

DuckDB & SQL

Lecture 19

Dr. Colin Rundel

SQL

Structures Query Language is a special purpose language for interacting with (querying and modifying) indexed tabular data.

- ANSI Standard but with dialect divergence (MySQL, Postgres, SQLite, etc.)
- This functionality maps very closely (but not exactly) with the data manipulation verbs present in dplyr.
- SQL is likely to be a foundational skill if you go into industry - learn it and put it on your CV

DuckDB

DuckDB is an open-source column-oriented relational database management system (RDBMS) originally developed by Mark Raasveldt and Hannes Mühleisen at the Centrum Wiskunde & Informatica (CWI) in the Netherlands and first released in 2019. The project has over 6 million downloads per month. It is designed to provide high performance on complex queries against large databases in embedded configuration, such as combining tables with hundreds of columns and billions of rows. Unlike other embedded databases (for example, SQLite) DuckDB is not focusing on transactional (OLTP) applications and instead is specialized for online analytical processing (OLAP) workloads.

From [Wikipedia](#) - DuckDB

DuckDB & DBI

DuckDB is a relational database just like SQLite and can be interacted with using DBI and the duckdb package.

```
1 library(DBI)
2 (con = dbConnect(duckdb::duckdb()))
```

```
<duckdb_connection fef40 driver=<duckdb_driver dbdir=':memory:' read_only=FALSE bigint
```

```
1 dbWriteTable(con, "flights", nycflights13::flights)
2 dbListTables(con)
```

```
[1] "flights"
```

```
1 dbGetQuery(con, "SELECT * FROM flights") |>
2   as_tibble()
```

```
# A tibble: 336,776 × 19
```

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time
	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>
1	2013	1	1	517	515	2	830
2	2013	1	1	533	529	4	850
3	2013	1	1	542	540	2	923
4	2013	1	1	544	545	-1	1004
5	2013	1	1	554	600	-6	812
6	2013	1	1	554	558	-4	740
7	2013	1	1	555	600	-5	913
8	2013	1	1	557	600	-3	709
9	2013	1	1	557	600	-3	838
10	2013	1	1	558	600	-2	753

```
# i 336,766 more rows
```

```
# i 12 more variables: sched_arr_time <int>, arr_delay <dbl>,  
# carrier <chr>, flight <int>, tailnum <chr>, origin <chr>,  
# dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,  
# minute <dbl>, time_hour <dtm>
```

```

1 library(dplyr)
2 tbl(con, "flights") |>
3   filter(month == 10, day == 29) |>
4   count(origin, dest) |>
5   arrange(desc(n))

```

```

# Source:      SQL [?? x 3]
# Database:    DuckDB 1.4.1 [root@Darwin 25.0.0:R 4.5.1/:memory:]
# Ordered by: desc(n)
   origin dest      n
   <chr>  <chr> <dbl>
1 JFK     LAX      32
2 LGA     ORD      30
3 LGA     ATL      29
4 JFK     SFO      23
5 LGA     CLT      21
6 EWR     ORD      18
7 EWR     SFO      16
8 LGA     DCA      16
9 LGA     BOS      16
10 JFK    BOS      16
# i more rows

```

DuckDB CLI

Connecting via CLI

```
1 > #| eval: false  
2 > duckdb employees.duckdb
```

v1.2.1 8e52ec4395

Enter ".help" for usage hints.

```
1 D
```

Table information

Dot commands are expressions that begins with `.` and are specific to the DuckDB CLI, some examples include:

```
1 D .tables
```

```
## employees
```

```
1 D .schema employees
```

```
## CREATE TABLE employees("name" VARCHAR, email VARCHAR, salary DOUBLE, dept VARCHAR);
```

```
1 D .indexes employees
```

```
1 D .maxrows 20
```

```
2 D .maxwidth 80
```

A full list of available dot commands can be found [here](#) or listed via `.help` in the CLI.

