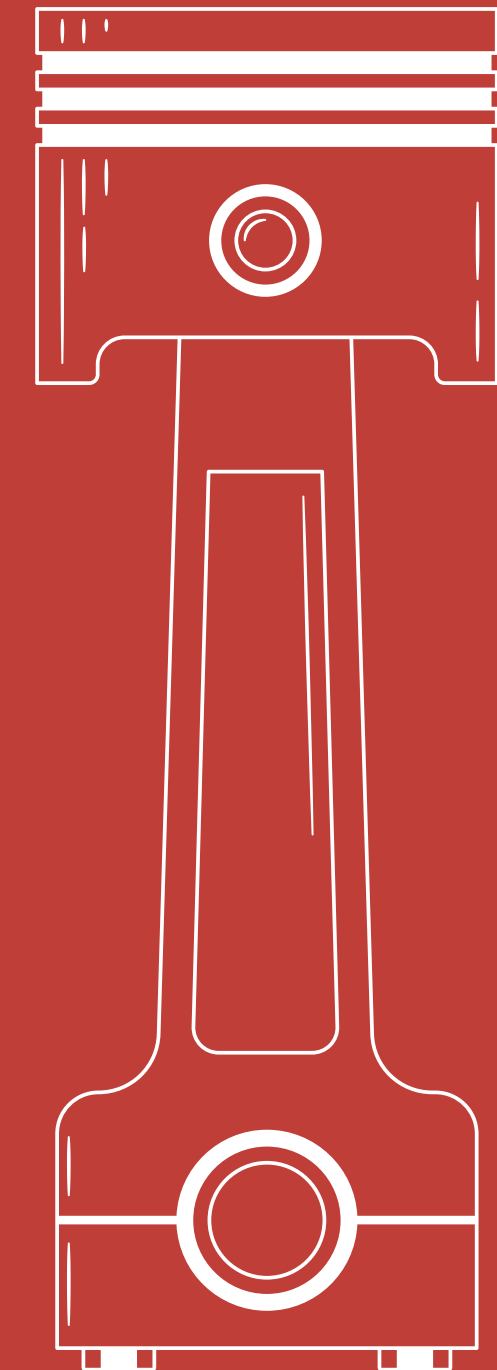
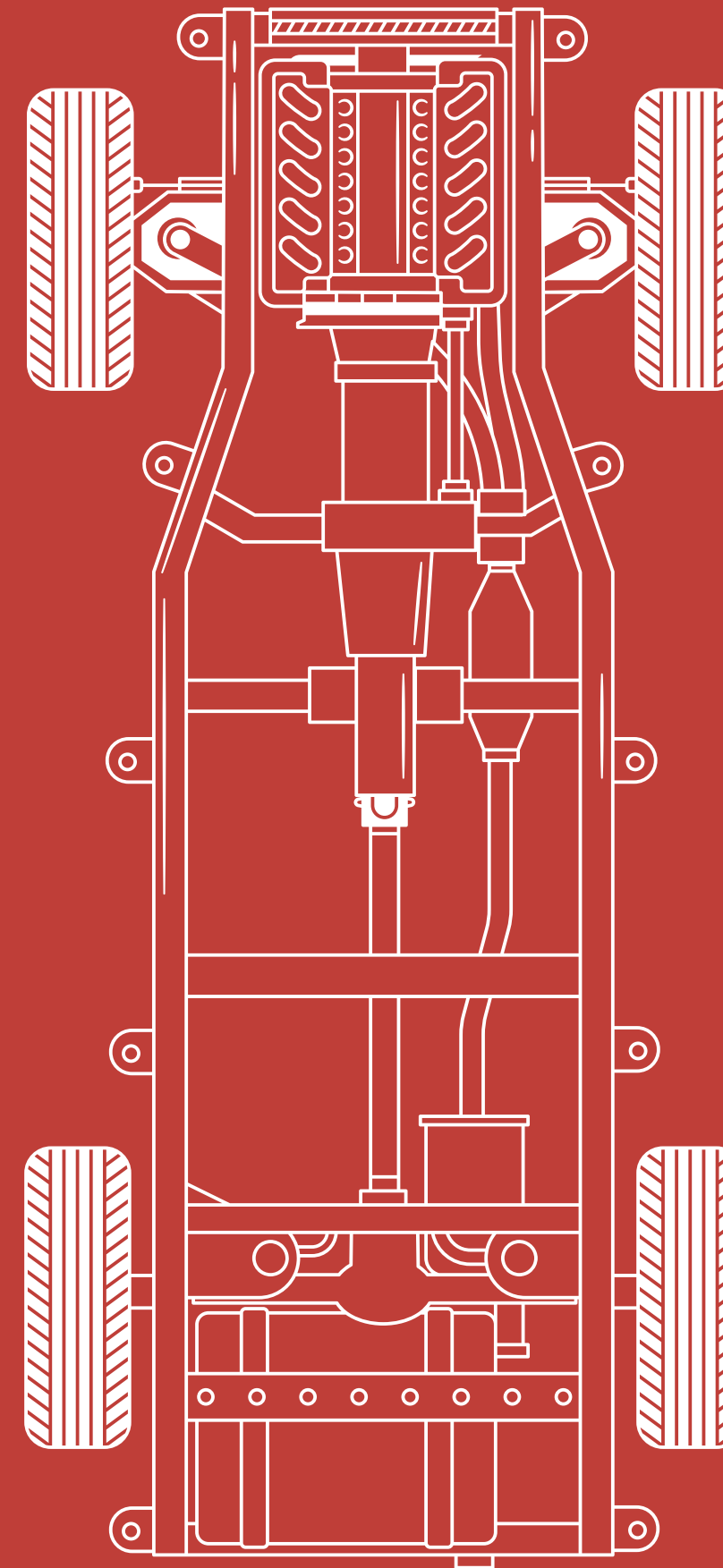


