TDH 数据平台认证工程师试题（三）

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、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

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、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

6、MapReduce 计算框架的特点包括（ D ）

A．自动化并行和分布式计算

B．出错容忍度高

C．优先数据本地化计算

D．以上都是

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

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、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、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、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

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、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、关于 Hue 对 hive server 的支持度描述正确的是（ B ）

A. 只支持 hive server1
B. 只支持 hive server2
C. 同时支持 hive server1 和 hive server2
D. 只支持开启 LDAP 的 hive server2

18、以下关于 oozie 三个编辑器说法正确的是（ CD ）

A. bundle 构建在 workflow 工作方式之上，提供定时运行和触发运行任务的功能。
B. bundle 将多个 workflow 管理起来，这样我们只需提供一个 bundle 提交即可
C. workflow 是最简单的一种工作方式
D. coordinator 可以包含一到多个 workflow

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、下面与 sqoop 做数据迁移有关的描述不正确的是（ B ）

A . sqoop 做数据迁移的主要瓶颈在网络带宽和 RDB 的 IO 限制

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 抽取数据分为全量抽取和增量抽取两种

D . 当-m 大于 1 时，就必须要设置--split-by 字段

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、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、以下属于 Guardian 的功能是（ABCD ）

A. 用户管理

B. 用户认证

C. 审计

D. 权限管理

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、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、假设集群的每个节点初始有 6 块硬盘，运行一段时间后，每个节点又加了 4 块新硬盘，为了使数据在所有硬盘上分布均匀，能否通过 hdfs balancer 达到效果，为什么？并列出能达到效果的两种措施。

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

4、 请描述高并发检索和综合搜索的场景特点，这两种场景应使用哪种技术来做支撑，并指出数据和索引各自的存储位置。