SELECT Statements

```
1 D SELECT * FROM employees;
```

##				
##	name	email	salary	dept
##	varchar	varchar	double	varchar
##				
##	Alice	alice@company.com	52000.0	Accounting
##	Bob	bob@company.com	40000.0	Accounting
##	Carol	carol@company.com	30000.0	Sales
##	Dave	dave@company.com	33000.0	Accounting
##	Eve	eve@company.com	44000.0	Sales
##	Frank	frank@comany.com	37000.0	Sales
##				

Output formats

The format of duckdb's output (in the CLI) is controlled via `.mode` - the default is `duckbox`, see possible [output formats](#).

```
1 D .mode csv
2 D SELECT * FROM employees;
```

```
## name,email,salary,dept
## Alice,alice@company.com,52000.0,Accounting
## Bob,bob@company.com,40000.0,Accounting
## Carol,carol@company.com,30000.0,Sales
## Dave,dave@company.com,33000.0,Accounting
## Eve,eve@company.com,44000.0,Sales
## Frank,frank@comany.com,37000.0,Sales
```

```
1 D .mode json
2 D SELECT * FROM employees;
```

```
## [{"name":"Alice","email":"alice@company.com","salary":52000.0
## {"name":"Bob","email":"bob@company.com","salary":40000.0,"dep
## {"name":"Carol","email":"carol@company.com","salary":30000.0,
## {"name":"Dave","email":"dave@company.com","salary":33000.0,"d
## {"name":"Eve","email":"eve@company.com","salary":44000.0,"dep
## {"name":"Frank","email":"frank@comany.com","salary":37000.0,"
```

```
1 D .mode markdown
2 D SELECT * FROM employees;
```

##	name	email	salary	dept
##	Alice	alice@company.com	52000.0	Accounting
##	Bob	bob@company.com	40000.0	Accounting
##	Carol	carol@company.com	30000.0	Sales
##	Dave	dave@company.com	33000.0	Accounting
##	Eve	eve@company.com	44000.0	Sales
##	Frank	frank@comany.com	37000.0	Sales

```
1 D .mode insert
2 D SELECT * FROM employees;
```

```
INSERT INTO "table"("name",email,salary,dept) VALUES('Alice','al
INSERT INTO "table"("name",email,salary,dept) VALUES('Bob','bob@
INSERT INTO "table"("name",email,salary,dept) VALUES('Carol','ca
INSERT INTO "table"("name",email,salary,dept) VALUES('Dave','dav
INSERT INTO "table"("name",email,salary,dept) VALUES('Eve','eve@
INSERT INTO "table"("name",email,salary,dept) VALUES('Frank','fr
```

A brief tour of SQL

select() using SELECT

We can subset for certain columns (and rename them) using **SELECT**

```
1 D SELECT name AS first_name, salary FROM employees;
```

```
##
```

first_name	salary
varchar	double
Alice	52000.0
Bob	40000.0
Carol	30000.0
Dave	33000.0
Eve	44000.0
Frank	37000.0

```
##
```

arrange() using ORDER BY

We can sort our results by adding **ORDER BY** to our **SELECT** statement and reverse the ordering by include **DESC**.

```
1 D SELECT name AS first_name, salary FROM employees
2 D ORDER BY salary;
```

```
##
##
```

first_name	salary
varchar	double
Carol	30000.0
Dave	33000.0
Frank	37000.0
Bob	40000.0
Eve	44000.0
Alice	52000.0

```
##
```

```
1 D SELECT name AS first_name, salary FROM employees
2 D ORDER BY salary DESC;
```

```
##
##
```

first_name	salary
varchar	double
Alice	52000.0
Eve	44000.0
Bob	40000.0
Frank	37000.0
Dave	33000.0
Carol	30000.0

```
##
```

filter() using WHERE

We can filter rows using a **WHERE** clause

```
1 D SELECT * FROM employees WHERE salary < 40000;
```

