Optimization of Common Table Expressions in MPP Database Systems

Amr El-Helw, Venkatesh Raghavan, Mohamed A. Soliman, George Caragea, Zhongxian Gu, Michalis Petropoulos

I. Motivation

Example Query:
WITH v AS (SELECT i_brand, i_color FROM item WHERE i_current_price < 1000)
SELECT * FROM v1, v2, v3 WHERE v1.i_brand = v2.i_brand AND v2.i_brand = v3.i_brand AND v3.i_color = 'red';

Index on i_color

Challenges:
- Deadlock hazard
- Inlining heuristics ➔ suboptimal plans
- CTEs optimized in isolation from context

Goals:
- Cost-based inlining approach
- Contextualized optimization
- Pushing filters, ordering into CTEs
- Deadlock-free execution

II. CTE Representation in Orca

Main Query Tree:

Possible alternative plans

CTEProducer(0)
CTEConsumer(0)

Join
TableScan(item)

3:
Join
TableScan(item)

4:
Join
TableScan(item)

5:
Join
TableScan(item)

6:
Join
TableScan(item)

7:
Join
TableScan(item)

8:
Join
TableScan(item)

Different combinations of operators can be used together to form different plans

Properties are used to ensure that only the valid plans are generated

III. Plan Enumeration

1. Main query tree inserted into compressed MEMO structure
   Each operator initializes a new group in the MEMO

2. Transformation rules are applied to generate equivalent alternatives, which are also added to the MEMO

3. Different combinations of operators can be used together to form different plans

IV. Predicate Push-down

Common expressions pulled out into CTE

- CTEs can be used to optimize the performance of queries with repeated expressions
- Similarly, they can be generated by the optimizer for expressions like window functions and distinct aggregates

V. Common Subexpression Elimination

Common expressions pulled out into CTE

VI. Execution in MPP Setting

- CTEConsumers are instantiated as SharedScans
- CTEProducers are instantiated as Materialize
- The Broadcast operator manages data exchange between the shown two active processes

VII. Experimental Results

Experimental setup:
- 5TB TPC-DS benchmark
- 8-node cluster, CPU: 3.33GHz, RAM: 48GB
- Compare Orca vs old Planner
- Compare Orca using different CTE inlining settings

Cost-based inlining up to 55% of execution time