SQL Next Steps: Optimization

Getting the most out of your database

By **Haki Benita**

About Me

Haki Benita

- DBA (Oracle and PostgreSQL)
- Full Stack Developer (Python / Django / Javascript)
- Team leader
- Data plumber
- Website: https://hakibenita.com/
- Twitter: @haki_be

What You'll Learn

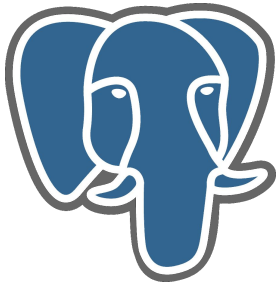
- How to avoid common mistakes in SQL
- How to improve performance of SQL
- How to be more productive writing SQL

SQL History

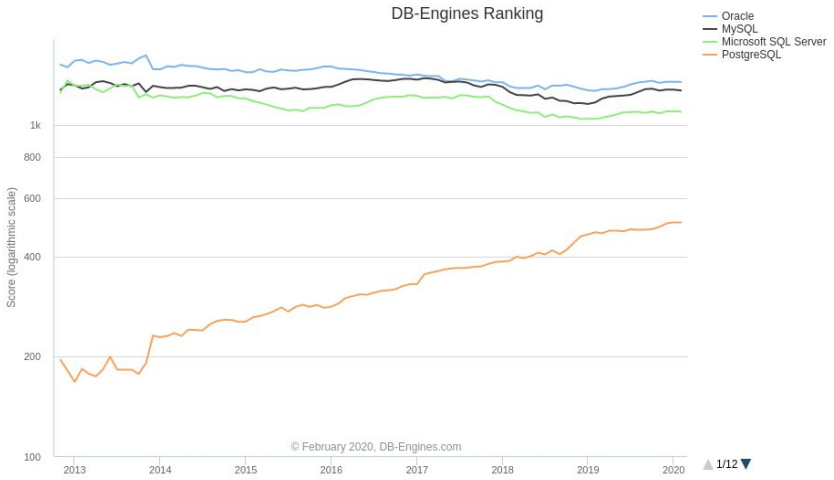
- SQL = **S**tructured **Q**uery **L**anguage
- Pronounced **S - Q - L** (not *SEQUEL*!)
- Used to interact with **relational databases** (RDBMS)
- Invented in the early 70s at IBM based on work by Edgar F. Codd
- Became a standard in 1986 (ANSI-86)
- Standard revised in 1992 (ANSI-92)

📖 Important Papers: Codd and the Relational Model
📖 Comparing Database Types: How Database Types Evolved to Meet Different Needs

PostgreSQL



- Based on the Berkeley POSTGRES project from 1986
- Released under the name Postgres95
- Open Source
- Free
- Growing fast!



Source: https://db-engines.com/en/ranking_trend

Anatomy of an SQL Query

```
SELECT
FROM
WHERE
GROUP BY
HAVING
ORDER BY
```

[Q] In which order are the parts of the query executed?

Anatomy of an SQL Query

In order of execution

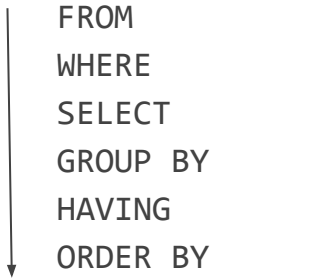
```
FROM
WHERE
SELECT
GROUP BY
HAVING
ORDER BY
```

Question

What's wrong with this query?

```
SELECT
   department,
   count(*) as number_of_employees
FROM
   employees
WHERE
   number_of_employees > 10
GROUP BY
   department
```

Order of execution:



Question

What's wrong with this query?

```
SELECT
   department,
   count(*) as number_of_employees
FROM
   employees
WHERE
   number_of_employees > 10
GROUP BY
   department
```

```
SELECT
   department,
   count(*) as number_of_employees
FROM
   employees
GROUP BY
   department
HAVING
   number_of_employees > 10
```

Aggregate results can only be used in the HAVING clause

When conditions in the WHERE clause are being evaluated, the result of the aggregation function is not yet available.

Common Mistakes In SQL

Writing correct SQL

Sales Database
Setup

- PostgreSQL 12
- Use DB Fiddle
- Use local database
- Follow along with presentation

Sales Database

```
select * from sale
```

id	branch	sold_at	customer	product	price	discount
1	NY	2020-03-31 20:15:00-07	Bill	Shoes	10000	1000
2	NY	2020-03-31 21:00:00-07		Shoes	5000	0
3	LA	2020-03-31 23:15:00-07	Lily	Shoes	15000	0
4	LA	2020-04-01 02:10:00-07	John	Shoes	5000	2500
5	NY	2020-03-31 20:15:00-07		Shirt	1500	0
6	NY	2020-03-31 19:07:00-07	John	Shirt	1850	0
7	LA	2020-03-31 02:55:00-07	Bill	Shirt	1250	0
8	LA	2020-03-31 03:45:00-07	Lily	Shirt	1850	100
9	NY	2020-03-31 00:45:00-07	Lily	Pants	5200	0
10	LA	2020-03-31 03:45:00-07	John	Pants	5200	0
11	LA	2020-04-01 00:01:00-07	David	Pants	4500	0
12	LA	2020-04-01 23:01:00-07		Hat	8000	8000
13	LA	2020-04-01 23:01:00-07	Bill	Give Away	0	0
14	NY	2020-03-31 10:01:00-07		Give Away	0	0
15	LA	2020-04-01 03:45:00-07		Give Away	0	0

Exercise

What is the discount rate on **Shoes?**

```
select * from sale where product = 'Shoes';

 id | branch |         sold_at         | customer | product | price | discount
----+--------+-------------------------+----------+---------+-------+----------
  1 | NY     | 2020-04-01 03:15:00+00 | Bill     | Shoes   | 10000 | 1000
  2 | NY     | 2020-04-01 04:00:00+00 |          | Shoes   |  5000 | 0
  3 | LA     | 2020-04-01 06:15:00+00 | Lily     | Shoes   | 15000 | 0
  4 | LA     | 2020-04-01 09:10:00+00 | John     | Shoes   |  5000 | 2500
```

Exercise

What is the discount rate on **Shoes?**

```
SELECT price, discount, discount / price * 100 as discount_rate
FROM sale
WHERE product = 'Shoes';

 price | discount | discount_rate
 ------+----------+---------------
10000 |     1000 |             0
 5000 |        0 |             0
15000 |        0 |             0
 5000 |     2500 |             0
```

How is this possible?