##

name varchar	email varchar	salary double	dept varchar
Carol	carol@company.com	30000.0	Sales
Dave	dave@company.com	33000.0	Accounting
Frank	frank@comany.com	37000.0	Sales

##

```
1 D SELECT * FROM employees WHERE salary < 40000 AND dept = 'Sales';
```

##

name varchar	email varchar	salary double	dept varchar
Carol	carol@company.com	30000.0	Sales
Frank	frank@comany.com	37000.0	Sales

##

group_by() and summarize() via GROUP BY

We can create groups for the purpose of summarizing using **GROUP BY**.

```
1 D SELECT dept, COUNT(*) AS n FROM employees GROUP BY dept;
```

```
##  
##  
##  
##  
##  
##  
##
```

dept varchar	n int64
Sales	3
Accounting	3

head() using LIMIT

We can limit the number of results we get by using **LIMIT**

```
1 D SELECT * FROM employees LIMIT 3;
```

```
##
##
##
##
##
##
##
##
```

name varchar	email varchar	salary double	dept varchar
Alice	alice@company.com	52000.0	Accounting
Bob	bob@company.com	40000.0	Accounting
Carol	carol@company.com	30000.0	Sales

Exercise 1

Using duckdb calculate the following quantities for `employees.duckdb`,

1. The total costs in payroll for this company
2. The average salary within each department

Reading from CSV files

DuckDB has a neat trick in that it can treat files as tables (for supported formats), this lets you query them without having to explicitly create a table in the database and .

We can also make this explicit by using the `read_csv()` function, which is useful if we need to use custom options (e.g. specify a different delimiter)

```
1 D SELECT * FROM 'phone.csv';
```

##		
##	name	phone
##	varchar	varchar
##		
##	Bob	919 555-1111
##	Carol	919 555-2222
##	Eve	919 555-3333
##	Frank	919 555-4444
##		

```
1 D SELECT * FROM  
2 D read_csv('phone.csv', delim = ',');
```

##		
##	name	phone
##	varchar	varchar
##		
##	Bob	919 555-1111
##	Carol	919 555-2222
##	Eve	919 555-3333
##	Frank	919 555-4444
##		

Tables from CSV

If we wanted to explicitly create a table from the CSV file this is also possible,

```
1 D .tables
```

employees

```
1 D CREATE TABLE phone AS
2 D SELECT * FROM 'phone.csv';
```

```
1 D .tables
```

employees phone

```
1 D SELECT * FROM phone;
```

##

##	name	phone
##	varchar	varchar
##	Bob	919 555-1111
##	Carol	919 555-2222
##	Eve	919 555-3333
##	Frank	919 555-4444
##		

Views from CSV

It is also possible to create a view from a file - this acts like a table but the data is not copied from the file.

```
1 D .tables
```

employees

```
1 D CREATE VIEW phone_view AS
2 D   SELECT * FROM 'phone.csv';
```

```
1 D .tables
```

employees phone phone_view

```
1 D SELECT * FROM phone_view;
```

##

##	name	phone
##	varchar	varchar
##	Bob	919 555-1111
##	Carol	919 555-2222
##	Eve	919 555-3333
##	Frank	919 555-4444
##		

Deleting tables and views

Tables and views can be deleted using **DROP**

```
1 D .tables
```

```
## employees  phone  phone_view
```

```
1 D DROP TABLE phone;  
2 D DROP VIEW phone_view;
```

```
1 D .tables
```

```
## employees
```

Joins - Default

If not otherwise specified the default join in DuckDB will be an inner join.

```
1 D SELECT * FROM employees JOIN phone;
```

```
## Parser Error: syntax error at or near ";"  
## LINE 1: SELECT * FROM employees JOIN phone;
```

this error occurs because duckdb requires an **ON** or **USING** clause unless using **NATURAL**.

```
1 D SELECT * FROM employees NATURAL JOIN phone;
```

```
##  
##  
##  
##  
##  
##  
##  
##  
##
```

name varchar	email varchar	salary double	dept varchar	phone varchar
Bob	bob@company.com	40000.0	Accounting	919 555-1111
Carol	carol@company.com	30000.0	Sales	919 555-2222
Eve	eve@company.com	44000.0	Sales	919 555-3333
Frank	frank@comany.com	37000.0	Sales	919 555-4444

