找到两个文件中彼此不相同的行（写出思路即可）。

第一个文件内容：	第二个文件内容：
A	B
C	C
F	X
F	H

请在答题纸作答！此处答题一律无效！

(二)综合题

小王大学毕业后入职一家大数据公司，在公司业务中遇到以下问题，希望你可以帮忙解决。

1.请分别阐述Hadoop生态系统中HDFS、HBase、Hive、Yarn、Zookeeper组件的功能。（5分）

请在答题纸作答！此处答题一律无效！

2.为什么说Spark在采用RDD这种设计方式以后，就具有高效的容错性？（5分）

请在答题纸作答！此处答题一律无效！

3.对于一个给定的文件file1.txt（绝对路径为："/usr/local/file1.txt"），需要对数据进行排序：根据第1列数据进行降序排序，如果第1列数据相等，则根据第2列数据进行降序排序。

输入文件file1.txt

5	3
1	6
4	9
8	3
4	7
5	6
3	2

输出结果

8	3
5	6
5	3
4	9
4	7
3	2
1	6

请编写一个Python程序，来完成上述功能。（10分）

请在答题纸作答！此处答题一律无效！