

# Hive 的入门级 Group By 全案例

原创 Lenis 有关SQL 2018-11-08

收录于话题

#数据技客回忆录

214个

之前总是用全家桶方式玩大数据栈，总觉得有点儿戏。

这两天把自己的 Hadoop/Hive/Spark 集群环境搭好了，准备正式的做点试验，写点文章。

**所以干货文章即将到来，小伙伴们，你们的赞准备好了嘛？**

我这里用到一张表，叫做 tblobj2. 熟悉 sql server 一定不陌生，其实就是从 sql server 导了一张系统表 sys.objects 到 Hive 里面。具体方法可以参考这里：

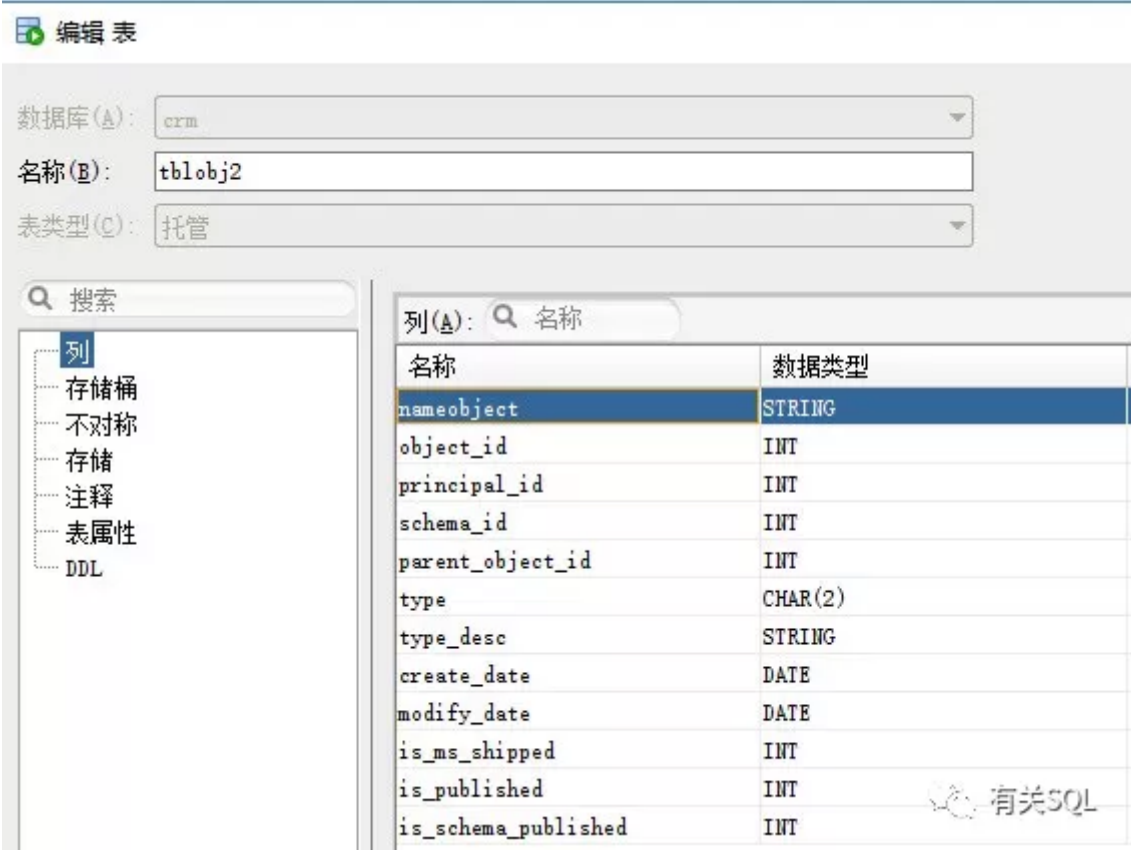
使用 Sqoop 将 30W+ MySQL 数据导入 Hive

**这是 Hive 的第一篇公开文，讲解 group by 用法。**

其余的文章存着，大家热情起来了，我再慢慢放。觉得小编嘚瑟的朋友，砖可以拍过来了。

扯远了，回归正题，这里是 5 道 Hive 的 group by 应用题，大家有兴趣先做着。我会在星球里公布正式答案。

已知表结构如下：



image

表的前 10 行数据 sample 如下：

select * from tblobj2 limit 10											
tblobj2.nameobject	tblobj2.object_id	tblobj2.principal_id	tblobj2.schema_id	tblobj2.parent_object_id	tblobj2.type	tblobj2.type_desc	tblobj2.create_date	tblobj2.modify_date	tblobj2.is_ms_shipped	tblobj2.is_published	tblobj2.is_schema_published
1 plan_persist_query	161575605	(null)	4	0 IT	INTERNAL_TABLE	(null)	(null)	(null)	(null)	(null)	(null)
2 NOTIFICATION_SEQUENCE	194099732	(null)	1	0 U	USER_TABLE	(null)	(null)	(null)	(null)	(null)	(null)
3 plan_persist_plan	197878742	(null)	4	0 IT	INTERNAL_TABLE	(null)	(null)	(null)	(null)	(null)	(null)
4 NOTIFICATION_SEQUENCE_PK	210099789	(null)	1	194099732 PK	PRIMARY_KEY_CONSTRAINT	(null)	(null)	(null)	(null)	(null)	(null)
5 plan_persist_runtime_stats	213878789	(null)	4	0 IT	INTERNAL_TABLE	(null)	(null)	(null)	(null)	(null)	(null)
6 IDMS_FK1	224099946	(null)	1	917878307 F	FOREIGN_KEY_CONSTRAINT	(null)	(null)	(null)	(null)	(null)	(null)