Inner Join - Explicit

```
1 D SELECT * FROM employees JOIN phone ON employees.name = phone.name;
```

##

##	name varchar	email varchar	salary double	dept varchar	name varchar	phone varchar
##	Bob	bob@company.com	40000.0	Accounting	Bob	919 555-1111
##	Carol	carol@company.com	30000.0	Sales	Carol	919 555-2222
##	Eve	eve@company.com	44000.0	Sales	Eve	919 555-3333
##	Frank	frank@comany.com	37000.0	Sales	Frank	919 555-4444
##						...

to avoid the duplicate `name` column we can specify `USING` instead of `ON`

```
1 D SELECT * FROM employees JOIN phone USING(name);
```

##

##	name varchar	email varchar	salary double	dept varchar	phone varchar
##	Bob	bob@company.com	40000.0	Accounting	919 555-1111
##	Carol	carol@company.com	30000.0	Sales	919 555-2222
##	Eve	eve@company.com	44000.0	Sales	919 555-3333
##	Frank	frank@comany.com	37000.0	Sales	919 555-4444
##					

Left Join - Natural

```
1 D SELECT * FROM employees NATURAL LEFT JOIN phone;
```

##					
##	name	email	salary	dept	phone
##	varchar	varchar	double	varchar	varchar
##					
##	Bob	bob@company.com	40000.0	Accounting	919 555-1111
##	Carol	carol@company.com	30000.0	Sales	919 555-2222
##	Eve	eve@company.com	44000.0	Sales	919 555-3333
##	Frank	frank@comany.com	37000.0	Sales	919 555-4444
##	Alice	alice@company.com	52000.0	Accounting	
##	Dave	dave@company.com	33000.0	Accounting	
##					

Left Join - Explicit

```
1 D SELECT * FROM employees LEFT JOIN phone ON employees.name = phone.name;
```

##	name	email	salary	dept	name	phone
##	varchar	varchar	double	varchar	varchar	varchar
##	Bob	bob@company.com	40000.0	Accounting	Bob	919 555-1111
##	Carol	carol@company.com	30000.0	Sales	Carol	919 555-2222
##	Eve	eve@company.com	44000.0	Sales	Eve	919 555-3333
##	Frank	frank@comany.com	37000.0	Sales	Frank	919 555-4444
##	Alice	alice@company.com	52000.0	Accounting		
##	Dave	dave@company.com	33000.0	Accounting		
##						

Duplicate **name** column can be avoided by more restrictive **SELECT**,

```
1 D SELECT employees.*, phone FROM employees LEFT JOIN phone ON employees.name = phone.name;
```

##	name	email	salary	dept	phone
##	varchar	varchar	double	varchar	varchar
##	Bob	bob@company.com	40000.0	Accounting	919 555-1111
##	Carol	carol@company.com	30000.0	Sales	919 555-2222
##	Eve	eve@company.com	44000.0	Sales	919 555-3333
##	Frank	frank@comany.com	37000.0	Sales	919 555-4444
##	Alice	alice@company.com	52000.0	Accounting	
##	Dave	dave@company.com	33000.0	Accounting	
##					

Other Joins

As you would expect all other standard joins are supported including **RIGHT JOIN**, **FULL JOIN**, **CROSS JOIN**, **SEMI JOIN**, **ANTI JOIN**, etc.

```
1 D SELECT employees.*, phone FROM employees NATURAL FULL JOIN phone;
```

