1. 下列选项中无法通过大数据技术实现的是？（A）[单选题] A.商业模式发现

B.信用评估C.商品推荐D.运营分析

1. FusionInsight Manager 的主要功能有以下哪些？（BCD）[多选题] A.数据集成
2. 系统管理
3. 安全管理D.服务治理
4. 假设每个用户最低资源保障设置为 yarn，scheduler，capacity，root， QueueA.minimum-user-limit-percent=24。则以下说法错误的是？（D）[单选题] A.第 3 个用户提交任务时，每个用户最多获得 33.33%的资源
5. 第 2 个用户提交任务时，每个用户最多获得 50%的资源
6. 第 4 个用户提交任务时，每个用户最多获得 25%的资源
7. 第 5 个用户提交任务时，每个用户最多获得 20%的资源
8. 华为大数据解决方案中平台架构包括以下哪些组成部分？（ABD）[多选题]
9. Hadoop 层
10. FusioInght Manager
11. GaussDB 200
12. DataFram
13. Spark 自带的资源管理框架是？（A）[单选题] A.Standal oneB.Mesos C.YARN

D.Docker

1. 关于 RDD，下列说法错误的是？（B）[单选题] A.RDD 具有血统机制（Lineage）B.RDD 默认存储在磁盘

C.RDD 是一个只读的，可分区的分布式数据集D.RDD 是 Spark 对基础数据的抽象

1. 关于 Spark SQL&Hive 区别与联系，下列说法正确的是？(BCD)[多选题] A.Spark SQL 依赖 Hive 的元数据
2. Spark SQL 的执行引擎为 Spark core，Hive 默认执行引擎为 MapReduce
3. Spark SQL 不可以使用 Hive 的自定义函数
4. Spark SQL 兼容绝大部分 Hive 的语法和函数
5. 在 FusionInsight 集群中，Spark 主要与以下哪些组件进行交互？(ABCD)[多选题]

A.HDFS B.YARN

C.Hive D.ZooKeeper

1. 关于 Hive 在 FusionInsight HD 中的架构描述错误的是？（A）[单选题] A.只要有一个 HiveServer 不可用，整个 Hive 集群便不可用
2. MotaStore 用于提供元数据服务，依赖于 DBService
3. 在同一时间点，HiveServer 只要一个处于 Active 状态，另一个则处于 Standby 状态D.HiveServer 负责接收客户端请求.解析.执行 HQL 命令并返回查询结果
4. 通常情况下，Hive 以文本文件存储的表会以回车作为其行分隔符，在华为 FusionInsight Hive 中，可以指定表数据的输入和输出格式处理。（A）[单选题]A.TRUE B.FALSE
5. 以下关于 HBase 中 HMaster 的功能描述哪些是正确的？（ABCD）[多选题] A.Region 负载均衡，Region 分裂以及分裂后的 Region 分配
6. 负责建表/修改表/删除表
7. 负责 RegionServer 的负载均衡D.RegionServer 失效后的 Region 迁移
8. Nimbus 在 Streaming 架构中的功能包括？（ABD）[多选题] A.监控任务执行状态
9. 任务调度
10. 启动/关闭工作进程D.资源分配
11. Topology 在任务完成后会自动结束运行。（B）[单选题] A.TRUEB.FALSE
12. 下面关于 flink 窗口的描述错误的是 （AC）[多选题] A.滚动窗口在时间上是重叠的
13. 滚动窗口在时间上是不重叠的
14. 滑动动窗口之间时间点不存在重叠D.滑动窗口之间时间点存在重叠
15. 在 FusionInsightManager 界面中，对 Loader 的操作不包括下列哪个选项？（A）[单选题] A.切换 Loader 主备节点
16. 启动 Loader 实例
17. 配置 Loader 参数
18. 查看 Loader 服务状态
19. 创建 Loader 作业中，可以在以下哪个步骤中设置过滤器类型？（B）[单选题] A.输入设置
20. 转换
21. 基本信息D.输出
22. 以下哪些数据源可以通过 Loader 实现与 FusionInghtHD 的数据交换？（ABCD）[多选题] A.NoSQLB.FTP Server C.SFTP Server D.MySQL
23. Kafka 日志的清理方式有两种：delete 和 compact。默认值是 delete。（A）[单选题] A.TRUEB.FALSE
24. kafka-cluster mirroring 工具可以实现以下那些功能？（A）[单选题] A.kafka 集群数据同步方案
25. kafka 单集群内数据备份
26. kafka 单集群内数据恢复D.以全部不对
27. fusioninsight HD 产品中，一个典型的 kafka 集群包含若干个 producer 若干个consummer 和一个 zookeeper 集群? （B）[单选题]A.true B.false
28. ZKFC 进程部署在 hdfs 中的以下那个节点上？（AB）[多选题] A.active namenodeB.standby namenode C.datanode

D.以上全部不对

1. 以下关于 Kafka Partition 偏移量的描述不正确的是？（D）[单选题] A.每条消息在文件中的位置称为 offset（偏移量）
2. 消费者通过（offset.partition.topic）跟踪记录
3. 唯一标记一条消息
4. Offset 是一个 String 型字符串
5. Channel 的作用类似队列，用于临时缓存进来的数据，当 Sink 成功地将数据发送到下一跳的 Channel 或最终目的，数据会从 Channel 移除（A）[单选题]A.TRUE B.FALSE
6. Zookeeper 中的数据节点 znode 分为哪几种类型？（BD）[多选题] A.semi-persistentB.ephemeral C.temporary D.persistent
7. HDFS 联邦环境下，NameSpace(命名空间)包含以下哪些内容？（ABC）[多选题] A.目录B.文件C.块

D.以上全不正确

1. HDFS 存储数据时，关键数据根据实际业务需要保存在具有高度可靠性的节点中，通过修改 DataNode 的存储策略，系统可以将数据强制保存在指定的节点组中。（A）[单选题]A.TRUE B.FALSE
2. RDD 有 Transformation 和 Action 算子，下列属于 Action 算子的是？（B）[单选题] A.mapB.saveAsTextFile C.Filter D.reduceByKey
3. Spark on Yarn-client 适合用于生产环境是因为可以更快的看到 APP 的输出。（B）[单选题]A.TRUE B.FALSE
4. 以下关于 Hive SQL 基本操作描述正确的是？（D）[单选题] A.创建外部表必须要指定 location 信息
5. 创建外部表使用 external 关键字，创建普通表需要指定 internal 关键字
6. 加载数据到 Hive 时源数据必须是 HDFS 的一个路径D.创建表时可以指定列分割符
7. 在 Streaming 中，消息可靠性级别中精确一次是通过 ACK 机制来实现的。（B）[单选题] A.TRUEB.FALSE
8. Flume 适用于以下哪些场景的数据收集？（ABC）[多选题] A.Thrift，Avro，Syslog，Kafka 等数据源上收集数据B.本地文件数据采集C.应用系统产生的日志采集D.大量数据的实时数据采集
9. 在 Zookeeper 和 Yarn 的协同工作中，当 ActiveResourceManager 产生故障时，Standby ResourceManager 会从以下哪些目录中获取 Application 相关信息？（B）[单选题]A.metastore B.Statestore C.Storeage D.Warehouse
10. ZooKeeper 所有节点都可以处理读请求。（A）[单选题] A.TRUEB.FALSE
11. HDFS 的副本放置策略中，同一机架不同的服务器之间的距离是（B）[单选题] A.3B.2

C.1

D.4

1. 下列哪些措施是为了保障数据的完整性（ABCD）[多选题] A.元数据可靠性保证B.重建失效数据盘的副本数据C.安全模式

D.集群数据均衡

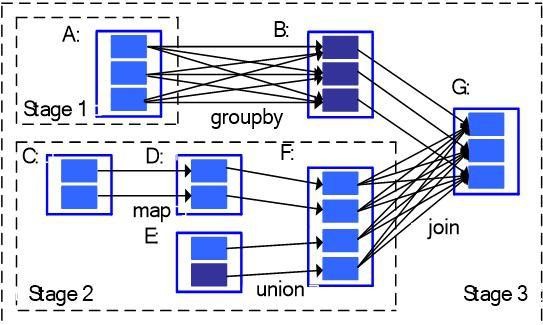
1. Zookeeper 的 Scheme 认证方式不包括以下哪项？（B）[单选题] A.digestB.sasl C.auth D.world
2. MapReduce 过程中，以下属于 Shuffle 机制的是？(AD)[多选题] A.CopyB.Partition C.Combine D.Sort/Merge
3. FusionInsight Manager 可以对哪些项目进行健康检查？（BC）[多选题] A.角色
4. 主机
5. 服务D.实例
6. FusionInsight Manager 与外部管理平台对接时，支持哪些接口？（BCD）[多选题] A.VPN

