2017年中国移动业务支撑
大数据序列人才选拔考试题

姓　名 _____

省公司 _____

部　门 _____

总　分 _____

考试注意事项	一、参加考试须带身份证或工作证，未带者不准进入考场。 二、书本、参考资料等物品一律放到考场指定位置。手机关闭或静音。 三、遵守考试纪律，不得有考场违纪或作弊行为。 四、请将答题内容做在试题答卷上，如考卷空白处不够，可以从监考人员处领取白纸。 五、注意填写完整的姓名等信息。

题号	一	二	三	四	五	六	七		总分
满分									
得分									

一. 单项选择题答案填写到下表中:

序号	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
答案										

二. 多项选择题答案填写到下表中:

序号	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
答案										
序号	11	12	13	14	15	16	17	18	19	20
答案										

三. 判断题答案填写到下表中：

序号	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
答案										
序号	11.	12.	13.	14.	15.					
答案										

一．单项选择题：(10 分，每题 1 分)

1. 下列不属于 Hadoop 生态系统的是（　　）。

A .Sqoop　　B. Avro　　C. Impala　　D .Openstack

2. 下列关于大数据的分析理念的说法中，错误的是（　　）

A. 在数据基础上倾向于全体数据而不是抽样数据

B. 在分析方法上更注重相关分析而不是因果分析

C. 在分析效果上更追究效率而不是绝对精确

D. 在数据规模上强调相对数据而不是绝对数据

3.（　　）反映数据的精细化程度，越细化的数据，价值越高。

A. 规模

B. 活性

C. 关联度

D. 颗粒度

4. 集中化经分对各省经分接口数据上传及时性要求，正确的是（　　）。

A. 在规定时限前，传至省端接口机

B. 在规定时限前，传至总部接口机

C. 通过文件级、记录级检查要求，且在规定时限前传至省端接口机

D. 在规定时限前，通过文件级、记录级检查要求

5. CART 算法属于以下哪类分析模型？（　　）

A.关联分析

B.神经网络分析

C.层次聚类分析

D.决策树分析

6. 下列关于聚类挖掘技术的说法中，错误的是（　　）。

A. 不预先设定数据归类类目，完全根据数据本身性质将数据聚合成不同类别

B. 要求同类数据的内容相似度尽可能小

C. 要求不同类数据的内容相似度尽可能小

D. 与分类挖掘技术相似的是，都是要对数据进行分类处理

7.　K-Means 聚类分析算法属于机器学习的哪种学习方式分类？（　　）

A. 监督式学习

B. 非监督式学习

C. 半监督式学习

D. 强化学习

8. 设某数据结构的二元组形式表示为 A=(D，R)，D={01，02，03，04，05，06，07，08，09}，R={r}，r={<01，02>，<01，03>，<01，04>，<02，05>，<02，06>，<03，07>，<03，08>，<03，09>}，则数据结构 A 是（　）。

(A) 线性结构　　(B) 树型结构　　(C) 物理结构　　(D) 图型结构

9. HDFS 分布式存储集群的后台进程不包括：（　　）。

A. SecondaryNameNode

B. DataNode

C. ResourceManager

D. NameNode

10．决策树构建过程中，属性选择的依据是：（　　）

A. 属性的取值范围

B. 属性的顺序

C. 属性的信息增益度

D. 属性的离散程度

二．多选题：(30 分，每题 1.5 分，多答、答错不得分，少答按比例得分)

