1、下列与 HDFS 有关的说法正确的是（　D　）

A．HDFS DataNode 节点上的磁盘需要做 RAID1，用来保证数据的可靠性

B．HDFS 可以在磁盘之间通过 balance 操作，平衡磁盘之间的负载情况

C．HDFS 建议 DataNode 之间的数据盘个数、容量大小不一致，以体现 HDFS 的负均衡能力

D．规划 HDFS 集群时，建议 Active NameNode 和 Standby NameNode 分配在不同机架上

2、NameNode 用于存储 HDFS 上数据块的元数据信息，它保存的数据形式是（BC　　）

A . block

B . fsimage

C . editlog

D . blockid

3、在集群中配置 HDFS 的副本数为 3，设置数据块大小为 128M，此时我们上传一份 64M 的数据文件，该数据文件占用 HDFS 空间大小为（ C ）

A . 64M

B . 128M

C . 384M

D . 192M

5、YARN 框架中，负责集群资源管理的组件是（　　A ）

A.　ResourceManager

B.　NodeManager

C.　Container

D.　JobTracker

7、以下关于外表和托管表描述正确的是（　　C ）

A、外表的数据存储在本地，托管表的数据存储在 hdfs 上

B、删除托管表只会删除 Inceptor 上的元数据不会删除数据文件，删除外表两者都会被删除

C、删除外表只会删除 Inceptor 上的元数据不会删除数据文件，删除托管表两者都会被删除

D、删除托管表或外表，incepotr 上的元数据和数据文件都会被删除

8、导入数据经常会用到 LOAD 命令，以下关于 LOAD 的描述错误的是（　A　）

A. 源数据文件存放于 hdfs 上，通过 load 命令加载数据文件，数据文件将被复制到表目录下

B. 目标表为分桶表时不能通过 load 命令加载数据

C. 目标表为分区表时不能通过 load 命令加载数据

D. 当元数据存放于本地时，需要通过指定 LOCAL 关键字

8、以下关于 Inceptor 数据倾斜场景正确的处理方式有（　CD　）

A. 对于数据倾斜的 SQL 重新跑一次即可解决

B. 剔除引起数据倾斜的数据，再重新执行 SQL

C. 导入数据期间格式转换出现错误引起 null 过多，可以通过重新清理数据解决

D. 将一起数据倾斜的数据和剩下的数据单独运行，再通过 union 合并的方式解决

13、有关 Minor Compact 的描述正确的是（　　D　）

A．一个 store 下的所有文件合并

B．删除过期版本数据

C．删除 delete marker 数据

D．把多个 HFile 合成一个

15、某公司有部门 A、部门 B…，各部门的源数据都取自于企业总线，要求部门内部共享数据源，部门间做到资源隔离，以下设计合理的有（　B　）

A. 部门里每个流任务起一个 application 管理 streamjob

B. 每个部门起一个 application 管理本部门的 streamjob

C. 公司起一个 application 管理所有的 streamjob

D. 每个部门起一个 streamjob 管理本部门的 application

15、某交通部门通过使用流监控全市过往 24 小时各个卡口数据，要求每分钟更新一次，原始流为 org_stream，以下实现正确的是（　　C　）

A. CREATE STREAMWINDOW traffic_stream AS SELECT * FROM original_stream STREAM w1 AS (length '1' minute slide '24' hour);

B. CREATE STREAM traffic_stream AS SELECT * FROM original_stream STREAMWINDOW w1 AS (length '1' minute slide '24' hour);
C. CREATE STREAM traffic_stream AS SELECT * FROM original_stream STREAMWINDOW w1 AS (length '24' hour slide '1' minute);
D. CREATE STREAM traffic_stream AS SELECT * FROM original_stream AS (length '24' second slide '1' minute);

19、有关使用 sqoop 抽取数据的原理的描述不正确的是（ B ）
A． sqoop 在抽取数据的时候可以指定 map 的个数，map 的个数决定在 hdfs 生成的数据文件的个数
B． sqoop 抽取数据是个多节点并行抽取的过程，因此 map 的个数设置的越多性能越好
C． sqoop 任务的切分是根据 split 字段的（最大值-最小值）/map 数
D． sqoop 抽取数据的时候需要保证执行当前用户有权限执行相应的操作

24、下列不属于 kafka 应用场景的是（ D ）
A．常规的消息收集
B．网站活动性跟踪
C．日志收集
D．关系型数据库和大数据平台之间的数据迁移

26、以下对各组件的运维页面描述不正确的是（ B ）
A．通过 Name Node 的 50070 页面对 HDFS 进行监控
B．通过 Resource Manager 的 8180 对 YARN 上运行的任务进行监控
C．通过 HMaster 的 60010 对 HBase 进行监控
D．通过 Hue Server 的 8888 页面登入 Hue

28、以下对 Hadoop 组件的应用场景描述正确的是（ ABCD ）
A. Hive 主要用于构建大数据数仓，主要做批处理、统计分析型业务
B. Hbase 主要用于检索查询的 OLTP 业务
C. ElasticSearch 主要用于全文检索的关键字查询业务
D. Spark Streaming 主要用于实时数据的业务场景

Hyperbase 使用哪种框架获得了强大的计算能力（）
Zookeeper

MapReduce

问答题

1.请列举 HDFS 中包含哪几种角色，并描述各自的功能。（5 分）

2.请描述将一个 100GB 文件以 BulkLoad 方式写入 HyperBase 的主要步骤（5 分）

3.请简述 Inceptor 分区（Partiton）的作用和分区键的选择标准，并说明分区的类型及其不同。（5 分）

4.请简述 Inceptor 分桶(Bucket)的含义和作用，以及分桶的数据存储方式。（5 分）

5 请简述 HyperBase 和 Search 的系统特点及数据模型，并分别描述一个适用场景。（10 分）