B.Syslog C.FTP D.SNMP

1. 下列选项中适合 MapReduce 的场景有？(D)[单选题] A.实时交互计算B.迭代计算C.流式计算D.离线计算
2. 以下关于 FusionInsight 中 CarbonData 说法正确的有？（ABCD）[多选题] A.使用 Carbon 的目的是对大数据即席查询提供超快速响应。
3. Carbon 使用轻量级压缩和重量级压缩的组合压缩算法压缩数据，可以减少 60%-80%数据存储空间，大大节省硬件存储成本。
4. Carbon 是一种新型的 Apache Hadoop 本地文件格式，使用先进的列式存储.索引.压缩和编码技术，以提高计算效率，有助于加速超过 PB 数量级的数据查询，可用于更换的交互查询。D.Carbon 也是一种将数据源与 Spark 集成的高性能分析引擎。
5. HDFS 联邦机制下，各 NameNode 间元数据是不共享的。（A）[单选题] A.TRUEB.FALSE
6. ApplicationMaster 采用轮询的方式通过 RPC 协议向 ResourceManager 申请和领取资源。（A）[单选题] A.TRUE B.FALSE
7. 在 YARN 的任务调度中，一旦 ApplicationMaster 申请到资源后，使与对应的ResourceManager 通信，要求它启动任务（B）[单选题]A.TRUE

B.FALSE

1. 如图所示，下列哪些选项属于宽依赖（AD）[多选题]



A.groupby：A->B B.map:C-

>D

C.union：D&E->F D.join：B&F->G

1. HFS 的出现解决了需要在 HDFS 中存储大量的小文件（10MB 以下）。同时也要存储一些大文件（10MB 以上）的混合的场景（B）[单选题]A.TRUE B.FALSE
2. 下列哪个命令是从 HDFS 下载目录/文件到本地的？(C)[单选题] A.dfs -put
3. dfs -cat
4. dfs -get
5. dfs -mkdir
6. 以下选项中，属于 HDFS 架构关键特性的是？(ABCD)[多选题] A.HA 高可靠性
7. 健壮机制
8. 元数据持久化机制D.多方式访问机制
9. HBase 的主 HMaster 是如何选举的？(C)[单选题] A.由 RegionServer 进行裁决B.HMaster 为双主模式，不需要进行裁决C.通过 ZooKeeper 进行裁决
10. 随机选举
11. 关于 Hive 与 Hadoop 其他组件的关系。以下描述错误的是？（D）[单选题] A.Hive 最终将数据存储在 HDFS 中

B.Hive 是 Hadoop 平台的数据仓库工具C.HQL 可以通过 MapReduce 执行任务D.Hive 对 HBase 有强依赖

1. Hive 不支持超时重试机制。（B）[单选题] A.TRUEB.FALSE
2. FusionInsight 中，HiveServer 将用户提交的 HQL 语句进行编译，解析成对应的 Yarn 任务.Spark 任务或者 HDFS 操作，从而完成数据转换，分析。（A）[单选题]
3. true B.False
4. HBase 的 Region 是由哪个服务进程来管理的？(A）[单选题] A.HRegionServer
5. ZooKeeper C.HMaster D.DataNode
6. Streaming 的处理节点 Bolt 中，可以完成以下哪些操作？(ABCD)[多选题] A.连接运算B.过滤（Filter） C.连接数据库

D.业务处理

1. Flume 传输的可靠性是指 Flume 在传输数据过程中，当下一跳的 Flume 节点故障或者数据接收异常时，可以自动切换到另外一路上继续传输？(A)[单选题]
2. True B.False
3. FusionInsight HD 集群中包含了多种服务，每种服务又由若干角色组成，下面哪些是服务的角色？(BC)[多选题]A.HDFS
4. NameNode C.DataNode D.HBase
5. FusionInsightHD 系统中使用 Streaming 客户端 Shell 命令提交了拓扑之后，使用 Storm UI 查看发现该拓扑长时间没有处理数据，可能原因有？(ABCD)[多选题]
6. 查看客户端异常堆栈，判断是否客户端使用问题
7. 查看主 Nimbus 的运行日志，判断是否 Nimbus 服务端异常
8. 查看 Supervisor 运行日志，判断是否 Supervisor 异常
9. 查看 Worker 运行日志
10. 假设 HDFS 在写入数据是只存 2 份，那么在写入过程中，HDFS Client 先将数据写入DataNode1，再将数据写入 DataNode2。(B)[单选题]
    1. True
    2. False
11. 容力量调度器在尽心资源分配，现有同级的 2 个队列 Q1 和 Q2，他们的容量均为 30，期中Q1 已使用 8. Q2 已使用 14，则会优先将资源分配 Q1。(A)[单选题]
    1. True
    2. False
12. HBase 中包含的一些典型的 Filter 有哪些？(ABCD)[多选题]
    1. SingleColumnValueFilter
    2. FilterList
    3. RowFilter
    4. KeyOnlyFilter
13. Flink 流式处理的数据源类型包括？(ABCD)[多选题]
    1. Collections
    2. JDBC
    3. Socket streams
    4. Files
14. server.channels.ch1.transactionCapacity 表示事务大小，即当前 channel 支持事务处理的事件个数，可和 source 的 batchsize 设置为同样大小，但不能小于 batchsize。(A)[单选题]
    1. True
    2. False
15. Flume 架构中，一个 sink 可以连接多个 channel。(B)[单选题]
    1. True
    2. False
16. 以下关于 Flink 关键特性描述不正确的是？（A）[单选题] A.SparkStreaming 与 Flink 相比，时延更低
17. Flink 流式处理引擎能够同时提供支持流处理和批处理应用的功能
18. 与 FusionInght HD 中的 Streaming 相比，Flink 具有更高的吞吐量D.checkpoint 实现了 Flink 的容错
19. FusionInsight 是华为企业级大数据存储，查询，分析的统一平台，能够帮助企业快速构建海量数据信息处理系统，通过对海量信息数据实时与非实时的分析挖掘，发现全新价值点和企业商机。(A)[单选题]
20. True B.False
21. Kafka Cluster Mirroring 工具可以实现以下哪项功能？（A）[单选题] A.Kafka 跨集群数据同步方式
22. Kafka 单集群内数据备份
23. Kafka 单集群内数据恢复D.以上全不正确
24. FusionInsightTool 是为技术支持工程师和维护工程师提供的一套健康检查工具，能够检查集群相关节点，服务的健康状态，中潜在的问题。并生成健康检查报告。方便技术支持工程和维护工程师快速了解系统的健康状况。（A）[单选题]A.TRUE B.FALSE
25. FusionInsight 产品中，关于 Kafka 说法不正确的是？(B)[单选题] A.Kafka 强依赖于 ZooKeeper
26. Kafka 的服务端可以产生消息
27. Kafka 部署的实例个数不得小于 2
28. Consumer 作为 kafka 的客户端角色专门进行消息的消费。
29. 为了提高 Kafka 的容错性，Kafka 支持 Partition 的复制策略，以下关于LeaderPartition 和 Follower Partition 的描述错误的是（C）[单选题]
    1. Kafka 针对 Partition 的复制需要选出一个 Leader。由该 Leader 负责 Partition 的读写操作。其他的副本节点只是负责数据同步
    2. 由于 LeaderServer 承载了全部的请求压力。因此从集群的整体考虑，Kafka 会将 Leader 均衡的分散在每个实例上，来确保数据均衡
    3. 一个 Kafka 集群各个节点间不可能互为 Leader 和 Flower
    4. 如果 Leader 失效。那么将会有其他 follower 来接管（成为新的 Leader）
30. Kafka 是一个高吞吐.分布式.基于发布订阅的消息系统，利用 Kafka 技术可在廉价 PC Server 上搭建起大规模消息系统。（A）[单选题]A.TRUE B.FALSE
31. flink 是一个批处理和流处理结合的统一计算框架，其核心是一个数据分发以及并行计算的流数据处理引擎（A）[单选题]A.True B.False
32. HBase 集群定时执行 Compaction 的目的是什么？(AC)[多选题] A.减少同一个 Region 同一个 ColumnFamily 下的文件数目B.减少同一个 Region 的文件数目C.提升数据读取性能

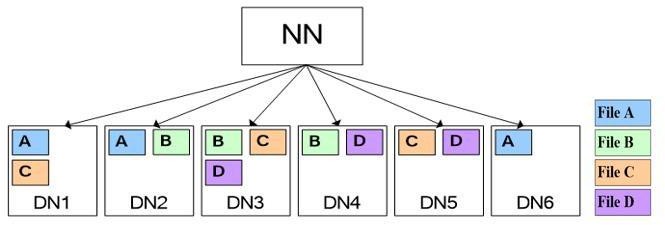
D.提升数据写入能力

1. 下列关于 Flink barrier 描述错误的是？（C）[单选题]
2. 一个 barrier 将本周期快照的数据与下一个周期快照的数据分隔开来
3. barrier 是 Flink 快照的核心
4. 在插入 barrier 的时候，会暂时阻断数据流
5. barrier 周期性插入到数据流中，并作为数据流的一部分随之流动
6. Flume 进程级联时，以下哪些 Sink 类型用于接收上一跳 Flume 发送过来的数据？（BCD） [多选题]
7. Null Sink
8. thrift sink C.HDFS sink D.avro sink
9. Flink 可以基于以下哪些窗口进行统计？（ABCD）[多选题] A.时间窗口B.滑动窗口C.会话窗口