1. 下列关于大数据的说法中，错误的是（　　）。

A. 大数据具有体量大、结构单一、时效性强的特征

B. 处理大数据需采用新型计算机架构和智能算法等新技术

C. 大数据的应用注重相关分析而不是因果分析

D. 大数据的应用注重因果分析而不是相关分析

E. 大数据的目的在于发现新的知识与洞察并进行科学决策

2. 以下哪些是 Yarn 的调度器（　　）。

A. FIFO Scheduler

B. Capacity Scheduler

C. Fair Scheduler

D. Task Scheduler

3. 国务院印发的《促进大数据发展行动纲要》中部署了三方面主要任务，包括（　　）。

A. 加快政府数据开放共享，推动资源整合，提升治理能力。

B. 推动产业创新发展，培育新兴业态，助力经济转型。

C. 强化安全保障，提高管理水平，促进健康发展。

D. 促进国际交流合作，建立完善国际合作机制。

4. MapReduce 与 HBase 的关系，哪些描述是正确的？（　　）。

A. 两者不可或缺，MapReduce 是 HBase 可以正常运行的保证

B. 两者不是强关联关系，没有 MapReduce，HBase 可以正常运行

C. MapReduce 可以直接访问 HBase

D. 它们之间没有任何关系

5. 以下哪些属于企业级大数据平台 1.0 的重点功能要求？（ ）

A. 统一数据采集

B. 统一数据中心

C. 基于多租户的开放框架

D. 统一运维管理

E. 数据治理

6. 考虑到企业级省大数据平台数据仓库的可扩展能力、投资成本和易于管理等多种因素，经营分析数据仓库逻辑数据模型基本上遵照第三范式进行设计。一个符合第三范式的关系必须满足的是哪些条件？（ ）