7 plan_persist_runtime_stats_interval	229878656	(null)	4	0 IT	INTERNAL_TABLE	(null)	(null)	(null)	(null)	(null)	(null)
8 IDMS_FK2	242099903	(null)	1	917878307 F	FOREIGN_KEY_CONSTRAINT	(null)	(null)	(null)	(null)	(null)	(null)
9 plan_persist_context_settings	249878913	(null)	4	0 IT	INTERNAL_TABLE	(null)	(null)	(null)	(null)	(null)	(null)
10 IDMS_FK3	250099960	(null)	1	917878307 F	FOREIGN_KEY_CONSTRAINT	(null)	(null)	(null)	(null)	(null)	(null)

image

需求得：

1. 按照 schema\_id, type\_desc 为分组的记录总数，如下：

	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	1246104

image

2. 按照 schema\_id, type\_desc 为分组的记录总数，以及按照 schema\_id 为分组的记录总数，且两个分组的记录总数需要合并到一个结果集，如下：

	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

3.按照 schema\_id, type\_desc 为分组的记录总数, 以及按照 type\_desc 为分组的记录总数, 且两个分组的记录总数需要合并到一个结果集, 如下:

	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	2359296

image

4. 按照schema\_id, type\_desc 各自为分组, 并汇总所有数据的总数, 最终结果展示在一个结果集, 如下:

	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

5. 按照 schema\_id + type\_desc, schema\_id 为分组依据求分组总数, 并合并所有数据总计到一个结果集:

	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	3359296

image

要求:

必须使用一个 SELECT ..Group by 求解, 而不是 union all/union

其实不仅仅是 Hive, SQL Server/Oracle 都有自己的 Group by 子选项案例。这里有篇旧文, 可供参考:

真以为自己懂 Group By 了?

想了解 Hadoop/Hive/Spark 集群搭建, 别求公司的 DevOps 大师们了, 他们是爷爷不会理你的。开玩笑啦, 其实他们才忙呢, 自个儿能解决的问题, 作为 IT 人别偷懒就是了。看这里:

Spark SQL 与 Hive 的第一场会师

Spark 高难度对话 SQL Server 后记

周末两三事儿: 大数据专栏以及百题SQL学习营

推荐刚认识的朋友写的号, 他们都非常了不起。向往 Python 编程路的你, 可以关注。本想自己写写 Python 的, 看了他们的号, 我都觉得没这个必要了。但后期还是会写的, 但方向肯定有所转变, 暂时保密。