D.countWindow

1. 关于 FusionInsight HD Streaming 的 Supervisor 描述正确的是？(B)[单选题] A.Supervisor 是在 Topology 中接受数据然后执行处理的组件
2. Supervisor 负责接受 Nimbus 分配的任务，启动和停止属于自己管理的 Worker 进程
3. Supervisor 负责资源分配和任务调度D.Supervisor 是运行具体处理逻辑的过程
4. Hadoop 中，如果 yarn.scheduler.capacity.root.QueueA.minim.m-user-limit-percent 设置为 50，下面说法错误的是？（D）[单选题]
5. 一个用户提交任务，可以使用 QueueA 的 100%的资源。
6. 如果 QueueA 中已经有 2 个用户的任务运行，这时第 3 个用户提交的任务需要等待释放资源。
7. QueueA 中必须保障每个用户至少得到 50%的资源。
8. QueueA 中的每个用户最多只能获得 50%的资源。
9. 关于 Flume，下列说法错误的是？（CD）[多选题] A.Flume 级联节点之间的数据传输支持加密
10. Flume 支持多级联和多路复用
11. Source 到 Channel 到 Sink 等进程内部有加密的必要
12. Flume 级联节点之间的数据传输不支持压缩
13. 以下关于 Kafka logs 中 segment file 的说法正确的是？（ABD）[多选题] A.通过索引信息可以快速定位 message
14. 稀疏存储，即将原来的完整数据，只间隔的选择多条进行存储
15. 是 index 元数据全部映射到 memory，可以避免 segment file 的 index 数据 IO 磁盘操作
16. 索引文件稀疏存储，可以大幅度降低 index 文件元数据占用空间大小
17. Streaming 主要通过 zookeeper 提供以下的哪项实现事件侦听？（B）[单选题] A.分布式锁机制B.Watcher C.Checkpoint D.ACK
18. Zookeeper 在分布式应用中的主要作用不包括一下哪些选项？（C）[单选题] A.选举 Master 节点B.保证各节点上数据的一致性C.分配集群资源

D.存储及群中服务器信息

1. Hadoop 集群规模很大时，数据的分布情况会非常关键，用户需要根据数据分布情况，决定集群是否扩容，数据是否需要做负载均衡等，fusionInsight 资源分布监控说法正确的是？（ABCD）[多选题]
2. 可以帮助用户快速聚焦在最关键的资源消耗上。
3. 通过每个服务主业的资源分布查看界面，看到关键的资源分布情况。
4. 通过 Dashboard 界面，可以查看到主机资源分不清哪个矿，例如内存占用率在 50-75%的主机列表，并提供链接跳转。
5. 帮助用户迅速找到资源消耗最高的点，采取适当的措施。
6. 以下关于华为大数据解决方案说法正确的是（ABCD）[多选题] A.GaussDB 是一款开源数据库产品
7. Farmer 是数据服务框架
8. FusionInsightManager 是一个分布式系统管理框架，管理员可以通过多接入点操控分布式集群
9. FusionInsight HD 是基于开源大数据软件 Hadoop 的增强版本
10. 下图展示了文件 A,B,C 的存放位置。其中文件 A 和 D 具有关联性，它们的存储位置符合Colocation 同分布策略（B）[单选题]

A.TRUE B.FALSE

1. HDFS 中 Namenode 的主备仲裁，是由哪个组件控制的（D）[单选题] A.HDFS ClientB.NodeManager C.ResourceManager

D.ZooKeeper Failover Controller

1. SparkStreaming 计算基于 DStream 将流式计算分解成一系列短小的批处理作业（A）[单选题]A.TRUE B.FALSE
2. 导入数据到 Hive 表时，不会检查数据合法性，只会在读取数据时候检查。（A）[单选题] A.TRUEB.FALSE
3. Topology 的处理逻辑都在 Bolt 中（A）[单选题] A.TRUEB.FALSE
4. 安装 FusionInsight HD 的 Streaming 组件是，Nimbus 角色要求安装几个节点？（C）[单选题]A.4

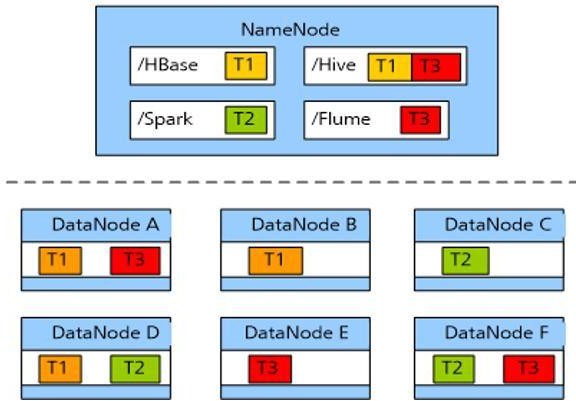
B.3

C.2

D.1

1. Flume 的 properties.protertises 配置文件中可以配置多个 channel 来传输数据（A） [单选题]
2. True B.False
3. 驱动型 source 是 flume 周期性主动去获取数据（B）[单选题] A.True
4. False
5. FusionInsight HD 数据高可靠性体现在以下哪些方面？（ABCD）[多选题] A.硬盘热插拔
6. 关键数据掉电保护
7. 跨数据中心容灾
8. 第三方备份系统集成
9. FusionInsight HD 系统审计日志不可以记录下面哪些操作？（C）[单选题] A.手动清除告警

B.启停服务实例C.查询历史监控D.删除服务实例

1. 下图展示了 HDFS 的标签存储策略，观察下图，HBase 的数据会被存储到哪些数据节点上（AB）？[多选题]
2. DataNode A
3. DataNode B
4. DataNode E
5. DataNode F
6. HBase 的数据文件 HFile 中一个 KeyValue 格式包含 Key，Value，TimeStamp，KeyType 等内容（A）[单选题]A.TRUE B.FALSE
7. 在有 N 个节点 FusionInsight HD 集群中部署 HBase 时，推荐部署（）个 HMaster 进程，（）个 RegionServer 进程（C）？[单选题] A.3，N

B.N，N C.2，N D.2，2

1. Flink 的数据转换操作在以下哪些环节中完成（B）？[单选题] A.channel
2. Transformation
3. sink
4. source
5. Kafka 消息传输保障通常有以下哪几种（ABD）？[多选题] A.仅有一次（Exactly Once）
6. 最少一次（At Least Once）
7. 最多三次（At Most Three Times）
8. 最多一次（At Most Once）
9. Kafka Logs segment 文件命名规则是 partion 全局的第一个 segment 从 0 开始，后续每个 segment 文件名为上一个全局 partion 事务最大 offset（偏移 message 数）。（A）[单选题]
10. TRUE
11. FALSE
12. HBase 的主要特点有哪些？（ABCD）[多选题] A.面向列B.高性能 C.可伸缩 D.高可靠性
13. YARN 中，“从”节点负责以下哪些工作？(BC)[多选题] A.集群中所有资源的统一管理和分配
14. 监督 container 的生命周期管理
15. 监控每个 Container 的资源使用（内存、CPU 等）情况D.管理日志和不同应用程序用到的附属服务
16. Spark 有哪些特点？(ABCD)[多选题] A.快
17. 灵
18. 轻
19. 巧
20. Hive 中“Groupby”指的是通过一定的规则将每一个数据集划分成若干个小的数据集， 然后针对若干个小的数据集进行数据分组处理（A）[单选题]
21. True B.False
22. Flink 适用于高并发处理数据、毫秒级时延的应用（A）[单选题] A.True
23. False
24. 与开源 sqoop 相比，loader 具有哪些增强特性？(ABCD)[多选题] A.高可靠性
25. 高性能
26. 安全性D.图形化
27. FusionInsight HD loader 可以将 HDFS 数据导出到以下哪些目标端？(ABCD)[多选题] A.orale 数据库B.SFTP 服务器

C.BD2 数据库D.FTP 服务器

1. FusionInsight HD 的 Loader 在创建作业时，连接器有以下哪些作业？（ABCD）[多选题] A.配置作业与内部数据源连接方式B.配置作业与外部数据源连接方式C.确定有转换

D.提供优化参数，提高数据导入导出性能

1. 当 zookeeper 集群的节点数为 5 节点时，请问集群的容灾能力和多少节点是等价的？（D） [单选题]A.3

B.4

C.5

D.以上全不正确

1. FusionInsight 家族包含下列哪些子产品？（ABCD）[多选题] A.FusionInsight FarmerB.FusionInsight Miner C.FusionInsight HD D.GaussDB 200
2. FusionInsight Manager 用户权限管理不支持哪个配置？(D)[单选题] A.给用户配置角色
3. 给用户组配置角色
4. 给角色配置权限 D.给用户组配置权限
5. Yarn 调度器分配资源的顺序，下面哪一个描述的正确？(C)[单选题] A.任意机器——>本地资源——>同机架B.任意机器——>同机架——>本地资源C.本地资源——>同机架——>任意机器D.同机架——>任意机器——>本地资源
6. 以下哪个不属于 Hadoop 中 MapReduce 组件的特点？(C)[单选题] A.高容错

