Hive练习10 排序

编程要求

在右侧编辑器补充代码，查询出`2013`年`7`月`22`日的哪三种股票买入量最多。

表名：`total`

col_name	data_type	comment
`tradedate`	`string`	交易日期
`tradetime`	`string`	交易时间
`securityid`	`string`	股票ID
`bidpx1`	`string`	买入价
`bidsize1`	`int`	买入量
`offerpx1`	`string`	卖出价
`bidsize2`	`int`	卖出量

部分数据如下所示：

```
20130724    145004    152896    2.62    6960    2.63    13000
20130724    145101    152896    2.86    13880    2.89    6270
20130724    145128    152896    2.85    327400    2.851    1500
20130724    145143    152896    2.603    44630    2.8    10650
```

数据说明：

```
（152896:每种股票id）
（20130724: 2013年7月24日）
（145004: 14点50分04秒）
```

代码

```
----------禁止修改----------
create database if not exists mydb;
use mydb;
create table if not exists total(
tradedate string,
tradetime string,
securityid string,
bidpx1 string,
bidsize1 int,
```

```
offerpx1 string,
bidsize2 int)
row format delimited fields terminated by ','
stored as textfile;
truncate table total;
load data local inpath '/root/files' into table total;
----------禁止修改----------

----------begin----------
SELECT securityid, sum(bidsize1) as size1 FROM total WHERE tradedate =
'20130722' GROUP BY securityid ORDER BY size1 desc LIMIT 3;

----------end----------
```

解析

① **order by**

- `order by`后面可以有多列进行排序，默认按字典排序(`desc`:降序，`asc`(默认):升序)；
- `order by`为全局排序；
- `order by`需要`reduce`操作，且只有一个`reduce`，无法配置(因为多个`reduce`无法完成全局排序)；
- 如果指定了`hive.mapred.mode=strict`（默认值是`nonstrict`）,这时就必须指定`limit`来限制输出条数。

表名：`student`

class	name	scores
A	xiaoming	89
A	xiaojun	72
B	xiaohong	88
C	xiaoqiang	92
C	xiaogang	84

按`scores`降序：

```
select * from student order by scores desc;
```

输出

```
C    xiaoqiang    92
A    xiaoming    89
B    xiaohong    88
C    xiaogang    84
A    xiaojun    72
```

② **sort by**

`Hive`中指定了`sort by`，那么在每个`reducer`端都会做排序，也就是说保证了局部有序（每个`reducer`出来的数据是有序的，但是不能保证所有的数据是有序的，除非只有一个`reducer`），好处是：执行了局部排序之后可以为接下去的全局排序提高不少的效率（其实就是做一次归并排序就可以做到全局排序了）。

按`scores`降序：

```
select * from student sort by scores desc;
```

输出：

```
C    xiaoqiang    92
A    xiaoming    89
B    xiaohong    88
C    xiaogang    84
A    xiaojun    72
```

③ **distribute by**

`distribute by`控制`map`输出结果的分发,相同字段的`map`输出会发到一个`reduce`节点去处理。`sort by`为每一个`reducer`产生一个排序文件，他俩一般情况下会结合使用。（这个肯定是全局有序的，因为相同的`class`会放到同一个`reducer`去处理。这里需要注意的是`distribute by`必须要写在`sort by`之前）。

按`scores`降序：

```
select * from student distribute by class sort by scores desc;
```

输出：

```
A    xiaojun    72
C    xiaogang    84
B    xiaohong    88
A    xiaoming    89
C    xiaoqiang    92
```

④ **cluster by**

如果`sort by`和`distribute by`中所用的列相同，可以缩写为`cluster by`以便同时制定两者所用的列`cluster by`的功能就是`distribute by`和`sort by`相结合（注意被`cluster by`指定的列只能是升序，不能指定`asc`和`desc`）。

以下两句`HQL`查询结果相同：

```
select * from student cluster by scores;select * from student distribute by
scores sort by scores desc;
```

输出：

```
A    xiaojun    72
C    xiaogang    84
B    xiaohong    88
A    xiaoming    89
C    xiaoqiang    92
```

limit

在`Hive`查询中要限制查询输出条数， 可以用`limit`关键词指定

只输出`2`条数据：

```
select * from student limit 2;
```

输出：

```
A    xiaoming    89
A    xiaojun    72
```