

如何解读四种 Oracle 优化器直方图

作者：包光磊 (Todd Bao)

<http://www.shoug.info/archives/todd-bao.html>



本文的适应读者： 初步理解直方图对优化器的作用。
预计阅读时间： 10 分钟。

数据库一共有四种直方图，它们是：频度型直方、顶端频度型直方、高度平衡型直方与混合型直方。

示例中的 DEPARTMENT_ID 字段相异值为 10、20、30、40、50、60、70、80、90、100、110，一共十一个部门。

-- Type 1: 频度型直方

测试（表内行数 82）：

```
begin
  dbms_stats.gather_table_stats(
    'HR',
    'EMPLOYEES',
    method_opt=>'for columns department_id size 100',
    estimate_percent=>dbms_stats.auto_sample_size);
end;
/

define owner='hr'
define tab='employees'
define col='department_id'

select column_name,num_distinct,num_buckets,histogram
from dba_tab_col_statistics
where owner=upper('&owner') and table_name=upper('&tab') and
column_name=upper('&col')
/
select column_name,endpoint_number,endpoint_value from
dba_tab_histograms
where owner=upper('&owner') and table_name=upper('&tab') and
column_name=upper('&col')
order by 1,2
/
undefine owner
undefine tab
undefine col
```

字段名 COLUMN_NAME	字段相异值 NUM_DISTINCT	直方数 NUM_BUCKETS	直方类型 HISTOGRAM
DEPARTMENT_ID	11	11	FREQUENCY

由于计划收集直方数（100）大于实际字段相异值数（11）所以实际直方数等于字段相异值数。在这种情况下，每个相异值分配一个直方（一柱）统计其所占行数，统计的结果在直方收集时是精确到每一个相异值的，其他的类型不具备这一特点。

字段名 COLUMN_NAME	行数 ENDPOINT_NUMBER	字段值 ENDPOINT_VALUE
DEPARTMENT_ID	1	10
DEPARTMENT_ID	3	20
DEPARTMENT_ID	5	30
DEPARTMENT_ID	6	40
DEPARTMENT_ID	31	50
DEPARTMENT_ID	36	60
DEPARTMENT_ID	37	70
DEPARTMENT_ID	71	80
DEPARTMENT_ID	74	90
DEPARTMENT_ID	80	100
DEPARTMENT_ID	82	110

由于行数是累积的，所以 DEPARTMENT_ID 等于 10 的行数是 1、为 20 的是 (3-1) 等于 2、以此类推。

-- Type 2: 顶端频度直方

测试（表内行数 82）：

```

begin
  dbms_stats.gather_table_stats(
    'HR', 'EMPLOYEES',
    method_opt=>'for columns department_id size 5',
    estimate_percent=>dbms_stats.auto_sample_size);
end;
/
define owner='hr'
define tab='employees'
define col='department_id'

```

```

select column_name,num_distinct,num_buckets,histogram
from dba_tab_col_statistics
where owner=upper('&owner') and table_name=upper('&tab') and
column_name=upper('&col')
/

```

```

select column_name,endpoint_number,endpoint_value
from dba_tab_histograms
where owner=upper('&owner') and table_name=upper('&tab') and
column_name=upper('&col')
order by 1,2
/

```

```

undefine owner
undefine tab
undefine col

```

字段名 COLUMN_NAME	字段相异值 NUM_DISTINCT	直方数 NUM_BUCKETS	直方类型 HISTOGRAM
DEPARTMENT_ID	11	5	TOP-FREQUENCY

字段名 COLUMN_NAME	行数 ENDPOINT_NUMBER	字段值 ENDPOINT_VALUE
DEPARTMENT_ID	1	10
DEPARTMENT_ID	26	50
DEPARTMENT_ID	60	80
DEPARTMENT_ID	66	100
DEPARTMENT_ID	67	110

在 11 个部门里选择 5 个值作为直方的统计有三个原因：a. 10 和 110 是两个极值，必须出现；b. 只能再加三个部门是因为直方数限制为 5；c. 50、80 和 100 选择这三个部门的原因是属于这些部门的行数总和最接近内部阈值： $1 - (1/\text{直方数})$ 与行基数的积，本例中为： $1 - (1/5)$ ，即 80% 的行数。

