TPC-H的22条查询语句分析(Q1-Q5)

2017-05-23 Prof.Guo DatabaseG

上一篇推文（**数据库基准测试TPC**）中说到TPC-H基准测试中有22 条查询分析。下面分4篇推文分析这22 个查询。同学们可以逐步在1G，10G....10T数据内分别进行测试。

使用TPC-H进行性能测试，需要有很多工作配合才能获得较高性能，如建立索引，表数据的合理分布（使用表空间和聚簇技术）等。

本文从查询优化技术的角度，对TPC-H的22条查询语句和主流数据库执行每条语句对应的查询执行计划进行分析，目的在于了解各个主流数据库的查询优化技术，以TPC-H实例进一步掌握查询优化技术，对比主流数据库的实现情况对查询优化技术融会贯通。

1 Q1: 价格统计报告查询

Q1语句是查询lineItems的一个定价总结报告。在单个表lineitem上查询某个时间段内，对已经付款的、已经运送的等各类商品进行统计，包括业务量的计费、发货、折扣、税、平均价格等信息。

解决的商业问题：价格摘要报告查询提供了给定日期的运送的所有行的价格摘要报告，这个日期在数据库包含的最大的运送日期的60 - 120天以内。查询列出了扩展价格、打折的扩展价格、打折的扩展价格加税收、平均数量、平均扩展价格和平均折扣的总和。这些统计值根据RETURNFLAG和LINESTATUS进行分组，并按照RETURNFLAG和LINESTATUS的升序排列。每一组都给出所包含的行数。

Q1语句的特点是：带有分组、排序、聚集操作并存的单表查询操作。这个查询会导致表上的数据有95%到97%行被读取到。

Q1的查询语句如下：

```
select
    l_returnflag, //返回标志
    l_linestatus,
    sum(l_quantity) as sum_qty, //总的数量
    sum(l_extendedprice) as sum_base_price, //聚集函数操作
    sum(l_extendedprice * (1 - l_discount)) as sum_disc_price,
    sum(l_extendedprice * (1 - l_discount) * (1 + l_tax)) as sum_charge,
    avg(l_quantity) as avg_qty,
    avg(l_extendedprice) as avg_price,
```

```
    avg(l_discount) as avg_disc,
    count(*) as count_order //每个分组所包含的行数
from
    lineitem
where
    l_shipdate <= date'1998-12-01' - interval '90' day //时间段是随机生成的
group by //分组操作
    l_returnflag,
    l_linestatus
order by //排序操作
    l_returnflag,
    l_linestatus;
```

2 Q2: 最小代价供货商查询

Q2语句查询获得最小代价的供货商。得到给定的区域内，对于指定的零件（某一类型和大小的零件），哪个供应者能以最低的价格供应它，就可以选择哪个供应者来订货。

Q2语句的特点是：带有排序、聚集操作、子查询并存的多表查询操作。查询语句没有从语法上限制返回多少条元组，但是TPC-H标准规定，查询结果只返回前100行（通常依赖于应用程序实现）。

Q2的查询语句如下：

```
select
     s_acctbal,
    s_name,
    n_name,
    p_partkey,
    p_mfgr,
    s_address,
    s_phone,
     s_comment /*查询供应者的帐户余额、名字、国家、零件的号码、生产者、供应者的地址、电话号码、备注信息 */
from
    part,
    supplier,
    partsupp,
```

```
    nation,
    region //五表连接
where
    p_partkey = ps_partkey
    and s_suppkey = ps_suppkey
    and p_size = [SIZE] //指定大小，在区间[1, 50]内随机选择
    and p_type like '%[TYPE]' //指定类型，在TPC-H标准指定的范围内随机选择
    and s_nationkey = n_nationkey
    and n_regionkey = r_regionkey
    and r_name = '[REGION]' //指定地区，在TPC-H标准指定的范围内随机选择
    and ps_supplycost = ( //子查询
        select
            min(ps_supplycost) //聚集函数
        from
            partsupp, supplier, nation, region //与父查询的表有重叠
        where
            p_partkey = ps_partkey
            and s_suppkey = ps_suppkey
            and s_nationkey = n_nationkey
            and n_regionkey = r_regionkey
            and r_name = '[REGION]'
    )
order by //排序
    s_acctbal desc,
    n_name,
    s_name,
    p_partkey;
```