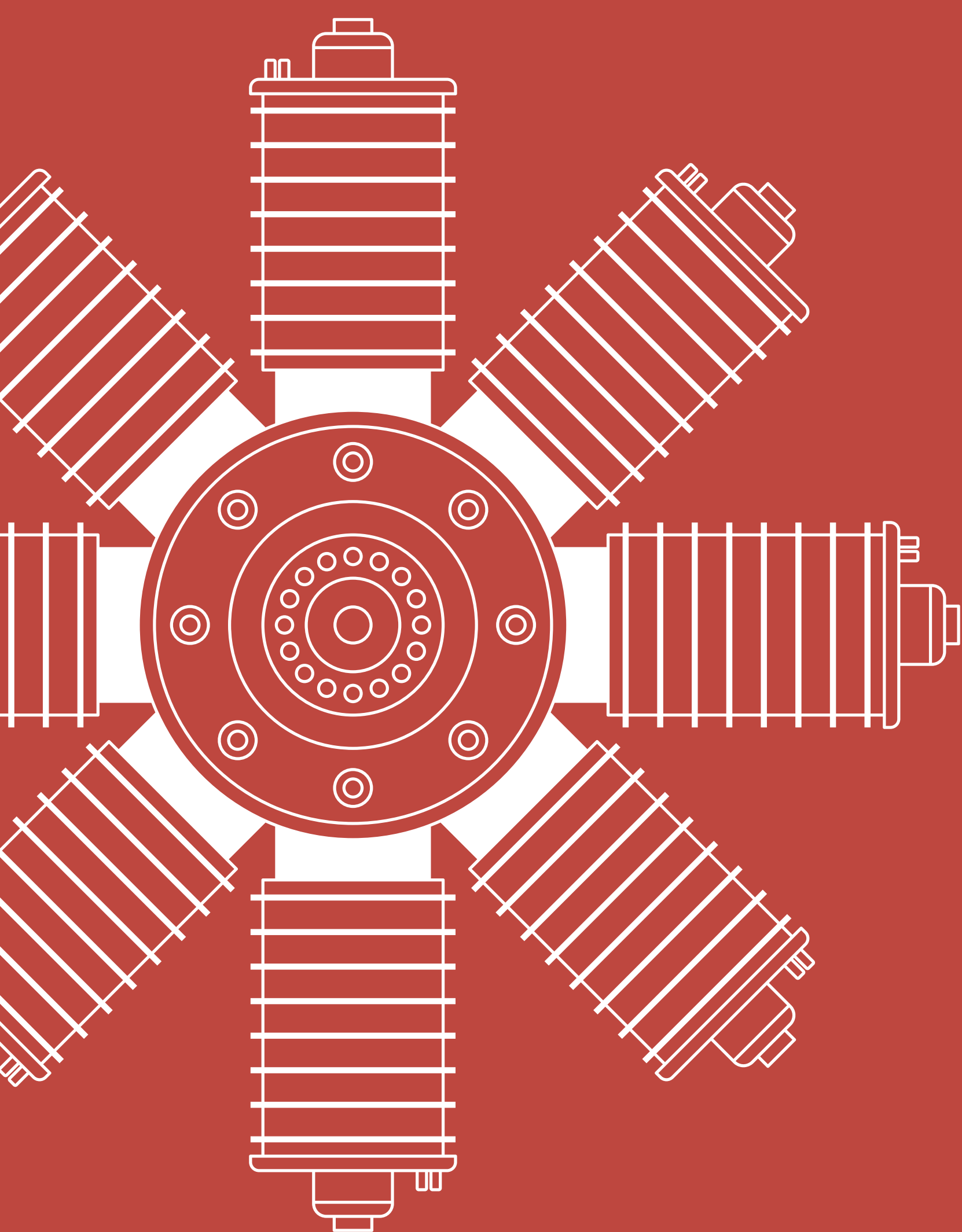


PROYECTO FINAL DATA ENGINEERING

F1 DATA STATS

Marta Díaz Artigot
Diciembre 2024

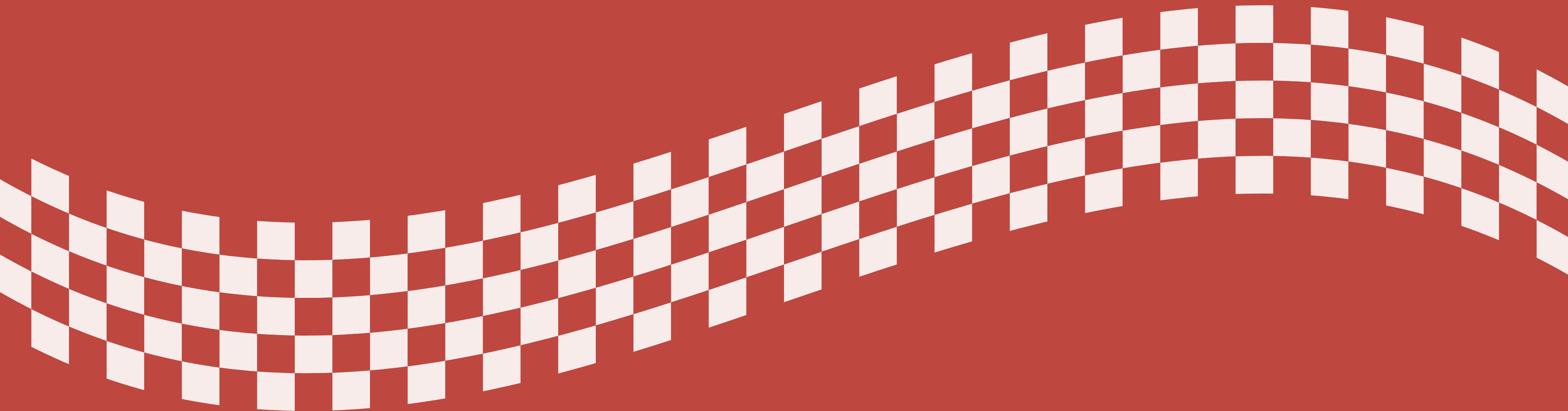




ETAPAS DEL PROYECTO

- 01 INGESTA
- 02 TRANSFORMACIÓN
 - BRONZE
 - SILVER
 - GOLD
- 03 ANÁLISIS

INGESTA



▲ 1763
New Notebook
Download

⋮

9 of 9 columns ▾

Reference name of driver

Valid	859	100%
Mismatched	0	0%
Missing	0	0%
Unique	859	
Most Common	hamilton	0%

Driver number

6	\N	93%	Valid <div><div></div></div>	859	100%
		0%	Mismatched <div><div></div></div>	0	0%
		0%	Missing <div><div></div></div>	0	0%
Other (55)		6%	Unique	47	
			Most Common	\N	93%

Version 23 (21.49 MB)

Lato

- ▶ 14 files
- ▶ 120 columns

KAGGLE

Q Search

DEV_ALUMNO14_PROYECTO_FINAL

- ALUMNO14_GOLD
- ALUMNO14_SILVER
- BRONZE
 - Tables
 - CIRCUITS
 - CONSTRUCTORS
 - CONSTRUCTOR_RESULTS
 - CONSTRUCTOR_STANDIN...
 - DRIVERS
 - DRIVER_STANDINGS
 - LAP_TIMES
 - PIT_STOPS
 - QUALIFYING
 - RACES
 - RESULTS
 - SPRINT_RESULTS
 - STATUS
 - Stages
 - ALUMNO14_STAGE

DEV_ALUMNO14_PROYECTO... / BRONZE / ALUMNO14_STAGE

Internal Stage CURSO_DATA_ENGINEERING 4 days ago

Stage Files Stage Details Lineage PREVIEW

ALUMNO14_STAGE (13 Files) Search WH_CURSO_DATA_ENGINEERING

NAME	SIZE	LAST MODIFIED ↓	
status.csv	2.1KB	4 days ago	...
results.csv	1.6MB	4 days ago	...
lap_times.csv	16.4MB	4 days ago	...
driver_standings.csv	856.0KB	4 days ago	...
sprint_results.csv	20.1KB	4 days ago	...
pit_stops.csv	418.5KB	4 days ago	...
constructor_results.csv	212.0KB	4 days ago	...
drivers.csv			
constructor_standings.csv			
circuits.csv			

Creo el esquema
bronze dentro
de mi base de
datos y el stage
para subir los
csv
←

Creo las tablas
de mi capa
bronze e
introduzco los
csv utilizando
COPY INTO

→

```
-- Tabla CONSTRUCTOR_RESULTS 2
CREATE OR REPLACE TABLE DEV_ALUMNO14_PROYECTO_FINAL.BRONZE.CONSTRUCTOR_RESULTS (
  constructorResultsId VARCHAR(250) NOT NULL,
