Índice

Estadísticas sobre la base de datos FBV

El primer paso pasa poder optimizar las consultas dadas es conocer el volumen total de la base de datos, el volumen de cada tabla individual, el tamaño promedio de cada fila, y las estadísticas apropiadas para cada consulta.

Estadísticas generales de la base de datos

En esta subsección se describen las estadísticas relacionadas con la base de datos en su totalidad, independientemente de las consultas que se vayan a optimizar.

Volumen total de los datos

Para determinar el volumen total de tuplas existentes en la base de datos se realizó la siguiente consulta:

```
select sum(n_live_tup)
from pg_stat_user_tables
where schemaname='original';
```

La consulta arrojó un valor de 8660591 tuplas.

Volumen total de cada tabla

Para calcular el volumen total de cada tabla utilizamos la siguiente consulta:

```
select
  c.relname,
  n_live_tup,
  relpages,
  floor(n_live_tup/relpages::float) as tuples_per_page
from
  pg_stat_user_tables b,
  pg_class c
where relnamespace = (select oid from pg_namespace where nspname='original')
  and schemaname = 'original'
  and b.relname= c.relname;
```

De esta consulta se obtuvieron los resultados de la tabla 1.

Tabla 1: Tuplas, páginas y tuplas por página de las tablas de FBV

Tabla	Tuplas	Páginas	Tuplas por pagina
`part`	200000	3715	53
`supplier`	10000	213	46
`partusupp`	800000	17451	45
`lineitem`	6001181	98544	60
`region`	5	1	5
`nation`	25	1	25
`customer`	150000	3566	42
`orders`	1500000	25196	59

Figura 1: Cantidad de tuplas en cada tabla

En la figura 1 podemos observar que la tabla de `lineitem` ocupa la mayor parte de los datos almacenados (aproximadamente un 69 % del total de tuplas). Cualquier consulta que requiera acceder a una buena parte de los datos almacenados en esta tabla exigirá mucha atención al momento de ser optimizada.

Para las tablas `nation` y `region`, dado que caben en una sola página, se puede concluir que no necesitan optimización alguna pues un acceso directo es siempre la mejor opción para recuperar sus registros.

Análisis de cada tabla

En el análisis de cada tabla tomarán en cuenta los siguientes aspectos:

- Tamaño promedio de la tupla
- Para cada atributo, su tamaño promedio, cantidad de elementos distintos y factor de reducción.
- Cualquier otra estadística que resulte últil para la optimización de alguna de las consultas propuestas.

Para el cálculo del tamaño promedio de cada tupla, en todas las tablas, se realizó la siguiente consulta:

```
select
  tablename,
  sum(avg_width) as tam_promedio_tuplas
from
  pg_stats
where
  tablename in (
    select
      relname
    from pg_stat_user_tables)
group by
  tablename;
```

De la cual se obtuvieron los siguientes resultados

Tabla 2: Tamaño promedio de cada tupla

Tabla	Tamaño promedio de tupla
part	114
supplier	137
partusupp	144
lineitem	98
region	78
nation	91
customer	158
orders	100

En el análisis de cada tabla, se realizó una consulta diseñada para extraer la siguiente información acerca de las tablas:

- El tamaño medio de cada atributo
- El número de valores distintos que tiene la columna. Si el valor es -1, es un valor único
- La correlación que existe entre el orden lógico y el orden físico (mientras más cercano a 1 o -1 mejor a la hora de que el manejador haga index scan)

- La frecuencia más alta hallada. Esta nos permite tener una cota superior para un determinado valor
- El factor reductor. Para conocer la selectividad de la columna.

La consulta es la siguiente:

```
prepare stats_table as
select
  attname,
  avg_width,
  ( case
      when n_distinct < 0  and n_distinct <> -1
        then -n_distinct * t.reltuples
        else n_distinct
      end) as ndistinct,
  correlation,
  most_common_freqs[1] as upper_bound,
  ( case
      when n_distinct < 0
        then -1 / (n_distinct * t.reltuples)
        else 1 / n_distinct
      end) as FR
from
  pg_stats,
  ( select
      relname,
      reltuples
    from
      pg_class
    where relnamespace = (
      select
        oid
      from
        pg_namespace
      where
        nspname='original')
    ) t
where t.relname = tablename
  and tablename = $1
  and schemaname = 'original'
order by
  fr;
```

Análisis de `lineitem`

Al ejecutar la consulta sobre lineitem se obtuvieron los resultados mostrados en la Tabla 3. De la tabla se puede decir que las columnas con peor selectividad son `l_linestatus`, `l_returnflag` y `l_shipsintruct`. Por otro lado, los atributos con mayor selectividad tenemos `l_comment`, `l_orderkey` y `l_extendedprice`.

Tabla 3: Resultados obtenidos para lineitem

atributo	Tam. prom.	# vals. dist.	Correlación	Cota sup.	Selectividad
`l_orderkey`	4	1206300	1	0.000016667	0.000000829
`l_partkey`	4	197029	0.00235099	0.00003	5.0753949E-06
`l_suppkey`	4	10000	-0.000106049	0.000173333	0.0001
`l_linenumber`	4	7	0.176068	0.250317	0.1428571429
`l_quantity`	5	50	0.0195651	0.0205067	0.02
`l_extendedprice`	8	767024	0.000341259	2.33333E-05	1.303740162E-06
`l_discount`	4	11	0.0868008	0.0922367	0.0909090909
`l_tax`	4	9	0.109181	0.11199	0.1111111111
`l_returnflag`	2	3	0.377041	0.506517	0.3333333333
`l_linestatus`	2	2	0.499747	0.500087	0.5
`l_shipdate`	4	2525	-0.00126623	0.000536667	0.0003960396
`l_comitdate`	4	2465	-0.00119272	0.000536667	0.0004056795
`l_receipdate`	4	2543	-0.00128363	0.000553333	0.0003932363
`l_shipinstruct`	13	4	0.250591	0.250767	0.25
`l_shipmode`	5	7	0.145059	0.143523	0.1428571429
`l_comment`	27	1763690	0.000151724	0.000193333	0.000000567