TDH 数据平台认证工程师试题（一）

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、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

2、Namenode 在启动时自动进入安全模式，在安全模式阶段，说法错误的是（ D ）

A. 安全模式目的是在系统启动时对数据有效性进行检查

B. 根据策略对数据块进行必要的复制或删除

C. 当数据块的上报数达到阈值时，会自动退出安全模式

D. 允许用户对文件系统进行读写操作

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、下列对 YARN 角色在集群中的作用描述正确的是（ AB ）

A . 集群资源管理

B . 集群任务调度与管理

C . 存储部分 HDFS 上的数据块

D . 以上都正确

5、在 Yarn 服务中，不包含以下哪种角色（ D ）

A . ResourceManager

B . NodeManager

C . ApplicationMaster

D . Contianer

PPT上好像都有?

难道是拼写错误?

6、下列计算框架中不属于分布式计算框架的是（ B ）

A . MapReduce

B . MATLAB MATLAB是一种用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境

C . SPARK

D . Tez 一个运行在YARN之上支持DAG作业的计算框架

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、以下对分桶表的描述正确的是（ A ）

A、分桶表通过改变数据的存储分布，对查询起到一定的优化作用

B、分桶键不能是表中的列 必须是表结构中的列

C、分桶数应为素数

D、事物表必须制定分桶，分桶字段可以被更新 分桶键和分桶数在建表时确定，不允许更改

9、以下关于 inceptor excutor 资源配置的说法正确的有（ ABC ）

A、Excutor 资源配置 fixed 和 ratio 两种模式

B、Excutor 内核数配置的是每个 excutor 所使用的逻辑 core 数量

C、Excutor 内核数和内存配置比例一般为 1 core:2G memory

D、 Excutor 分布可以指定每个节点运行的 excutor 数量或 executor 在集群上运行的总数量，但是不能指定运行的节点

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、以下为 Hyperbase 分布式存储的最小单元的是（ B ）

A、Region server

B、Region

C、StoreFile

D、Store

Key由Row Length、Row、Column Family Length、Column Family、Column Qualifier、Time Stamp、Key Type七部分组成

13、以下有关 Hyperbase 说法正确的是（ D ）

A、数据类型丰富，支持 String、Int、Char 等类型

B、Key/value 系统，key 由 Row,Column Family,Column Qualifier 组成

C、Hyperbase 表中 rowkey 有序，按字典序降序排列

D、以上说法都不正确

14、以下关于 StreamSQL 的概念描述正确的是（ AC ）

A. Stream 是数据流

B. Streamjob 是对一个或多个 stream 进行计算并将结果写进一个流的任务

C. Application 是一个或多个 streamjob 的集合

D. 以上说法都不正确

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、以下不是 Zookeeper 的功能是（ D ）

A. 配置管理
B. 集群管理
C. 分布式锁
D. 存储大量数据

17、以下服务需要与 zookeeper 进行通信的是（D ）

A. HMaster
B. Active NameNode
C. InceptorSQL
D. Active ResourceManager

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

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

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、sqoop 抽取数据时需要做一些数据转换的工作，下面说法不正确的是（ B ）

A . --fields-terminated-by '\\01' 用来设置在 hdfs 生成的文件的分割符
B . --hive-drop-import-delims 用来设置在 hdfs 生成的文件的存储形式为列存储
C . --null-string '\\N' 用来把所有的 String 类型的空值 转换成 hive 的 NULL 值
D . --null-non-string '\\N' 用来把非 String 类型的空值 转换成 hive 的 NULL 值

21、下列有关 flume 的描述不正确的是（ C ）

A . flume 是 Apache 的一个子项目

B . flume 主要是一个日志采集，传输系统

C . flume 和 sqoop 功能相似，因此可以相互替代

D . flume 可以同时采集集群内部和集群外部的日志数据

22、下列 sink 中哪些是 flume 不支持的 sink（ C ）

A . HDFS sink

B . kafka sink

C . memory sink

D . file roll sink

23、以下对 ElasticSearch 描述不正确的是（ C ）

A. ElasticSearch 是分布式全文搜索引擎

B. ElasticSearch 集群中分 master 和 data 节点

C. ElasticSearch 数据存储在 HDFS 上

D. ElasticSearch 数据可以按 Shard 分布在不同的节点上

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、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

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、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、以下不属于管理角色的是（ D ）

A . Name Node
B . HMaster
C . Resource Manager
D . Node Manager

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、 集群有 8 个节点，每个节点有 8 块硬盘（默认 3 副本）。如果某个节点有 3 块盘损坏，是否可能存在数据块丢失情况；如果有 3 个节点发生故障，是否可能存在数据块丢失情况；并简述原因。

答：

2、 请描述 TDH 平台中在 Yarn 上可以使用哪几种调度策略，并分别阐述各调度策略的特点。

3、 请简述 bulkload 的作用和操作步骤（包括原理的步骤和使用 sqlbulkload 的步骤）。

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