Beyond Joins and Indexes

BRUCE MOMJIAN

As a follow up to the presentation, *Explaining the Postgres Query Optimizer*, this talk shows the non-join and non-index operations that the optimizer can choose.

https://momjian.us/presentations  
Creative Commons Attribution License

Last updated: April 2023
My previous talk, *Explaining the Postgres Query Optimizer*, covered:

- Query optimization basics
- Optimizer statistics
- Join methods
- Scan methods, including indexes
- Limit

https://momjian.us/main/presentations/performance.html#optimizer
This Presentation Covers Everything Else

1. Result
2. Values Scan
3. Function Scan
4. Incremental Sort
5. Unique
6. Append
7. Merge Append
8. Subquery Scan
9. HashSetOp
10. SetOp
11. Materialize
12. Memoize
13. Group
14. Aggregate
15. GroupAggregate
16. HashAggregate
17. MixedAggregate
18. WindowAgg
19. Parallel Seq Scan
20. Partial Aggregate
21. Gather
22. Finalize Aggregate
23. Gather Merge
24. Parallel Append
25. Parallel Hash
26. Parallel Hash Join
27. CTE Scan
28. WorkTable Scan
29. Recursive Union
30. ProjectSet
31. LockRows
32. Sample Scan
33. Table Function Scan
34. Foreign Scan
35. Tid Scan
36. Insert
37. Update
38. Delete
39. Merge
40. Semi Join
41. Anti Join
42. SubPlan
43. Others

https://www.pgmustard.com/docs/explain
Controls

My previous talk covered:
- enable_seqscan
- enable_bitmapscan
- enable_indexscan
- enable_indexonlyscan
- enable_nestloop
- enable_hashjoin
- enable_mergejoin
- enable_sort

This talk will cover:
- enable_incremental_sort
- enable_material

- enable_memoize
- enable_hashagg
- enable_gathermerge
- enable_parallel_append
- enable_parallel_hash
- enable_tidscan

Not covered:
- enable_async_append
- enable_partition_pruning
- enable_partitionwise_join
- enable_partitionwise_aggregate

https://www.postgresql.org/docs/current/runtime-config-query.html
* https://momjian.us/main/presentations/performance.html#partitioning
1. Result

```
-- This disables EXPLAIN cost output
\set EXPLAIN 'EXPLAIN (COSTS OFF)'

:EXPLAIN SELECT 1;
```

<table>
<thead>
<tr>
<th>QUERY PLAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>Result</td>
</tr>
</tbody>
</table>

All the queries used in this presentation are available at https://momjian.us/main/writings/pgsql/beyond.sql.
2. Values Scan

:EXPLAIN VALUES (1), (2);

QUERY PLAN

-----------------------------
Values Scan on "*VALUES*"

Causes are in blue, optimizer choices are in red.
3. Function Scan

```
EXPLAIN SELECT * FROM generate_series(1,4);
```

**QUERY PLAN**

```
| Function Scan on generate_series |
```
CREATE TABLE large (x) AS SELECT generate_series(1, 1000000);
ANALYZE large;
CREATE INDEX i_large ON large (x);

ALTER TABLE large ADD COLUMN y INTEGER;

EXPLAIN SELECT * FROM large ORDER BY x, y;

QUERY PLAN

-----------------------------------------
| Incremental Sort                      |
| Sort Key: x, y                         |
| Presorted Key: x                       |
| -> Index Scan using i_large on large   |
Incremental Sort

```
<table>
<thead>
<tr>
<th>3</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
```
5. Unique, First Example

```
:EXPLAIN SELECT DISTINCT * FROM generate_series(1, 10) ORDER BY 1;
QUERY PLAN

---------------------------------------------
Unique
  -> Sort
    Sort Key: generate_series
  -> Function Scan on generate_series
```

The query and its query plan are shown above.
-- not UNION ALL

:EXPLAIN SELECT 1 UNION SELECT 2;

QUERY PLAN

--------------------------
Unique
   -> Sort
      Sort Key: (1)
      -> Append
         -> Result
         -> Result
Unique