  raceId VARCHAR(250),
  constructorId VARCHAR(250),
  points FLOAT,
  status VARCHAR(250),
  primary key (constructorResultsId)
);


COPY INTO DEV_ALUMNO14_PROYECTO_FINAL.BRONZE.CONSTRUCTOR_RESULTS
FROM @DEV_ALUMNO14_PROYECTO_FINAL.BRONZE.ALUMNO14_STAGE/constructor_results.csv
FILE_FORMAT = (
  TYPE = 'CSV',
  SKIP_HEADER = 1,
  FIELD_OPTIONALLY_ENCLOSED_BY = ''
)
ON_ERROR = 'CONTINUE';
```

circuits	
circuitId 	INT
circuitRef	VARCHAR
name	VARCHAR
location	VARCHAR
country	VARCHAR
lat	FLOAT
lng	FLOAT
alt	FLOAT
url	VARCHAR

results	
resultId 	INT
raceId	INT
driverId	INT
constructorId	INT
number	INT
grid	INT
position	INT
positionText	VARCHAR
positionOrder	INT
points	FLOAT
laps	INT
time	VARCHAR
milliseconds	INT
fastestLap	INT
rank	INT
fastestLapTime	VARCHAR
fastestLapSpeed	FLOAT
statusId	INT


constructors	
constructorId 	INT
constructorRef	VARCHAR
name	VARCHAR
nationality	VARCHAR
url	VARCHAR

sprint_results	
resultId 	INT
raceId	INT
driverId	INT
constructorId	INT
number	INT
grid	INT
position	INT
positionText	VARCHAR
positionOrder	INT
points	FLOAT
laps	INT
time	VARCHAR
milliseconds	INT
fastestLap	INT
fastestLapTime	VARCHAR
statusId	INT

status	
statusId 	INT
status	VARCHAR


drivers	
driverId 	INT
driverRef	VARCHAR
number	INT
code	VARCHAR
forename	VARCHAR
surname	VARCHAR
dob	DATE
nationality	VARCHAR
url	VARCHAR


lap_times	
raceId	INT
driverId	INT
lap	INT
position	INT
time	VARCHAR
milliseconds	INT

constructor_results	
constructorResultsId 	INT
raceId	INT
constructorId	INT
points	FLOAT
status	VARCHAR

races	
raceId 	INT
year	INT
round	INT
circuitId	INT
name	VARCHAR
date	DATE
time	TIME
url	VARCHAR
fp1_date	DATE
fp1_time	TIME
fp2_date	DATE
fp2_time	TIME
fp3_date	DATE
fp3_time	TIME
quali_date	DATE
quali_time	TIME
sprint_date	DATE
sprint_time	TIME

pit_stops	
raceId	INT
driverId	INT
stop	INT
lap	INT
time	VARCHAR
duration	FLOAT
milliseconds	INT

