Part 2

Tuesday, January 18, 2022 5:26 PM

Tyrone Lagore V00995698

1. Using EXPLAIN ANALYZE, record the number of tuples that postgres estimates each query will return and the actual number of tuples returned.

EXPLAIN ANALYZE select * from productions where year > 1994 and year <= 1996 and productiontype = 'short';

```
Gather  (cost=1000.00..150028.02 rows=13366 width=72) (actual
time=0.433..344.389 rows=5327 loops=1)
   Workers Planned: 2
   Workers Launched: 2
   ->  Parallel Seq Scan on productions  (cost=0.00..147691.42
rows=5569 width=72) (actual time=0.580..333.838 rows=1776 loops=3)
         Filter: ((year > 1994) AND (year <= 1996) AND (productiontype
= 'short'::text))
         Rows Removed by Filter: 2563151
Planning time: 0.152 ms
Execution time: 344.642 ms
(8 rows)
```

2. Using only the information in pg stats and pg class compute the selectivity of the WHERE clause of this query

Obtain histogram fraction from null frac and most common frequency sum

```
SELECT 1-null_frac-mcfs_sum AS hist_sum, null_frac, mcfs_sum
       FROM (SELECT
    (SELECT CAST(null_frac AS float) FROM pg_stats AS pgs WHERE
pgs.tablename = 'productions' AND attname = 'year') AS null_frac,
    SUM(mcfs) AS mcfs_sum FROM
        (SELECT UNNEST(most_common_freqs) AS mcfs
        FROM pg_stats WHERE tablename = 'productions' AND attname =
'year') as t) as p;
```

```
     hist_sum      |     null_frac     | mcfs_sum
-------------------+-------------------+----------
0.192266747355461 | 0.111733332276344 |   0.695999997667968
```

Null frac and histogram were not populated for 'productiontype', therefore just using MCF values.

Obtain most common frequencies for productiontype and year

```
SELECT most_common_vals, most_common_freqs FROM pg_stats WHERE
tablename = 'productions' AND attname = 'productiontype';
```

```
--
```

```
{tvEpisode,short,movie,video,tvSeries,tvMovie,tvMiniSeries,tvSpecial,v
ideoGame,tvShort}
--
{0.736133,0.105767,0.0713667,0.0292667,0.0271333,0.0165333,0.00476667,
0.0045,0.0034,0.00113333}
```

```
 SELECT most_common_vals, most_common_freqs FROM pg_stats WHERE
tablename = 'productions' AND attname = 'year';
```

```
most_common_vals

----------------------------------------------------------------------
--------------------------------------------------------------
{2017,2018,2016,2019,2015,2020,2014,2013,2012,2011,2010,2009,2008,2007
,2006,2005,2021,2004,2003,2002,2001,2000,1999,1998,1997,1995}


most_common_freqs

----------------------------------------------------------------------
--------------------------------------------------------------
{0.0501,0.0489333,0.0479333,0.0473333,0.0453667,0.0423,0.0422333,0.039
5667,0.0369667,0.032,0.0276,0.0268,0.0237333,0.0235667,0.0206,0.0189,0
.0179333,0.0154333,0.0137333,0.0129333,0.0127667,0.0117,0.01,0.0095,0.
00933333,0.00873333}
```

Obtain histogram_bounds for year

```
SELECT histogram_bounds FROM pg_stats WHERE tablename = 'productions'
and attname = 'year';
```

```
histogram_bounds



-------------------------------
{1896,1905,1910,1912,1913,1915,1916,1920,1924,1929,
1934,1939,1944,1948,1951,1952,1954,1955,1956,1957,
1959,1960,1961,1962,1963,1964,1964,1965,1966,1966,
1967,1967,1968,1968,1969,1969,1970,1971,1971,1972,
1972,1973,1973,1974,1974,1975,1976,1976,1977,1977,
1978,1978,1979,1979,1980,1980,1981,1981,1981,1982,
1982,1983,1983,1984,1984,1984,1985,1985,1986,1986,
1986,1987,1987,1987,1988,1988,1989,1989,1989,1990,
1990,1990,1991,1991,1991,1992,1992,1992,1993,1993,
1993,1993,1994,1994,1994,1994,1996,1996,1996,1996,2023}
```

Calculating Specificity

> 1994 AND <= 1996 is values for 1995 and 1996.

- This is 4 buckets in histogram, with histogram representing 0.192266747355461 of the tuples
- MCF for 1995 is 0.008733333, with no MCF for 1996

productiontype='short'

- MCF is 0.105767 with no histogram_bounds, and null_frac = 0 (null)

Therefore the formula is:

((4/100) * 0.192266747355461 + 0.008733333)*0.105767 = **0.001737118**

The specificity of this query is **0.001737118**

3. using this selectivity, compute the expected number of matching tuples of this query.

Obtain number of tuples

SELECT reltuples FROM pg_class WHERE relname = 'productions';

```
reltuples
------------
7.6944e+06
(1 row)
```

7694400 tuples in the relation

0.001737118*7694400 = **13366.07700059**

Rounded: **13366**

```
Gather  (cost=1000.00..150028.02 rows=13366 width=72) (actual
time=0.433..344.389 rows=5327 loops=1)
```

Calculation matches that calculated by PSQL ANALYZE