2024 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（A 场）

（本试卷考试时间 150 分钟）

一、单选题（ 本大题 15 道小题 ，每小题 1 分，共 15 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.关于数据和大语言模型的关系，说法正确的是____。

A.用于大语言模型的训练数据需要经过抽取、清洗、转换等加工过程

B.互联网上的数据是公开的，所以从互联网获取大语言模型的训练数据是没有成本的

C.用于训练大语言模型的数据必须是结构化数据

D.因为训练大语言模型需要海量数据，所以不用关心数据质量的问题

2.基础 IT 资源的管理，属于____服务的范畴。

A.IaaS

B.PaaS

C.SaaS

D.DaaS

3.分布式计算中的 ACID 原则是指____。

A.原子性、一致性、隔离性、持久性

B.原子性、一致性、隔离性、可用性

C.原子性、一致性、隔离性、多节点性

D.一致性、隔离性、多节点性、可用性

4.VMware 中的____组件可以实现全面的虚拟网络功能。

A.NSX

B.ESX

C.ESXi

D.GSX

5.DevOps 来源于 Development 和____两个单词的组合。

A.Operations

B.Optimums

C.Operators

D.Open-Source

6.用来显示容器日志的 Docker 命令是 docker____。

A.logs

B.files

C.docs

D.text

7.大数据处理的核心功能是____。

A.分析和预测数据
B.绘制数据图形
C.减少数据量
D.批量修改数据

8.从数据类型角度来说，大多数的大数据都是____类型的。
A.非结构化或半结构化
B.非结构化
C.结构化
D.半结构化

9.数据清洗的方法不包括____。
A.修改原始数据
B.一致性检查
C.缺失值处理
D.噪声数据清除

10.HDFS 默认 Block Size 的大小是____。
A.64MB
B.32MB
C.128MB
D.256MB

11.HDFS 分布式文件系统适合的读写任务是____。
A.一次写入，多次读
B.多次写入，少次读
C.多次写入，多次读
D.一次写入，少次读

12.HBase 依靠____存储底层数据。
A.HDFS
B.Hadoop
C.Memory
D.MapReduce

13.Hadoop 集群搭建中常用的有 4 个配置文件为 core-site.xml，hdfs-site.xml，mapred-site.xml 和____。
A.yarn-site.xml
B.application.xml
C.mapreduce.xml
D.config.xml

14.通过 Flink 和 Kafka 配合使用可以对流式数据进行处理、分析，当 Kafka 的分区数为 16 时，Flink 的并行度应该设置成____较为合适。

A. 12
B. 20
C. 28
D. 36

15. 关于 Zookeeper 的描述，错误的是____。
A. Zookeeper 被设计用来实现大容量数据存储
B. Zookeeper 的数据访问具有原子性
C. Zookeeper 维护着一个树形的层次结构
D. Zookeeper 被设计用来实现协调服务

二、多选题（ 本大题 10 道小题 ，每小题 2 分，共 20 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. 如果出现磁盘 I/O 的性能问题，排查故障的正确思路包括____。
A. 查看磁盘的使用空间
B. 观察磁盘的 IOPS、吞吐量和响应时间
C. 跟踪块设备的 I/O 事件
D. 跟踪进程的 I/O 系统调用

2. 关于 VPC（Virtual Private Cloud）的功能，描述正确的包括____。
A. 一个 VPC 可以包含多个子网
B. 一个子网中可以存在多台云主机
C. 可以为子网中的云主机手动分配内网 IP
D. 同一个子网中的一个内网 IP 可以分配给多台云主机

3. 对象存储的优点包括____。
A. 对象存储简化了数据管理
B. 对象存储系统可以支持多个用户或组织使用
C. 使用公有云上的对象存储，早期不需要有过多的成本投入
D. 对象存储对外提供 SCSI 接口和文档

4. DevOps 体系的出现存在一定的必然性，原因包括____。
A. 用户需求多变要求软件系统快速演化
B. 开发团队与运维团队之间的协作更加紧密
C. 软件系统部署环境越来越错综复杂
D. 要求软件的交付过程更加自动化和标准化

5. 关于 HBase 的架构以及数据读写，描述正确的包括____。
A. HBase 采用主从架构，由 Master 节点统一管理所有 RegionServer
B. 表中的每个单元格都有一个时间戳来标识其版本
C. 写入的数据首先存放在内存的 MemStore 中，并同步更新到 WAL 日志
D. 读取数据时先检查 BlockCache，如果缓存命中则直接返回结果

6. 关于 Hive、高性能数据库和非关系数据库，描述正确的包括____。

A.Hive 可以将 SQL 语句转换为 MapReduce 任务进行执行
B.高性能数据库通常采用优化的索引和分区策略以提升查询效率
C.NoSQL 数据库支持灵活的数据模型，但不适合大规模数据分析
D.Hive 主要用于在线事务处理（OLTP）操作

7.Flink 作为流处理框架，其特性包括____。
A.提供批处理和流处理的统一接口
B.基于事件时间窗口进行计算
C.弱一致性保证
D.只支持无状态计算

8.关于 Zookeeper 和 YARN ，描述正确的包括____。
A.Zookeeper 使用 ZAB 协议保证数据一致性
B.Zookeeper 提供类似文件系统的操作模式
C.Node Manager 在 YARN 中负责单个节点上的任务执行和资源管理
D.YARN 中的 Application Master 负责全局资源管理和调度

9.关于 Spark 框架，描述正确的包括____。
A.Spark 是开源的大数据计算框架
B.Spark 是基于内存的大数据计算框架
C.Spark 集群可以部署在廉价的硬件之上
D.Spark 不具有可扩展性