除两个极值以外，大约占总行数 20% 左右的 DEPARTMENT_ID 字段值在此例中没有展现，它们被优化器认为选择度偏高。

与频度直方比较很容易发现这种直方把“矮楼”（选择度高）铲平了，只对“高楼”（选择度低）感兴趣，自带强拆队，有做房地产商的潜质。

×注意：是否测试不成功？是否收集不到这种直方？是否收到的是 HEIGHT BALANCED-高度平衡型的？那么请期待下个版本的 Oracle 数据库。如果下个版本



还是没有这种直方, 请期待下下个版本的 Oracle 的数据库。

— Type 3: 高度平衡型直方

一旦刚才提到的房地产商出现, 我们就可以和高度平衡型直方说再见了。为了忘却的纪念让我再介绍一下它。

测试 (表内行数 107)

省略收集...

字段名 COLUMN_NAME	字段相异值 NUM_DISTINCT	直方数 NUM_BUCKETS	直方类型 HISTOGRAM
DEPARTMENT_ID	11	5	HEIGHT BALANCED

字段名 COLUMN_NAME	行单位数 ENDPOINT_NUMBER	字段值 ENDPOINT_VALUE
DEPARTMENT_ID	0	10
DEPARTMENT_ID	2	50
DEPARTMENT_ID	4	80
DEPARTMENT_ID	5	110

查看此类直方必须了解表的行基数, 此例为 107。

一个行单位等于行数除以直方数=107/5 约 21.

所以, 估计各行数是

- 10 号部门: $21 * 0 =$ 大约 0, 当然不可能是 0 行, 极少的意思
- 20-50 号部门: $21 * (2-0) =$ 大约 42
- 60-80 号部门: $21 * (4-2) =$ 大约 42
- 90-110 号部门: $21 * (5-4) =$ 大约 21

与实际比较一下：

```
SYS@fmw//Scripts> select count(*) from hr.employees where department_id  
between 11 and 50;
```

```
  COUNT(*)  
-----  
         54
```

```
SYS@fmw//Scripts> select count(*) from hr.employees where department_id  
between 51 and 80;
```

```
  COUNT(*)  
-----  
         40
```

```
SYS@fmw//Scripts> select count(*) from hr.employees where department_id  
between 81 and 110;
```

```
  COUNT(*)  
-----  
         11
```

-- Type 4: 混合型直方

Oracle 多加一个字段来描述这种直方：ENDPOINT_REPEAT_COUNT。

测试（表内行数 82）

字段名	字段相异值	直方数	直方类型
COLUMN_NAME	NUM_DISTINCT	NUM_BUCKETS	HISTOGRAM
-----	-----	-----	-----
DEPARTMENT_ID	11	4	HYBRID

字段名 COLUMN_NAME	行数 ENDPOINT_NUMBER	字段值 ENDPOINT_VALUE	字段值重复度 ENDPOINT_REPEAT_COUNT
DEPARTMENT_ID	1	10	1
DEPARTMENT_ID	31	50	25
DEPARTMENT_ID	71	80	34
DEPARTMENT_ID	82	110	2

10 号部门的行数: 1;
 20-50 号部门的行数: 31-1=30 行, 其中 50 号部门 25 行;
 60-80 号部门的行数: 71-31=40 行, 其中 80 号部门 34 行;
 90-110 号部门的行数: 82-71=11 行, 其中 110 号部门 2 行。



什么, 没有这种直方? ← 请参考 Type 2.

四种直方的解读至此介绍完毕。

参考资料:

www.oracle-base.com/blog/2012/10/06/oracle-openworld-2012-day-5/

郑重声明: 本文仅代表个人观点, 不代表 Oracle 公司。

包光磊

新浪微博 <http://weibo.com/toddbao>

我主持的国内 ADF/SOA 论坛: <http://www.databi.cn/forum-57-1.html>

SHOUG 的网址: www.shoug.info 找我们:

The screenshot shows a Google search interface with the query '上海Oracle用户组'. The search results show 'About 234,000 results (0.34 seconds)'. The top result is '上海Oracle用户组 | SHOUG, 走近全系Oracle技术和数据库专家' with the URL 'www.shoug.info/'. Below it is another result 'SHOUG成员 | 上海Oracle用户组' with the URL 'www.shoug.info/archives/category/shoug成员'. The search tools bar includes 'Web', 'Images', 'Maps', 'More', and 'Search tools'.