6
3
2
4
12
5
8
6
11
3
2
3
4
5
6
6
8
11
12

Sort

Unique

2
3
5
6
8
11
12

2
3
4
5
6
8
11
12
6. Append

```sql
:EXPLAIN SELECT 1 UNION ALL SELECT 2;
```

```
QUERY PLAN
--------------
 Append
  -> Result
  -> Result
```
Append
7. Merge Append

:\texttt{EXPLAIN (VALUES (1), (2) ORDER BY 1)}
\texttt{UNION ALL}
\texttt{(VALUES (3), (4) ORDER BY 1)}
\texttt{ORDER BY 1;}

\texttt{QUERY PLAN}
-----------------------------------------
\texttt{Merge Append}
\quad \texttt{Sort Key: "VALUES".column1}
\quad \rightarrow \texttt{Sort}
\quad \quad \texttt{Sort Key: "VALUES".column1}
\quad \quad \rightarrow \texttt{Values Scan on "VALUES"}
\quad \rightarrow \texttt{Sort}
\quad \quad \texttt{Sort Key: "VALUES\_1".column1}
\quad \quad \rightarrow \texttt{Values Scan on "VALUES\_1"}
Merge Append

Sort

Merge Append
CREATE TABLE small (x) AS
SELECT generate_series(1, 1000);
ANALYZE small;

:EXPLAIN SELECT * FROM small EXCEPT SELECT * FROM small;

QUERY PLAN
---------------------------------------------
HashSetOp Except
  -> Append
    -> Subquery Scan on "*SELECT* 1"
      -> Seq Scan on small
    -> Subquery Scan on "*SELECT* 2"
      -> Seq Scan on small small_1
HashSetOp

Logically UNION, EXCEPT, and INTERSECT without ALL remove duplicates in joined queries and results. For example, the query VALUES (1), (1), (2), (2) EXCEPT VALUES (1) returning 2 shows joined query removal, while UNION shows result removal.

The count is useful for efficiently processing duplicates when ALL is used.
10. SetOp

-- table has to be too large to hash
:EXPLAIN SELECT * FROM large INTERSECT SELECT * FROM large;

QUERY PLAN

---------------------------------------------------------------
SetOp          Intersect
  ->  Sort
       Sort Key: "*SELECT* 1".x, "*SELECT* 1".y
  ->  Append
       ->  Subquery Scan on "*SELECT* 1"
       ->  Seq Scan on large
       ->  Subquery Scan on "*SELECT* 2"
       ->  Seq Scan on large large_1
SetOp

Append with "1" Label

Append with "2" Label

Sort

SetOp with INTERSECT

Count
11. Materialize

:EXPLAIN SELECT * FROM small s1, small s2 WHERE s1.x != s2.x;

QUERY PLAN

-----------------------------
Nested Loop
  Join Filter: (s1.x <> s2.x)
  -> Seq Scan on small s1
  -> Materialize
    -> Seq Scan on small s2
Materialize

---

Outer

- aag
- aay
- aar
- aai

Local Memory

- aai
- aag
- aas
- aar
- aay
- aaa
- aag

Inner

Shared Buffers

- aai
- aag
- aas
- aar
- aay
- aaa
- aag

---

Materialize

---
-- needs duplicates and too small for a hash join
CREATE TABLE small_with_dups (x) AS
SELECT generate_series(1, 1000)
FROM generate_series(1, 10);

-- unique and too big for a hash join
CREATE TABLE medium (x) AS
SELECT generate_series(1, 100000);

-- index required for this memoize example
CREATE INDEX i_medium ON medium (x);
ANALYZE;
Memoize

:EXPLAIN SELECT * FROM small_with_dups JOIN medium USING (x);

QUERY PLAN

-------------------------------------
Nested Loop
  ->  Seq Scan on small_with_dups
  ->  Memoize
      Cache Key: small_with_dups.x
      Cache Mode: logical
      ->  Index Only Scan using i_medium on medium
          Index Cond: (x = small_with_dups.x)

Only happens in nested loops; supported in Postgres 14 and later.

https://blog.jooq.org/postgresql-14s-enable-memoize-for-improved-performance-of-nested-loop-joins/
Inner-side lookups that return no rows are also recorded in the cache.
13. Group

