# Hive 入门 Group By 全案例【附代码】

原创 Lenis 有关SQL 2018-11-09

收录于话题
#数据技客回忆录 214个

不明就里的读者可以看上一篇：

Hive 的入门级 Group By 全案例

昨晚发文之后，有读者陆陆续续在星球发问了，脚本到底该怎么写？

当然也有星友在第一时间拿出了自己的方案，工工整整，让我好生钦佩。

不废话了，下面是大家想看的具体实现。

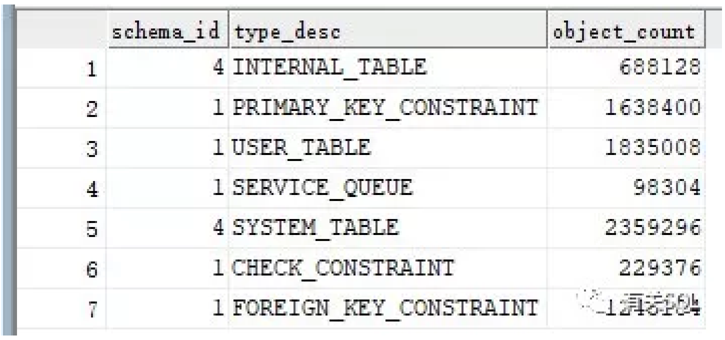
**环境：**

```
Hive: 2.7.7
Oracle SQL Developer
Cloudera JDBC Driver
```

## 案例 - 1 ： Group by 的常规化应用

```
select  schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by schema_id,type_desc
```

结果：



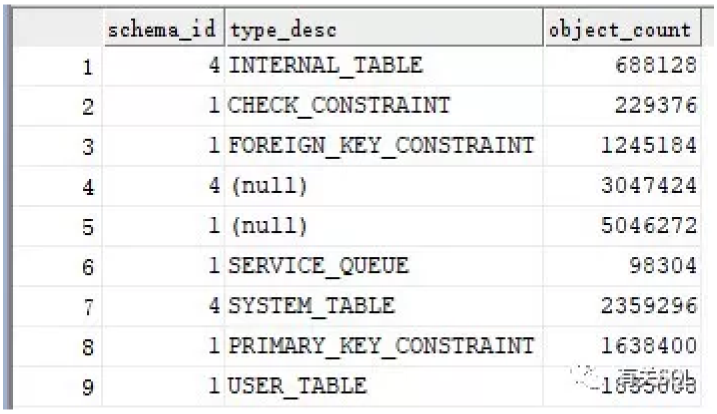
|   | schema_id | type_desc | object_count |
|---|---|---|---|
| 1 | 4 | INTERNAL_TABLE | 688128 |
| 2 | 1 | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 3 | 1 | USER_TABLE | 1835008 |
| 4 | 1 | SERVICE_QUEUE | 98304 |
| 5 | 4 | SYSTEM_TABLE | 2359296 |
| 6 | 1 | CHECK_CONSTRAINT | 229376 |
| 7 | 1 | FOREIGN_KEY_CONSTRAINT | 1245184 |

image

## 案例 - 2： Group by 之 Grouping Sets 应用

```
select  schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by schema_id,type_desc
grouping sets((schema_id,type_desc),schema_id)
```

结果:

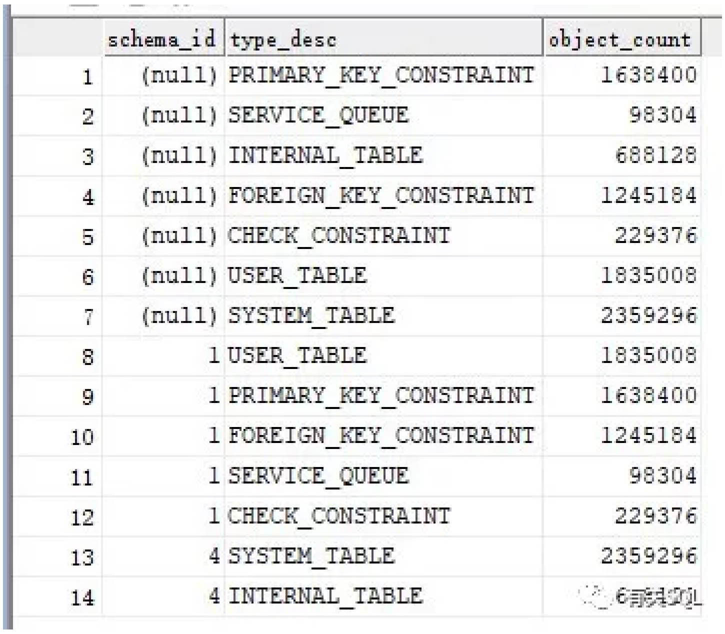


| | schema_id | type_desc | object_count |
|---|---|---|---|
| 1 | 4 | INTERNAL_TABLE | 688128 |
| 2 | 1 | CHECK_CONSTRAINT | 229376 |
| 3 | 1 | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 4 | 4 | (null) | 3047424 |
| 5 | 1 | (null) | 5046272 |
| 6 | 1 | SERVICE_QUEUE | 98304 |
| 7 | 4 | SYSTEM_TABLE | 2359296 |
| 8 | 1 | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 9 | 1 | USER_TABLE | 1835008 |

image

```
select  schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by schema_id,type_desc
grouping sets((schema_id,type_desc),type_desc)
```

结果:

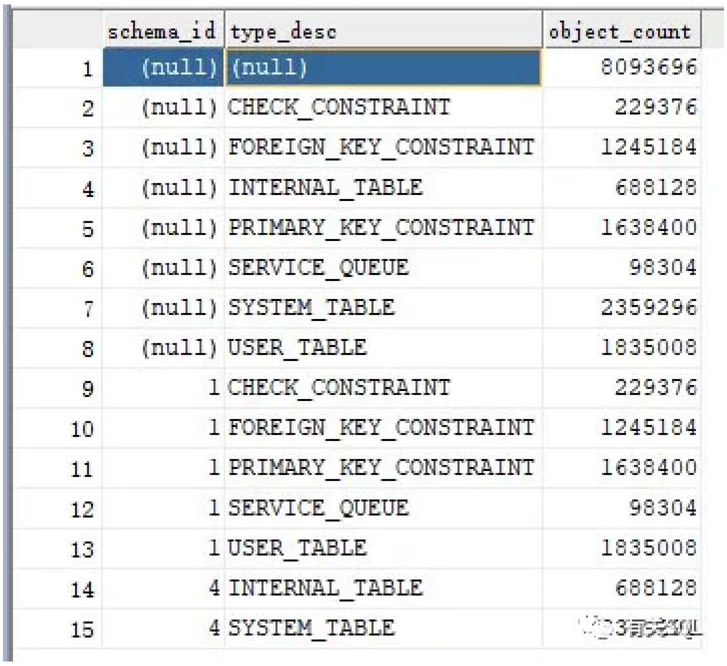


| | schema_id | type_desc | object_count |
|---|---|---|---|
| 1 | (null) | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 2 | (null) | SERVICE_QUEUE | 98304 |
| 3 | (null) | INTERNAL_TABLE | 688128 |
| 4 | (null) | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 5 | (null) | CHECK_CONSTRAINT | 229376 |
| 6 | (null) | USER_TABLE | 1835008 |
| 7 | (null) | SYSTEM_TABLE | 2359296 |
| 8 | 1 | USER_TABLE | 1835008 |
| 9 | 1 | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 10 | 1 | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 11 | 1 | SERVICE_QUEUE | 98304 |
| 12 | 1 | CHECK_CONSTRAINT | 229376 |
| 13 | 4 | SYSTEM_TABLE | 2359296 |
| 14 | 4 | INTERNAL_TABLE | 6[illegible] |

image

```
select  schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by schema_id,type_desc
grouping sets((schema_id,type_desc),type_desc,())
order by schema_id ,type_desc
```

