# 日期生成序列

CREATE TABLE TEMP\_DX\_TELECOM\_DATA\_SEQ AS

SELECT B.USER\_CODE,USE\_DATE,

CASE

WHEN USE\_DATE = 'w1d1' THEN

1

WHEN USE\_DATE = 'w1d2' THEN

2

WHEN USE\_DATE = 'w1d3' THEN

3

WHEN USE\_DATE = 'w1d4' THEN

4

WHEN USE\_DATE = 'w1d5' THEN

5

WHEN USE\_DATE = 'w1d6' THEN

6

WHEN USE\_DATE = 'w1d7' THEN

7

WHEN USE\_DATE = 'w2d1' THEN

8

WHEN USE\_DATE = 'w2d2' THEN

9

WHEN USE\_DATE = 'w2d3' THEN

10

WHEN USE\_DATE = 'w2d4' THEN

11

WHEN USE\_DATE = 'w2d5' THEN

12

WHEN USE\_DATE = 'w2d6' THEN

13

WHEN USE\_DATE = 'w2d7' THEN

14

WHEN USE\_DATE = 'w3d1' THEN

15

WHEN USE\_DATE = 'w3d2' THEN

16

WHEN USE\_DATE = 'w3d3' THEN

17

WHEN USE\_DATE = 'w3d4' THEN

18

WHEN USE\_DATE = 'w3d5' THEN

19

WHEN USE\_DATE = 'w3d6' THEN

20

WHEN USE\_DATE = 'w3d7' THEN

21

WHEN USE\_DATE = 'w4d1' THEN

22

WHEN USE\_DATE = 'w4d2' THEN

23

WHEN USE\_DATE = 'w4d3' THEN

24

WHEN USE\_DATE = 'w4d4' THEN

25

WHEN USE\_DATE = 'w4d5' THEN

26

WHEN USE\_DATE = 'w4d6' THEN

27

WHEN USE\_DATE = 'w4d7' THEN

28

WHEN USE\_DATE = 'w5d1' THEN

29

WHEN USE\_DATE = 'w5d2' THEN

30

WHEN USE\_DATE = 'w5d3' THEN

31

WHEN USE\_DATE = 'w5d4' THEN

32

WHEN USE\_DATE = 'w5d5' THEN

33

WHEN USE\_DATE = 'w5d6' THEN

34

WHEN USE\_DATE = 'w5d7' THEN

35

WHEN USE\_DATE = 'w6d1' THEN

36

WHEN USE\_DATE = 'w6d2' THEN

37

WHEN USE\_DATE = 'w6d3' THEN

38

WHEN USE\_DATE = 'w6d4' THEN

39

WHEN USE\_DATE = 'w6d5' THEN

40

WHEN USE\_DATE = 'w6d6' THEN

41

WHEN USE\_DATE = 'w6d7' THEN

42

WHEN USE\_DATE = 'w7d1' THEN

43

WHEN USE\_DATE = 'w7d2' THEN

44

WHEN USE\_DATE = 'w7d3' THEN

45

WHEN USE\_DATE = 'w7d4' THEN

46

WHEN USE\_DATE = 'w7d5' THEN

47

WHEN USE\_DATE = 'w7d6' THEN

48

WHEN USE\_DATE = 'w7d7' THEN

49

ELSE

0

END USE\_SEQ,

VIDEO\_TYPE,

USE\_COUNT

FROM TEMP\_DX\_TELECOM\_DATA\_LOAD A,TEMP\_DX\_TELECOM\_DATA\_USER B

WHERE A.USER\_ID=B.USER\_ID(+)

# 生成用户与日期的全对应

CREATE TABLE TEMP\_DX\_TELECOM\_USER\_SEQ AS

SELECT \* FROM TEMP\_DX\_TELECOM\_DATA\_USER T, TEMP\_DX\_DATE\_SEQ T1

# 生成v1/2/3…10的数据序列

CREATE TABLE TEMP\_DX\_TELECOM\_DATA\_SEQ1 AS

SELECT A.USER\_CODE,

A.USE\_DATE,

A.USE\_SEQ,

'1' VIDEO\_TYPE,

NVL(B.USE\_COUNT, 0) USE\_COUNT

FROM (SELECT \* FROM TEMP\_DX\_TELECOM\_DATA\_SEQ WHERE VIDEO\_TYPE = 'v1') B,

(SELECT \*

FROM TEMP\_DX\_TELECOM\_USER\_SEQ

WHERE USER\_CODE IN (SELECT DISTINCT USER\_CODE

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ

WHERE VIDEO\_TYPE = 'v1')) A

WHERE A.USER\_CODE = B.USER\_CODE(+)

AND A.USE\_SEQ = B.USE\_SEQ(+)

R能处理500万以内数据，根据数据量处理，v1,4,5,10直接剔除少于3个的预测

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **视频编码** | **完整对应表** | **记录数** | **全部预测用户数** | **是否直接采用** | **记录数>3,预测用户数** | **修正后预测用户数，记录数>3或记录数<=3且w7d4后 记录数>=2** | **预测后数据记录，剔除首行初始化** |
| V1 | Seq1 | 9998156 | 204044 | 否 | 99719 | 100219 | 701534 |
| 2 | 2 | 1336181 | 27269 | 否 | 1334 | 1389 | 9724 |
| 3 | 3 | 3984778 | 81322 | 否 | 12439 | 12446 | 87123 |
| 4 | 4 | 12345599 | 251951 | 否 | 142884 | 143297 | 1003080 |
| 5 | 5 | 8122485 | 165765 | 否 | 29787 | 31281 | 218968 |
| 6 | 6 | 4612517 | 94133 | 否 | 9885 | 10024 | 70169 |
| 7 | 7 | 6289003 | 128347 | 否 | 16763 | 17615 | 123306 |
| 8 | 8 | 837900 | 17100 | 否 | 211 | 257 | 1800 |
| 9 | 9 | 5547388 | 113212 | 否 | 7826 | 8572 | 60005 |
| 10 | 10 | 12913754 | 263546 | 否 | 162910 | 163631 | 1145418 |