-- must be small enough not to trigger HashAggregate
-- removing WHERE and adding ORDER BY x does the same

:EXPLAIN SELECT x FROM large WHERE x < 0 GROUP BY x;

QUERY PLAN

GROUP
  Group Key: x
  -> Index Only Scan using i_large on large
     Index Cond: (x < 0)
GROUP BY without aggregates is similar to SELECT DISTINCT, except duplicate detection can consider more columns than those selected for output.
Group, Single Columns

Sorted

Group, Single Column

1
2
3
14. Aggregate

:EXPLAIN SELECT COUNT(*) FROM medium;

QUERY PLAN

--------------------------
Aggregate
  -> Seq Scan on medium
15. GroupAggregate

```
:EXPLAIN SELECT x, COUNT(*) FROM medium GROUP BY x ORDER BY x;
QUERY PLAN

-----------------------------------------------
GroupAggregate
  Group Key: x
  -> Index Only Scan using i_medium on medium
```
16. HashAggregate

:EXPLAIN SELECT DISTINCT x FROM medium;

QUERY PLAN

--------------------------
HashAggregate
  Group Key: x
  -> Seq Scan on medium
HashAggregate

HashAggregate

8.4
3.2
6.2
3.1
11,5
6.7
8.9
6.2
11,4
3.1

8
3
6
11
17. MixedAggregate

:EXPLAIN SELECT x FROM medium GROUP BY ROLLUP(x);

QUERY PLAN

--------------------------
MixedAggregate
  Hash Key: x
  Group Key: ()
  -> Seq Scan on medium
MixedAggregate
18. WindowAgg

```sql
:EXPLAIN SELECT x, SUM(x) OVER ()
FROM generate_series(1, 10) AS f(x);
```

QUERY PLAN

```
WindowAgg
  -> Function Scan on generate_series f
```
WindowAgg

<table>
<thead>
<tr>
<th>Sorted</th>
<th>WindowAgg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,1</td>
<td>1,1,Agg</td>
</tr>
<tr>
<td>1,2</td>
<td>1,2,Agg</td>
</tr>
<tr>
<td>1,2</td>
<td>1,2,Agg</td>
</tr>
<tr>
<td>1,4</td>
<td>1,4,Agg</td>
</tr>
<tr>
<td>2,1</td>
<td>2,1,Agg</td>
</tr>
<tr>
<td>2,5</td>
<td>2,5,Agg</td>
</tr>
<tr>
<td>3,2</td>
<td>3,2,Agg</td>
</tr>
<tr>
<td>3,2</td>
<td>3,2,Agg</td>
</tr>
<tr>
<td>3,2</td>
<td>3,2,Agg</td>
</tr>
<tr>
<td>3,2</td>
<td>3,2,Agg</td>
</tr>
<tr>
<td>3,4</td>
<td>3,4,Agg</td>
</tr>
</tbody>
</table>

Window functions allow aggregates across rows while the individual rows remain.
:EXPLAIN SELECT * FROM large;

QUERY PLAN

----------------------------------------------
Finalize Aggregate
  -> Gather
    Workers Planned: 2
  -> Partial Aggregate
    -> Parallel Seq Scan on large

https://www.postgresql.org/docs/current/parallel-plans.html
Parallel Seq Scan uses background workers to scan different parts of a table in parallel.

Gather collects results from background workers.

Finalize Aggregate
CREATE TABLE huge (x) AS SELECT generate_series(1, 100000000);
ANALYZE huge;

:EXPLAIN SELECT * FROM huge ORDER BY 1;

QUERY PLAN

---------------------------------------
Gather Merge
  Workers Planned: 2
  -> Sort
    Sort Key: x
    -> Parallel Seq Scan on huge
Gather Merge collects ordered results from background workers, retaining their ordering.
24. Parallel Append

```sql
EXPLAIN SELECT * FROM huge UNION ALL SELECT * FROM huge ORDER BY 1;
QUERY PLAN
```

Gather Merge
  Workers Planned: 2
  -> Sort
    Sort Key: huge.x
    -> Parallel Append
      -> Parallel Seq Scan on huge
      -> Parallel Seq Scan on huge huge_1
Parallel Append

