数据库专题训练 · Lab5

计01　容逸朗　2020010869

实验目的

1. 初步了解数据库系统的查询计划优化过程；
2. 学习简单的基数估计算法和连接顺序选择算法，加强对逻辑优化过程的理解。

基础实验内容

1. 补全直方图的构建和估计函数

- 初始化时根据桶数量 `num_buckets_` 分桶，计算每个区间的长度 `width_`；
- 然后把数据按照公式找到对应的桶并插入即可：

$$tar = \min\left(\left\lfloor \frac{val - min}{width} \right\rfloor, num_{buckets}\right)$$

- 对于估计函数同理。

2. 补全读取表的数据并构建直方图的过程

- 利用 `TableScanNode` 和对应的 `Next` 方法取得表中所有数据项，然后创建新的直方图并传入数据，同时维护表项和直方图的对应关系 `stats_map_`。

3. 补全 Filter 算子的基数估计过程

- 首先判断 `cond_` 是否为空，若为空则返回子结点的 `Cost`；
- 然后判断 `cond_` 是 `AlgebraCondition` 还是 `AndCondition`，
- 若结点是 `AlgebraCondition`，则先调用 `UpdateBound` 取得对应数据的上下界，并根据上下界情况调用对应的直方图估计函数，得到的结果与子结点 `Cost` 的乘积为本结点的估计值；
- 若结点为 `AndCondition`，那么可以忽略所有不是 `AlgebraCondition` 的子结点，然后剩余结点按上面的方式处理即可，得到结果的乘积再和子结点的 `Cost` 相乘即为 `AndCondition` 的估计值。

4. 完成 Optimizer 中连接顺序优化过程

- 首先利用 `UndirectedGraph` 类构建无向图，具体方式是遍历 `table_filter_` 中所有含有两个表的项，然后在无向图中标记两个表的连接关系；
- 下一步是取得上一步得到的所有表格的对应的估计值，找出最小者并将其加入小根堆中；
- 然后不断从小根堆中取出最小元素，并把与之相连（且未曾加入过堆）的结点加入小根堆；（注意标记新加入结点和旧结点的连接关系，这样可以方便得到最终的连接顺序）
- 不断取值，直至小根堆为空，此时得到了完整的连接顺序。

总结

- 基础功能 Commit ID：`0aa076fad26f7861629f245ac6d44e8690ff79a0`（位于 `ch5` 分支）
- 合计用时 5 小时。