hive 模拟update 操作

笔记本： 大数据

创建时间： 2015/7/30 14:48 **更新时间：** 2015/7/30 14:50

在使用hive构建数据仓库的时候，经常会遇到从mysql或者oracle数据导入到hive中。

一般情况下，每天导一次数据，有些数据需要更新操作，最典型的例子就是订单数据，比如：

订单创建时间	订单导入hive时间	订单更新时间
2015-07-01 12:09:11	2015-07-02 00:30:00	2015-07-02 20:07:12

这个例子就很明显，如果按照订单生成时间导数据的话，订单状态更新后的数据就无法同步到hive 的数据仓库。

因为hive 不支持update操作，那么如何合并订单历史和更新数据呢？

那就使用hive模拟update操作。

hortonworks有一篇文章：http://zh.hortonworks.com/blog/four-step-strategy-incremental-updates-hive/

这篇文章就是说如何使用hive现有的操作来达到update数据的效果。

按照这么文章的操作，也会遇到一些问题：

1、join容易产生多条记录

2、jion操作比较耗时

这里我们推荐一种更轻快的解决办法，使用rank函数

以下面代码为例：

```
[java] view plain copy print ?
01. create view order_view as
02. select t2.* from (
03.       select t1.*,rank() over (partition by order_id order by update_time desc,coalesce(update_time,0),rand()) as order
04.       from (select * from order_his
05.             union all
06.            select * from order_daily) t1
07. ) t2  where order_rank=1;
08. drop table order_all;
09. create table order_all as select * from order_view;
10. insert overwrite table order_his select * from order_all;
```

order_his： 订单历史表

order_dail：每天新导入的订单数据，这里不仅要考虑创建时间是前一天的订单，还有将update_time 是前一天的订单导出来

对同一个order_id 的数据按照更新时间排序，取最晚的一条，也就是最新的数据。

这种方式即简单又安全，因为在实际应用中，经常遇到jion数据出现重复数据问题