Be Careful When Dividing Integers

Integer division truncates the result

```
SELECT 1000 / 10000;
?column?
----------
         0

SELECT 1000 / 10000::float;
?column?
----------
       0.1

SELECT 1000 / 10000::float * 100;
?column?
----------
        10
```

Casting the denominator to float produces the expected result

📖 Mathematical Functions and Operators

Exercise

What is the discount rate on **Shoes?**

Put it together:

```
SELECT price, discount, discount / price::float as discount_rate
FROM sale
WHERE product = 'Shoes';

price | discount | discount_rate
-------+----------+---------------
10000 |     1000 |           0.1
 5000 |        0 |             0
15000 |        0 |             0
 5000 |     2500 |           0.5
```

Exercise

What is the discount rate on **Shoes?**

Multiple by 100:

```
SELECT price, discount, discount / price::float * 100 as discount_rate
FROM sale
WHERE product = 'Shoes';

price | discount | discount_rate
-------+----------+---------------
10000 |     1000 |            10
 5000 |        0 |             0
15000 |        0 |             0
 5000 |     2500 |            50
```

Exercise

Find the **average** discount rate **by product**

Exercise

Find the average discount rate **by product**

```
SELECT
   product,
   AVG(discount / price::float) * 100 as discount_rate
FROM sale
GROUP BY product;

ERROR:  division by zero
```

Guard Against "division by zero" Errors

```
SELECT id, product, price, discount
FROM sale
WHERE price = 0;

id |  product   | price | discount
---+------------+-------+----------
13 | Give Away |     0 |        0
14 | Give Away |     0 |        0
15 | Give Away |     0 |        0
```

Product “Give Away” price is zero and it causes the division to fail

Guard Against "division by zero" Errors

NULLIF(*value1, value2*)
The NULLIF function returns a null value if *value1* equals *value2*; otherwis it returns *value1*.

📖 NULLIF Documentation

```
SELECT
   id, product, price, discount,
   NULLIF(price, 0) AS adjusted_price
FROM sale
WHERE price = 0;

id |  product   | price | discount | adjusted_price
---+-----------+-------+----------+----------------
13 | Give Away |     0 |        0 |
14 | Give Away |     0 |        0 |
15 | Give Away |     0 |        0 |
```

The adjusted price is null when the price equals zero

Dividing by null produces null, not an error

No error!

```
SELECT
   id,
   product,
   discount / NULLIF(price, 0) AS discount_rate
FROM sale
WHERE price = 0;

id |  product   | discount_rate
---+-----------+---------------
13 | Give Away |
14 | Give Away |
15 | Give Away |
```

Result or division is now null.

Exercise

Find the average discount rate **by product**

```
SELECT
   product,
   AVG(discount / NULLIF(price, 0)::float) * 100 AS discount_rate
FROM sale
GROUP BY product;
```

```
 product   |    discount_rate
-----------+--------------------
Shirt      | 1.3513513513513513
Pants      |                  0
Hat        |                100
Shoes      |                 15
Give Away  |
```

What is the real discount here?

Exercise

Find the average discount rate **by product**

`COALESCE(`*`value1, ..., valueN`*`)`
Returns the first of its arguments that is not null.

📖 COALESCE Documentation

Make discount rate zero for products with price zero.

```
SELECT
   product,
   COALESCE(AVG(discount / NULLIF(price, 0)::float), 0) * 100
FROM sale
GROUP BY product;
```

```
 product   |    discount_rate
-----------+--------------------
Shirt      | 1.3513513513513513
Pants      |                  0
Hat        |                100
Shoes      |                 15
Give Away  |                  0
```

Exercise

How many **unique users** purchased **each product**?

Exercise

How many unique users purchased each product?

```
SELECT
   product,
   COUNT(DISTINCT customer) AS customers
FROM sale
GROUP BY product;
```

```
  product   | customers
------------+-----------
Give Away |         1
Hat       |         0
Pants     |         3
Shirt     |         3
Shoes     |         3
```

Is this correct?

Be Careful When Aggregating Nullable Column

Aggregate functions ignore null values!

```
select * from sale where product = 'Hat';
id |         sold_at          | customer | product | price | discount
---+--------------------------+----------+---------+-------+----------
12 | 2020-04-01 18:01:00+03 |          | Hat     |  8000 |     8000
```

Why was this customer not counted?

Be Careful When Aggregating Nullable Column

Aggregate functions ignore null values!

```
SELECT
   product,
   COUNT(*) AS cnt,
   COUNT(customer) AS cnt_customer
FROM sale
GROUP BY product;
```

```
 product   | cnt | cnt_customer
-----------+-----+--------------
Shirt      |   4 |            3
Pants      |   3 |            3
Hat        |   1 |            0
Shoes      |   4 |            3
Give Away  |   3 |            1
```

This can also be useful!

Exercise

How many **known customers** purchased each product?

Hint: Use the fact that aggregate functions ignore null!