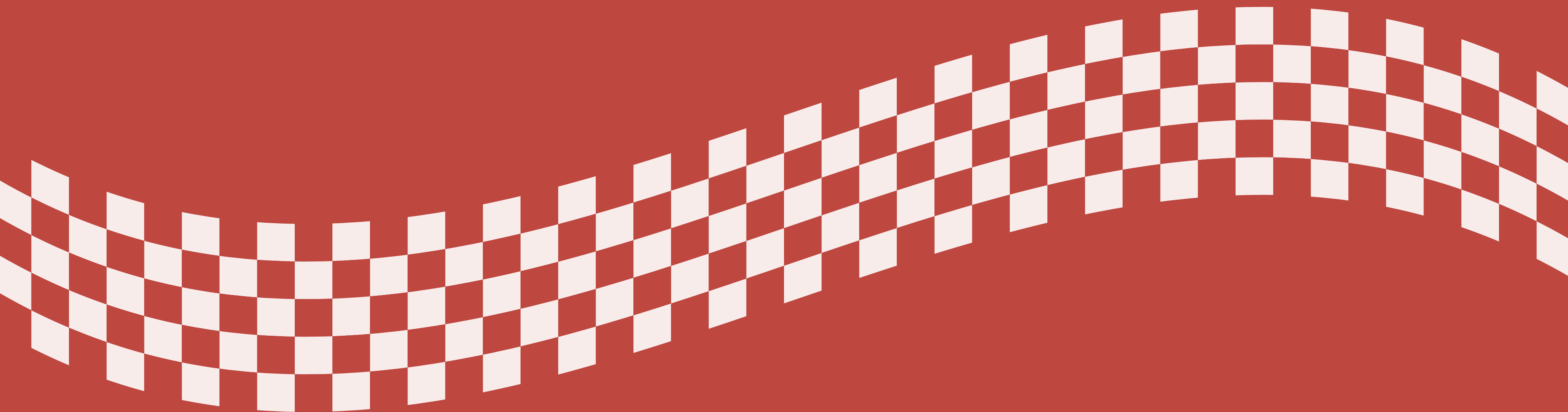
constructor_standings	
constructorStandingsId 	INT
raceId	INT
constructorId	INT
points	FLOAT
position	INT
positionText	VARCHAR
wins	INT

driver_standings	
driverStandingsId 	INT
raceId	INT
driverId	INT
points	FLOAT
position	INT
positionText	VARCHAR
wins	INT

qualifying	
qualifyId 	INT
raceId	INT
driverId	INT
constructorId	INT
number	INT
position	INT
q1	VARCHAR
q2	VARCHAR
q3	VARCHAR

TRANSFORMACIÓN

CAPA BRONZE



TEST

Ejecutados en el fichero de sources para comprobar que los datos son correctos antes de llevarlos a la capa silver haciendo las tranformaciones necesarias.

1

NOT NULL

- Viene por defecto en DBT
- Comprueba que no haya valores nulos en una columna
- 7 columnas no lo pasaron

2

RELATIONSHIPS

- Viene por defecto en DBT
- Comprueba la integridad referencial entre dos tablas.
- Todas las columnas lo pasaron

3

POSITIVE_VALUES

- Macro creada para comprobar que los valores de una columna sean positivos.
- Todas las columnas lo pasaron.

4

UNIQUE

- Viene por defecto en DBT
- Comprueba que los valores de una columna sean únicos
- Todas las columnas lo pasaron



Version control

5

Commit and sync



Changes

dim_date.sql

A

File explorer



marts

M

staging

M

bronze

M

_bronze__models.yml

_bronze__sources.yml

stg_bronze__circuits.sql

stg_bronze__constructor_resu...

stg_bronze__constructor_stan...

stg_bronze__constructors.sql

stg_bronze__driver_standings...

stg_bronze__drivers.sql

stg_bronze__lap_times.sql

M

stg_bronze__pit_stops.sql

M

stg_bronze__qualifying.sql

stg_bronze__races.sql

stg_bronze__results.sql

stg_bronze__sprint_results.sql

stg_bronze__status.sql

_bronze__sources.yml



models / staging / bronze / _bronze__sources.yml

Save

```
1 version: 2
2
3 sources:
4   - name: bronze
5     schema: BRONZE
6     database: DEV_ALUMNO14_PROYECTO_FINAL
7
8   quoting:
9     database: false
10    schema: false
11    identifier: false
12
13   tables:
14     Generate model
15     - name: circuits
16       description: "Details of F1 circuits."
17       columns:
18         - name: circuitId
19           description: "Unique ID for each circuit."
20           tests:
21             - unique
22             - not_null
23         - name: circuitRef
24           description: "Reference name of the circuit."
25           tests:
26             - unique
27             - not_null
28         - name: name
29           description: "Full name of the circuit."
30           tests:
31             - unique
32             - not_null
```

