作业题

1. 找出全部夺得3连贯的队伍
2. team,year
3. 活塞,1990
4. 公牛,1991
5. 公牛,1992
6. 公牛,1993
7. 火箭,1994
8. 火箭,1995
9. 公牛,1996
10. 公牛,1997
11. 公牛,1998
12. 马刺,1999
13. 湖人,2000
14. 湖人,2001
15. 湖人,2002
16. 马刺,2003
17. 活塞,2004
18. 马刺,2005
19. 热火,2006
20. 马刺,2007
21. 凯尔特人,2008
22. 湖人,2009
23. 湖人,2010

create table t1(

team string,

year int

)row format delimited fields terminated by ',';

load data local inpath "/root/data/t1.dat" into table t1;

答案：

1.求连续值的则想到使用排序函数row\_number()

2.根据队伍来分组，根据年份来排序

3.使用年份year -num 得到结果可以作为后面分组计算的依据

4.根据上面的第到的值去分组，并且统计其个数大于三的就是最终结果了

SQL：

select team,num from (select team,year, (year - row\_number() over(partition by team order by year)) num from t1) t2 group by num,team having count(\*) >= 3;

team num

公牛 1990

公牛 1992

湖人 1999

2、找出每个id在在一天之内所有的波峰与波谷值

id,time,price

sh66688,9:35,29.48

sh66688,9:40,28.72

sh66688,9:45,27.74

sh66688,9:50,26.75

sh66688,9:55,27.13

sh66688,10:00,26.30

sh66688,10:05,27.09

sh66688,10:10,26.46

sh66688,10:15,26.11

sh66688,10:20,26.88

sh66688,10:25,27.49

sh66688,10:30,26.70

sh66688,10:35,27.57

sh66688,10:40,28.26

sh66688,10:45,28.03

sh66688,10:50,27.36

sh66688,10:55,26.48

sh66688,11:00,27.41

sh66688,11:05,26.70

sh66688,11:10,27.35

sh66688,11:15,27.35

sh66688,11:20,26.63

sh66688,11:25,26.35

sh66688,11:30,26.81

sh66688,13:00,29.45

sh66688,13:05,29.41

sh66688,13:10,29.10

sh66688,13:15,28.24

sh66688,13:20,28.20

sh66688,13:25,28.59

sh66688,13:30,29.49

sh66688,13:35,30.45

sh66688,13:40,30.31

sh66688,13:45,30.17

sh66688,13:50,30.55

sh66688,13:55,30.75

sh66688,14:00,30.03

sh66688,14:05,29.61

sh66688,14:10,29.96

sh66688,14:15,30.79

sh66688,14:20,29.82

sh66688,14:25,30.09

sh66688,14:30,29.61

sh66688,14:35,29.88

sh66688,14:40,30.36

sh66688,14:45,30.88

sh66688,14:50,30.73

sh66688,14:55,30.76

sh88888,9:35,67.23

sh88888,9:40,66.56

sh88888,9:45,66.73

sh88888,9:50,67.43

sh88888,9:55,67.49

sh88888,10:00,68.34

sh88888,10:05,68.13

sh88888,10:10,67.35

sh88888,10:15,68.13

sh88888,10:20,69.05

sh88888,10:25,69.82

sh88888,10:30,70.62

sh88888,10:35,70.59

sh88888,10:40,70.40

sh88888,10:45,70.29

sh88888,10:50,70.53

sh88888,10:55,70.92

sh88888,11:00,71.13

sh88888,11:05,70.24

sh88888,11:10,70.37

sh88888,11:15,69.79

sh88888,11:20,69.73

sh88888,11:25,70.52

sh88888,11:30,71.23

sh88888,13:00,72.85

sh88888,13:05,73.76

sh88888,13:10,74.72

sh88888,13:15,75.48

sh88888,13:20,75.80

sh88888,13:25,76.74

sh88888,13:30,77.22

sh88888,13:35,77.12

sh88888,13:40,76.90

sh88888,13:45,77.80

sh88888,13:50,78.75

sh88888,13:55,78.30

sh88888,14:00,78.68

sh88888,14:05,78.99

sh88888,14:10,78.35

sh88888,14:15,78.37

sh88888,14:20,78.07

sh88888,14:25,78.80

sh88888,14:30,79.78

sh88888,14:35,79.72

sh88888,14:40,80.71

sh88888,14:45,79.92

sh88888,14:50,80.49

sh88888,14:55,80.44

最终结果与此类似：

id time price feature

sh66688 10:05 27.09 波峰

sh66688 10:15 26.11 波谷

sh66688 10:25 27.49 波峰

sh66688 10:30 26.7 波谷

sh66688 10:40 28.26 波峰

sh66688 10:55 26.48 波谷

sh66688 11:00 27.41 波峰

sh66688 11:05 26.7 波谷

create table t2(

id string,

time string,

price double

)row format delimited fields terminated by ',';

load data local inpath "/root/data/t2.dat" into table t2;

答案：

select id,time,price,

case when price >p1 and price >p2 then '波峰'

when price <p1 and price <p2 then '波谷' end wave from

(select id,time,price, LAG(price,1,price) over (partition by id order by time) p1, LEAD(price,1,price) over (partition by id order by time) p2 from t2) t3

where (price>p1 and price >p2) or (price<p1 and price<p2);