A. 每个属性的值唯一，不具有多义性

B. 每个属性，包括主属性或非主属性，都完全依赖于候选键，并且不存在传递依赖情况

C. 每个非主属性必须完全依赖于整个主键，而非主键的一部分

D. 关系模式中不存在传递依赖

7. Spark 支持的分布式部署方式包括哪些 ( )。

A. Standalone

B. Spark on local

C. Spark on YARN

D. Spark on mesos

8. 常用组件部署原则以下正确的包括（ ）。

A. ZooKeeper 每个集群内配置 2 个到节点上。如需扩展，请保持数量为偶数个。

B. HDFS 的 DataNode 至少部署 3 个在数据节点上。

C. Yarn 的 NodeManager 分别部署在 2 个节点上，主备配置。

D. Spark 的 SparkResource 所有节点上都要部署。

E. HBase 的 HMaster 分别部署在 2 个节点上，主备配置。

9. 在网络爬虫的爬行策略中，应用最为基础的是（ ）。

A. 深度优先遍历策略

B. 广度优先遍历策略

C. 高度优先遍历策略

D. 反向链接策略

E. 大站优先策略

10. 元数据分为技术元数据、业务元数据和管理元数据三类，其中以下哪个属于技术元数据？（　　）

A. 数据库表定义　　B. 业务术语　　C. 字段类型　　D. 主键信息　　E.用户权限

11. 以下哪个场景适合利用 Hive 数据仓库（　　）

A. 日志分析系统

B. 机票预订系统

C. 银行交易系统

D. 广告分析系统

12. 以下哪些是 HDFS 适合的场景（　　）？

A. 存储并管理 PB 级数据

B. 处理非结构化数据

C. 随机读

D. 高吞吐对延迟不敏感的数据处理

E. 一次写，多次读

13. 数据模型设计分哪三个阶段？（　　）

A. 物理模型

B. 概念模型

C. 关系模型

D. 逻辑模型

14. HBase 表具有哪些特点?（　　）

A. 多维度

B. 稀疏表

C. 每行有相同的列（族）属性

D. 多版本数据记录

15. 在关联分析模型中，常用于分析关联规则的度量指标包括：（　　）

A. 支持度

B. 置信度

C. 召回率

D. 查准率

16. HBase 的 LSM 树更能保证哪种操作的性能？（ ）

A. 读

B. 写

C. 随机读

D. 合并

17. Spark 大数据分析处理中的内存数据结构是：（ ）

A. RDD(弹性分布式数据集)

B. Key-Value（键值对结构）

C. Record（记录结构）

D. Document（文档对象结构）

18. 以下属于关联分析算法的包括（ ）

A. K-Means 算法

B. Apriori 算法

C. FP-Growth 算法

D. Random Forest

19. 常用的机器学习数据挖掘模型包括:（ ）

A. 分类分析模型

B. 聚类分析模型

C. 回归预测模型

D. 关联分析模型

20.Hadoop 分布式分建系统可以对文件进行哪些操作？（ ）

A. 追加

B. 删除

C. 流式读取

D. 随机修改

三．判断题（15 分，每空 1 分）（正确填写 T， 错误填写 F）

1. 公司将大数据列为“大连接”战略中一项重点工作，是因为“大连接”是立足核心优势的战略，大数据是重要优势之一。大连接要面向数字化服务转型， 大数据是大连接（云、物、数）三大基础技术之一。（ ）
2. 在中国移动大数据平台建设初期，实现“整合”和“开放”能力是平台建设的首要工作重点。（）
3. 中国移动企业级省大数据平台的建设原则是“统一性、开放性、先进性、渐进性、安全性、易用性、利旧性”。（ ）
4. 企业级省大数据平台的元数据管理，要改变以往事前元数据管理的模式，向事后元数据管理模式演进。（ ）
5. Scale Out是架构扩展的一种方式，是目前大数据时代主流的扩展模式，通过为计算节点增加更多的CPU Cores，存储设备和内存，提升计算效率从而达到扩展的目标。（ ）
6. 决策树是一种基于树形结构的预测模型，每一个树形分叉代表一个分类条件，叶子节点代表最终的分类结果，其优点在于易于实现，决策时间短，并且适合处理非数值型数据。（ ）
7. NameNode全权管理数据块的复制，它周期性的从集群中的每个Datanode接收心跳信号和块状态报告。（ ）
8. 如果NameNode意外终止，SecondaryNameNode会接替它使集群继续工作。（ ）
9. Hadoop默认调度器策略为FIFO，并支持多个Pool提交Job。（ ）
10. 信息安全涉及信息的5个方面，即可用性、机密性、完整性、可控性、不可抵赖性。（ ）
11. 决策树算法属于无监督式学习算法。（ ）
12. 贝叶斯算法属于分类算法，可用于分类预测场景。（ ）
13. K-means算法是典型的关联规则挖掘算法。（ ）
14. 推荐算法包括基于用户的推荐算法和基于商品的推荐算法。（ ）
15. 在Hbase中，Hmaster的高可用性由Zookeeper负责。（ ）

四．填空题（15 分，每空 1 分）

1. 企业级省大数据平台的统一数据采集是为平台提供了汇集数据的主要功能，按采集方式上来分，包含下述主要采集功能：_____、_____、_____。
2. 计算引擎是处理大数据的主要功能部件，它为各类场景提供所需的各种计算能力，主要包括_____、_____以及_____等计算方式。
3. 在企业级省大数据平台中，面向多租户应提供哪些服务能力：_____开放、_____开放、_____开放。
4. 企业级省大数据平台的数据开放服务通过 _____形式提供给数据消费者使用。
5. VoLTE的信令和媒体经EPC路由至_____网络，由其提供会话控制和业务逻辑。
6. 4G位置信令是从网络的_____接口采集的。除此之外，还可以从哪些途径获得位置信

息？(列举一种)_____。

7. Flume是分布式、可靠、高可用的海量日志采集系统。每个Flume Agent包含至少一个_____（决定从哪里取数据）和 _____（将数据送到哪里）。

五．简答题（20 分，每题 5 分）

1、 请简述 CAP 原理。

2、 请简述 DPI 原理，描述 DPI 日志结构。

3、 为解决或排查数据异常，两级经分系统间建立了工单处理流程，各省可以申请三种方式排查或解决数据问题。请简述有哪三种方式，以及各自适用场景。

4、 列举企业级省大数据平台各层的安全威胁场景及应对策略？

六．论述题（10 分）

结合实践工作，描述如何用大数据分析应用支撑四轮驱动（可从移动市场、家庭市场、集客市场、新业务市场（含物联网）选取一个应用场景，分析当前存在的问题或市场需求，从数据接口、数据建模、分析应用或触点营销互动等几个方面介绍技术方案。

加分题（10 分）

请简述如何理解十三五发展愿景：成为数字化创新的领先运营商？ 十三五战略和大数据的关系？在其中大数据应发挥哪些作用？