Exercise

How many **known customers** purchased each product?

```
SELECT
   product,
   COUNT(customer) as known_customers,
   COUNT(*) - COUNT(customer) as unknown_customers
FROM sale
GROUP BY product;
 product   | known_customers | unknown_customers
-----------+-----------------+-------------------
Shirt      |               3 |                 1
Pants      |               3 |                 0
Hat        |               0 |                 1
Shoes      |               3 |                 1
Give Away  |               1 |                 2
```

Exercise

Write a query to find the sales made by the customer **Bill**

Exercise

Write a query to find the sales made by the customer **Bill**

```
SELECT * FROM sale WHERE customer = 'Bill';

id | branch |        sold_at         | customer |  product   | price | discount
---+--------+------------------------+----------+------------+-------+----------
 1 | NY     | 2020-03-31 20:15:00-07 | Bill     | Shoes      | 10000 |     1000
 7 | LA     | 2020-03-31 02:55:00-07 | Bill     | Shirt      |  1250 |        0
13 | LA     | 2020-04-01 23:01:00-07 | Bill     | Give Away  |     0 |        0
```

Exercise

Write a query to find the sales made by an **unknown customer**

Exercise

Write a query to find the sales made by an **unknown customer**

```
SELECT * FROM sale WHERE customer IS NULL;

 id | branch |        sold_at         | customer |  product   | price | discount
----+--------+-------------------------+----------+------------+-------+----------
  2 | NY     | 2020-03-31 21:00:00-07 |          | Shoes      |  5000 |        0
  5 | NY     | 2020-03-31 20:15:00-07 |          | Shirt      |  1500 |        0
 12 | LA     | 2020-04-01 23:01:00-07 |          | Hat        |  8000 |     8000
 14 | NY     | 2020-03-31 10:01:00-07 |          | Give Away  |     0 |        0
 15 | LA     | 2020-04-01 03:45:00-07 |          | Give Away  |     0 |        0
```

Exercise

Write a query to find the sales made by a customer, **using a parameter**

```
SELECT * FROM sale WHERE customer = :name;
```

Using psql \set command to set the values of this parameter

📖 psql Documentation

Exercise

Write a query to find the sales made by a customer, **using a parameter**

```
\set name '\''Bill'\''

SELECT * FROM sale WHERE customer = :name;
id | branch |        sold_at         | customer |  product   | price | discount
---+--------+-------------------------+----------+------------+-------+----------
 1 | NY     | 2020-03-31 20:15:00-07 | Bill     | Shoes      | 10000 |     1000
 7 | LA     | 2020-03-31 02:55:00-07 | Bill     | Shirt      |  1250 |        0
13 | LA     | 2020-04-01 23:01:00-07 | Bill     | Give Away |     0 |        0
```

Exercise

Write a query to find the sales made by a customer, **using a parameter**

```
\set name null
SELECT * FROM sale WHERE customer = :name;
id | branch | sold_at | customer | product | price | discount
----+--------+---------+----------+---------+-------+----------
```

Is this correct?

Comparing NULL values

To compare null, we need to use **IS**

```
SELECT NULL IS NULL;
?column?
----------
t

SELECT NULL = NULL;
?column?
----------

SELECT (NULL = NULL) IS NULL;
?column?
----------
t
```

Comparing NULL values

How to compare both null and literal values?

```
SELECT *
FROM sale
WHERE
   (:name IS NULL AND customer IS NULL)
   OR
   (:name IS NOT NULL AND customer = :name);
```

Handle NULL case

Handle Literal values

Comparing NULL values

How to compare both null and literal values?

```
\set name '\''Bill'\''
SELECT * FROM sale
WHERE (:name IS NULL AND customer IS NULL)
OR (:name IS NOT NULL AND customer = :name);

id | branch |        sold_at         | customer |  product   | price | discount
---+--------+-------------------------+----------+------------+-------+----------
 1 | NY     | 2020-03-31 20:15:00-07 | Bill     | Shoes      | 10000 |     1000
 7 | LA     | 2020-03-31 02:55:00-07 | Bill     | Shirt      |  1250 |        0
13 | LA     | 2020-04-01 23:01:00-07 | Bill     | Give Away |     0 |        0
```

Comparing NULL values

How to compare both null and literal values?

```
\set name NULL
SELECT * FROM sale
WHERE (:name IS NULL AND customer IS NULL)
OR (:name IS NOT NULL AND customer = :name);

id | branch |        sold_at         | customer |  product   | price | discount
---+--------+-------------------------+----------+------------+-------+----------
 2 | NY     | 2020-03-31 21:00:00-07 |          | Shoes      |  5000 |        0
 5 | NY     | 2020-03-31 20:15:00-07 |          | Shirt      |  1500 |        0
12 | LA     | 2020-04-01 23:01:00-07 |          | Hat        |  8000 |     8000
14 | NY     | 2020-03-31 10:01:00-07 |          | Give Away  |     0 |        0
15 | LA     | 2020-04-01 03:45:00-07 |          | Give Away  |     0 |        0
```

Comparing NULL values

There must be a better way!

There is!