Two Tables
- Parallel Seq Scan
  - Background Worker
    - 6
    - 3
    - 2
  - Background Worker
    - 4
    - 12
- Parallel Seq Scan
  - Background Worker
    - 5
    - 8
    - 6
  - Background Worker
    - 11
    - 3

One Result
- Parallel Append
  - Background Worker
    - 6
    - 3
    - 2
    - 4
    - 12
- Parallel Append
  - Background Worker
    - 5
    - 8
    - 6
    - 11
    - 3
- Sort
- Gather Merge
- Background Worker
  - 2
  - 3
  - 4
  - 6
  - 12
- Background Worker
  - 2
  - 3
  - 4
  - 5
  - 6
  - 8
  - 11
  - 12
::EXPLAIN SELECT * FROM huge h1 JOIN huge h2 USING (x);
QUERY PLAN

Gather
  Workers Planned: 2
  -> Parallel Hash Join
     Hash Cond: (h1.x = h2.x)
     -> Parallel Seq Scan on huge h1
     -> Parallel Hash
        -> Parallel Seq Scan on huge h2
Parallel Hash, Parallel Hash Join

- Parallel Hash
- Parallel Hash Join
- Shared Hash
- Inner
- Outer
- Background Worker
- Parallel Seq Scan
- Parallel Hash
27. CTE Scan

`:EXPLAIN WITH source AS MATERIALIZED (  
   SELECT 1  
)  
SELECT * FROM source;

QUERY PLAN

-------------------
  CTE Scan on source
  CTE source
  -> Result

https://momjian.us/main/presentations/sql.html#cte`
CTE Scan

Materialized Common Table Expressions

CTE Source

6
3
2
4
12
5
8
6
11
3

CTE Scan
WorkTable Scan, Recursive Union

```sql
:EXPLAIN WITH RECURSIVE source (counter) AS ( 
    SELECT 1 
    UNION ALL 
    SELECT counter + 1 
    FROM source 
    WHERE counter < 10
  )
SELECT * FROM source;
```

**QUERY PLAN**

```
CTE Scan on source
CTE source
  -> Recursive Union
     -> Result
        -> WorkTable Scan on source source_1
            Filter: (counter < 10)
```
WITH RECURSIVE source AS ( 
    SELECT 1
    UNION ALL
    SELECT 1 FROM source
)
SELECT * FROM source;
WorkTable is cleared before every iteration. Recursion stops when the recursive CTE returns no rows.
:EXPLAIN SELECT generate_series(1,4);

QUERY PLAN

-------------

ProjectSet
  -> Result
31. LockRows

:EXPLAIN SELECT * FROM small FOR UPDATE;

QUERY PLAN

-------------------------

LockRows

-> Seq Scan on small
32. Sample Scan

:EXPLAIN SELECT * FROM small TABLESAMPLE SYSTEM(50);

QUERY PLAN

---------------------------------
Sample Scan on small
Sampling: system ('50'::real)
33. Table Function Scan

```sql
:EXPLAIN SELECT *
FROM XMLTABLE('/ROWS/ROW'
PASSING $$
  <ROWS>
    <ROW id="1">
      <COUNTRY_ID>US</COUNTRY_ID>
    </ROW>
  </ROWS>
$$
COLUMNS id int PATH '@id',
_id FOR ORDINALITY);
```

---

Table Function Scan on "xmltable"
CREATE EXTENSION postgres_fdw;

CREATE SERVER postgres_fdw_test
FOREIGN DATA WRAPPER postgres_fdw
OPTIONS (host 'localhost', dbname 'fdw_test');

CREATE USER MAPPING FOR PUBLIC
SERVER postgres_fdw_test
OPTIONS (password '');

CREATE FOREIGN TABLE other_world (greeting TEXT)
SERVER postgres_fdw_test
OPTIONS (table_name 'world');

:EXPLAIN SELECT * FROM other_world;

QUERY PLAN
-----------------------------
Foreign Scan on other_world
35. Tid Scan

```sql
:EXPLAIN SELECT * FROM small WHERE ctid = '(0,1)';
    QUERY PLAN

-----------------------------------
  Tid Scan on small
    TID Cond: (ctid = '(0,1)::tid)
```

