无窗口、窗口 两类SQL的业务含义

概念

窗口函数：

Group Window Aggregation 是一种在窗口上的聚合。Group Window Aggregation 是每个窗口结束发出一条结果数据（无early fire时），有点类似 micro batch。 最常见的是几种时间窗口，如 TUMBLE（滚动窗口），HOP（滑动窗口）, SESSION（会话窗口）。例如有用户想统计在过去的1分钟内有多少用户点击了某个的网页。在这种情况下，我们可以定义一个窗口，用来收集最近一分钟内的数据，并对这个窗口内的数据进行计算。

案例

测试数据：

username (VARCHAR) click_url (VARCHAR) ts (TIMESTAMP)
Jark http://taobao.com/xxx 2017-10-10 10:00:00.0
Jark http://taobao.com/xxx 2017-10-10 10:00:10.0
Jark http://taobao.com/xxx 2017-10-10 10:00:49.0
Jark http://taobao.com/xxx 2017-10-10 10:01:05.0
Jark http://taobao.com/xxx 2017-10-10 10:01:58.0
Timo http://taobao.com/xxx 2017-10-10 10:02:10.0

```
INSERT INTO tumble_output
SELECT
TUMBLE_START(ts, INTERVAL '1' MINUTE),
TUMBLE_END(ts, INTERVAL '1' MINUTE),
username,
COUNT(click_url)
FROM window_input
GROUP BY TUMBLE(ts, INTERVAL '1' MINUTE), username
```

测试结果：

window_start (TIMESTAMP) window_end (TIMESTAMP) username (VARCHAR) clicks (BIGINT)
2017-10-10 10:00:00.0 2017-10-10 10:01:00.0 Jark 3
2017-10-10 10:01:00.0 2017-10-10 10:02:00.0 Jark 2
2017-10-10 10:02:00.0 2017-10-10 10:03:00.0 Timo 1

Group Aggregation：

是一种在无限流上的聚合，由于没有窗口，是无限大窗口上的聚合，所以计算模式是每到达一条数据就会增量计算一次，并发出更新后的结果。关于状态保存，流计算分三层来保存状态：内存为第一层，它的主要作用其实更

多的作为缓存来使用。硬盘位第二层，流计算使用是SSD做存储。HDFS为第三层。所以不存在内存爆掉的的问题存在。

测试数据

Customer OrderPrice
Bush 1000
Carter 1600
Bush 700
Bush 300
Adams 2000
Carter 100
SELECT Customer,SUM(OrderPrice) FROM XXXX
GROUP BY Customer;
测试结果

Customer SUM(OrderPrice)
Bush 2000
Carter 1700
Adams 2000