Comparing NULL values
IS DISTINCT FROM

```
SELECT *
FROM sale
WHERE customer IS NOT DISTINCT FROM :name;
```

a IS DISTINCT FROM b
a IS NOT DISTINCT FROM b
Treating null like an ordinary value

📖 Comparison Functions and Operators
📖 The Many Faces of DISTINCT in PostgreSQL

Exercise

Write a query to find the sales made by a customer, **using a parameter**

```
\set name '\''Bill'\''
SELECT * FROM sale WHERE customer IS NOT DISTINCT FROM :name;
id | branch |        sold_at         | customer |  product   | price | discount
---+--------+-------------------------+----------+------------+-------+----------
 1 | NY     | 2020-03-31 20:15:00-07 | Bill     | Shoes      | 10000 |     1000
...

\set name NULL
SELECT * FROM sale WHERE customer IS NOT DISTINCT FROM :name;
id | branch |        sold_at         | customer |  product   | price | discount
---+--------+-------------------------+----------+------------+-------+----------
 2 | NY     | 2020-03-31 21:00:00-07 |          | Shoes      |  5000 |        0
...
```

IS DISTINCT FROM

All cases

```
WITH t AS (
   SELECT 1    as a, 1    as b UNION ALL
   SELECT 1    as a, 2    as b UNION ALL
   SELECT 1    as a, NULL as b UNION ALL
   SELECT NULL as a, NULL as b
)
SELECT
   a,
   b,
   a = b as equal,
   a IS NOT DISTINCT FROM b AS is_not_distinct_from
FROM
   t;
```

```
 a    |  b    | equal | is_not_distinct_from
------+-------+-------+----------------------
    1 |     1 | t     | t
    1 |     2 | f     | f
    1 |  NULL | NULL  | f
 NULL |  NULL | NULL  | t
```

Common Mistakes In SQL

Working with dates and times

Exercise

Find the amount of sales during each month

```
  month     | total_sales
------------+-------------
2020-03-01  |      ?
2020-04-01  |      ?
```

Hint: Use date_trunc to get the month

Working with dates and times

How is date represented in the database?

- **EPOCH**: 01/01/1970 00:00:00 UTC
- **Unix time**: Seconds since **EPOCH**

```
SELECT extract('epoch' FROM now());

    date_part
--------------------
1583585129.063067
```

Working with dates and times
The year 2038 problem (Y2k38)

- **On 32-bit systems, signed integer can only go as far as 2038-01-19 03:14:07 UTC**
- Remember Y2K?

```
SET TIME ZONE UTC;
SET

SELECT
 '1970-01-01 UTC'::timestamptz
 + interval '1 second' * (pow(2, 31) - 1) AS y28k;

y28k | 2038-01-19 03:14:07+00
```

📖 Year 2038 Problem

IMAGE CREDIT

Working with dates and times

Time zones

- **UTC, Coordinated Universal Time:** the primary time standards, not adjusted to daylight savings.
- **GMT, Greenwich Mean Time**: Synonym for UTC (time at Greenwich, England).

```
SELECT
   now() at time zone 'UTC' as utc,
   now() at time zone 'GMT' as gmt;

utc | 2020-03-12 14:07:03.139617
gmt | 2020-03-12 14:07:03.139617
```

📖 Timezones: Documentation

Time Zone

Represented as an offset from UTC

```
select *
from pg_timezone_names
where name ~* 'Sydney|Tel_Aviv|Paris|London|New_York'
and name not like 'posix%'
order by utc_offset;

      name        | abbrev | utc_offset | is_dst
------------------+--------+------------+--------
America/New_York  | EST    | -05:00:00  | f
Europe/London     | GMT    | 00:00:00   | f
Europe/Paris      | CET    | 01:00:00   | f
Asia/Tel_Aviv     | IST    | 02:00:00   | f
Australia/Sydney  | AEDT   | 11:00:00   | t
```

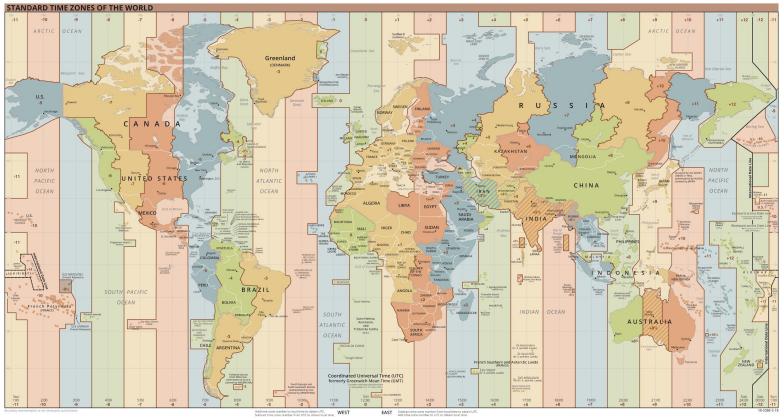


IMAGE CREDIT

📖 IANA: Database of all the time zones

Exercise

Find your time zone

💡 Use the table `pg_timezone_names`

Working with dates and times

Using time zones

- Specify the time zone of a timestamp

```
SELECT '2020-03-22 17:00:00 America/New_York'::timestamptz;
```

- Convert a timestamp to a different time zone

```
SELECT '2020-03-22 17:00:00 America/New_York'::timestamptz AT TIME ZONE 'Australia/Sydney';
     timezone
---------------------
2020-03-23 08:00:00

SELECT now() AT TIME ZONE 'Australia/Sydney'
  timezone
----------------------------
2020-03-08 00:13:37.826917
```

Exercise

What is the time in **your local time zone**?

Use `NOW() at time zone 'YOUR LOCAL TIME ZONE NAME'`

Working with dates and times
Daylight Saving

- **DST, Daylight saving:** Changes to the clock to extend daylight (in warm seasons)
- **For example**: In the US, clock moves on the second Sunday in March (8/3/2020):

```
-- "Standard Time"
SELECT '2020-03-01 12:00:00 America/New_York' AT TIME ZONE 'UTC';
     timezone
---------------------
2020-03-01 17:00:00

-- "Daylight Time" (AKA "Summer Time"):
SELECT '2020-03-22 12:00:00 America/New_York' AT TIME ZONE 'UTC';
     timezone
---------------------
2020-03-22 16:00:00
```

Working with dates and times
Daylight Saving

- Daylight saving can produce surprising results:

```
select '2020-03-07 12:00:00 America/New_York'::timestamptz -
'2020-03-08 12:00:00 America/New_York'::timestamptz as diff;
  diff
-----------
-23:00:00
```

The day the clock moves there's only 23 hours!