##

##	name varchar	email varchar	salary double	dept varchar	phone varchar
##	Bob	bob@company.com	40000.0	Accounting	919 555-1111
##	Carol	carol@company.com	30000.0	Sales	919 555-2222
##	Eve	eve@company.com	44000.0	Sales	919 555-3333
##	Frank	frank@comany.com	37000.0	Sales	919 555-4444
##	Alice	alice@company.com	52000.0	Accounting	
##	Dave	dave@company.com	33000.0	Accounting	
##					

```
1 D SELECT employees.*, phone FROM employees NATURAL RIGHT JOIN phone;
```

##

##	name varchar	email varchar	salary double	dept varchar	phone varchar
##	Bob	bob@company.com	40000.0	Accounting	919 555-1111
##	Carol	carol@company.com	30000.0	Sales	919 555-2222
##	Eve	eve@company.com	44000.0	Sales	919 555-3333
##	Frank	frank@comany.com	37000.0	Sales	919 555-4444
##					

Subqueries

Tables can be nested within tables for the purpose of queries,

```
1 D SELECT * FROM (  
2 D   SELECT * FROM employees NATURAL LEFT JOIN phone  
3 D ) combined WHERE phone IS NULL;
```

##

name varchar	email varchar	salary double	dept varchar	phone varchar
Alice	alice@company.com	52000.0	Accounting	
Dave	dave@company.com	33000.0	Accounting	

##

```
1 D SELECT * FROM (  
2 D   SELECT * FROM employees NATURAL LEFT JOIN phone  
3 D ) combined WHERE phone IS NOT NULL;
```

##

name varchar	email varchar	salary double	dept varchar	phone varchar
Bob	bob@company.com	40000.0	Accounting	919 555-1111
Carol	carol@company.com	30000.0	Sales	919 555-2222
Eve	eve@company.com	44000.0	Sales	919 555-3333
Frank	frank@comany.com	37000.0	Sales	919 555-4444

##

Exercise 2

Lets try to create a table that has a new column - `abv_avg` which contains how much more (or less) than the average, for their department, each person is paid.

Hint - This will require joining a subquery.

Query plan

Setup

To give us a bit more variety (and data), we have created another SQLite database `flights.duckdb` that contains both `nycflights13::flights` and `nycflights13::planes`, the latter of which has details on the characteristics of the planes in the dataset as identified by their tail numbers.

```
1 db = DBI::dbConnect(duckdb::duckdb(), "flights.duckdb")
2 dplyr::copy_to(db, nycflights13::flights, name = "flights", temporary = FALSE, over
3 dplyr::copy_to(db, nycflights13::planes, name = "planes", temporary = FALSE, overv
4 DBI::dbDisconnect(db)
```

All of the following code will be run in the DuckDB command line interface, make sure you've created the database and copied both the flights and planes tables into the db or use the version provided in the [exercises/](#) repo.

Opening `flights.duckdb`

The database can then be opened from the terminal tab using,

```
1 > #| eval: false  
2 > duckdb flights.duckdb
```

As before we set a couple of configuration options so that our output is readable. We also include `.timer on` so that we get timings for our queries.

```
1 D .maxrows 20  
2 D .maxwidth 80  
3 D .timer on
```

flights

```
1 D SELECT * FROM flights LIMIT 10;
```

##

year	month	day	...	distance	hour	minute	time_hour
int32	int32	int32		double	double	double	timestamp
2013	1	1	...	1400.0	5.0	15.0	2013-01-01 10:00:00
2013	1	1	...	1416.0	5.0	29.0	2013-01-01 10:00:00
2013	1	1	...	1089.0	5.0	40.0	2013-01-01 10:00:00
2013	1	1	...	1576.0	5.0	45.0	2013-01-01 10:00:00
2013	1	1	...	762.0	6.0	0.0	2013-01-01 11:00:00
2013	1	1	...	719.0	5.0	58.0	2013-01-01 10:00:00
2013	1	1	...	1065.0	6.0	0.0	2013-01-01 11:00:00
2013	1	1	...	229.0	6.0	0.0	2013-01-01 11:00:00
2013	1	1	...	944.0	6.0	0.0	2013-01-01 11:00:00
2013	1	1	...	733.0	6.0	0.0	2013-01-01 11:00:00
10 rows							19 columns (7 shown)