All

148

Pass

141

Warn

0

Error

7

Skip

0

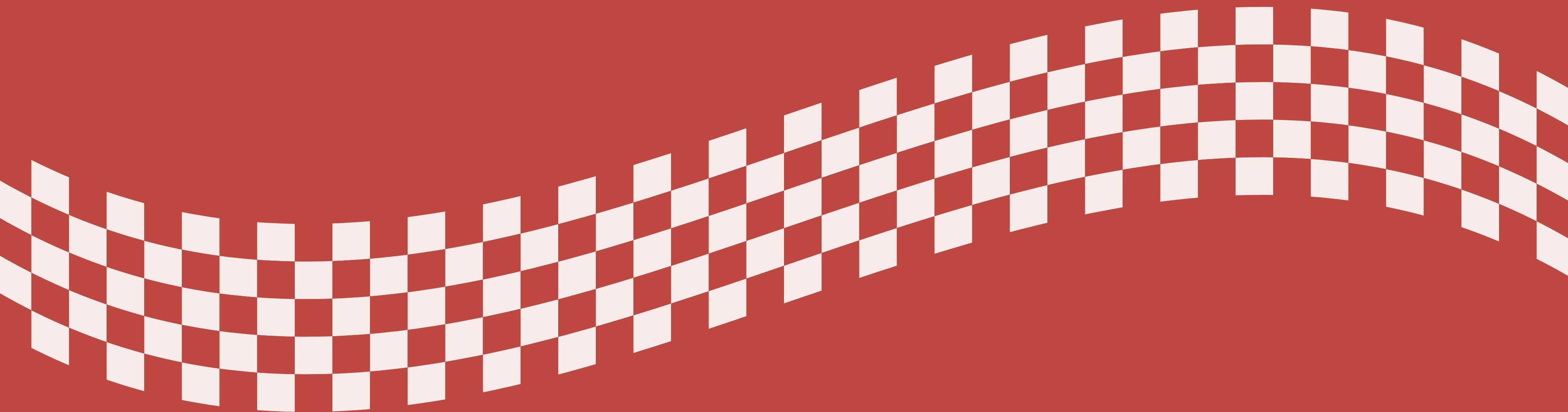
Running

0

>	✖ source_not_null_bronze_qualifying_q1	0.23s
>	✖ source_not_null_bronze_qualifying_q2	0.20s
>	✖ source_not_null_bronze_qualifying_q3	0.18s
>	✖ source_not_null_bronze_races_time	0.21s
>	✖ source_not_null_bronze_results_number	0.25s
>	✖ source_not_null_bronze_results_position	0.19s
>	✖ source_not_null_bronze_sprint_results_position	0.24s

TRANSFORMACIÓN

CAPA SILVER



Transformaciones

Transformaciones básicas aplicadas en los modelos para limpiar los datos y adecuarlos a la primera forma normal (1FN).



1

Crear claves primarias en lap_times y pit_stops utilizando generate_surrogate_key de dbt_utils

2

Eliminar columnas con un alto porcentaje de valores nulos o que no aportan información útil

3

Cambiar nombre de columnas por otros más explicativos

4

Cambiar valores nulos por cadenas vacías en columnas tipo texto y por código numérico en columnas tipo numérico