Working with dates and times

Database types

- **date:** date without time
- **timestamp:** date and time without time zone
- **timestamp with time zone (timestamptz):** date and time with time zone

```
\d sale
       Table "public.sale"
  Column   |           Type
-----------+--------------------------
 id        | integer
 branch    | text
 sold_at   | timestamp with time zone
 customer  | text
 product   | text
 price     | integer
 discount  | integer
```

📖 Date/Time Types Documentation

Working with dates and times
Interval

```
SELECT '2020-03-22 17:00 UTC'::timestamptz + INTERVAL '3 hours';
2020-03-22 20:00:00+00


SELECT '2020-03-22 17:00 UTC'::timestamptz + INTERVAL '3 hours 2 minutes 55 seconds';
2020-03-22 20:02:55+00


SELECT '2020-03-22 17:00 UTC'::timestamptz - INTERVAL '1 days' * 3;
2020-03-19 17:00:00+00
```

📖 Interval Documentation

Working with dates and times

Useful functions

```
SELECT now();
now | 2020-03-12 14:41:18.910296+00

SELECT now()::date;
now | 2020-03-12

SELECT date_trunc('month', now());
date_trunc | 2020-03-01 00:00:00+00

SELECT date_trunc('hour', now());
date_trunc | 2020-03-12 14:00:00+00
```

```
SELECT date_part('month', now());
date_part | 3

SELECT date_part('dow', now());
date_part | 4

SELECT extract('month' from now());
date_part | 3

SELECT extract('day' from now());
date_part | 12
```

Exercise

Find the amount of sales during each month

What went wrong...

```
SET TIME ZONE 'America/New_York';

SELECT date_trunc('month', sold_at) AS month,
sum(price) AS total_sales
FROM sale
GROUP BY month;

        month           | total_sales
------------------------+-------------
2020-04-01 00:00:00-04 |       37500
2020-03-01 00:00:00-05 |       26850
```

```
SET TIME ZONE 'America/Los_Angeles';

SELECT date_trunc('month', sold_at) AS month,
sum(price) AS total_sales
FROM sale
GROUP BY month;

        month           | total_sales
------------------------+-------------
2020-04-01 00:00:00-07 |       17500
2020-03-01 00:00:00-08 |       46850
```

Take Away

Unless explicitly mentioned, timezone is usually set by the client application

```
show time zone;

    TimeZone
------------------
America/New_York
```

Exercise

Find the amount of sales during each month

***Assuming that billing for all branches is according to time zone "America/New_York"**

Hint: Explicitly set the time zone for the sold_at date.

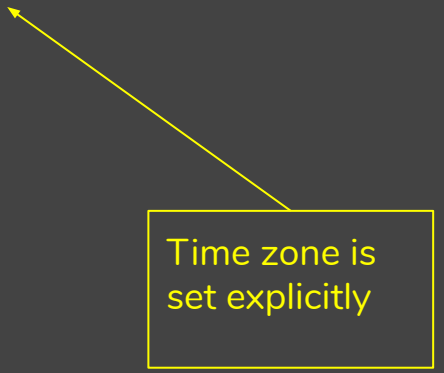
Exercise

Find the amount of sales during each month

***Assuming that billing for all branches is according to time zone "America/New_York"**

```
SELECT
   date_trunc('month', sold_at at time zone 'America/New_York') AS month,
   SUM(price) AS total_sales
FROM sale
GROUP BY month;
```

```
       month         | total_sales
---------------------+-------------
2020-04-01 00:00:00 |       37500
2020-03-01 00:00:00 |       26850
```



Exercise

What is the busiest **hour of the day** in all branches?

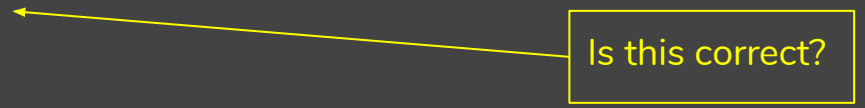
Hint: Which time zone should you use to extract the hour?

Exercise

What is the busiest **hour of the day** in all branches?

```
SELECT extract('hour' FROM sold_at) AS hour_of_day, COUNT(*) AS sales
FROM sale
GROUP BY hour_of_day
ORDER BY sales desc;

hour_of_day | sales
------------+-------
         10 |     3
          6 |     3
          9 |     2
          3 |     2
          7 |     2
...
```



Exercise

What is the busiest **hour of the day** in all branches?

```
SELECT extract('hour' FROM sold_at AT TIME ZONE CASE
       WHEN branch = 'NY' then 'America/New_York'
       WHEN branch = 'LA' then 'America/Los_Angeles'
   END) AS hour_of_day,
   count(*) AS sales
FROM sale
GROUP BY hour_of_day
ORDER BY sales DESC;
```

```
hour_of_day | sales
------------+-------
         23 |     5
          3 |     4
          0 |     2
          2 |     2
         22 |     1
         13 |     1
```

When working with dates and times be explicit about the time zone

Wrong!	Right!
`date_part('month', sold_at)`	`date_part('month', sold_at at time zone 'America/New_York')`
`extract('month' from sold_at)`	`extract('month' from sold_at at time zone 'America/New_York')`
`sold_at::date`	`(sold_at at time zone 'America/New_York')::date`
`‘2020-03-22 11:00’`	`‘2020-03-22 11:00 America/New_York’`

Exercise

How many sales were there in **March**?

***In "America/New_York" time**

Exercise

How many sales were there in **March**?

***In "America/New_York" time**

```
SELECT count(*)
FROM sale
WHERE sold_at BETWEEN '2020-03-01 America/New_York' AND '2020-04-01 America/New_York';

count
-------
 9
```

Exercise

How many sales were there in **April**?