结果:

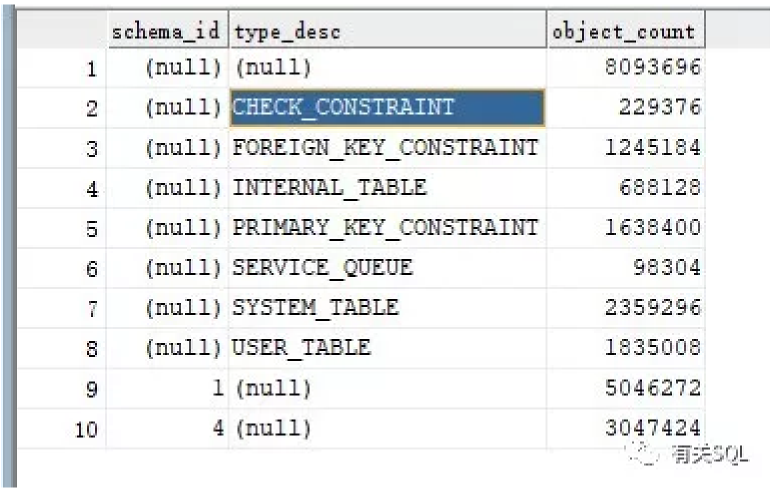


| | schema_id | type_desc | object_count |
|---|---|---|---|
| 1 | (null) | (null) | 8093696 |
| 2 | (null) | CHECK_CONSTRAINT | 229376 |
| 3 | (null) | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 4 | (null) | INTERNAL_TABLE | 688128 |
| 5 | (null) | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 6 | (null) | SERVICE_QUEUE | 98304 |
| 7 | (null) | SYSTEM_TABLE | 2359296 |
| 8 | (null) | USER_TABLE | 1835008 |
| 9 | 1 | CHECK_CONSTRAINT | 229376 |
| 10 | 1 | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 11 | 1 | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 12 | 1 | SERVICE_QUEUE | 98304 |
| 13 | 1 | USER_TABLE | 1835008 |
| 14 | 4 | INTERNAL_TABLE | 688128 |
| 15 | 4 | SYSTEM_TABLE | [illegible] |

image

```
select  schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by schema_id,type_desc
grouping sets(schema_id,type_desc,())
order by schema_id ,type_desc
```

结果:



| | schema_id | type_desc | object_count |
|---|---|---|---|
| 1 | (null) | (null) | 8093696 |
| 2 | (null) | CHECK_CONSTRAINT | 229376 |
| 3 | (null) | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 4 | (null) | INTERNAL_TABLE | 688128 |
| 5 | (null) | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 6 | (null) | SERVICE_QUEUE | 98304 |
| 7 | (null) | SYSTEM_TABLE | 2359296 |
| 8 | (null) | USER_TABLE | 1835008 |
| 9 | 1 | (null) | 5046272 |
| 10 | 4 | (null) | 3047424 |

image

**结论:**

grouping sets 的作用就是将选定的分组字段，再分子组进行汇总。

(schema_id,type_desc) 用来指定细分字段组合;

单个字段，比如 schema_id, type_desc 用来指定细分的单个字段;