运行结果：

id time price wave

sh66688 10:05 27.09 波峰

sh66688 10:15 26.11 波谷

sh66688 10:25 27.49 波峰

sh66688 10:30 26.7 波谷

sh66688 10:40 28.26 波峰

sh66688 10:55 26.48 波谷

sh66688 11:00 27.41 波峰

sh66688 11:05 26.7 波谷

sh66688 11:25 26.35 波谷

sh66688 13:00 29.45 波峰

sh66688 13:20 28.2 波谷

sh66688 13:35 30.45 波峰

sh66688 13:45 30.17 波谷

sh66688 13:55 30.75 波峰

sh66688 14:05 29.61 波谷

sh66688 14:15 30.79 波峰

sh66688 14:20 29.82 波谷

sh66688 14:25 30.09 波峰

sh66688 14:30 29.61 波谷

sh66688 14:45 30.88 波峰

sh66688 14:50 30.73 波谷

sh66688 14:55 30.76 波峰

sh66688 9:50 26.75 波谷

sh88888 10:10 67.35 波谷

sh88888 10:30 70.62 波峰

sh88888 10:45 70.29 波谷

sh88888 11:00 71.13 波峰

sh88888 11:05 70.24 波谷

sh88888 11:10 70.37 波峰

sh88888 11:20 69.73 波谷

sh88888 13:30 77.22 波峰

sh88888 13:40 76.9 波谷

sh88888 13:50 78.75 波峰

sh88888 13:55 78.3 波谷

sh88888 14:05 78.99 波峰

sh88888 14:10 78.35 波谷

sh88888 14:15 78.37 波峰

sh88888 14:20 78.07 波谷

sh88888 14:30 79.78 波峰

sh88888 14:35 79.72 波谷

sh88888 14:40 80.71 波峰

sh88888 14:45 79.92 波谷

sh88888 14:50 80.49 波峰

sh88888 9:40 66.56 波谷

Time taken: 1.681 seconds, Fetched: 44 row(s)

3、写SQL

3.1、每个id浏览时长、步长 3.2、如果两次浏览之间的间隔超过30分钟，认为是两个不同的浏览时间；再求每个id浏览时长、步长

备注：请仔细阅读计算规则

-- 建表语句

create table t3(

id string,

dt string,

browseid string

)row format delimited fields terminated by '\t';

-- 导入数据

load data local inpath "/root/data/t3.dat" into table t3;

测试数据略，共98条：

934e8bee978a42c7a8dbb4cfa8af0b4f 2020/05/28 17:02 https://www.lagou.com/jobs/9590606.html?show=IEEE1FIJ3106A1H062HA

934e8bee978a42c7a8dbb4cfa8af0b4f 2020/05/28 17:23 <https://www.lagou.com/jobs/998375.html?show=EC1JGEC8G3HJC82JIHCD>

id dt browseid

id：唯一的用户id

dt：用户在这个时间点点击进入了一个页面

browseid：用户浏览了哪个页面

简化数据(以下为某个用户，在一天内的浏览记录)：

1 08:20 1.html

1 08:23 2.html

1 08:24 3.html

1 08:40 4.html

1 09:33 5.html

1 09:40 6.html

1 09:30 7.html

1 09:36 8.html

1 09:37 9.html

1 09:41 a.html

3.1、每个id浏览时长、步长

用户1的浏览时长 = 09:41 - 08:20 = 81分钟

用户1的浏览步长 = count数 = 10次

3.2、如果两次浏览之间的间隔超过30分钟，认为是两个不同的浏览时间；再求每个id浏览时长、步长

用户1在 8:40 - 09:30 的间隔超过了30。生产中认为：

用户1在 08:20 - 08:40 浏览一次网站。这次浏览时长为20分钟，步长为4

用户1在 09:30 - 09:41 又浏览一次网站。这次浏览时长为11分钟，步长为6

对于测试数据SQL1的结果：

934e8bee978a42c7a8dbb4cfa8af0b4f 104.0 13

对于测试数据SQL2的结果：

934e8bee978a42c7a8dbb4cfa8af0b4f 32.0 6

934e8bee978a42c7a8dbb4cfa8af0b4f 35.0 7

答案：

1.按照id分组，使用max(),min()函数计算出当天时间的最大最小值，并且将时间使用UNIX\_TIMESTAMP(）函数按照格式转换为时间戳

2.使用最大值时间戳减去最小的值再去除以60就是分钟差了（时长），如果需要指定格式的话还可以使用round()指定保留小数点

SQL:

select id,(UNIX\_TIMESTAMP(max(dt),"yyyy/MM/dd HH:mm")-UNIX\_TIMESTAMP(min(dt),"yyyy/MM/dd HH:mm"))/60 ,count(\*) as bc from t3 group by id;

结果如下：

id \_c1 bc

022f86d4533740ad914f233cbd9c4430 51.0 8

307d9dce3b7f495ab8ad6033f8c54930 650.0 17

32258fe7130844399859aec54b6df5ff 162.0 13

80ea80b2e5a64cbebfaf34aa797125f0 51.0 5

934e8bee978a42c7a8dbb4cfa8af0b4f 104.0 13

95273392ab1a4579914273cdd1f3a3ae 1242.0 19

de0096ad04ec4273b0462c7da7d79653 658.0 17

f5ae36c6cdda40d5954e08a2d14954a7 40.0 6

Time taken: 1.592 seconds, Fetched: 8 row(s)

判分标准：

1.查询语句逻辑正确，无语法错误，结果正常。（80%）

2.查询方法符合高效执行并符合作业要求。（10%）

3.验证资料完整，能够正确演示结果，代码/语句有注释（10%）