***In "America/New_York" time**

```
SELECT count(*)
FROM sale
WHERE sold_at BETWEEN '2020-04-01 America/New_York' AND '2020-05-01 America/New_York';

count
-------
 7
```

Exercise

What can possibly go wrong?

```
SELECT count(*) FROM sale;

count
-------
 15
```

March (9) + April (7) = 16 😭😭😭

BETWEEN is Inclusive!

One sale is counted twice

```
SELECT * FROM sale WHERE sold_at = '2020-04-01 America/New_York';

id | branch |        sold_at         | customer | product | price | discount
---+--------+-------------------------+----------+---------+-------+----------
 2 | NY     | 2020-03-31 21:00:00-07 |          | Shoes   |  5000 |        0
```

Use Half Open Ranges

Ranges that don't overlap

```
SELECT count(*)
FROM sale
WHERE sold_at >= '2020-03-01 America/New_York'
AND sold_at < '2020-04-01 America/New_York';
count
-------
    8

SELECT count(*)
FROM sale
WHERE sold_at >= '2020-04-01 America/New_York'
AND sold_at < '2020-05-01 America/New_York';
count
-------
    7
```

March (8) + April (7) = 15 

BETWEEN is Inclusive!

Not only timestamps...

- **Timestmaps**
  - Unlikely if data is created by users (odd times)
  - More likely when data is created automatically (by batch jobs etc.)
- **Integers**
  - Binning
  - Dividing to buckets
  - Search for ranges...
- **Date** (with no time)
  - Very common when filtering on range of dates

```
-- Search for 90's babies...
SELECT * FROM birthdates WHERE birthdate BETWEEN '1990-01-01' AND '2000-01-01';
```

SUMMARY

Mistakes you can now avoid!

- **Be Careful when dividing integers**
  Dividing by an integer truncates the result
- **Guard against "division by zero" errors**
  Use `NULLIF` and `COALESCE`
- **Be Careful when aggregating nulls**
  Aggregate functions ignore `NULL` values
- **Be Careful when comparing nulls**
  Use `IS DISTINCT FROM` to treat `NULL` like a value
- **Dates and times in PostgreSQL**
  How to work with timestamps, time zones, intervals and common functions
- **Be careful with date and times**
  Always be explicit about the time zone
- **BETWEEN is inclusive**
  Use half open ranges to avoid overlap

Writing Faster SQL

Performance Tips

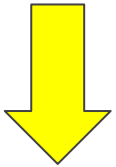
The Path of a Query

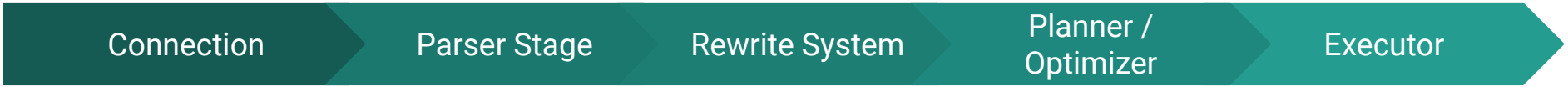
📖 Official Documentation

Connection	Parser Stage	Rewrite System	Planner / Optimizer	Executor
Client application creates a connection to the database server.	Check the query for syntax errors.	Make adjustments to the query for the database internal needs (for example, inline views). This stage does not change the logic of the query.	Evaluate the query and find the **best execution plan**.	Execute the query according to the execution plan.

The Optimizer

📖 Official Documentation





- **Generate all possible execution plans**
  - This can take some time, depending on the query.
- **Estimate the cost for each plan**
  - Cost is measured in arbitrary units
  - Cost mostly measures disk page fetches (IO)
  - Using statistics obtained from analysing tables and indexes
  - Cost can be used to compare execution plans
- **Choose the plan with the lowest cost**
  - The lower the cost the faster is execution is (exepcted) to be

Generate Some Data

Can’t optimize for a table with 15 rows…

```
INSERT INTO sale (id, branch, sold_at, customer, product, price, discount)
SELECT
   (SELECT MAX(id) FROM sale) + generate_series(1, 10000) as id,
   (ARRAY['NY', 'LA'])[ceil(random() * 2)] AS branch,
   '2020-03-01 00:00:00 UTC'::timestamptz + interval '1 hour' * random() * 24 * 30 * 6 AS sold_at,
   (ARRAY['Bill', 'David', 'John', 'Lily'])[ceil(random() * 30)] AS customer,
   (ARRAY['Shoes', 'Shirt', 'Pants', 'Hat', 'Give Away'])[ceil(random() * 4)] AS product,
   round(random() * 150 * 100)::integer / 10 * 10 as price,
   0 as discount;
```

Execution Plan

Using the EXPLAIN command

- View execution plan (will not execute the query)

```
EXPLAIN SELECT * FROM sale;
                        QUERY PLAN
-----------------------------------------------------------
Seq Scan on sale  (cost=0.00..176.15 rows=10015 width=33)
```

- Execute the query, view execution plan and timing

```
EXPLAIN (ANALYZE ON, TIMING ON) SELECT * FROM sale;
                                QUERY PLAN
-------------------------------------------------------------------------------------------
Seq Scan on sale  (cost=0.00..176.15 rows=10015 width=33) (actual time=0.023..2.539 rows=10015 loops=1)
Planning Time: 0.092 ms
Execution Time: 3.587 ms
```

Execution Plan

A closer look

```
EXPLAIN SELECT * FROM sale;
                        QUERY PLAN
-------------------------------------------------------------
Seq Scan on sale  (cost=0.00..176.15 rows=10015 width=33)
```

Database did a **sequential scan** on the entire sale table

The estimated cost is 176 (roughly equal making 176 fetches from disk)

Database estimated that the query will return 10,015 rows

📖 Using EXPLAIN

Execution Plan

A closer look

```
EXPLAIN SELECT * FROM sale WHERE id = 1000;
                                QUERY PLAN
-------------------------------------------------------------------------
Index Scan using sale_pkey on sale  (cost=0.29..8.30 rows=1 width=33)
  Index Cond: (id = 1000)
```

Database **used the index** on the primary key to access only the relevant blocks

The estimated cost is 8

Database estimated that the query will return only one row.