FICHEROS

📁 staging M

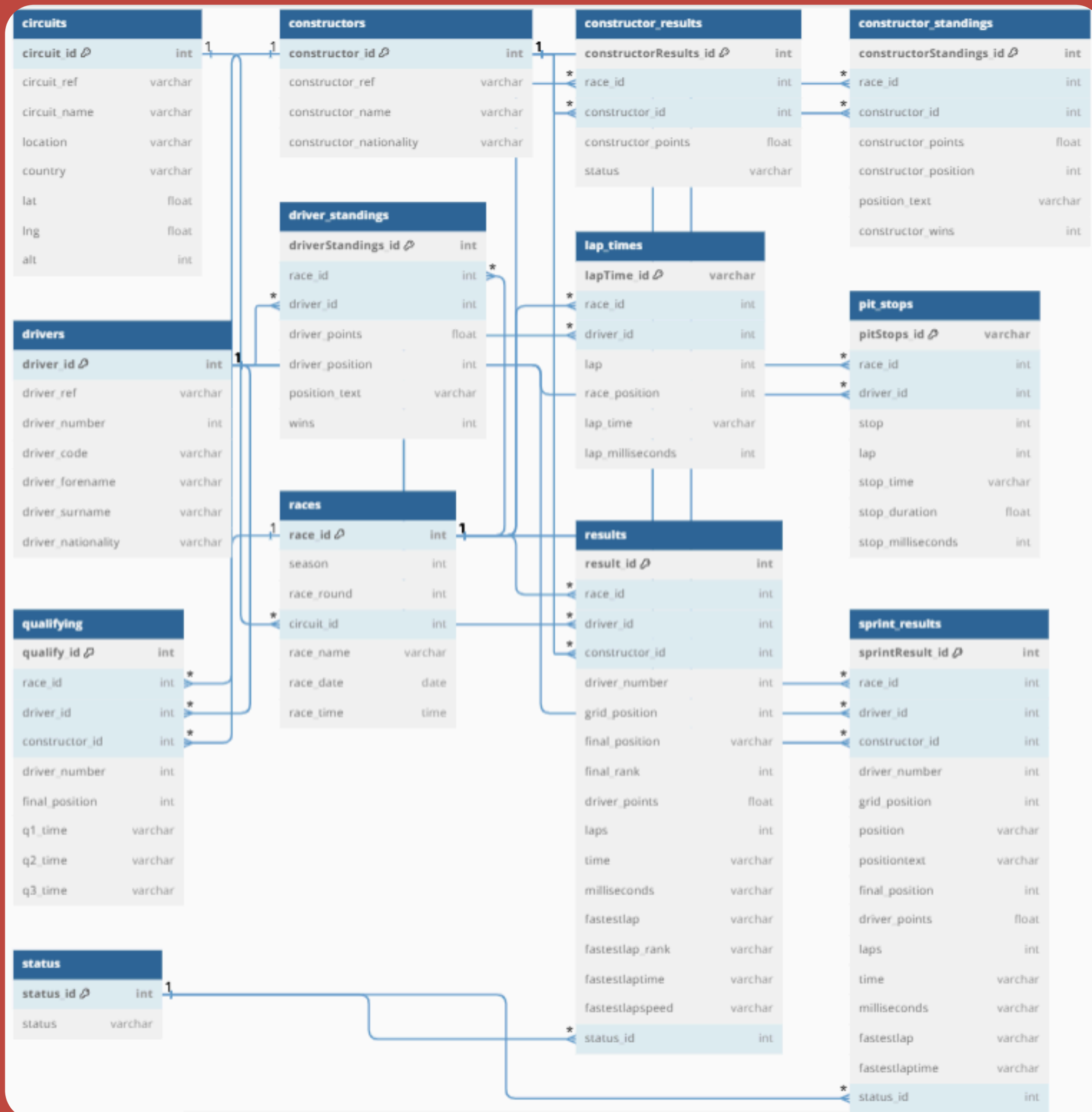
📁 bronze M

- 📄 _bronze__models.yml
- 📄 _bronze__sources.yml
- 📦 stg_bronze__circuits.sql
- 📦 stg_bronze__constructor_results.sql
- 📦 stg_bronze__constructor_standings.sql
- 📦 stg_bronze__constructors.sql
- 📦 stg_bronze__driver_standings.sql
- 📦 stg_bronze__drivers.sql
- 📦 stg_bronze__lap_times.sql M
- 📦 stg_bronze__pit_stops.sql M
- 📦 stg_bronze__qualifying.sql
- 📦 stg_bronze__races.sql
- 📦 stg_bronze__results.sql
- 📦 stg_bronze__sprint_results.sql
- 📦 stg_bronze__status.sql

EJEMPLOS DE TRANSFORMACIÓN

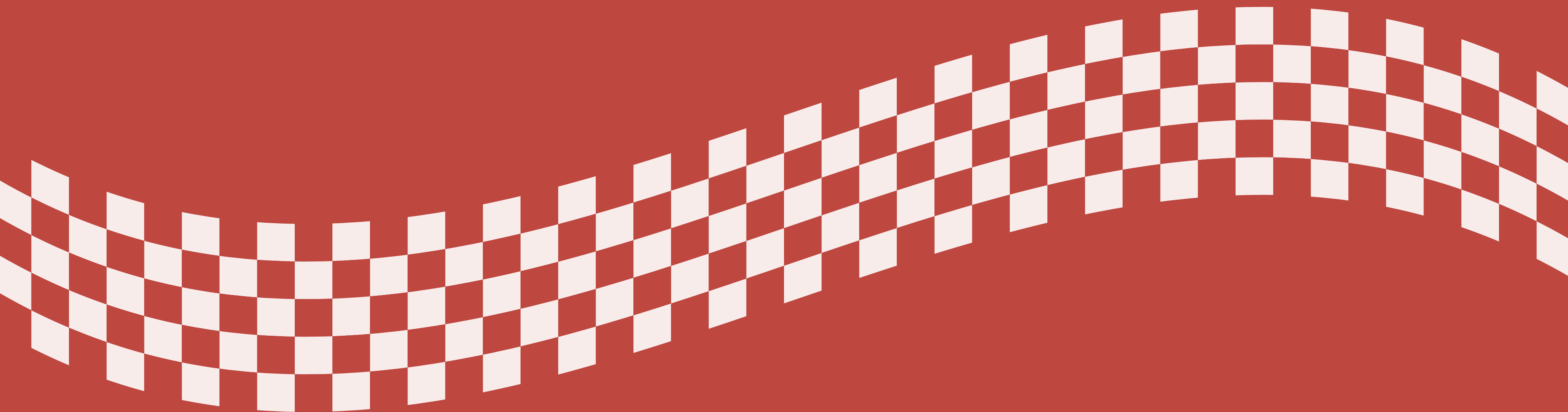
```
select
    driverid AS driver_id,
    driverref AS driver_ref,
    COALESCE(number, 111) AS driver_number,
    COALESCE(code, '') AS driver_code,
    forename AS driver_forename,
    surname AS driver_surname,
    nationality AS driver_nationality
from source
```

```
select
    {{ dbt_utils.generate_surrogate_key(['raceid', 'driverid', 'lap']) }} AS lapTime_id,
    raceid AS race_id,
    driverid AS driver_id,
    lap,
    position AS race_position,
    time AS lap_time,
    milliseconds AS lap_miliseconds
from source
```



TRANSFORMACIÓN

CAPA GOLD



MONTH_NAME	MONTH_NAME_SHORT	MONTH_START_DATE
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01
January	Jan	1950-01-01

Creación DIM_DATE

Creamos una tabla llamada `dim_fecha` que va a aportarnos datos extra sobre las fechas de nuestras tablas de hechos que no tenemos

Para crearla utilizamos el paquete `dbt_date` y creamos el modelo:

```
models / marts / dim_date.sql

1  {{
2    |    config(
3    |        materialized = "table"
4    |    )
5  }}
6
7  {{ dbt_date.get_date_dimension("1950-01-01", "2050-12-31") }}
8
9
```


