TDH 数据平台认证工程师试题（二）

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、在 HDFS 服务中，为了保证 Name Node 高可用性的角色不包括（ A ）

A . Data Node

B . Journal Node

C . ZKFC

D . Zookeeper

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、下列有关 YRAN 中角色的描述不正确的是（ C ）

A . ResourceManager 控制整个集群并管理基础计算资源的分配

B . NodeManager 管理每个节点的资源，管理抽象容器

C . NodeManager 负责调度当前节点的所有 ApplicationMaster

D . ApplicationMaster 管理一个 YARN 内运行的应用程序的实例

6、Spark 与 MapReduce 对比，突出的优势不包括（ D ）

A . 基于内存的计算，效率更高

B . Spark 能支持比 MapReduce 更多的应用场景

C . Spark 支持多种编程语言接口，框架开销更低

D . Spark 可以运行在 YARN 之上而 MapReduce 不能

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、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

9、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

10、假设使用场景中有如下查询语句

```
SELECT Sex, Region, COUNT(ID), AVG (Salary)
FROM Employee
WHERE Department = 'IT'
GROUP BY Sex, Region
ORDER BY Sex, Region;
```

通过 holodesk 的 cube 和 index 手段对这种过滤率和聚合率高的业务进行优化，以下建表正确的是（ A ）

A.
```
CREATE TABLE Employee
TBLPROPERTIES (
'cache' = 'RAM',
'holodesk.index' = 'Department',
'holodesk.dimension' = 'Sex, Region'
)
```

B. CREATE TABLE Employee
TBLPROPERTIES (
'cache' = 'RAM',
'holodesk.index' = 'Sex, Region'
'holodesk.dimension' = 'Department'
)

C. CREATE TABLE Employee
TBLPROPERTIES (
'cache' = "Department",
'holodesk.index' = 'Department',
'holodesk.dimension' = 'Sex, Region'
)

D. CREATE TABLE Employee
TBLPROPERTIES (
'cache' = 'RAM',
'holodesk.index' = 'Department',
'holodesk.dimension' = 'Sex'
)

11、关于 Hyperbase 全局索引的描述，哪些是正确的？（ ABD ）

A．核心是倒排表
B．全局索引概念是对应 Rowkey 这个“一级”索引
C．全局索引使用平衡二叉树
D．全局索引使用 B+树检索数据

12、以下不属于 Hyperbase 存储模型单位的是（ B ）

A．table
B．region
C．StoreFile
D．block

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

14、以下的 stream 的描述不正确的是（ C ）

A. Input 定义了如何从数据源读取数据
B. Derived stream 是对 stream 转换而来的，可分为单 batch 变形和多 batch 变形
C. 定义 Derived stream 后 stream 当即根据转换规则进行变形
D. 窗口变形的长度必须是当前流的整数倍

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

16、Zookeeper 服务描述正确的是（ C ）

A. Zookeeper 可以存储文件，所以它是用于存储大量数据信息的文件系统。
B. 它是集群的管理服务，总控节点间所有通信。
C. 它是分布式应用程序协调服务。
D. 它是保存所有集群服务的元数据库。

17、我们可以通过 hue 图形化的操作 HDFS，hue 可以实现 hdfs 的（ ABCD）

A. 创建目录
B. 上传文件
C. 直接查看文件
D. 更改权限

18、通过 oozie workflow 调度 sqoop 任务，以下说法正确的是（ BC ）

A. 必须使用 sudo 用户
B. 确保对应的 jdbc 驱动正确上传到 hdfs 上
C. Sqoop 导入的 hdfs 目录必须前提不存在
D. 以上说法都对

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、有关 sqoop 的参数说法不正确的是（ C ）

A . --username 是必需参数
B .--m 大于 1 时，--split-by 参数是必需参数
C . --query 是执行 sqoop 操作的必需参数
D .--field-terminated-by 用来指定在 hdfs 生成数据文件时的列分隔符

21、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

22、以下不属于 Flume 的 Source 类型的是（ B ）

A . exec source

B . file source

C . spooling directory source

D . kafka source

23、有关 Elasticsearch 特性描述有误的一项是（ D ）

A. 分布式实时文件存储，可将每一个字段存入索引

B. 实时分析的分布式搜索引擎。

C. 支持插件机制，分词插件、同步插件

D. 以上都不正确

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

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、在安装有 kerberos 服务的集群中如何切换用户（ B ）

A. 不需要切换，所有用户都为服务公用用户，可以直接使用。

B. 直接使用 kinit 用户名称方式进行切换

C. 必须先 destroy ，才能再使用 kinit 用户名称 方式登录

D. 以上都不正确

27、以下对 Transwarp Manager 描述不正确的是（ C ）

A. Transwarp Manger 是 TDH 的管理运维平台

B. 通过 Transwarp Manager 的 8180 界面登入

C. 在 Transwarp Manager 上能启动和停止 Transwarp Agent 角色

D. 在 Transwarp Manager 上能对 Inceptor 表进行赋权操作

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

29、某电信部门有 100 亿条用户过往使用通讯记录，现需要提供客户终端根据电话号精确查询历史通讯，满足用户同时并发访问，则该表应该设计为（ A ）

A. Hyperbase 表+全局索引
B. Hyperbase 表+es 索引
C. Es 表+es 索引
D. 以上方式都可以

30、可以安装 TDH 的操作系统有？（ ACD ）

A. SUSE SP2-SP3。
B. Win7/Win10。
C. CentOS 6.3-6.5。
D. REHL 6.3-6.5。

【客观简答题（每题 10 分，共 40 分）】

1、 请描述一个 100GB 文件写入 HDFS 的整个过程：

答：

2、 请以 WordCount 为例描述 MapReduce 的运行过程，并列出 Spark 相较 MapReduce 的优势：

答：

3、 写出一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

4、 请列出 TDH 下的 4 大组件（Inceptor、Hyperbase、StreamSQL、Discover）的特性以及适用场景。