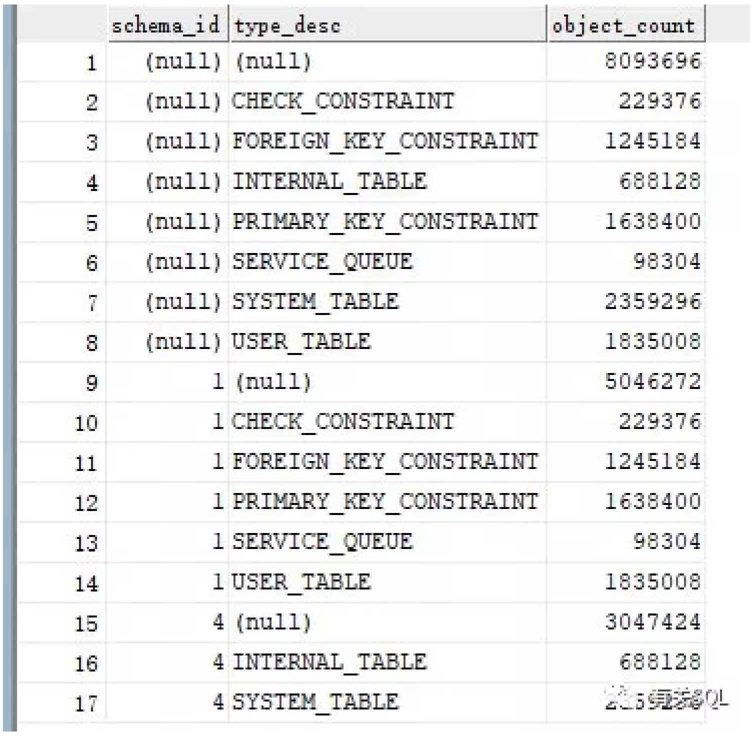
()用来计算总和，总计等，目标对象是符合条件的所有数据，即相当于没有使用字段做 group by 的聚合计算。

最终将这些 grouping sets 里面指定的细分字段聚合得到的结果联合在一个结果集而展现出来。

### 案例 - 3： Group by 之 with cube

```
select schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by schema_id,type_desc
with cube
order by schema_id ,type_desc
```

结果：



|  | schema_id | type_desc | object_count |
|---|---|---|---|
| 1 | (null) | (null) | 8093696 |
| 2 | (null) | CHECK_CONSTRAINT | 229376 |
| 3 | (null) | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 4 | (null) | INTERNAL_TABLE | 688128 |
| 5 | (null) | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 6 | (null) | SERVICE_QUEUE | 98304 |
| 7 | (null) | SYSTEM_TABLE | 2359296 |
| 8 | (null) | USER_TABLE | 1835008 |
| 9 | 1 | (null) | 5046272 |
| 10 | 1 | CHECK_CONSTRAINT | 229376 |
| 11 | 1 | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 12 | 1 | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 13 | 1 | SERVICE_QUEUE | 98304 |
| 14 | 1 | USER_TABLE | 1835008 |
| 15 | 4 | (null) | 3047424 |
| 16 | 4 | INTERNAL_TABLE | 688128 |
| 17 | 4 | SYSTEM_TABLE | 2359296 |

image

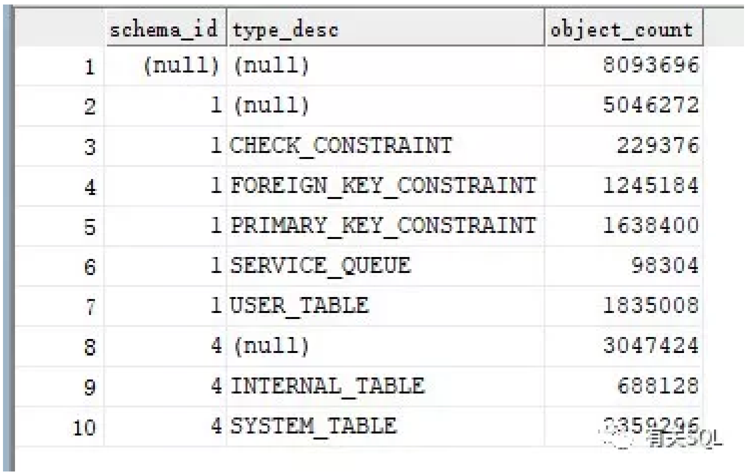
相当于是以下 grouping sets 的简化版本

```
select  schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by schema_id,type_desc
grouping sets((schema_id,type_desc),schema_id,type_desc,())
order by schema_id ,type_desc
```