FICHEROS

📁 models

📁 marts

📁 constructor_stats

📦 dim_constructors.sql

📦 fct_constructor_standings.sql

📁 core

📦 dim_date.sql

📦 dim_race_circuit.sql

📦 fct_qualifying_results.sql

📦 fct_results.sql

📦 fct_sprint_results.sql

📁 driver_stats

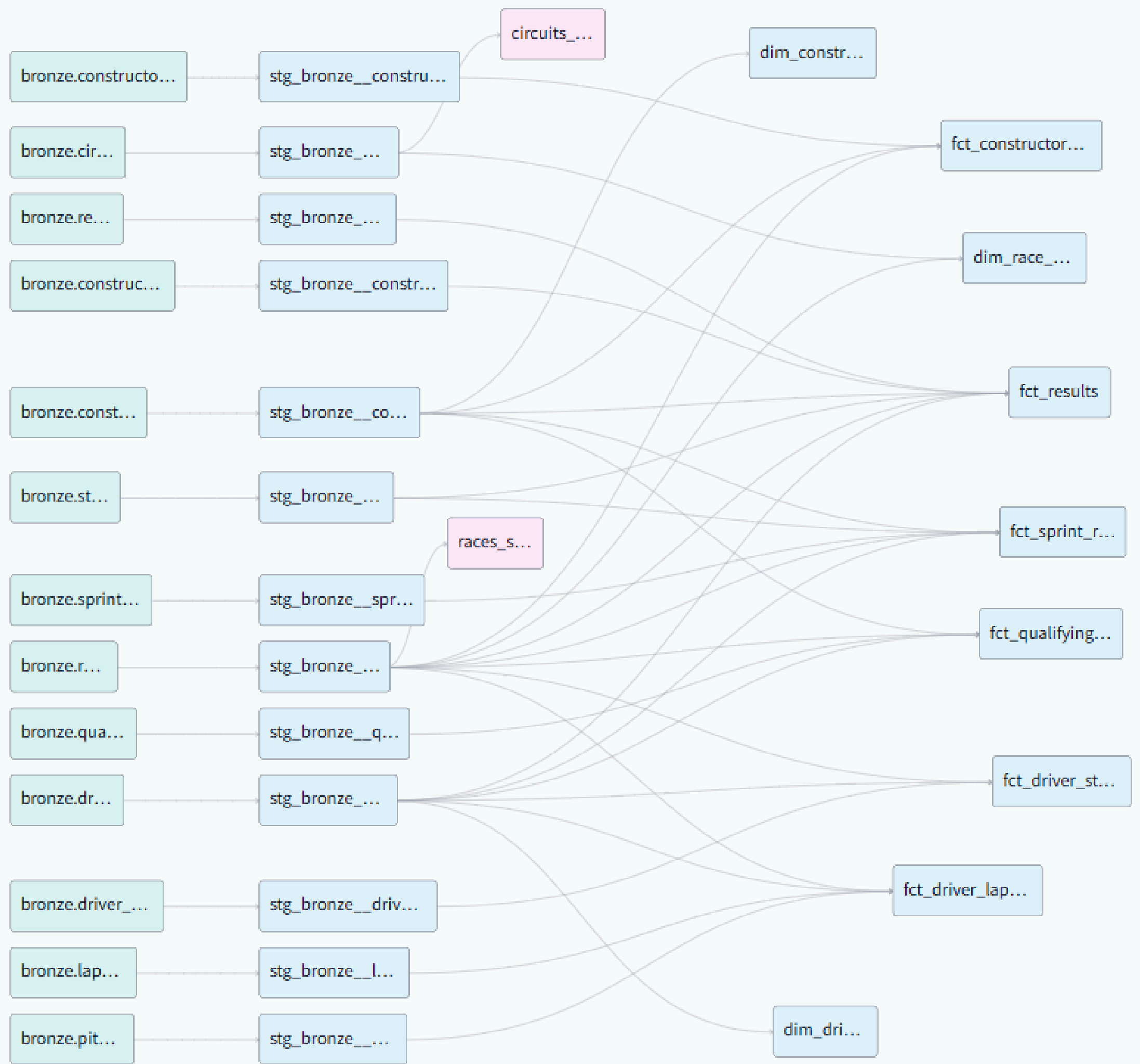
📦 dim_drivers.sql

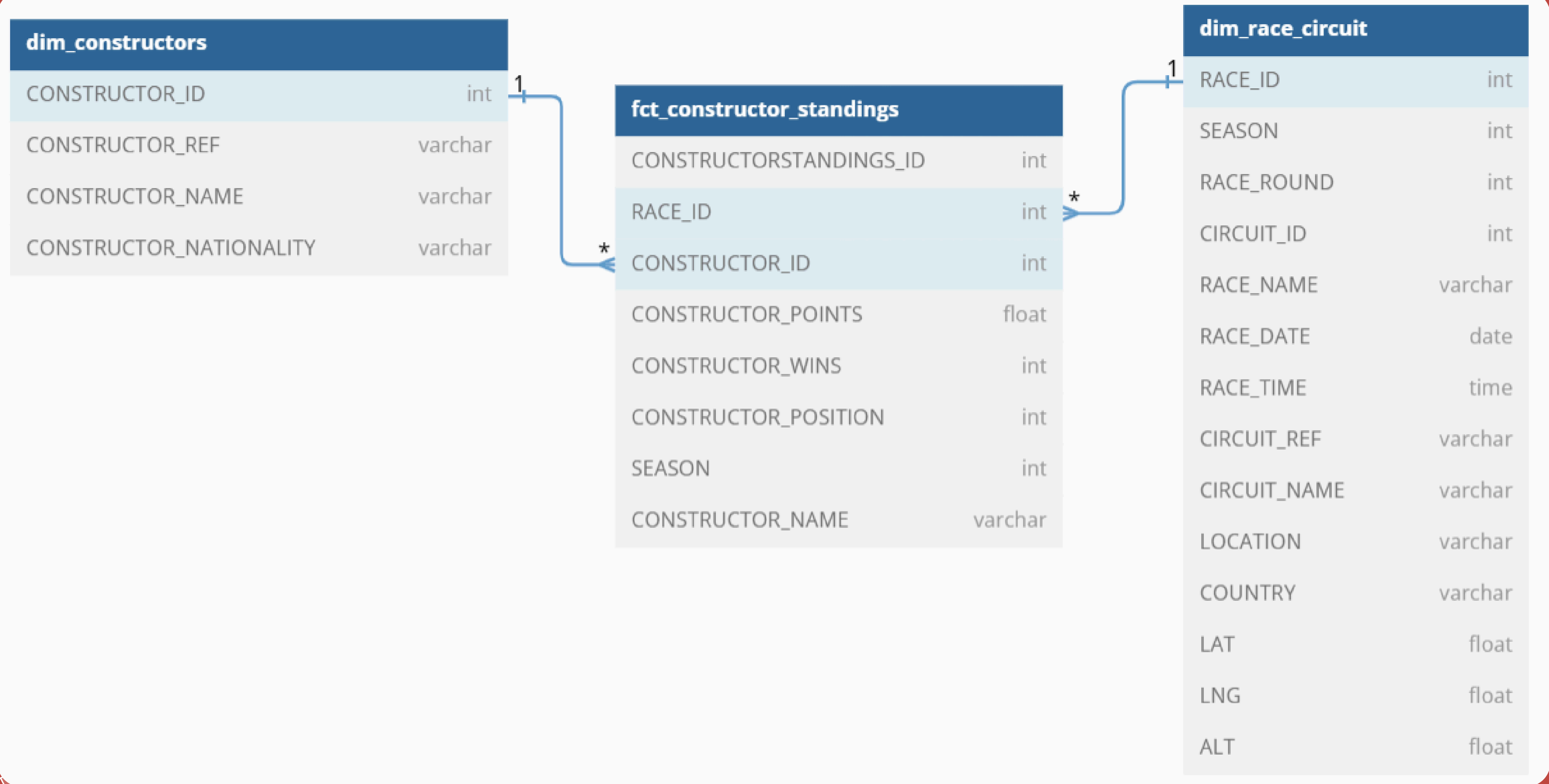
📦 fct_driver_lap_analysis.sql

📦 fct_driver_standings.sql

📁 staging

📁 bronze

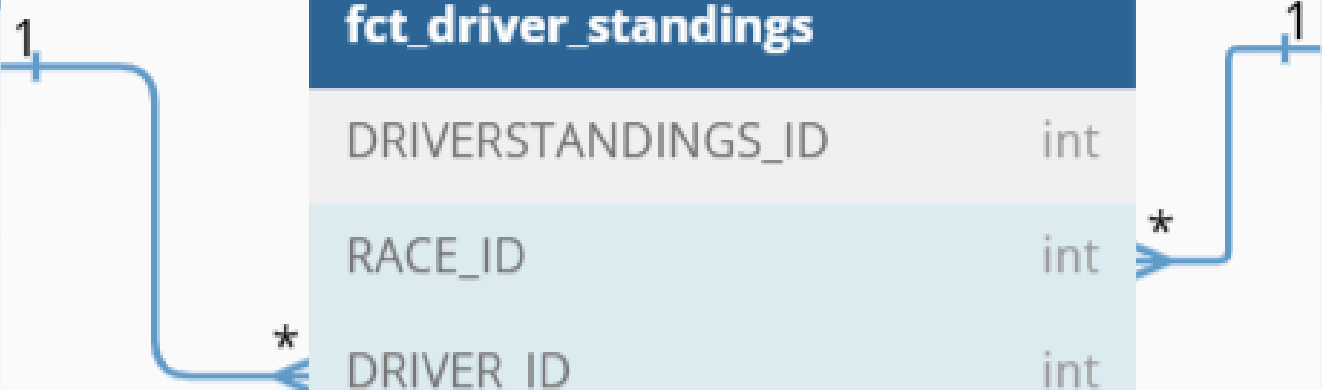




