大数据学科hadoop阶段考试（二）

考试时间：18：00-20：00　考试分数：100

注意：本连接中不支持复制操作，最好能用自己概述出来。另外，望同学们使用电脑作答，考试时间较长，避免误触，导致重新作答。

基本信息：

姓名：

张超

班级：

SH200317

1. 请用Hive完成
已知一个表order_tab，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)
建表语句如下：
create table order_tab(dt string,order_id string,user_id string,amount decimal(10,2)) row format delimited fields terminated by '\t';

无需建表，请直接填写sql。

```
select count(user_id) from (select user_id from order_tab group by user_id);
select count(order_id) order_count, count(distinct(user_id)) user_count,
sum(amount) all, substring(dt,1,7) month from order_tab
where substriing(dt, 1,4)='2017'
group by month;
```

2. 请列举常用Linux命令，并说明命令用途。

```
ssh ssh-keygen ssh-copy-id ssh相关命令
> >> 输出重定向 覆盖写/追加写
head 显示文件头部信息 tail 显示文件尾部信息
chmod 改变文件权限 chown 改变文件所有者 chgrp 改变所属组
ps 查看系统进程 kill 终止进程 pstree 以树的形式查看进程 top 查看系统状态和资源占用情况
```

3. 请列举hadoop常用端口号。

```
web:
9870 namenode web访问
9868 2nn web访问
8088 resourcemanager web访问
19888 jobhistoryserver web访问
```

4. 简述hadoop的MapReduce的Shuffle过程(文字描述)。

1. 在 map 方法之后，reduce 方法之前的处理过程就是 shuffle 过程.
2. map 方法写出去的 k-v，会被一个收集线程收集到缓冲区中.
3. 缓冲区大小默认是 100M，达到 80% 发生溢写，缓冲区记录了 k-v，k-v 的下标，k-v 的分区等信息.
4. 溢写的时候，是按照 key 的分区进行排序，采用快排对索引进行排序，再按照分区进行溢写，从而完成

5. 简述Flume组成（三个组件）及每个组件的常用类型（两个），并说明其特点。

Sink:
hdfs 写入hdfs
avro 写入avro端口
file 写入某文件
logger 以日志方式输出

6. 结合数仓项目说明HDFS存储大量小文件造成的影响，以及HDFS Sink如何避免生成大量小文件。

可以通过调整 hdfs sink 的参数：
hdfs.rollInterval=3600 每小时将临时文件滚动成正式文件
hdfs.rollSize=134217728 tmp 文件到达128m时才生成正式文件
hdfs.rollCount=0 不按照事务的数量滚动生成正式文件

7. 请简单说明Kafka消费者的分区分配策略。

一个消费者 group 中有许多的消费者，一个 topic 有多个分区，消费者组内每个消费者负责不同分区的数据，一个分区只能由一个组内消费者消费，但消费者之间互不影响.

kafka 有 roundrobin 和 range 两种分区分配策略.
roundrobin 即轮询，多个分区依次分给组内的消费者；range 即先划分再分配，先按照同一组内消费者

8. 请对Hive的内部表和外部表做出说明。

内部表 MANAGED TABLE也叫管理表. hive 会控制内部表数据的生命周期. 当我们删除一个内部表时，hive 也会删除相应的这个表中的数据，所以内部表并不适合和其他工具共享数据.
与之对应的是外部表 EXTERNAL TABLE，hive 并非认为其完全拥有这份数据. 删除外部表并不会删除相应的数据，只会删除元数据信息.
内部表和外部表可以相互转换

9. 简述使用sqoop进行hive与mysql的导入导出时应该注意哪些问题？如何解决?

需要对不同种类的数据表区别处理，存储完整数据的全量表，存储新增加的数据的增量表，存储新增加的数据和变化的数据的新增及变化表，只需要存储一次的特殊表等.
全量同步策略：每天存出一份完整的数据作为一个分区. 适合数据量不打且每天既会有新数据的插入，也会有旧数据的修改的场景.
增量同步策略：每天存储一份增量数据，作为一个分区，适合表数据量大，且每天只会有新数据插入的场景

对于近期的学习你有什么想说的?

提交

举报