📖 Using EXPLAIN

Execution Plan

Plan-reading is an art that requires some experience to master...

📖 Using EXPLAIN

Execution Plan

What you need to know

- The lower the cost, the faster the query is expected to be.
- Cost is estimated based on statistics, so it can be inaccurate.

Use cost to compare execution plans

📖 pg_stats \ 📖 pg_class

```
SELECT attname, null_frac, n_distinct, correlation
FROM pg_stats
WHERE tablename = 'sale';

attname   | null_frac | n_distinct   | correlation
----------+-----------+--------------+-------------
branch    |         0 |            2 |   0.5007372
sold_at   |         0 |  -0.9997004 | 0.010782066
id        |         0 |           -1 |           1
customer  | 0.8650025 |            4 |  0.21757069
product   |         0 |            5 |  0.24206425
price     |         0 | -0.14947578 | 0.008074116
discount  |         0 |            5 |  0.99760634
```

Setup

1. Create an index on the field `customer`

```
CREATE INDEX sale_customer_ix ON sale(customer);
```

2. Create an index on the field `sold_at`

```
CREATE INDEX sale_sold_at_ix ON sale(sold_at);
```

3. Inspect the table

```
\d sale
 Column   |           Type           | Nullable |
----------+--------------------------+----------+
id        | integer                  | not null |
branch    | text                     |          |
sold_at   | timestamp with time zone |          |
customer  | text                     |          |
product   | text                     |          |
price     | integer                  |          |
discount  | integer                  |          |

Indexes:
    "sale_pkey" PRIMARY KEY, btree (id)
    "sale_customer_ix" btree (customer)
    "sale_sold_at_ix" btree (sold_at)
```

📖 PostgreSQL Indexes: Documentation
📖 Indexes in PostgreSQL
📖 Indexes in PostgreSQL: B-Tree

Nulls are Not Indexed

Null values are not index by a B-Tree index

```
EXPLAIN SELECT * FROM sale WHERE customer = 'Bill';
------------------------------------------------------------------------------
Bitmap Heap Scan on sale  (cost=6.83..86.95 rows=329 width=33)
  Recheck Cond: (customer = 'Bill'::text)
  ->  Bitmap Index Scan on sale_customer_ix  (cost=0.00..6.75 rows=329 width=0)
        Index Cond: (customer = 'Bill'::text)


EXPLAIN SELECT * FROM sale WHERE customer IS NULL;
------------------------------------------------------------
Seq Scan on sale  (cost=0.00..176.15 rows=8663 width=33)
  Filter: (customer IS NULL)
```

Avoid Transformations on Indexed Fields

Indexes cannot be used with transformations

```
EXPLAIN SELECT * FROM sale WHERE lower(customer) = 'bill';


------------------------------------------------------------
Seq Scan on sale  (cost=0.00..226.22 rows=50 width=33)
  Filter: (lower(customer) = 'bill'::text)
```

📖 Avoid transformations on indexed fields

Avoid Transformations on Indexed Fields

Common mistakes

Simple arithmetics on indexed field (id):

```
EXPLAIN SELECT * FROM sale WHERE id + 1 = 100;
------------------------------------------------------------------------
Seq Scan on sale  (cost=0.00..226.22 rows=50 width=33)
  Filter: ((id + 1) = 100)


EXPLAIN SELECT * FROM sale WHERE id = 99;
------------------------------------------------------------------------
Index Scan using sale_pkey on sale  (cost=0.29..8.30 rows=1 width=33)
  Index Cond: (id = 99)
```

Avoid Transformations on Indexed Fields

Common mistakes

Apply timezone on indexed field in comparison:

```
EXPLAIN SELECT * FROM sale WHERE sold_at at time zone 'America/New_York' > '2021-01-01';
------------------------------------------------------------------------
Seq Scan on sale  (cost=0.00..226.22 rows=3338 width=33)
  Filter: (timezone('America/New_York'::text, sold_at) > '2021-01-01 00:00:00'::timestamp
without time zone)


EXPLAIN SELECT * FROM sale WHERE sold_at  > '2021-01-01 America/New_York';
------------------------------------------------------------------------------
Index Scan using sale_sold_at_ix on sale  (cost=0.29..8.30 rows=1 width=33)
  Index Cond: (sold_at > '2021-01-01 05:00:00+00'::timestamp with time zone)
```

Avoid Transformations on Indexed Fields

Common mistakes

Date arithmetics on the indexed field:

```
EXPLAIN SELECT * FROM sale WHERE sold_at - interval '1 day' > '2021-01-01 America/New_York'::timestamptz;
------------------------------------------------------------------------------------------
Seq Scan on sale  (cost=0.00..226.22 rows=3338 width=33)
  Filter: ((sold_at - '1 day'::interval) > '2021-01-01 05:00:00+00'::timestamp with time zone)



EXPLAIN SELECT * FROM sale WHERE sold_at > '2021-01-01 America/New_York'::timestamptz + interval '1 day';
------------------------------------------------------------------------------------------
Index Scan using sale_sold_at_ix on sale  (cost=0.29..8.30 rows=1 width=33)
  Index Cond: (sold_at > ('2021-01-01 05:00:00+00'::timestamp with time zone + '1 day'::interval))
```

Avoid Transformations on Indexed Fields

Common mistakes

Change string case (lower / upper)

Dont:

```
SELECT * FROM users WHERE lower(email) = 'me@hakibenita.com'
```

Do:

```
SELECT * FROM users WHERE email = lower('ME@HakiBenita.com')
```

📖 Indexes on Expressions (if you really have to)

Avoid Transformations on Indexed Fields

Common mistakes

String concatenation

Dont:

```
SELECT * FROM users WHERE first_name || ' ' || last_name = 'Haki Benita'
```

Do:

```
SELECT * FROM users WHERE first_name = 'Haki' AND last_name = 'Benita'
```

UNION vs. UNION ALL

When to use each one...