B.良好的扩展性C.实时计算

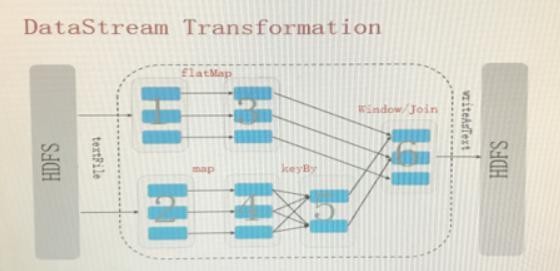
D.易于编程

1. Hbase 的最小存储单元是 Region。(A)[单选题] A.TRUEB.FALSE
2. Hbase 的某张表的 RowKey 划分 Splitkey 为 9.E.a.z.请问表里面有几个 Region?(C)[单选题]A.6

B.3

C.5

D.4

1. 根据数据流如何在两个 TransFormation 之间传输数据，数据流可以分为那些类型？ (AC)[多选题]

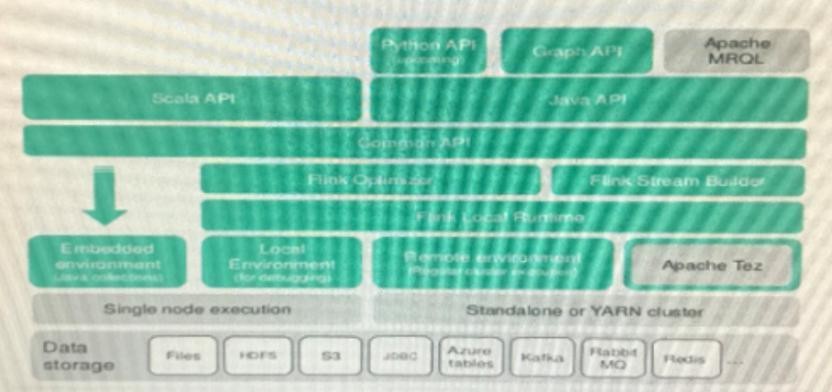
A.一对一流B.distributing 流C.一对多流D.redistributig 流

1. 为了保障流应用的快照存储的可靠性，快照主要存储在哪里？(D)[单选题] A.jobManager 的内存中B.可靠性高的单机数据库中C.本地文件系统中

D.hdfs 中

1. 在 fusionlnsight 产品中，关于 kafka 的 topic.以下描述不正确的是？(B)[单选题] A.topic 的 partition 数量可以创建时配置
2. 每个 topic 只能被分成一个 partition 区
3. 每条发布到 kafka 的消息都有一个类别，这个类别被称为 topic.也可以理解为一个存储消息的队列
4. 每个 partition 在存储层面对应一个 log 文件，log 文件中记录了所有的消息数据
5. Kafa 集群在运行期间，直接依赖于下面哪些组件？(B)[单选题] A.spark
6. zookeeper C.hdfs D.hbase
7. 关于 fusioninsight manager 界面 hive 日志收集的描述中，哪个不对？(C)[单选题] A.可指定实例进行日志收集，比如指定收集 metaStore 的日志B 可指定节点 ip 进行日志收集，例如仅下载某个 ip 的日志
8. 可指定特定用户进行日志收集，例如仅下载 user 用户的日志
9. 可指定时间端进行日志收集，不如只收集 2016-1-1 到 2016-1-10 的日志
10. fusioninsight manager 对于管理操作，下列错误的是？(C)[单选题] A.可对服务进行启停重启
11. 可以添加和卸载服务
12. 可设置不常用服务隐藏或显示D.可查看服务的当前状态
13. 为减少写入磁盘的数据量，mapreduce 支持对 MOF（mapoutfile）进行压缩后再写入。(A)[单选题]A.True B.False
14. ResourceManager 采用高可用方案，当 Active resourcemanager 发现故障时，只能通过内置的 zookeeper 来启动 standby 的 resourcemanager，将其状态切换为 active。(B)[单选题]A.True B.False
15. Hive 是一种数据仓库处理工具，使用类 SQL 的 HiveQL 语言实现数据查询功能，所有Hive 的数据都存储在 HDFS 中。(A)[单选题]
16. True
17. False
18. 在 WebHCat 架构中，用户能够通过安全的 HTTPS 协议执行以下哪些操作?(D)[单选题] A.执行 Hive DDL 操作。
19. 运行 MapReduce 任务
20. 运行 Hive HQL 任务D.以上全都正确
21. Flink 的兼容性体现在以下哪些方面?(ABCD)[多选题]

A.YARN 能够作为 F1ink 集群的资源调度管理器

1. 能够与 Hadoop 原有的 Mappers 和 Reducers 混合使用
2. 能够使用 Hadoop 的格式化输入和输出口D.能够从本地获取数据
3. Kafka Consumer 写数据总体流程是，Consumer 连接指定 TopicPartition 所在的LeaderBroker.用主动获取方式从 Kafka 中获取消息。(B)[单选题]
4. true
5. false
6. 图形化的健康巡检工具 FusionInsightTool 由 FusionCare 和 SysChecker 两部分组成。(A)[单选题]A.true B.false
7. FusionInsight Manager 界面 上，当收到 Kafka 磁盘容量不足告警，且该告警的原因已经排除硬硬件故障时，系统管理员需要考虑扩容解决此问题。(A)[单选题]A.true B.false
8. 如图所示，Flink 流式理接口 DataStream API 支持的语言包括？(AC)[多选题]A.Java

B.C 语言C.Scala D.Python

1. 以下关于 HBase 的文件存储模块（HBase FileStream，简称 HFS）描述正确的有？ (ABCD)[多选题]A.应用在 FusionInsight HD 的上层应用

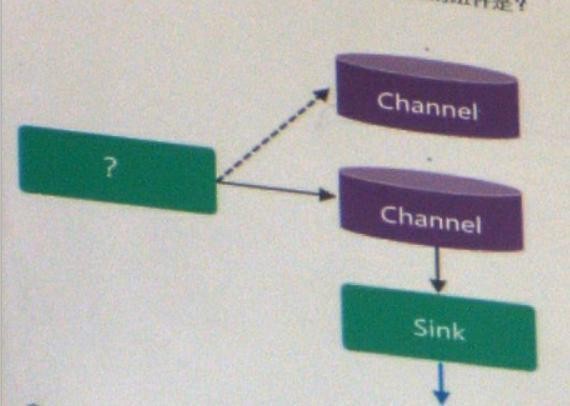
B.HFS 封装了 HBase 与 HDFS 的接口

C.为上层应用提供文件的存储.读取.删除等功能D.HFS 是 HBase 的独立模块

1. 若在消息处理过程中允许部分信息丢失，关闭消息可靠性处理机制的方式有以下哪些？

(ACD)[多选题]

* 1. 将参数 Config.Topology\_ACKKRS 设置为 0
  2. 将参数 Config.Topology\_ACKKRS 设置为 False
  3. Spout 发送消息时，使用不指定消息 messageID 的接口进行发送
  4. Bolt 发送消息时使用 Unanchor 方式发送

1. 在 Flink 技术架构中，以下哪项是流处理和批处理的计算引擎？(B)[单选题] A.StandaloneB.Runtime C.FlinkCore D.DataSteam
2. 下图为 Flume 数据传输架构，图中“？”号处的组件是？(C)[单选题]A.Interceptor B.Channel Processor

C.Channel Selector D.以上全不正确

1. 采用 Flume 传输数据过程中，为了防止因 Flume 进程重启而丢失数据，可以使用以下哪种 Channel 类型？(BC)[多选题]A.Memory Channel B.JDBC Channel C.File Channel D.HDFS Channel
2. 华为 FusionInsight HD 是国内首家符合国家金融等级保护的大数据平台，其安全性现在以下哪些方面？(D)[单选题]

