工业和信息化人才培养工程大数据分析师

复习题库（二）

1、 在网络爬虫的爬行策略中，应用最为基础的是（AB ）。
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
2、 当前，大数据产业发展的特点是（ACE ）。
A：规模较大 B：规模较小 C：增速很快
D：增速缓慢 E：多产业交叉融合
3、 下列关于数据生命周期管理的核心认识中，正确的是（ABC ）。
A：数据从产生到被删除销毁的过程中，具有多个不同的数据存在阶段
B：在不同的数据存在阶段，数据的价值是不同的
C：根据数据价值的不同应该对数据采取不同的管理策略
D：数据生命周期管理旨在产生效益的同时，降低生产成本
E：数据生命周期管理最终关注的是社会效益
4、 下列关于基于大数据的营销模式和传统营销模式的说法中，错误的是（ AB）
A：传统营销模式比基于大数据的营销模式投入更小
B：传统营销模式比基于大数据的营销模式针对性更强
C：传统营销模式比基于大数据的营销模式转化率低
D：基于大数据的营销模式比传统营销模式实时性更强
E：基于大数据的营销模式比传统营销模式精准性更强
5、 下列关于脏数据的说法中，正确的是（ ABCDE）。
A：格式不规范 B：编码不统一 C：意义不明确 D：与实际业务关系不大 E：数据不完整
6、 数据再利用的意义在于（ ABC）。
A：挖掘数据的潜在价值 B：实现数据重组的创新价值 C：利用数据可扩展性拓宽业务领域
D：优化存储设备，降低设备成本 E：提高社会效益，优化社会管理
7、 按照涉及自变量的多少，可以将回归分析分为（CD ）。
A：线性回归分析 B：非线性回归分析 C：一元回归分析
D：多元回归分析 E：综合回归分析
8、 传统数据密集型行业积极探索和布局大数据应用的表现是（ BCE）。
A：投资入股互联网电商行业 B：打通多源跨域数据
C：提高分析挖掘能力 D：自行开发数据产品 E：实现科学决策与运营
9、 大数据人才整体上需要具备（ABE ）等核心知识。
A：数学与统计知识 B：计算机相关知识 C：马克思主义哲学知识
D：市场运营管理知识 E：在特定业务领域的知识
10、 下列关于大数 据的说法中，错误的是（AD ）。
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