10.关于数据挖掘、机器学习和数据建模这几种概念，描述正确的包括____。
A.数据挖掘可以使用机器学习技术
B.机器学习可以用于预测和分类
C.数据挖掘可以包括无监督的学习任务，如聚类
D.数据建模仅用于创建线性模型以进行预测

三、是非题（本大题 15 道小题，每小题 1 分，共 15 分）。
1.Linux 云主机仅支持通过用户名、密码方式登录。

2.云主机的安全防护包括云主机间的安全隔离、访问控制、安全监控、安全漏洞加固等。

3.分布式系统中，节点之间可以通过消息传递进行通信，也可以通过共享内存进行通信。

4.在 OpenStack 框架中，镜像管理是 Nova 组件所具有的功能。

5.下载任何容器镜像之前，必须通过 docker login 命令进行登录。

6.在微服务框架中，当下游服务、接口出现故障时，为了防止整个服务出现雪崩，返回一组提前定义好的静态消息或数据的手段叫做熔断。

7.华为的 OpenGauss 是云数据仓库产品。

8.数据湖以自然或原始格式存储所有规模的结构化和非结构化数据。

9.Hive 性能优化方法只包括 SQL 语句优化和数据存储优化两方面。

10.Hive 数据仓库平台除了支持基本数据类型如整型、浮点和字符串外，还支持复杂数据类型如映射和结构体等。

11.在 HBase 的使用中，通过合理地设计行键，就能使数据尽可能均匀分布，从而达到提高性能的目的。

12.分布式架构、水平分片和多副本存储等技术的合理运用保证了 HBase 的高可用性和伸缩性。

13.时间序列分析是一种重要的数据挖掘技术，只能用于金融市场序列数据分析。

14.决策树是一种常见的分类模型，通常使用 C4.5 算法构建，但无法使用 MapReduce 框架实现决策树的训练。

15.在深度学习中，Transformer 模型最初是为自然语言处理（NLP）设计的，但随后应用到了图像、视频和音频处理等多个领域。

四、操作题

以下（一）基础实践题、（二）场景设计与应用题请在文件"C:\KS\大数据与云计算-答题纸.docx"中作答！

（一）基础实践题（共2题，每题10分，共20分）

1. Docker是一个用于构建、发布和运行应用程序的开放平台，它简化了应用程序的打包、分发和运行过程，提高了开发、部署和管理应用程序的效率，越来越多的应用部署从虚拟机逐步迁移到了容器环境。我们可以通过Docker的操作命令来管理Docker容器的创建、运行、停止、删除等全生命周期，请写出以下操作的Docker命令。

1）从中央仓库中拉取一个叫作my_image的镜像，标签（版本）为v1。

2）以镜像my_image在后台启动一个新的容器，命名为my_container，并将容器的80端口映射到主机的8080端口。

3）停止容器my_container，并删除镜像my_image。

2. 某电商平台积累了大量客户购买行为数据，现在拟通过机器学习的建模方式进行客户分

类，并对同类客户进行商品推荐。请回答以下问题。

1）请简述什么是ETL以及ETL的作用。

2）请从数据是否有标签的角度说明机器学习算法的种类。

3）请列举2种常见的分类算法。

4）请列举2种常见的商品推荐算法。

5）针对本场景，请选择一种无监督的客户聚类算法，并对此算法进行简述。

（二）场景设计与应用题（共3题，每题5个答题点，每个答题点2分，共30分）

一家中型科技公司正在经历快速的业务增长，随着客户基础的扩大，数据量和计算需求不断增加。为了应对这些挑战，公司决心进行技术升级，以保持竞争优势。主要目标是利用云计算的灵活性来优化基础设施，同时通过大数据分析来深入洞察客户的行为，从而推动更智能的决策。

1．公司正在将基础设施迁移到云端，以提高可扩展性和成本效率。他们考虑使用基础设施即服务（IaaS）和平台即服务（PaaS）的组合来运行他们的应用程序。

1）请简述 IaaS 和 PaaS 的区别。

2）公司决定将其传统数据中心迁移到云端，请列出迁移过程中需要考虑的2个主要挑战，并提供可能的解决方案。

3）请简述在云迁移过程中，如何确保数据的安全性和合规性，请列出至少2种安全措施。

4）请简述云计算中可扩展性一词的含义，并举例说明它在实际应用中的意义。

5）请简述云原生应用程序与传统应用程序的区别，并说明云原生应用程序更适合在云环境中运行的原因。

2．公司决定使用 Hadoop 的 MapReduce 编程模型来存储和处理客户行为数据。

1）请简述 Hadoop 的架构，并说明 HDFS 和 MapReduce 是如何协同处理大数据的。

2）请简述 NameNode 和 DataNode 在 HDFS 中的作用，并说明Hadoop 如何确保数据的可靠性和容错性。

3）请用 Java 或 Python 编写一个简单的 MapReduce 程序，统计给定文本文件中每个单词出现的次数。

4）请简述MapReduce 模型的局限性，并举例说明可以通过哪些其他大数据处理框架来解决该问题。

5）请简述YARN组件对Hadoop MapReduce框架的作用。

3．在 Hadoop 部署成功后，公司计划通过集成实时数据处理和改善数据治理来增强其数据处理能力。

1）请比较实时数据处理和批处理的异同，并举例说明它们的使用场景。

2）请简述 Zookeeper 在大数据生态系统中的作用，并说明它在协调和配置管理中的重要性。

3）请使用 Python 或 Java 编写一个简单的 ETL 脚本，从数据源提取数据，对数据进行转换，并加载到数据库中。

4）请简述使用数据湖的优势，并说明它与传统数据仓库的区别。

5）请简述数据治理的概念以及它在大数据环境中的重要性，并说明公司如何确保有效的数据治理。