A.系统安全B.权限认证C.数据安全

D.以上全都正确

1. 在华为大数据解决方案中，hadoop 底层包含以下那些组件？（ABD） [多选题] A.fink
2. hive
3. miner D.spark
4. spark 的核心模块是？（B）[单选题] A.spark streamingB.spark core C.mapreduce D.spark sql
5. 关于 hbase 存储模型的描述正确的是？（ABCD）[多选题]
6. 即使是 key 值相同，qualifier 也相同的多个 keyvalue.也可能有多个，此时使用时间戳来区分
7. 同一个 key 值可以关联多个 value
8. keyvalue 中期有时间戳，类型等关键信息
9. 每一个 keyvalue 都有一个 qualifier 标识
10. hbase 的底层数据以（）的形式存在的？（A）[单选题] A.keyvalueB.列存储 C.行存储 D.实时存储
11. 二级索引为 hbase 提供了按照某些列的值进行索引的能力，二级索引先查索引表，再定位到数据表中的位置，不用全表扫描（A）[单选题]A.true B.false
12. colocation(同分布)文件级的同分布实现文件的快速访问，避免了因数据迁移带来的大 量网络开销（A）[单选题]A.true B.false
13. fusioninsight HD 集群包含多种服务，每种服务又有若干个角色组成，下面那些是角色?（AC）[多选题] A.namenode B.hbase C.datanode
14. fusininsight HD 平台中，那些组件支持对列表加密？（CD） [多选题] A.HDFS
15. Flink
16. hbase D.hive
17. Hadoop 系统中，如果 HDFS 文件系统的备份因子是 3，那么 MapReduce 每次允许 task 都是从 3 个有副本的机器上传输需要处理的文件.（B） [单选题]
18. True B.False
19. YARN 容量调度器的主要特点有哪些？(ABCD)[多选题] A.容量保证
20. 动态更新配置文件C.灵活性D.多重租赁
21. Kafka Cluster Mirroring 工具可以实现以下哪些功能？（A）[单选] A.Kafka 跨集群数据同步方案
22. Kafka 单集群内数据备份
23. Kafka 但集群内数据恢复D.以上全不正确
24. 硬件故障被认为是常态，为了解决这个问题，HDFS 设计了副本机制。默认情况下，一份文件，HDFS 会存（A）份？[单选题]
25. 3
26. 5
27. 2
28. 4
29. 现有 3 个机架，有一个文件需要存 3 份，其中副本 1 和副本 2 存放在与 client 相同的机架且不同的服务器上。根据 HDFS 的副本放置策略一定要存放在其他机架。（B）[单选题] A.TRUEB.FALSE
30. Flume 架构中，一个 Source 可以连接多个 Channel。（A）[单选题]
31. TRUE
32. FALSE
33. FusionInsight HD 的 HBase 中保存了一张用户信息表 msg\_table，RowKey 为用户 id，其中一类为用户昵称，现在按先后顺序往这列写 keyValue：001：Li，001：Mary，001：Lily， 请问 scan‘msg\_table’，（VERSIONS=>2）会返回哪几条数据？（D）[单选题]
    1. 1：Li
    2. 1：Lily C.001：Li，001：Mary，001：Lily D.001：Mary，001：Lily
34. Kafka 中 partition replication 之间同步数据，从 partition 的 leader 复制数据到follower 需要线程（replicationFetcherThread），Follower（一个 follower 相当于 consumer）主动从 leader 批量拉取消息的，这极大提高了吞吐量。（A）[单选题]A.True B.False
35. 以下哪些选项属于 FusionInsight 数据安全性？（BC）[多选] A.操作系统安全加固B.组件数据加密 C.数据完整性校验

D.用户权限认证管理

1. NodeManager 的内存和 CPU 的数量，是通过下列哪些选项进行配置？（BCD）[多选题] A.Yarn.scheduler.capacity.root.QueueA.maximum-capacity B.Yarn.nodemanager.resource.cpu-vcore C.Yarn.nodemanager.vmem-pmom-ratio

D.Yarn.modemanager.resource.memory-mb

1. 下列选项中，哪些是 MapReduce 一定会有的过程？（BCD）[多选题] A.CombineB.Map C.Reduce D.Partition
2. 下列选项中，哪些是 Spark 的重要角色（ABD）[多选题]
   1. Nodemanager
   2. ResourceManager
   3. DataNode
   4. Driver
3. 以下属于 Hive SQL 中 DDL（数据定义语言）的是？（ABD）[多选题] A.修改表B.删除表 C.数据导入D.建表
4. 关于 HIVE 的描述不正确的是？（B）[单选题]
   1. Hive 最佳使用场景是大数据集的批处理作业
   2. Hive 可以实现在大规模数据集上实现低延迟快速的查询
   3. Hive 构建在基于静态批量处理的 Hadoop 之上，Hadoop 通常有较高的延迟并且在作业提交和调度的时候需要大量的开销
   4. Hive 查询操作过程严格遵循 HadoopMapReduce 的作用执行模型，Hive 将用户的 HiveQL 语句通过解释器转换为 MapReduce Hadoop 集群上
5. 在 Flink 中，checkpoint 机制能够保证应用在运行过程中出现失效时，从某一个检查点恢复，在此过程中，流快照是根据数据流入建立的（A）[单选题]
6. TRUE
7. FALSE
8. Hive 中的“Group by”指的是通过一定规则将一个数据集划分成若干个小的数据集，然后针对若干个小的数据集进行数据分组处理。(A)[单选题]
   1. True
   2. False
9. Hive 中~UNION ALL~操作符用于合并两个或多个 select 语句的结果集，结果集中，不允许有重复的值。(B)[单选题]
   1. True
   2. False
10. Structured Streaming 在 OutPut 阶段可以定义不同的数据写入方式，包括以下哪些方式？(BCD)[多选题]
11. General Mode
12. Append Mode
13. Complete Mode
14. Update Mode
15. Flume 数据采集的过程中，下列选项中能对数据进行过滤和修饰的是？（D）[单选题]
    1. Sink
    2. Channel
    3. Channel selector
    4. Interceptor
16. 以下关于 FusionInsight CTBase 的描述不正确的是？（B）[单选题]
17. CTBase 的读写数据接口，统一封装了行定义的接口，自动进行冷字段的合并和解析，不需要在应用程序中进行合并和解释
18. CTBase 是基于 HBase 的聚簇表开发框架
19. CTBase 提供了一套 WebUI 进行元数据定义，提供了只管医用的表设计工具，降低表设计的难度
20. CTBase 的 javaAPI 提供一套 HBase 连接池管理的接口，内部进行连接共享，减少客户端应用开发难度。
21. Kafka 是一个高吞吐，分布式，基于发布订阅的消息系统，利用 Kafka 技术可在廉价PC Server 上搭建起大规模消息系统。（A）[单选题]
22. TRUE
23. FALSE
24. YARN 中设置队列 QueueA 的最大使用资源量，需要配置哪个参数？(B)[单选题] A.yarn.scheduler.capacity.root.QueueA.minimum-user-limit-percent B.yarn.scheduler.capacity.root.QueueA.maximum-capacity C.yarn.scheduler.capacity.root.QueueA.minimum.user-limit-factor D.yarn.scheduler.capacity.root.QueueA.state
25. 如果某些 Containers 的物理内存利用率超过了配置的内存阈值，但所有 Containers 的总内存利用率并没有超过设置的 NodeManager 内存阈值,那么内存使用过多的 containers 仍可以继续运行。(A)[单选题]
    1. True
    2. False
26. 下图为 Spark&MapReduce 的性能对比数据，可以得出结论，与 MapReduce 计算相比， Spark 用 的资源，获得 倍性能？(D)[单选题]

A.1/10,4 B.1/8,3 C.1/8,4 D.1/10.3

1. HBase 中以下哪种场景会出发 Flush 操作？(ABCD)[多选题]
   1. HBase 定期刷新 Memstore，默认周期为 1 小时
   2. 当 WALs 中文件数量达到阈值时
   3. Region 中 MemStore 的总大小，达到了预设的 Flush Size 阈值
   4. MemStore 占用内存的总量和 RegionServer 总内存比值超出了预设的阈值大小
2. FusionInsight HD 系统中，HBase 支持动态扩展列。(A)[单选题] A.TrueB.False
3. 以下哪些选项属于 Hive 的数据存储模型？(E)[单选题] A.桶B.数据库C.分区 D.表

E.以上全都正确

1. Hive 中的这条命令”ALTER TABLE employeel ADD columns（columnl string）.”是什么含义？(C)[单选题]A.创建表B.删除表C.添加列

D.修改文件格式

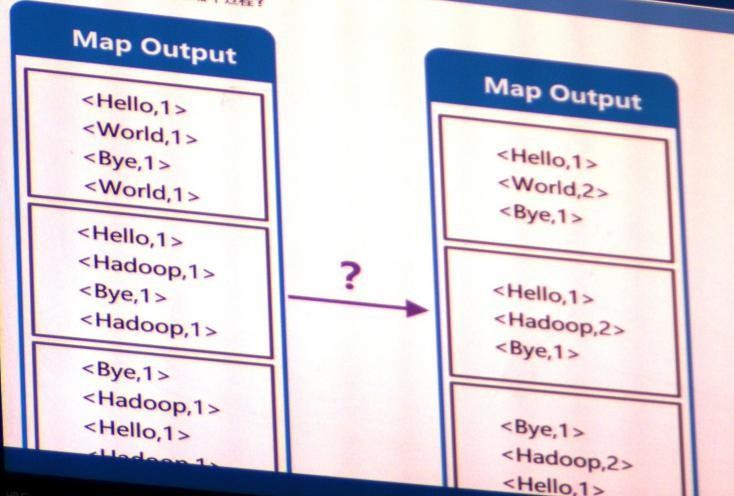
1. 下列关于 SparkStreaming 和 Streaming 比较说法不正确的是？(B)[单选题] A.SparkStreaming 是一个微批处理框架，事件需要积累到一定量时才进行处理B.Streaming 的执行逻辑是即时启动，运行完后再回收C.SparkStreaming 的吞吐量大约是 Streaming 的 2-5 倍D.SparkStreaming 事件处理时延比 Streaming 更高
2. 以下关于 Flink 与其它组件交互说法正确的是？(ABCD)[多选题] A.Flink 任务的运行依赖 Yarn 来进行资源的调度管理
3. Flink 的 checkpoint 的实现依赖于 Zookeeper
4. Flink 可以将接收的组件发送给 Kafka
5. Flink 在 HDFS 文件系统中读写数据
6. 创建 Loader 作业时，可以在以下哪个步骤中设置 Map 数？(A)[单选题] A.输出B.输入设置C.转换