11.大数据作为一种数据集合，它的含义包括（ ）。

☑ A.数据很大 ☐ B.很有价值 ☑ C.构成复杂 ☑ D.变化很快

12.大数据处理流程可以概括为以下哪几步？

☑ A.挖掘 ☑ B.采集 ☑ C.统计和分析 ☑ D.导入和预处理

13.宁家骏委员指出，（）主导了 21 世纪。

☑ A.云计算 ☐ B.移动支付 ☑ C.大数据 ☑ D.物联网

14.大数据的主要特征表现为（）。

☑ A.数据容量大 ☑ B.商业价值高 ☑ C.处理速度快 ☑ D.数据类型多

15.大数据作为一种数据集合，当我们使用这个概念的时候，实际包含有哪几层含义？

☑ A.数据很大 ☑ B.构成复杂 ☑ C.变化很快 ☑ D.蕴含大价值

16.贵州发展大数据的顶层设计是要逐步建成三个中心，即（）。3 分

☐ A.大数据人才中心 ☑ B.大数据金融中心 ☑ C.大数据内容中心 ☑ D.大数据服务中心

17.云计算的特点包括以下哪些方面？

☑ A.服务可计算 ☑ B.高性价比 ☑ C.服务可租用 ☐ D.低使用度

18.下列选项中，属于贵州发展大数据的先天优势的是（）。

☑ A.空气清新 ☑ B.远离地震带 ☑ C.气候凉爽 ☑ D.电力资源充沛

19.下列各项表述中正确的有哪些？

☐ A.我国中央网络安全和信息化领导小组宣告成立是在 2013 年。

☑ B.中央网络安全和信息化领导小组组长是习近平。

☑ C.我国中央网络安全和信息化领导小组宣告成立是在 2014 年。

☐ D.中央网络安全和信息化领导小组组长是李克强。

20.“十二五”以来我国信息化发展的亮点包括以下哪些方面？

☑ A.信息产业的支撑性、保障性、带动性作用进一步增强

☑ B.信息基础设施建设取得长足进步，为信息化全面深化发展提供了有力保障

21.贵州发展大数据的“八个一”建议包括（）。

☑ A.制定一个工作计划、建立一个领导机构 ☑ B.培养一批干部、出台一批政策

C.引入一批人才、聚集一批创客 D.谋划一批产业、引进一批项目

22.云计算使得使用信息的存储是一个（）的方式，它会大大地节约网络的成本，使得网络将来越来越泛在、越来越普及，成本越来越低。

A.分布式 B.密闭式 C.密集式 D.共享式

23.郭永田副主任指出，物联网在大田作物生产中的应用体现在以下哪些方面？

A.农作物病虫害监测 B.农业精准生产控制 C.农田环境监测 D.农作物长势苗情监测

24.医疗领域如何利用大数据？

A.临床决策支持 B.个性化医疗 C.社保资金安全 D.用户行为分析

25.2012 年“中央 1 号文件”提出，要全面推进农业农村信息化，着力提高（）的信息服务水平。

A.农业生产经营 B.质量安全控制 C.文化交流 D.市场流通

26.20 世纪中后期至今的媒介革命，以（ ）的出现为标志。

A.互联网 B.自动化 C.计算机 D.数字化

27.大数据的应用能够实现一场新的革命，提高综合管理水平的原因是

A.从柜台式管理走向全天候管理 B.从粗放化管理走向精细化管理

C.从被动反应走向主动预见型管理 D.从单兵作战走向联合共享型管理

28.建立大数据需要设计一个什么样的大型系统？

A.能够把应用放到合适的平台上 B.能够开发出相应应用

C.能够处理数据 D.能够存储数据

29.大数据的应用能够实现一场新的革命，提高综合管理水平的原因是（ ）。

A.从被动反应走向主动预见型管理 B.从粗放化管理走向精细化管理

C.从单兵作战走向联合共享型管理 D.从柜台式管理走向全天候管理

30.下列哪些国家已经将大数据上升为国家战略？

A.英国 B.日本 C.美国 D.法国

31、目录上的读取锁可以防止目录被怎样？

A、访问 B、删除 C、改名 D、快照

32、在 Hadoop 集群启用 HDFSHighAvailability(HA)是为了以实现下哪些目的？

A、如果 NameNode 宕机，将会自动故障转移

B、维护其中一个 NameNode 无需中断整个集群运行

C、配置无限个热备 NameNode

D、实现同时向两个集群写入数据

33、一个 300MB 的文件，写入块大小配置为 128MB 的 HDFS 中，其它所有 Hadoop 默认值都未改变，以下描述不正确的是？

A、该文件将消耗 1152MB 的集群空间　　B、第三块是 64MB

C、第三块初始块将为 44MB　　D、最初的两个块将为 128MB

34、以下说法正确的是

A、最新版的 hive 可以针对每行进行增删改查

B、hive 可以通过 serde 支持多种数据存储格式

C、最新版的 hive 还不支持索引　　D、hive 可以完全替代关系数据库

35、以下方法错误的是

A、hive 是一个数据仓库系统　B、hive 的数据必须存储在 HDFS 上

C、hive 的 metastore 必须是关系数据库

D、hive 可以完全替代关系数据库

36、hive 可以将 SQL 查询语言转化为执行步骤并在以下哪些引擎上运行？

A、mr (MapReduce)　　B、Tez　　C、spark　　D、Stream

37、对 hive 的 parquet 存储格式与 ORC 存储格式的说法正确的是

A、parquet 存储格式是列式存储，对数据仓库的事实表具有很好的压缩效果

B、ORC 格式是行存储格式，没有办法进行压缩

C、用 parquet 格式存储的表和用 ORC 格式存储的表之间没有办法 join

D、列式存储更适合统计分析，行式存储更适合数据加载

38、当存储空间比较紧缺时，可以通过以下哪种方式加快空间回收速率？

A、手动删除已被删除的文件　B、调整回收策略，缩短空间回收时间

C、添加 datanode 数量　　D、添加副本数

39、关于 MapReduce 与 Yarn 的关系以下说法正确的是？

A、yarn 负责资源分配，MapReduce 负责计算过程

B、在 yarn 上可以跑各种不同的任务，MapReduce 只是其中一种

C、在 hadoop 版本 2 上 MapReduce 可以不依赖 yarn 直接运行

D、yarn 是常驻服务，mapreduce 在运行时才启动

40、HDFS 采用 NameNode HA 技术之后，以下说法正确的是？

A、在 NameNode 失去响应时 SecondaryNameNode 会接管集群

B、系统中会出现多个 NameNode

C、可以有多个 NameNode 提供服务

D、已经运行的 job 不会因为 NameNode 节点宕机而退出

41、用户配置记录存储在 OLTP 数据库中，Web 服务器日志已经被加载到 HDFS 中，如果需要联合查询用户配置记录和 web 服务器日志，在 HDFS 中获取用户配置记录的最佳方式是以下哪些项？

A、使用 Hadoopstreaming 获取　　B、使用 ApacheFlume 获取

C、使用 Hive 的 LOADDATA 命令获取　D、使用 Sqoop 获取

42、关于 Hadoop 中配置机架感知，以下描述正确的是？

A、如果一个机架出问题，会影响数据读写

B、写入数据的时候会写到不同机架的 DataNode 中

C、MapReduce 会根据机架获取离自己比较近的网络数据

D、对数据进行自由分配

43、MapReducev2(MRv2/YARN)的设计是用来解决以下哪些问题？

A、JobTracker 的资源压力　B、HDFS 的延迟问题

C、运行 MapReduce 的框架之外的其它框架，例如 MPI

D、减少 MapReduceAPI 的复杂性

44、Web 服务器日志文件通常生成为什么格式，以及为了在 Hadoop 中被分析它们需要被转换成什么格式？

A、生成为 XML 文件格式，需要转换成 JSON 格式用于分析

B、生成为 Text 文件格式，需要解析出有用的字段用于分析

C、生成为 CSV 文件格式，需要解析出有用的字段用于分析

D、生成为 HTML 文件格式，需要转换为平面文本或 CSV 格式用于分析