## 案例 - 4： Group by 之 with rollup

这是一个上卷的操作，唯一一个有方向性的分组聚合操作

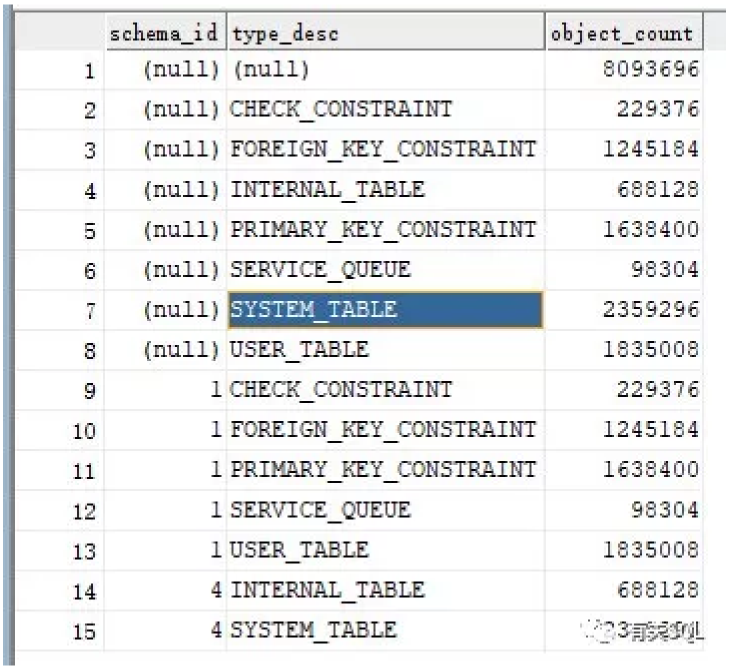
```
select schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by schema_id,type_desc
with rollup
order by schema_id ,type_desc
```



| | schema_id | type_desc | object_count |
|---|---|---|---|
| 1 | (null) | (null) | 8093696 |
| 2 | 1 | (null) | 5046272 |
| 3 | 1 | CHECK_CONSTRAINT | 229376 |
| 4 | 1 | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 5 | 1 | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 6 | 1 | SERVICE_QUEUE | 98304 |
| 7 | 1 | USER_TABLE | 1835008 |
| 8 | 4 | (null) | 3047424 |
| 9 | 4 | INTERNAL_TABLE | 688128 |
| 10 | 4 | SYSTEM_TABLE | 2359296 |

image

```
select schema_id
    ,   type_desc
    ,   count(object_id) as object_count
from tblobj2
group by type_desc,schema_id
with rollup
order by schema_id ,type_desc
```



| | schema_id | type_desc | object_count |
|---|---|---|---|
| 1 | (null) | (null) | 8093696 |
| 2 | (null) | CHECK_CONSTRAINT | 229376 |
| 3 | (null) | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 4 | (null) | INTERNAL_TABLE | 688128 |
| 5 | (null) | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 6 | (null) | SERVICE_QUEUE | 98304 |
| 7 | (null) | SYSTEM_TABLE | 2359296 |
| 8 | (null) | USER_TABLE | 1835008 |
| 9 | 1 | CHECK_CONSTRAINT | 229376 |
| 10 | 1 | FOREIGN_KEY_CONSTRAINT | 1245184 |
| 11 | 1 | PRIMARY_KEY_CONSTRAINT | 1638400 |
| 12 | 1 | SERVICE_QUEUE | 98304 |
| 13 | 1 | USER_TABLE | 1835008 |
| 14 | 4 | INTERNAL_TABLE | 688128 |
| 15 | 4 | SYSTEM_TABLE | [illegible] |

image

按照分组字段从右到左的上卷汇总，最后汇总所有符合条件的数据到一个结果集。

---