D.基本信息

1. Channels 支持事务，提供较弱的顺序保证，可以连接任何数量的 Source 和 Sink。(A)[单选题]A.True B.False
2. Kafka 集群中，Kafka 服务端的角色是？(A)[单选题] A.BrokerB.Consumer C.ZooKeeper D.Producer
3. KafkaProducer 读数据总体流程是，Producer 连接任意存活的 Broker，请求制定 Topic、Partition 的 Leader 元数据信息，然后直接与对应的 Broker 直接连接，发布数据。(B)[单选题]

A.True B.False

1. 下图所示的 Map 阶段，是描述哪个过程？（D ）[单选题]



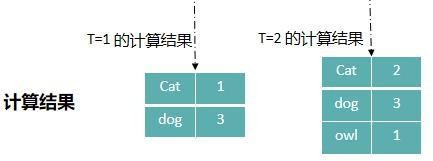
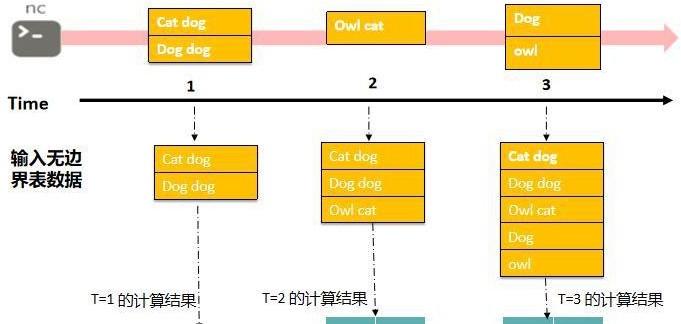
A.Partition B.Sort C.Splill/Merge D.combine

1. Spark 根据 RDD 的依赖关系来划分 Stage，调度器从 DAG 图末端出发，逆向遍历整个依赖关系链，遇到窄依赖就断开，遇到宽依赖就将其加入当前 Stage。(B)[单选题]
2. TRUE
3. FALSE
4. 关于 DataSet，下列说法不正确的是？（D）[单选题]
5. DataSet 不需要反序列化就可执行大部分操作
6. DataSet 是一个由特定域的对象组成的强类型集合
7. DataSet 与 RDD 高度类似，性能比 RDD 好
8. DataSet 执行 sort，filter，shuffle 等操作需要进行反序列化
9. Hbase 中 OpenScanner 的过程，会创建两种不同的 Scanner 来读取 HFile 和 MemStore 的数据，HFile 对应的 Scanner 为 StoreFileScanner，MemStore 对应的 Scanner 为MemStoreScanner。( A)[单选题]
10. TRUE
11. FALSE
12. 执行 HBase 读取数据业务，需要读取哪几部分数据？（BD）[多选题]
13. HMaster
14. HFile
15. Hlog
16. MemStore
17. 使用 HBase 客户端批量写入 10 条数据，某个 HRegionServer 节点上包含该表的 2 个Region，分别为 A 和 B，10 条数据中有 6 条属于 A，4 条属于 B，请问写入这 10 条数据需要向该 HRegionServer 发送几次 RPC 请求？(D)[单选题]A.10 B.6

C.2

1. Hive 不适用于以下哪个场景 （D）[单选题]
2. 非实时分析，例如日志分析，统计分析
3. 数据挖掘，例如用户行为分析，兴趣分析，区域展示
4. 数据汇总，例如每天，每用户点击数，点击排行
5. 实时在线数据分析
6. 关于 Hive 建表基本操作，描述正确的是（C）[单选题]
7. 一旦表建好，不可再修改表名
8. 一旦表建好，不可再增加新列
9. 创建外部表时需要制定 external 关键字
10. 一旦表创建好，不可再修改列名
11. FusionInsight HD 系统中 Hive 支持的存储格式包括？（ABC）[多选题]
12. TextFile
13. SequenceFile
14. RCFile
15. HFile
16. 在 Flink 的运行流程中，负责申请资源的角色是？（B）[单选题]
17. ResourceManager
18. JobManager
19. Client
20. TaskManager
21. 在 FuisonInsightHD 中，创建 Loader 作业的进行数据转换的正确步骤是？（A）[单选题]
22. 输入设置，转换，输出
23. 抽取，转换，输出
24. 加载，转换，输出
25. 加载，转换，抽取
26. Flume 用于收集数据，其传输的数据基本单位是？（C）[单选题]
27. Split
28. Block
29. Event
30. Packet
31. FusionnInsiget HD 系统中，Flume 数据流在节点内不需要经过哪个组件？（C）[单选题]
32. Source
33. Sink
34. Topic
35. Channel
36. 传统数据处理的数据规模的单位是？（D）[单选题]
37. TB
38. EB
39. PB
40. GB
41. 下列哪个命令是删除文件的？(C)[单选题]
42. dfs -clear
43. dfs -ls
44. dfs -rm
45. dfs -del
46. 下图为 HDFS 分级存储的配置，如果 Block 的副本数为 4，则以下说法错误的是？（D） [单选题]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 策略 ID | 名称 | Block 放置位置  （副本书） | 备选存储策略 | 副本的存储策略 |
| 15 | LAZY\_PERSITST | RAM\_DISK:1,DIS K:n-1 | DISK | DISK |

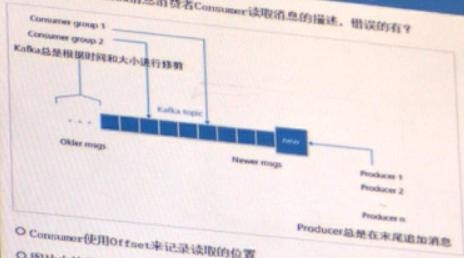
1. 第 1 个 Block 副本将写入 RAM\_DISK,其余副本写入 DISK
2. 如果是第 1 个副本之外的其他副本写入失败，则尝试写入“副本的备选存储策略”指定的存储策略
3. 如果第 1 个 Block 副本写入 RAM\_DISK 失败，则尝试写入”备选存储策略”的存储类型
4. 4 个 Block 都将写入 RAM\_DISK
5. Hadoop 平台中，要查看 YARN 服务中一个 application 的信息，通常需要使用什么命令？（D）[单选题]
6. container
7. jar
8. application-attempt
9. Application
10. Spark 适用于以下哪些场景？（ABCD）[多选题]
11. 交互式查询
12. 实时流处理
13. 批处理
14. 图计算
15. 下图展示了 Structured Streaming 的计算模型。通过观察可以得出最终 T3 的计算结果是（B）[单选题]
16. Dog 1 ,owl 1
17. Cat 2,dog 4,owl 2
18. Cat 2,dog 3,owl 1
19. Cat 1,cat 1,dog 2,dog 2,owl 2
20. SparkStreaming 容错机制是指 RDD 中任意的 Partition 出错，都可以根据其父 RDD 重新计算生成，如果父 RDD 丢失，则需要去磁盘中查找原始数据。（A）[单选题]
21. TRUE
22. FALSE
23. HBase 元数据 Meta Region 路由信息保存在哪里？（A）[单选题] A.ZookeeperB.Meta 表

