TDH 数据平台认证工程师试题（四）

姓名：_____ 分数：_____

【说明】

a) 客观题 30 题，每题 2 分，总计 60 分

b) 主观题 4 题，每题 10 分，总计 40 分

c) 满分 100 分。

【不定项选择题（每题 2 分共 60 分）】

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 的负载均衡能力

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

2、以下哪个服务作为 HDFS 高可靠协调服务的共享存储？（ B ）

A. ZooKeeper

B. JournalNodes

C. NameNode

D. ZKFailoverController

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

4、在 Yarn 服务中，不包含以下哪种角色（ D ）

A . ResourceManager

B . NodeManager

C . ApplicationMaster

D . Contianer

5、ResourceManager 是 YARN 的主要组成部分，有关其功能描述不正确的是（ A ）

A．它直接将集群所拥有的资源按需分配给运行在 YARN 上的应用程序

B．它负责将集群中的所有资源进行统一管理和分配

C．它接受各个节点的资源汇报信息

D．它把资源按照策略分配给各应用

6、当前用户提交了一个 wordcount 词频统计的任务，最后任务执行失败，可能的原因有哪些（ D ）

A．当前集群中没有足够的资源，不足以满足当前 wordcount 任务的需求

B．执行该任务的用户没有权限访问 HDFS 上的数据文件

C．用户在执行任务之前在 HDFS 相应的目录下创建了提交任务时指定的输出目录

D．以上原因都有可能

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、SQL 运行中如果出现 maptask 数据特别多，执行时间又很短时可以通过小文件合并来进行优化，以下是合并参数有（ ABC ）

A. SET ngmr.partition.automerge = TRUE;

B. SET ngmr.partition.mergesize = n;

C. SET ngmr.partition.mergesize.mb = m;

D. SET mapred.reduce.tasks = N;

9、以下关于 inceptor 日志信息描述正确的有（ABCD）

A. Inceptor server 日志存放于各节点的/var/log/inceptorsql[x]/hive-server.log

B. 可以通过 inceptor server 4040 查看 SQL 错误日志

C. Excutor 日志存放于 excutor 节点的/var/log/inceptorsql[x]/spark-excutor.log

D. ExcutorGC 日志存放于 excutor 节点的/var/log/inceptorsql[x]/spark-excutor.gc.log

10、tableA 有 10G 的数据，tableB 有 100G 的数据，两个表通过共有的 id 列做关联查询 name 列，以下方式可以优化计算效率的是（ C ）

A. select /*+MAPJOIN(a)*/ a.name,b.name from tableA a join tableB b on a.id=b.id

B. select /*+MAPJOIN(b)*/ a.name,b.name from tableA a join tableB b on a.id=b.id

C. 建表时将 tableA 和 tableB 根据 id 字段分相同数量的桶

D. 建表时将 tableA 和 tableB 根据 name 字段分相同数量的桶

11、以下属于 HMaster 功能的是（ AD ）

A．为 Region Server 分配 region

B．存储数据元信息

C．对 region 进行 compact 操作

D．管理用户对 table 的增删改查操作

12、Hyperbase 与 Inceptor 的关系，描述正确的是（ CD ）

A. 两者不可或缺，Inceptor 保证 Hyperbase 的服务的正常运行
B. 两者没有任何关系
C. Inceptor 可以访问 Hyperbase
D. 两者相辅相成

13、下列创建全局索引的语句，正确的是（ B ）

A. add_index 't1', 'index_name',
'COMBINE_INDEX|INDEXED=f1:q1:9|rowKey:rowKey:10,UPDATE=true'
B. add_global_index 't1', 'index_name',
'COMBINE_INDEX|INDEXED=f1:q1:9|rowKey:rowKey:10,UPDATE=true'
C. add_fulltext_index 't1', 'index_name',
'COMBINE_INDEX|INDEXED=f1:q1:9|rowKey:rowKey:10,UPDATE=true'
D. create_global_index 't1', 'index_name',
'COMBINE_INDEX|INDEXED=f1:q1:9|rowKey:rowKey:10,UPDATE=true'

14、以下对流处理计算框架描述不正确的是（ D ）

A. Spark Streaming 是基于微批（batch）对数据进行处理的
B. Apache Storm 是基于时间（event）对数据进行处理的
C. Transwarp StreamSQL 可基于微批或事件对数据进行处理
D. 以上说法都不对

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);

16、Zookeeper 服务描述正确的为（ D ）

A . Zookeeper 中每一个 server 互为 leader。
B . Zookeeper 中只有一个 leader，并通过备份机制产生。
C . Zookeeper 中不存在 leader,所有 server 共同提供服务。
D . Zookeeper 通过选举机制确定 leader，有且仅有一个。

