尚硅谷大数据技术之大数据基础阶段考试题（二）

(作者：尚硅谷大数据研发部)

官网：www.atguigu.com

版本：V1.0

一 Zookeeper

1 请简述 ZooKeeper 的选举机制

2 ZooKeeper 的监听原理是什么？

3 ZooKeeper 的部署方式有哪几种？集群中的角色有哪些？集群最少需要几台机器？

4 ZooKeeper 的常用命令

二 Hive

1 Hive 表关联查询，如何解决数据倾斜的问题？

2 请谈一下 Hive 的特点，Hive 和 RDBMS 有什么异同？

3 请说明 Hive 中 Sort By，Order By，Cluster By，Distrbute By 各代表什么意思？

4 Hive 有哪些方式保存元数据，各有哪些特点？

5 Hive 内部表和外部表的区别？

6 Hive 的 HSQL 转换为 MapReduce 的过程？

7 请把下面语句用 Hive 实现

```
SELECT a.key,a.value
FROM a
WHERE a.key not in (SELECT b.key FROM b)
```

8 写出将 text.txt 文件放入 Hive 中 test 表‘2016-10-10’ 分区的语句，test 的分区字段是 l_date。

9 Hive 自定义 UDF 函数的流程?

10 对于 Hive，你写过哪些 udf 函数，作用是什么？

11 Hive 中的压缩格式 TextFile、SequenceFile、RCfile 、ORCfile 各有什么区别？

12 Hive join 过程中大表小表的放置顺序？

13 Hive 的两张表关联，使用 MapReduce 怎么实现？

14 所有的 Hive 任务都会有 MapReduce 的执行吗？

15 Hive 的函数：UDF、UDAF、UDTF 的区别？

16 说说对 Hive 桶表的理解？

17 Hive 可以像关系型数据库那样建立多个库吗？

18 Hive 实现统计的查询语句是什么？

19 Hive 优化措施

20 Hive 数据分析面试题

场景举例.北京市学生成绩分析.
成绩的数据格式:时间,学校,年纪,姓名,科目,成绩
样例数据如下:
2013,北大,1,裘容絮,语文,97
2013,北大,1,庆眠拔,语文,52
2013,北大,1,乌洒筹,语文,85
2012,清华,0,钦尧,英语,61
2015,北理工,3,冼殿,物理,81
2016,北科,4,况飘索,化学,92
2014,北航,2,孔须,数学,70
2012,清华,0,王脊,英语,59
2014,北航,2,方部盾,数学,49
2014,北航,2,东门雹,数学,77
问题:

情景题：分组 TOPN

```
# 1.分组 TOPN 选出 今年每个学校,每个年级,分数前三的科目.
```

```
北大    0    英语    95    1
北大    0    英语    77    2
北大    0    英语    50    3
北大    2    数学    80    1
北大    2    数学    62    2
北大    2    数学    56    3
北大    3    物理    82    1
北大    3    物理    61    2
北大    3    物理    57    3
北大    4    化学    72    1
北大    4    化学    56    2
北大    4    化学    52    3
北大    5    生物    84    1
北大    5    生物    62    2
北大    5    生物    57    3
北理工  0    英语    93    1
北理工  0    英语    71    2
北理工  0    英语    59    3
北理工  1    语文    79    1
北理工  1    语文    62    2
北理工  1    语文    49    3
北理工  2    数学    93    1
北理工  2    数学    85    2
北理工  2    数学    73    3
北理工  3    物理    78    1
北理工  3    物理    49    2
北理工  4    化学    97    1
北理工  4    化学    89    2
北理工  4    化学    78    3
北理工  5    生物    86    1
北理工  5    生物    77    2
北理工  5    生物    57    3
```

情景题：where 与 having

```
-- 今年 清华 1 年级 总成绩大于 200 分的学生 以及学生数
```

```
OK
清华    1    枚欧涧    231.0    4
清华    1    禹钥蜗    276.0    4
清华    1    苏俘      219.0    4
清华    1    顾藻娘    285.0    4
```

情景题：数据倾斜

今年加入进来了 10 个学校,学校数据差异很大计算每个学校的平均分。简述需要注意的点。

情景题：分区表

假设我创建了一张表，其中包含了 2016 年客户完成的所有交易的详细信息：CREATE TABLE transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’ ;

现在我插入了 100 万条数据，我想知道每个月的总收入。

问：如何高效的统计出结果。写出步骤即可。

21 Hive 有哪几层？

22 解释一下 sql 运行步骤，是否有优化空间，如果有，如何优化：

SELECT a.id, b.name FROM a LEFT OUTER JOIN b ON a.id = b.id WHERE a.dt = ‘2016-01-01’ AND b.dt = ‘2016-01-01’;