C.HMaster D.Root 表

1. Hbase 的 BloomFiliter 是用来过滤数据的？（B）[单选题]
2. TRUE
3. FALSE
4. 执行 HBase 读数据业务，需要读取哪几部分数据？（CD）[单选题]
5. HLog
6. HMaster
7. HFile
8. MemStore
9. FusionInsight 中 Hive 包含 2 个角色，HiverServer 和 MetaStore?（A）[单选题]
10. TRUE
11. FALSE
12. Colocation 同分布文件级的同分布实现文件的快速访问，避免了因数据迁移带来的大量网络开销（A）[单选题]
13. TRUE
14. FALSE
15. 关于 Hive 与传统数据仓库的对比，以下描述错误的是？（D）[单选题]
16. Hive 元数据存储独立于数据存储之外，从而解耦合元数据和数据，灵活性高，而传统数据仓库数据应用单一，灵活性低
17. Hive 基于 HDFS 存储，理论上存储量可无限扩展，而传统数据仓库存储量会有上限
18. 由于 Hive 的数据存储在 HDFS 中，所以可以保证数据的高容错，高可靠
19. 由于 Hive 基于大数据平台，所以查询效率比传统数据仓库快
20. YARN 的基于标准调度，是对下列选项中的哪个进行标签化？（ C）[单选题]
    1. AppMaster
    2. ResourceManager
    3. NodeManager
    4. Container
21. 对于 Hive 中关于普通表和外部表描述不正确的是？（B）[单选题]
    1. 默认创建普通表
    2. 删除外部表时，只删除外部表数据而不删除元数据
    3. 外部实质是将已存在于 HDFS 上的文件路径跟表关联起来
    4. 删除普通表时，元数据和数据同时被删除
22. 在 FusionInsight 产品中，关于创建 Kafka 的 Topic，以下哪些描述是正确的？（C） [单选题]
23. 在创建 Kafka 的 Topic 时，必须设置 Partition 个数
24. 在创建 Kafka 的 Topic 时，必须设置 Partition 副本个数
25. 设置多副本可以增强 Kafka 服务的容灾能力
26. 以上全都正确
27. Zookeeper 的增强性包括在审计日志中添加 ephemeralnode 被删除的审计日志（A）[单选题]
28. TRUE
29. False
30. 下列选项中，关于 Zookeeper 可靠性含义说法正确的是？（D）[单选题]
    1. 可靠性通过主备部署模式实现
    2. 可靠性是指更新更新只能成功或者失败，没有中间状态
    3. 可靠性是指无论哪个 server，对外展示的均是同一个视图
    4. 可靠性是指一个消息被一个 Server 接收，它将被所有的 Server 接受
31. 华为 FusionInsight HD 行业成功案例都有哪些？（ABCD）[多选题]
    1. 数字政府
    2. 智慧园区
    3. 智慧交通
    4. 金融
32. FusionInsight Manager 会定时备份哪些元数据？（ABCD）[多选题]
    1. OMS
    2. LDAP Server
    3. DB Service
    4. Namenode
33. YARN 中默认的资源调度器是？（B）[单选题]
    1. FIFO 调度器
    2. 容量调度器
    3. Fair 调度器
    4. 以上全不是
34. 假设在写入数据时只存 2 份，那么写入过程中，HDFSClient 现将数据写入 DataNode1， 再将数据写入 Datanode2（B）[单选题]
35. TRUE
36. FALSE
37. Yarn-Client 和 Yarn-Cluster 主要区别是 ApplicationMaster 进程的区别（A）[单选题]
    1. TRUE
    2. FALSE
38. 部署 FusionInsightHD 时，同一集群内的 FlumeServer 节点建议至少部署几个？（D） [单选题]A.1

B.3

C.4

1. 以下哪些是 Kafka 实际的应用场景？（ABC）[多选题] A.网站活性跟踪
2. 日志收集
3. 聚合统计系统运营数据（监控数据） D.资源管理
4. FusionInsightHD 系统中，下面哪个方法不能查看到 Loader 作业执行的结果（C）[单选题]
5. 通过 Loader UI 界面查看
6. 通过 YARN 任务管理查看
7. 通过 NodeManager 查看
8. 通过 Manager 的告警查看
9. 如图，以下关于 Kafka 消息消费者 Consumer 读取消息的描述，错误的有？（D）[单选题]
10. Consumer 使用 Offset 来记录读取的位置
11. 图片中的蓝色框为 Kafka 的一个 Topic，即可以理解为一个队列，每个格子代表一条消息。
12. 生产者产生的消息逐条放到 Topic 的末尾
13. 消费者从右至左顺序读取消息
14. 在大数据时代，企业所面临的挑战有以下哪些？（ABCD）[多选题]
15. 企业各部门间数据分散，相同数据在各部门内部存储格式不一致。
16. 数据结构多样化。
17. 竞争对手的技术进步。
18. 数据存在噪音、缺失、存储类型不规范等问题，需要进行大量的数据预处理工作。
19. FusionInsight HD 中 HBase 默认使用什么组件作为其底层文件存储系统？（C）[单选题] A.FileB.Kafka C.HDFS

D.Memory 220.在很多小文件场景下，Spark 会起很多 Task，当 SQL 逻辑中存在 Shuffle 操作时，会大大增加 hash 分桶数，严重影响性能。FusionInsight 中，针对小文件的场景通常采

用（）算子，来对 Table 中的小文件生成的 partition 进行合并，减少 partition 数，从而避免在 shuffle 的时候，生成过多的 hash 分桶，提升性能？(D)[单选题]

1. group by
2. coalosce
3. connect
4. join
5. 下列关于 Worker（工作进程）、Executor（线程）、Task（任务）说法正确的是？（ABD） [多选题]
6. 每个 Executor（线程）可以运行多个 Task（任务）
7. 每个 Worker 可以运行多个 Executor（线程）
8. 每个 Worker 只能为一个拓扑运行 Executor（线程）
9. 每个 Executor（线程）可以运行不同组件（Spout 或 Bolt）的 Task（任务）
10. 以下关于 Kafka 的描述错误的有？（C）[单选题]
11. 使用 Scala、Java 语言实现
12. 用作活动流和运营数据处理管道的基础
13. 由 Apache Hadoop 开发，并于 2011 年开源
14. 具有消息持久化、高吞吐、实时等特性
15. 查看 kafka 某 Topic 的 partition 详细信息时，使用如下那个命令？（D） [单选题] A bin/kakfa-topics.sh --createB bin/kakfa-topics.sh --delete C bin/kakfa-topics.sh --list

D bin/kakfa-topics.sh --describe

1. 用户权限管理基于角色的访问控制(RBAC),提供可视化的多组统一的集群中用户权限管理。(A) [单选题] A.true B.false
2. 对高价值高度聚合的信息和知识的批次处理是大数据行业主要商业诉求（A）[单选题] A.trueB.false
3. Mapreduce 过程中，默认情况下，一个分片就是一个块，也是一个 mapTask.（A）[单选题]A.true B.false
4. FusionInsight 家族包含以下哪些子产品（ABCD）？[多选题]
   1. GaussDB 200
   2. FusionInsight Miner
   3. FusionInsight HD
   4. FusionInsight Farmer
5. 以下关于 ZooKeeper 关键特性中的原子性说法正确的是（B）？[单选题]
6. 客户端发送的更新会按照他们被发送的顺序进行应用
7. 更新只能全部完成或失败，不会部分完成
8. 一条消息被一个 server 接收，将被所有 server 接收D.集群中无论哪台服务器，对外展示的均是同一视图
9. Flume 支持监控并传输目录下新增的文件，可实现实时数据传输，以下描述的是哪一类Source（D）？[单选题]
10. http source
11. exec source
12. syslog source
13. spooling directory source
14. 以下哪一项不属于创建 Loader 作业时必选项（A）？[单选题]A. 优先级B.名称
15. 连接
16. 类型
17. 为了保证流应用的快照存储的可靠性，快照主要存储在（C）？[单选题] A.本地文件系统中
18. JobManager 的内存中
19. HDFS 中
20. 可靠性高的单机数据库中
21. FusionInsightHD 系统使用 Streaming 客户端 Shell 命令查看拓扑或者提交拓扑失败， 以下哪些定位手段是正确的（ABCD）？[多选题]
22. 查看 Supervisor 运行日志，判断是否 Supervisor 异常
23. 查看 Worker 运行日志
24. 查看主 Nimbus 的运行日志，判断是否 Nimbus 服务端异常D.查看客户端异常堆栈，判断是否客户端使用问题
25. 如果需要由数据生产者决定数据发送给目标 Bolt 的某一个确定的 Task，应选择以下哪种消息发布策略（C）？[单选题]
26. 局部字段分组
27. 广播分组C.直接分组D.全局分组
28. 华为大数据产品 FusionInsight HD 中的 Streaming 具有哪些关键特性（ABC）？[多选题] A.容灾能力
29. 灵活性
30. 消息可靠性D.可伸缩性
31. 以下关于 Hive SQ1 基本操作描述正确的是（D）？[单选题] A.加载数据到 Hive 时源数据必须是 HDFS 的一个路径
32. 创建外部表必须要指定 location 信息
33. 创建外部表使用 external 关键字，创建普通表需要指定 internal 关键字D.创建表示可以指定列分隔符
34. Flink 采用 checkpoint 机制保障应用程序运行中的容错性。（A）[单选题]
35. TRUE
36. FALSE
37. Flink 支持的时间操作类型包括以下哪些选项？（BD）[多选题]
38. 结束时间
39. 处理时间
40. 采集时间
41. 事件时间
42. Flink 中的（）接口用于流数据处理，（）接口用于批处理？（C）[单选题]
43. DataBatch API，DataStream API
44. Stream API，Batch API
45. DataStream API，DataSet API
46. Batch API， Stream API
47. 下列属于 Flume 数据监控的指标是？（ACD）[多选题]
48. Sink 写入数据量
49. DataNode 数量
50. Channel 缓存数据量
51. Source 接受数据量
52. 调用 Zookeeper 客户端命令中包含哪些信息？（ABCD）[多选题]
53. Ip 地址
54. 端口号
55. 服务器名称
56. 用户名
57. 以下关于 Zookeeper 的 Leader 节点在收到数据变更请求后的读写流程说法正确的是？（D）[单选题]
58. 仅写入内存
59. 同时写入磁盘和内存
60. 先写入内存再写入磁盘
61. 先写磁盘再写内存
62. 以下关于 Zookeeper 的 Leader 选举说法正确的是？（AB）[多选题]
63. 当实例 n 为奇数时，假定 n=2x+1，则成为 leader 节点需要 x+1 票
64. Zookeeper 选举 leader 时，需要半数以上的票数
65. 当实例数为 8 时，则成为 leader 节点需要 5 票，容灾能力为 4
66. 当实例数 n 为奇数时，假定 n=2x+1，则成为 leader 节点需要 x 票
67. 传统数据处理的性能瓶颈有？（AC）[多选题]
68. 数据存储成本高
69. 流式数据处理性能不足
70. 扩展能力有限
71. 批量数据处理缺失
72. HDFS 不适用于以下哪些场景？（BD）[多选题]
73. 流式数据访问
74. 大量小文件存储
75. 大文件存储与访问
76. 随机写入
77. HDFS 中的主备仲裁，是由哪个组件控制的？（A）[单选题]
78. Zooleeper Failover Controller
79. NodeManager
80. ResourceManager
81. HDFS Client
82. 在 MapReduce 编程中，代码不仅要描述做什么，还要描述具体怎么做。（B）[单选题]
83. TRUE
84. FALSE
85. 在 Loader 历史作业记录中，可以查看以下哪些内容？（ABCD）[多选题] A.作业状态
86. 脏数据链接
87. 错误行/文件数量 D.作业开始/运行时间
88. 关于 Kafka 磁盘容量不足的告警，对于可能的原因以下分析不正确的是？（C）[单选题] A.业务规划不合理，导致数据分配不均，使部分磁盘达到使用率上限
89. 数据保存时间配置过长，数据累计达到磁盘使用率上限
90. Broker 节点故障导致
91. 用于存储 Kafka 数据的磁盘配置（如磁盘数目，磁盘大小等），无法满足当前业务数据流量，导致磁盘使用率达到上限
92. FusionInsight HD 中 Loader 从 SFTP 服务器导入文件时，不需要做编码转换和数据转换且速度最快的文件类型是以下哪项？（B）[单选题]
    1. graph-file
    2. binary-file
    3. text-file
    4. sequence-file
93. Zookeeper 可以为 FusionInsight HD 中哪些组件提供分布式管理支持？（ACD）[多选题] A.HiveB.Loader C.HBase D.Spark
94. 以下关于 FusionInsight 多租户管理的描述正确的有？（ABCD）[多选题] A．可视化的多级租户管理，与企业组织结构相匹配，简化系统资源分配与管理B．一站式管理租户资源，计算资源（CPU/内存/IO）、存储资源（HDFS）、服务资源（HBase…） C.与企业组织结构相匹配的多级的租户模型，不同部门对应不同的租户，按需动态增删租户