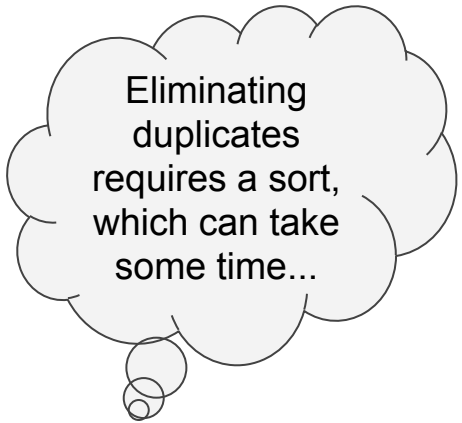
`UNION ALL`: Concatenate results

```
SELECT 1 UNION ALL SELECT 1;
?column?
----------
       1
       1
```

`UNION`: Concatenate results **and removes duplicates**

```
SELECT 1 UNION SELECT 1;
?column?
----------
       1
```

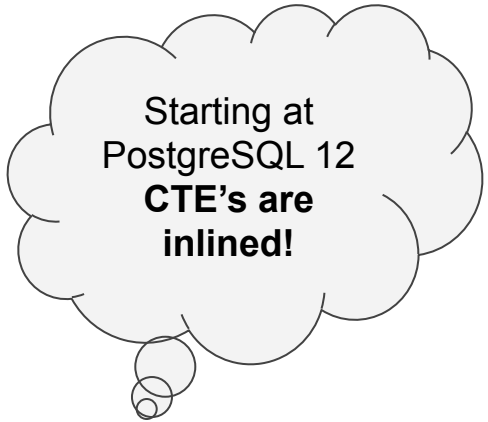
📖 Know the Difference Between UNION and UNION ALL



Common Table Expressions
CTE materialization

```
EXPLAIN SELECT * FROM sale WHERE id = 1;
------------------------------------------------------------------
Index Scan using sale_pkey on sale (cost=0.15..8.17 rows=1 width=49)
   Index Cond: (id = 1)


EXPLAIN WITH cte as (SELECT * FROM sale) SELECT * FROM cte WHERE id = 1;
------------------------------------------------------------------
CTE Scan on cte (cost=20.60..44.45 rows=5 width=49)
   Filter: (id = 1)
   CTE cte
     -> Seq Scan on sale (cost=0.00..20.60 rows=1060 width=49)
```



📖 Be Careful With CTE in PostgreSQL

SUMMARY

Make the most of your DB!

- **How the database processes a query**
  Parse, rewrite, plan and execute
- **How to produce and compare execution plans**
  Using `EXPLAIN`
- **Nulls are not indexed**
  by B-TREE indexes
- **Avoid transformations on indexed fields**
  With examples using arithmetics, date and string manipulation
- **UNION vs. UNION ALL**
  When to use each one...
- **Common Table Expressions**
  When to be careful

Writing SQL Faster

Productivity Tips

CASE Can Take Many Forms

Avoid repetition

```
SELECT
   CASE
       WHEN fruit = 'apple'  THEN 'red'
       WHEN fruit = 'pear'   THEN 'green'
       WHEN fruit = 'orange' THEN 'orange'
       ELSE '?'
   END AS color
FROM
   fruit;
```



```
SELECT
   CASE fruit
       'apple'  THEN 'red'
       'pear'   THEN 'green'
       'orange' THEN 'orange'
       ELSE '?'
   END AS color
FROM
   fruit;
```

Reference Column in GROUP BY & ORDER BY

Use position or alias to avoid repetition

```sql
SELECT
   first_name || ' ' || last_name as full_name,
   count(*) as sales_by_user
FROM
   sale
GROUP BY
   first_name || ' ' || last_name
ORDER BY
   count(*) DESC
```



```sql
SELECT
   first_name || ' ' || last_name as full_name,
   count(*) as sales_by_user
FROM
   sale
GROUP BY
   1
ORDER BY
   sales_by_user DESC
```

It's best to avoid positional column reference in code, and use it only for ad-hoc queries.

Use Selective Aggregates

Where you used to use CASE


```sql
SELECT
   COUNT(*),
   SUM(CASE WHEN customer IS NULL THEN 1 ELSE 0 END),
   SUM(CASE WHEN customer IS NOT NULL THEN 1 ELSE 0 END)
FROM
   sale;
```

→

```sql
SELECT
   COUNT(*) AS sales,
   COUNT(*) FILTER (WHERE customer IS NULL),
   COUNT(*) FILTER (WHERE customer IS NOT NULL)
FROM
   sale;
```

📖 Aggregate Expressions: Documentation

Use DISTINCT ON

Get the first / last **row** in a group

One of the most useful features of PostgreSQL!

```
SELECT DISTINCT ON (customer)
    *
FROM
    sale
ORDER BY
    customer,
    sold_at
```

- Can be used instead of `RANK` / `ROW_NUMBER`
- `DISTINCT ON` clause can accept multiple fields
- Field in `DISTINCT ON` must be in `ORDER BY`
- Can control first / last using sort order (`ASC` / `DESC`)

📖 SELECT: Documentation

📖 The many faces of distinct in PostgreSQL: `DISTINCT ON`

What You've Learned

- How to avoid common mistakes in SQL
- How to write faster SQL
- How to write SQL faster

THANK YOU!

Haki Benita