**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：[矩阵文本题] \*

|  |  |
| --- | --- |
|  |  |
| 姓名： | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 班级： | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

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'; [填空题] \*

无需建表，请直接填写sql。

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**1)select  
    date\_format(dt,'yyyy-MM'),  
    count(order\_id),  
    count(distinct user\_id),  
    sum(amount)  
from  
    order\_tab  
group by date\_format(dt,'yyyy-MM');  
2)select  
    tmp.mn,  
    count(\*)  
from  
    (  
    select  
        date\_format(min(dt),'yyyy-MM') mn,  
        user\_id  
    from  
        order\_tab  
    group by user\_id  
    having date\_format(min(dt),'yyyy-MM')='2017-11'  
    ) tmp  
group by tmp.mn;

2. 请列举常用Linux命令，并说明命令用途。 [填空题] \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**命令 命令解释  
top 查看内存  
df -h 查看磁盘存储情况  
iotop 查看磁盘IO读写(yum install iotop安装）  
iotop -o 直接查看比较高的磁盘读写程序  
netstat -tunlp | grep 端口号 查看端口占用情况  
uptime 查看报告系统运行时长及平均负载  
ps aux 查看进程

3. 请列举hadoop常用端口号。 [填空题] \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**dfs.namenode.http-address: 50070/9870  
dfs.datanode.http-address:50075/9864  
SecondaryNameNode辅助名称节点端口号：50090/9868  
dfs.datanode.address:50010/9866  
fs.defaultFS:8020 /9000/9820  
yarn.resourcemanager.webapp.address:8088  
历史服务器web访问端口：19888

4. 简述hadoop的MapReduce的Shuffle过程(文字描述)。 [填空题] \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**1）MapTask收集我们的map()方法输出的kv对，放到内存缓冲区中  
2）从内存缓冲区不断溢出本地磁盘文件，可能会溢出多个文件  
3）多个溢出文件会被合并成大的溢出文件  
4）在溢出过程及合并的过程中，都要调用Partitioner进行分区和针对key进行排序  
5）ReduceTask根据自己的分区号，去各个MapTask机器上取相应的结果分区数据  
6）ReduceTask会取到同一个分区的来自不同MapTask的结果文件，ReduceTask会将这些文件再进行合并（归并排序）  
7）合并成大文件后，Shuffle的过程也就结束了，后面进入ReduceTask的逻辑运算过程（从文件中取出一个一个的键值对Group，调用用户自定义的reduce()方法）

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。 [填空题] \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**一、（1）Taildir Source：断点续传、多目录。Flume1.6以前需要自己自定义Source记录每次读取文件位置，实现断点续传。（2）Avro Source：Avro端口监听并接收来自外部的Avro客户流的事件。（3）Exec Source：Exec Source的配置就是设定一个Unix(linux)命令，然后通过这个命令不断输出数据。如果进程退出，Exec Source也一起退出，不会产生进一步的数据。（4）Spooling Directory Source：Spooling Directory Source监测配置的目录下新增的文件，并将文件中的数据读取出来。  
二、（1）File Channel：数据存储在磁盘，宕机数据可以保存。但是传输速率慢。适合对数据传输可靠性要求高的场景，比如，金融行业。（2）Memory Channel：数据存储在内存中，宕机数据丢失。传输速率快。适合对数据传输可靠性要求不高的场景，比如，普通的日志数据。（3）Kafka Channel：减少了Flume的Sink阶段，提高了传输效率。   
三、(1)HDFS Sink：当需要将事件消息写入到Hadoop分布式文件系统（HDFS）时，可以使用HDFS Sink。 （2）Avro Sink：和 Avro Source一起工作，用于构建Flume分层收集数据消息结构。 （3）Kafka Sink：通过该Sink可将事件消息数据发布到Kafka topic 上。其目标是将Flume与Kafka集成，以便基于拉式的处理系统可以处理来自各种Flume Source的数据。 目前支持Kafka 0.9.x以上系列版本。

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。 [填空题] \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**（1）HDFS存入大量小文件，有什么影响？  
元数据层面：每个小文件都有一份元数据，其中包括文件路径，文件名，所有者，所属组，权限，创建时间等，这些信息都保存在Namenode内存中。所以小文件过多，会占用Namenode服务器大量内存，影响Namenode性能和使用寿命  
计算层面：默认情况下MR会对每个小文件启用一个Map任务计算，非常影响计算性能。同时也影响磁盘寻址时间。  
（2）HDFS小文件处理  
官方默认的这三个参数配置写入HDFS后会产生小文件，hdfs.rollInterval、hdfs.rollSize、hdfs.rollCoun,基于以上hdfs.rollInterval=3600，hdfs.rollSize=134217728，hdfs.rollCount =0，hdfs.roundValue=10，hdfs.roundUnit= second几个参数综合作用，效果如下：  
（1）tmp文件在达到128M时会滚动生成正式文件  
（2）tmp文件创建超10秒时会滚动生成正式文件  
举例：在2018-01-01 05:23的时侯sink接收到数据，那会产生如下tmp文件：  
/atguigu/20180101/atguigu.201801010520.tmp  
即使文件内容没有达到128M，也会在05:33时滚动生成正式文件

7. 请简单说明Kafka消费者的分区分配策略。 [填空题] \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**在 Kafka内部存在两种默认的分区分配策略：Range和 RoundRobin。  
Range是默认策略。Range是对每个Topic而言的（即一个Topic一个Topic分），首先对同一个Topic里面的分区按照序号进行排序，并对消费者按照字母顺序进行排序。然后用Partitions分区的个数除以消费者线程的总数来决定每个消费者线程消费几个分区。如果除不尽，那么前面几个消费者线程将会多消费一个分区。  
RoundRobin：前提：同一个Consumer Group里面的所有消费者的num.streams（消费者消费线程数）必须相等；每个消费者订阅的主题必须相同。

8. 请对Hive的内部表和外部表做出说明。 [填空题] \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**1）管理表：当我们删除一个管理表时，Hive也会删除这个表中数据。管理表不适合和其他工具共享数据。  
2）外部表：删除该表并不会删除掉原始数据，删除的是表的元数据

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？ [填空题] \*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**答案解析：**Hive中的Null在底层是以“\N”来存储，而MySQL中的Null在底层就是Null，为了保证数据两端的一致性。在导出数据时采用--input-null-string和--input-null-non-string两个参数。导入数据时采用--null-string和--null-non-string。  
–staging-table一致性问题

对于近期的学习你有什么想说的? [填空题]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_