23 Hive 中，建的表为压缩表，但是输入文件为非压缩格式，会产生怎样的现象或者结果？

24 已知表 a 是一张内部表，如何将它转换成外部表？请写出相应的 hive 语句。

25 HiveSQL 语句中 select from where group by having order by 的执行顺序

26 Hive 中 mapjoin 的原理和实际应用？

27 数据仓库的整体架构是什么，其中最重要的是哪个环节？

28 订单详情表 ord_det(order_id 订单号，sku_id 商品编号，sale_qtty 销售数量，dt 日期分区)任务计算 2016 年 1 月 1 日商品销量的 Top100，并按销量降级排序

29 某日志的格式如下：

pin|-|request_tm|-url|-|sku_id|-|amount

分隔符为‘|-|’，

数据样例为:

张三|-|q2013-11-23 11:59:30|-|www.jd.com|-|100023|-|110.15

假设本地数据文件为 sample.txt,先将其导入到 hive 的 test 库的表 t_sample 中，并计算每个用户的总消费金额，写出详细过程包括表结构。

30 有一张很大的表：TRLOG，该表大概有 2T 左右

```
CREATE TABLE TRLOG
(     PLATFORM string,
      USER_ID int,
      CLICK_TIME string,
      CLICK_URL string)
      row format delimited fields terminated by ‘\t’;
```

数据:

PLATFORM　　USER_ID　　CLICK_TIME　　CLICK_URL

WEB	12332321	2013-03-21 13:48:31.324	/home/
WEB	12332321	2013-03-21 13:48:32.954	/selectcat/er/
WEB	12332321	2013-03-21 13:48:46.365	/er/viewad/12.html
WEB	12332321	2013-03-21 13:48:53.651	/er/viewad/13.html
……	……	……	……

把上述数据处理为如下结构的表 ALLOG:

```
CREATE TABLE ALLOG
    (PLATFORM string,
    USER_ID int,
    SEQ int,
    FROM_URL string,
    TO_URL string)
    row format delimited fields terminated by '\t';
```

整理后的数据结构：

PLATFORM	USER_ID	SEQ	FROM_URL	TO_URL
WEB	12332321	1	NULL	/home/
WEB	12332321	2	/home/	/selectcat/er/
WEB	12332321	3	/selectcat/er/	/er/viewad/12.html
WEB	12332321	4	/er/viewad/12.html	/er/viewad/13.html
WEB	12332321	1	NULL	/m/home/
WEB	12332321	2	/m/home/	/m/selectcat/fang/

PLATFORM 和 USER_ID 还是代表平台和用户 ID:SEQ 字段代表用户按时间排序后的访问顺序，FROM_URL 和 TO_URL 分别代表用户从哪一页跳转到哪一页。某个用户的第一条访问记录的 FROM_URL 是 NULL（空值）。

实现基于纯 Hive SQL 的 ETL 过程，从 TRLOG 表生成 ALLOG 表:（结果是一套 SQL）

31 已知一个表 STG.ORDER，有如下字段:Date，Order_id，User_id，amount。请给出 sql 进行统计:数据样例:2017-01-01,10029028,1000003251,33.57。

1) 给出 2017 年每个月的订单数、用户数、总成交金额。

2) 给出 2017 年 11 月的新客数(指在 11 月才有第一笔订单)

三 Flume

1 flume 有哪些组件，flume 的 source、channel、sink 具体是做什么的

2 你是如何实现 flume 数据传输的监控的

3 flume 的 source,sink,channel 的作用？你们 source 是什么类型？

4 你们的 Flume 怎么做数据监听？有没有做 ETL?

四 Kafka

1 kafka 的 balance 是怎么做的

2 kafka 的消费者有几种模式

3 为什么 kafka 可以实现高吞吐？单节点 kafka 的吞吐量也比其他消息队列大，为什么？

4 kafka 的偏移量 offset 存放在哪儿，为什么？

5 Kafka 消费过的消息如何再消费

6 Kafka 里面用的什么方式 拉的方式还是推的方式？如何保证数据不会出现丢失或者重复消费的情况？做过哪些预防措施，怎么解决以上问题的？Kafka 元数据存在哪？

7 kafka 支不支持事物，

8 Kafka 的原理

五 Hbase

1 hbase 查询一条记录的方法是什么？Hbase 写入一条记录的方法是什么？

2 Hbase 优化及 rowkey 设计原则，

3 Hbase 架构和各个组件的作用？在插入大量数据和删除数据会发生什么？

六 Oozie

有没有使用 OZ 调度 hadoop 任务

七 Hadoop HA

1 HAnamenode 是如何工作的?