---
Tid Scan

/data /base /16385 /24692

Postgres
Postgres
Postgres

Tuple
Tuple Tuple Special
8K
8k 8k 8k 8k

Page Header Item Item Item

Tupple
Tupple Tupple Special
36. Insert

:EXPLAIN INSERT INTO small VALUES (0);

QUERY PLAN

-----------------
Insert on small
  ->  Result
37. Update

:EXPLAIN UPDATE small SET x = 1 WHERE x = 0;

QUERY PLAN

-------------------------
Update on small
  -> Seq Scan on small
    Filter: (x = 0)
38. Delete

```sql
:EXPLAIN DELETE FROM small;

QUERY PLAN

-------------------------
Delete on small
   -> Seq Scan on small

-- You cannot run EXPLAIN on utility commands like TRUNCATE.

:EXPLAIN TRUNCATE small;
ERROR: syntax error at or near "TRUNCATE"
LINE 1: EXPLAIN (COSTS OFF) TRUNCATE small;`
```
CREATE TABLE mergetest (x, y) AS VALUES (1, NULL), (3, NULL), (5, NULL);

:EXPLAIN MERGE INTO mergetest
USING (VALUES (1), (2), (3), (4), (5), (6)) m (x)
ON mergetest.x = m.x
WHEN NOT MATCHED THEN
    INSERT (x) VALUES (m.x)
WHEN MATCHED THEN
    UPDATE SET y = TRUE;

QUERY PLAN
-------------------------------------------------------
Merge on mergetest
  -> Hash Right Join
      Hash Cond: (mergetest.x = "VALUES".column1)
  -> Seq Scan on mergetest
  -> Hash
      -> Values Scan on "VALUES"
40. Semi Join, First Example

```
:EXPLAIN SELECT *
FROM small
WHERE EXISTS (SELECT * FROM medium WHERE medium.x = small.x);
```

QUERY PLAN

```
-----------------------------------
Hash Semi Join
  Hash Cond: (small.x = medium.x)
  ->  Seq Scan on small
  ->  Hash
    ->  Seq Scan on medium

Stop scan after first inner match.
```
:EXPLAIN SELECT *
FROM small
WHERE small.x IN (SELECT medium.x FROM medium);

QUERY PLAN

-----------------------------------
Hash Semi Join
    Hash Cond: (small.x = medium.x)
    -> Seq Scan on small
    -> Hash
        -> Seq Scan on medium

EXISTS and IN are equivalent in handling of NULLs because EXISTS only checks for row existence while IN logically does OR comparisons that can ignore non-true results from NULL comparisons.
41. Anti Join

```sql
:EXPLAIN SELECT *
FROM medium
WHERE NOT EXISTS (SELECT * FROM small WHERE small.x = medium.x);
```

**QUERY PLAN**

```
Hash Anti Join
    Hash Cond: (medium.x = small.x)
    -> Seq Scan on medium
    -> Hash
        -> Seq Scan on small
```

Stop scan after first inner match; negate result.
42. SubPlan

```sql
:EXPLAIN SELECT *
FROM small
WHERE small.x NOT IN (SELECT medium.x FROM medium);
```

**QUERY PLAN**
```
------------------------------------
Seq Scan on small
  Filter: (NOT (hashed SubPlan 1))
SubPlan 1
  -> Seq Scan on medium
```

**NOT IN** and **NOT EXISTS** are not equivalent for NULLs because **NOT IN** logically does repeated not-equal AND comparisons which must all be true to return true; NULL affects this.
43. Others: Outer Join Removal

-- UNIQUE index guarantees at most one right row match
CREATE UNIQUE INDEX i_small ON small (x);

-- LEFT JOIN guarantees every left row is returned
:EXPLAIN SELECT medium.x FROM medium LEFT JOIN small USING (x);

QUERY PLAN
-------------------
Seq Scan on medium
• Named Tuplestore Scan: after triggers
• Custom Scan: custom scan providers
Conclusion

https://momjian.us/presentations

https://www.flickr.com/photos/glassholic/