# 剔除只有3条及以下记录用户

SELECT \*

FROM (SELECT USER\_CODE, COUNT(1)

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ4

WHERE USE\_COUNT > 0

GROUP BY USER\_CODE

HAVING COUNT(1) <= 3) A,

(SELECT USER\_CODE, COUNT(1)

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ4

WHERE USE\_COUNT > 0

AND USE\_DATE >= 'w7d4'

GROUP BY USER\_CODE

HAVING COUNT(1) >= 2) B

WHERE A.USER\_CODE = B.USER\_CODE

针对前7周总记录<=3条的用户，若在w7d4之后有两条以上记录的为可能新加入视频观看用户，所以加入考虑，每种视频增加人数如下：

*-- 1:500,2: 55, 3:7, 4: 413, 5:1494, 6:139, 7: 852, 8:46, 9: 746 ,10: 721*

*调整数据集2-10*

***修正数据集语句:***

CREATE TABLE TEMP\_DX\_TELECOM\_DATA\_SEQ101 AS

SELECT \*

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ10

WHERE USER\_CODE IN (SELECT USER\_CODE

FROM (SELECT USER\_CODE, COUNT(1)

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ10

WHERE USE\_COUNT > 0

GROUP BY USER\_CODE

HAVING COUNT(1) > 3)

UNION ALL

SELECT A.USER\_CODE

FROM (SELECT USER\_CODE, COUNT(1)

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ10

WHERE USE\_COUNT > 0

GROUP BY USER\_CODE

HAVING COUNT(1) <= 3) A,

(SELECT USER\_CODE, COUNT(1)

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ10

WHERE USE\_COUNT > 0

AND USE\_DATE >= 'w7d4'

GROUP BY USER\_CODE

HAVING COUNT(1) >= 2) B

WHERE A.USER\_CODE = B.USER\_CODE)

大数据拆分：

V10 –p2/p1/p0

CREATE TABLE temp\_dx\_telecom\_data\_seq10\_p2 AS

SELECT \*

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ101

WHERE USER\_CODE IN (SELECT USER\_CODE

FROM (SELECT CASE

WHEN USER\_DAYS >= 10 THEN

2

WHEN USER\_DAYS >= 7 THEN

1 *--[7,10)*

ELSE

0 *--[2,7)*

END USER\_ORD,

USER\_CODE,

USER\_DAYS

FROM TEMP\_DX\_TELECOM\_USER\_SEQ101)

WHERE USER\_ORD = 2)

**V4: - p1/p0**

CREATE TABLE temp\_dx\_telecom\_data\_seq4\_p0 AS

SELECT \*

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ41

WHERE USER\_CODE IN (SELECT USER\_CODE

FROM (SELECT CASE

WHEN USER\_DAYS >= 7 THEN

1

ELSE

0 *--[2,7)*

END USER\_ORD,

USER\_CODE,

USER\_DAYS

FROM TEMP\_DX\_TELECOM\_USER\_SEQ41)

WHERE USER\_ORD = 0)

V1**: - p1/p0**

CREATE TABLE temp\_dx\_telecom\_data\_seq1\_p0 AS

SELECT \*

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ11

WHERE USER\_CODE IN (SELECT USER\_CODE

FROM (SELECT CASE

WHEN USER\_DAYS >= 7 THEN

1

ELSE

0 *--[2,7)*

END USER\_ORD,

USER\_CODE,

USER\_DAYS

FROM TEMP\_DX\_TELECOM\_USER\_SEQ11)

WHERE USER\_ORD = 0)

# 训练集表名以2结尾

训练集只有前6周数据，预测第7周数据,调整参数

CREATE TABLE TEMP\_DX\_TELECOM\_DATA\_SEQ12 AS

SELECT \* FROM TEMP\_DX\_TELECOM\_DATA\_SEQ11

WHERE substr(use\_date,1,2)<>'w7'

# V1/V2/…/用户ID

预测用户集

CREATE TABLE TEMP\_DX\_TELECOM\_USER\_SEQ11 AS

SELECT USER\_CODE, SUM(CASE WHEN USE\_COUNT>0 THEN 1 ELSE 0 END) USER\_DAYS

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ11

GROUP BY USER\_CODE

# 预测数据集

create table TEMP\_DX\_TELECOM\_PREDS\_V3

(

ORD\_ID NUMBER,

USER\_CODE VARCHAR2(10),

USE\_SEQ NUMBER,

USE\_COUNT NUMBER

)

;

# 前6周没有观看第7周转粉的各视频用户数

SELECT COUNT(1)

FROM (SELECT USER\_CODE, SUM(USE\_COUNT)

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ6

WHERE USE\_DATE < 'w7d1'

GROUP BY USER\_CODE

HAVING SUM(USE\_COUNT) = 0) A,

(SELECT USER\_CODE, SUM(USE\_COUNT)

FROM TEMP\_DX\_TELECOM\_DATA\_SEQ6

WHERE USE\_DATE LIKE 'w7%'

GROUP BY USER\_CODE

HAVING SUM(USE\_COUNT) > 0) B

WHERE A.USER\_CODE = B.USER\_CODE