D.基于容器机制的租户资源隔离，为租户 SLA 保驾护航

1. 以下关于 Kafka Partition 说法正确的有?(ABCD)[多选题] A.引入 Partition 机制，保证了 Kafka 的高吞吐能力
2. 每个 Partition 都是有序且不可变的消息队列
3. Partition 数量决定了每个 consumer group 中并发消费者的最大数量D.每个 Partition 在存储层面对应一个 log 文件
4. hive 在 load 是不检索数据是否符合 schema 的，hive 遵循的是 schema on read(读时模式)只有在读时模式的时候才检查 hive 的数据字段，schema。（A） [单选题]A.TRUE B.FALSE
5. Loader 可以实现以下哪些转换规则（ABCD）？[多选题] A.长整型时间转换B.增量转换C.拼接转换D.空值转换
6. FusionInsight HD 系统中使用 Streaming 客户端 shell 命令提交拓扑之后，使用 Storm UI 查看发现该拓扑长时间没有处理数据，可能原因（ABCD）[多选题]
   1. 拓扑业务存在逻辑错误，提交之后无法正常运行
   2. Supervisor 是在 Topology 中接收数据然后执行处理的组件
   3. 拓扑过于复杂或者并发太大，导致 worker 启动时间过长，超过 supervisor 的等待时间
   4. Supervisor 的 slots 资源被耗尽，拓扑提交上去后分不到 slot 去启动 worker 进程
7. FusionInsightManager 界面显示 Hive 服务状态为 Bad 时，可能的原因有哪些？（BD） [多选题]
8. DBService 服务不可用
9. HDFS 服务不可用
10. HBase 服务不可用
11. Metastore 实例不可用
12. 以下哪些选项属于 FusionInsight SparkSQL 的优点?（E）[单选题]
13. 稳定和高性能
14. 长稳测试
15. SQL 兼容性
16. 支持数据插入，更新和删除
17. 以上全都正确
18. 下列关于 Flink 中 Transformation 的说法正确的是？（ABD）[多选题]
19. 可以通过 window 设定时间窗口
20. Filtor 操作是对每个元素执行 boolean 函数
21. flatMap 可以对文本进行切分
22. keyBy 是将源头数据按照 key 进行分组，以保证同一个 key 的元数据分到同样的组中
23. FusionInsight Manager 对服务的管理操作，下面说法错误的是？（A）[单选题]
24. 可设置不常用的服务隐藏显示
25. 可以添加和卸载服务
26. 可对服务进行启停重启操作
27. 可查看服务的当前状态
28. 下面关于 ZooKeeper 特性的描述错误的是？(B)[单选题] A.消息更新只能成功或者失败，没有中间状态
29. ZooKeeper 节点数必须为奇数个
30. 一条消息要被超过半数的 Server 接收，它将可以成功写入磁盘D.客户端所发送的更新会按照它们被发送的顺序进行应用
31. Flume 的数据流可以根据 headers 的信息发送到不同的 Channel 中。(A)[单选题]
    1. True
    2. False
32. HBase 中数据存储的文件格式是什么？（C）[单选题]
33. HLog
34. TextFile
35. HFile
36. SequenceFile
37. HBase 中的 Region 是由哪个服务进程来管理的？（A）[单选题] A.HRegionServerB.Zookeeper C.HMaster D.DataNode
38. 在 Flink 中，checkpoint 机制能够保证应用在运行过程中出现失败时，从某一个检查点恢复，在此过程中，流快照是根据数据流入顺序依次创建的。（A）[单选题]A.TRUE B.FALSE
39. 以下关于 FusionInsight 网络安全可靠性，说法正确的有？（ABCD）[多选题] A.阻止外部攻击者通过管理通道入侵实际业务数据
40. FusionInsight 支持网络划分为三级：集群业务平面、集群管理平面和集群外维护网络。彼此之间实施物理隔离
41. 避免业务平面的高负载阻塞集群管理通道
42. 网络平面隔离，避免管理与业务带宽抢占，相互干扰
43. 在 FusionInsight 产品中，关于 Kafka 组件说法正确的是？（ABCD）[多选题]
44. 在创建 Topic 时，副本数不得大于当前存活的 Broker 实例个数，否则创建 Topic 将会失败
45. Kafka 的 Producer 发送消息时，可以指定该消息被哪个 Consumer 消费
46. Kafka 会将元数据信息存放到 Zookeeper 上
47. Kafka 安装完成后就不能再配置数据存放目录
48. Kafka 所有消息都会被持久化到硬盘中，同时 Kafka 通过对 Topic Partition 设置Replication 来保障数据可靠（A）[单选题]A.TRUE

B.FALSE

1. 在 FusionInsight 产品中，关于 Kafka Topic，以下描述不正确的有？（A）[单选题] A.每个 Topic 只能被分成一个 partition（区）
2. Topic 的 partition 数量可以在创建时配置
3. 每个 Partition 的存储层面对应一个 log 文件，log 文件中记录了所有的信息数据
4. 每条发布到 Kafka 的消息都有一个类别，这个类别被称为 Topic，也可以理解为一个存储消息的队列
5. 在 hive 中，以下关于分桶的描述正确的是( BCD )多选A.桶内不可排序
6. 建表时可以指定桶的个数
7. 数据可以根据桶的方式将不同数据放入不同的桶中
8. 分桶的好处是可以获取更高的查询处理效率，使抽样更高效
9. 以下属于 Streaming 特点是(ABCD )多选A.延迟低
10. 属于事件驱动
11. 数据先存储在计算D.可连续查询
12. FusionInsight HD 系统中使用Streaming 客户端 Shell 命令查看拓扑或提交拓扑失败，以下哪些定位手段是正 确的？(ABCD)
13. 查看客户端异常堆栈，判断是否客户端使用问题
14. 查看主 Nimbus 的运行日志，判断是否 Nimbus 服务端异常
15. 查看 Supervisor 运行日志，判断是否 Supervisor 异常
16. 查看 Worker 运行日志
17. Kafka 作为一个分布式消息系统，支持在线和离线消息处理，并提供了 Java API 以便其他组件对接使用。在 FusionInsight 解决方案中，Kafka 属于 FusionInsight HD 模块. ( 对 )
    1. True B.False