##

##

Run Time (s): real 0.020 user 0.000784 sys 0.002284

planes

```
1 D SELECT * FROM planes LIMIT 10;
```

##	##	##	##	##	##	##	##
##	tailnum	year	type	...	seats	speed	engine
##	varchar	int32	varchar		int32	int32	varchar
##	N10156	2004	Fixed wing multi e...	...	55		Turbo-fan
##	N102UW	1998	Fixed wing multi e...	...	182		Turbo-fan
##	N103US	1999	Fixed wing multi e...	...	182		Turbo-fan
##	N104UW	1999	Fixed wing multi e...	...	182		Turbo-fan
##	N10575	2002	Fixed wing multi e...	...	55		Turbo-fan
##	N105UW	1999	Fixed wing multi e...	...	182		Turbo-fan
##	N107US	1999	Fixed wing multi e...	...	182		Turbo-fan
##	N108UW	1999	Fixed wing multi e...	...	182		Turbo-fan
##	N109UW	1999	Fixed wing multi e...	...	182		Turbo-fan
##	N110UW	1999	Fixed wing multi e...	...	182		Turbo-fan
##	10 rows						9 columns (6 shown)
##	Run Time (s): real 0.003 user 0.000819 sys 0.000018						

Exercise 3

Write a query that determines the total number of seats available on all of the planes that flew out of New York in 2013.

A Solution?

Does the following seem correct?

```
1 D SELECT sum(seats) FROM flights NATURAL LEFT JOIN planes;
```

```
##
```

##	sum(seats)
##	int128
##	614366

```
## Run Time (s): real 0.012 user 0.016061 sys 0.002386
```

Why?

Correct solution

Join and select:

```
1 D SELECT sum(seats) FROM flights LEFT JOIN planes USING (tailnum);
```