3 Q3: 运送优先级查询

Q3语句查询得到收入在前10位的尚未运送的订单。在指定的日期之前还没有运送的订单中具有最大收入的订单的运送优先级（订单按照收入的降序排序）和潜在的收入（潜在的收入为l_extendedprice * (1-l_discount)的和）。

Q3语句的特点是：带有分组、排序、聚集操作并存的三表查询操作。查询语句没有从语法上限制返回多少条元组，但是TPC-H标准规定，查询结果只返回前10行（通常依赖于应用程序实现）。

Q3的查询语句如下：

```
select
    l_orderkey,
    sum(l_extendedprice*(1-l_discount)) as revenue, //潜在的收入，聚集操作
    o_orderdate,
    o_shippriority
from
    customer,
    orders,
    ineitem //三表连接
where
    c_mktsegment = '[SEGMENT]' //在TPC-H标准指定的范围内随机选择
    and c_custkey = o_custkey
    and l_orderkey = o_orderkey
    and o_orderdate < date '[DATE]' //指定日期段，在在[1995-03-01, 1995-03-31]中随机选择
    and l_shipdate > date '[DATE]'
group by //分组操作
    l_orderkey, //订单标识
    o_orderdate, //订单日期
    o_shippriority //运输优先级
order by //排序操作
    revenue desc, //降序排序，把潜在最大收入列在前面
    o_orderdate;
```

4 Q4: 订单优先级查询

Q4语句查询得到订单优先级统计值。计算给定的某三个月的订单的数量，在每个订单中至少有一行由顾客在它的提交日期之后收到。

Q4语句的特点是：带有分组、排序、聚集操作、子查询并存的单表查询操作。子查询是相关子查询。

Q4的查询语句如下：

```
select
    o_orderpriority, //订单优先级
    count(*) as order_count //订单优先级计数
from
    orders //单表查询
where
    o_orderdate >= date '[DATE]'
    and o_orderdate < date '[DATE]' + interval '3' month //指定订单的时间段--某三个月，DATE是在1993年1月和1997年10月之间随机选择的一个月的第一天
    and exists ( //子查询
        select *
        from
        lineitem
        where
            l_orderkey = o_orderkey
            and l_commitdate < l_receiptdate
)
group by //按订单优先级分组
    o_orderpriority
order by //按订单优先级排序
    o_orderpriority;
```

5 Q5: 供货商为公司带来的收入查询

Q5语句查询得到通过某个地区零件供货商而获得的收入（收入按sum(l_extendedprice * (1 - l_discount))计算）统计信息。可用于决定在给定的区域是否需要建立一个当地分配中心。

Q5语句的特点是：带有分组、排序、聚集操作、子查询并存的多表连接查询操作。

Q5的查询语句如下：

```
select   n_name,
    sum(l_extendedprice * (1 - l_discount)) as revenue //聚集操作
```

```
from
    customer,
    orders,
    lineitem,
    supplier,
    nation,
    region //六表连接
where
    c_custkey = o_custkey
    and l_orderkey = o_orderkey
    and l_suppkey = s_suppkey
    and c_nationkey = s_nationkey
    and s_nationkey = n_nationkey
    and n_regionkey = r_regionkey
    and r_name = '[REGION]' //指定地区，在TPC-H标准指定的范围内随机选择
    and o_orderdate >= date '[DATE]' //DATE是从1993年到1997年中随机选择的一年的1月1
日
    and o_orderdate < date '[DATE]' + interval '1' year
group by //按名字分组
    n_name
    order by //按收入降序排序，注意分组和排序子句不同
    revenue desc;
```

喜欢就点赞，爱就转发。

长按“识别图中二维码”关注