17、通过 Hue 修改 HDFS 目录或文件的权限可以通过以下哪些方式实现（ AC ）

A. Hdfs 相应的权限

B. 通过 Hue 超级用户 hue 登录

C. 以 hdfs 用户登录

D. 以上都可以

18、通过 Oozie 使用 ssh，必须满足以下条件（ BC ）

A. 以 root 用户登录各个节点

B. Oozie 用户可以免密钥登录

C. Oozie 用户必须要有 bash 权限

D. 所访问必须是集群的节点

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

20、在使用 sqoop 连接关系型数据时，下面哪个命令可以查看关系型数据库中有哪些表？（ D ）

A. sqoop list-databases
--username root
--password 111111
--connect jdbc:mysql://192.168.164.25:3306/

B. sqoop list-databases
--username root
-P
--connect jdbc:mysql://192.168.164.25:3306/

C. sqoop list-databases
--username root
--password-file file:/root/.pwd
--connect jdbc:mysql://192.168.164.25:3306/

D. sqoop list-tables
--username root
--password 111111
--connect jdbc:mysql://192.168.164.25:3306/test

21、要将采集的日志数据作为 kafka 的数据源，则 flume sink 需要设置为下列哪项参数（ C ）

A . hdfs

B . kafka

C . org.apache.flume.sink.kafka.KafkaSink

D . {topicname}

22、下列是关于 flume 和 sqoop 对比的描述，不正确的是（ C ）

A . flume 主要用来采集日志而 sqoop 主要用来做数据迁移

B . flume 主要采集流式数据而 sqoop 主要用来迁移规范化数据

C . flume 和 sqoop 都是分布式处理任务

D . flume 主要用于采集多数据源小数据而 sqoop 用来迁移单数据源数据

23、有关 Elasticsearch 描述有误的一项是（ C ）

A. 它会利用多播形式发现节点。

B. 主节点(master node) 通过选举方式产生。

C. 主节点(master node)进行集群的管理，只负责集群节点添加和删除。

D. 主节点会去读集群状态信息，必要的时候进行恢复工作。

24、下面措施中，不能保证 kafka 数据可靠性的是（ D ）

A . kafka 会将所有消息持久化到硬盘中保证其数据可靠性

B . kafka 通过 Topic Partition 设置 Replication 来保证其数据可靠性

C . kafka 通过设置消息重发机制保证其数据可靠性

D . kafka 无法保证数据可靠性

25、TDH 提供哪几种认证模式？（ ABC）

A. 所有服务使用简单认证模式——所有服务都无需认证即可互相访问

B. 所有服务都启用 Kerberos 认证，用户要提供 Kerberos principal 和密码（或者 keytab）来访问各个服务

C. 所有服务都启用 Kerberos 同时 Inceptor 启用 LDAP 认证

D. 所有服务都启用 LDAP 认证

26、开启 LDAP 后，应该使用哪个命令连接 Inceptor（ B ）

A. transwarp -t -h $ip。

B. beeline -u jdbc:hive2://$ip:10000 -n $username -p $password。

C. beeline -u "jdbc:hive2://$ip:10000/default;principal=hive/node1@TDH"。

D. beeline -u "jdbc:hive2://$ip:10000/default;principal=user1@TDH"。

27、Inceptor server 服务无法启动时，该如何查看日志是（ B ）

A. 查看 TDH manager 所在节点/var/log/inceptorsql*/目录下的 hive-server2.log 日志
B. 查看 Inceptor server 所在节点/var/log/inceptorsql*/目录下的 hive-server2.log 日志
C. 查看 Resource Manager 所在节点 /var/log/Yarn*/ 目录下的 yarn-yarn-resourcemanager-poc-node1.log 日志
D. 查看任意节点/var/log/inceptorsql*/目录下的 hive-server2.log 日志

28、现有一批数据需要进行清洗，要求对其中 null 通过 update 转换为 0，删除重复的记录，添加部分新的记录，则该表应该设计为（ C ）

A. Tex 表
B. Orc 表
C. Orc 事务表
D. Holodesk 表

29、现有一个表数据要存储在 hyperbase 上，并创建全文索引，原表数据 10GB，HDFS 配置为 3 副本，hyperbase 压缩比例按 1:3 计算，索引数据量为 20GB，ES 副本数为 1，ES 压缩比按 1:3 计算，则该表需要多大的存储空间存储（ B ）

A. 16.67GB
B. 23.33GB
C. 30GB
D. 70GB

30、下面哪些工作不属于集群预安装工作（ D ）

A．为集群中每个节点的安装操作系统
B．选一个节点作为管理节点，修改其 /etc/hosts 文件
C．安装 Transwarp Manager 管理界面
D．配置集群安全模式

【客观简答题（每题 10 分，共 40 分）】

1、请描述 HDFS 的高可用性实现机制：

答：

2、请列举出平台支持的 5 种存储格式/引擎的表，并详细描述各自的存储特点、使用场景、支持的操作以及是否支持分区分桶。

答：

Text 表：

ORC 表：

事务表：

HoloDesk 表：

Hyperbase 表：

3、 请描述一个 100GB 文件写入 Hyperbase 表的整个过程（使用 bulkload 方式实现）

4、写出以下场景下的优化思路

（1）、假设在 Inceptor 上执行任务，发现 Map Task 数量多、执行时间短，应采取哪种措施来提升性能？

（2）、请简述在 Inceptor 中大表与大表做 join、大表与小表做 join 时分别有哪些优化手段