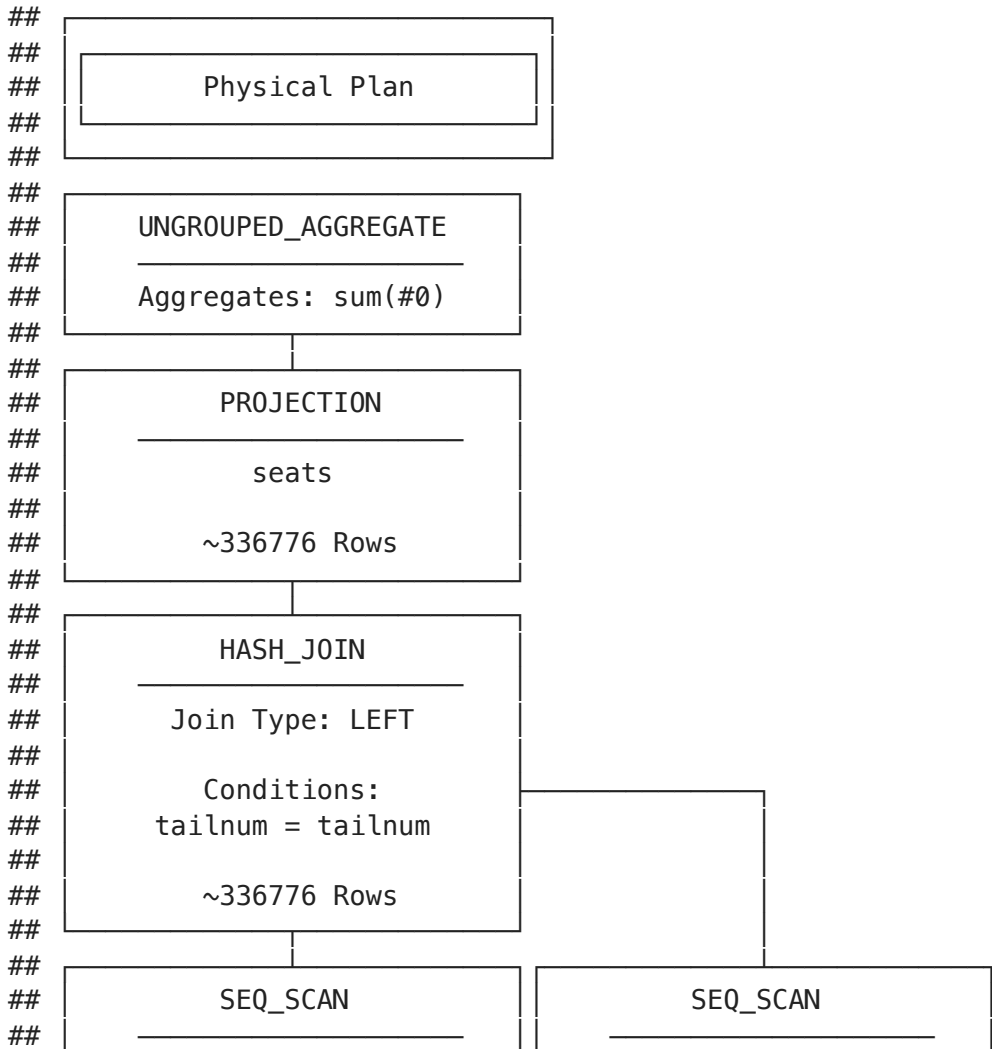
##

##	sum(seats)
##	int128
##	38851317
##	

Run Time (s): real 0.005 user 0.010150 sys 0.000291

EXPLAIN

```
1 D EXPLAIN SELECT sum(seats) FROM flights LEFT JOIN planes USING (tailnum);
```

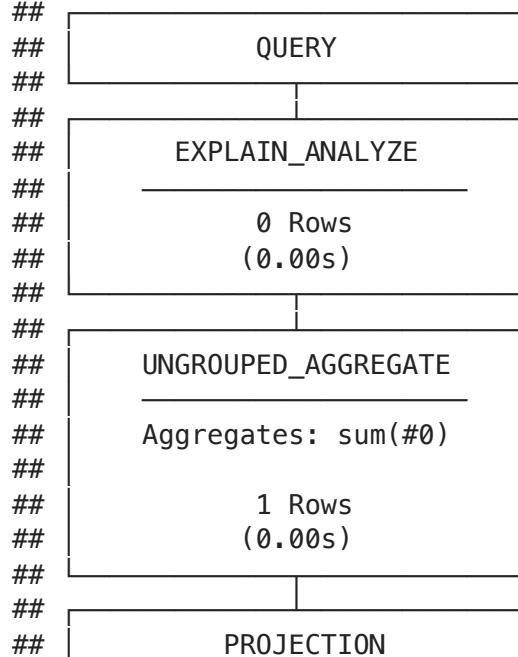


EXPLAIN ANALYZE

```
1 D EXPLAIN ANALYZE SELECT sum(seats) FROM flights LEFT JOIN planes USING (tailnum);
```

```
##
##
##  Query Profiling Information
##
##
## EXPLAIN ANALYZE SELECT sum(seats) FROM flights LEFT JOIN planes USING (tailnum);
```

Total Time: 0.0045s



dplyr

```
1 library(dplyr)
2 flights = nycflights13::flights
3 planes = nycflights13::planes
4
5 system.time({
6   flights |>
7     left_join(planes, by = c("tailnum" = "tailnum")) |>
8     summarise(total_seats = sum(seats, na.rm = TRUE))
9 })
```

user	system	elapsed
0.042	0.003	0.045

NYC Taxi Demo