dim_drivers	
DRIVER_ID	int
DRIVER_REF	varchar
DRIVER_NUMBER	int
DRIVER_CODE	varchar
DRIVER_FORENAME	varchar
DRIVER_SURNAME	varchar
DRIVER_NATIONALITY	varchar

fct_driver_standings	
DRIVERSTANDINGS_ID	int
RACE_ID	int
DRIVER_ID	int
DRIVER_POINTS	float
WINS	int
DRIVER_POSITION	int
SEASON	int

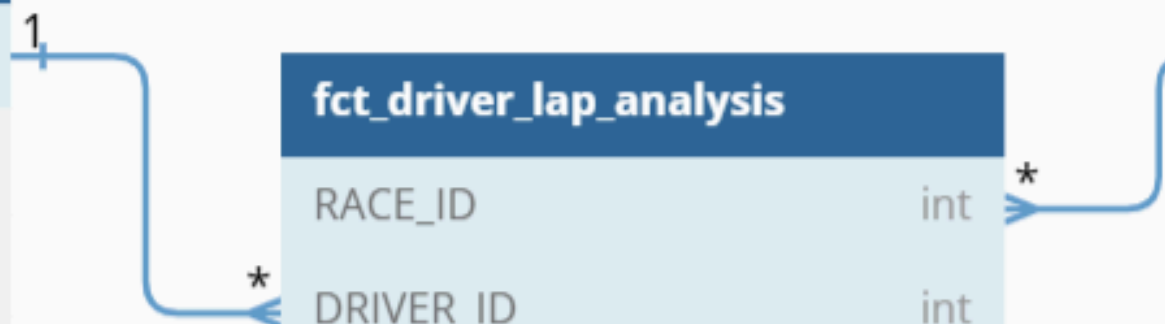
dim_race_circuit	
RACE_ID	int
SEASON	int
RACE_ROUND	int
CIRCUIT_ID	int
RACE_NAME	varchar
RACE_DATE	date
RACE_TIME	time
CIRCUIT_REF	varchar
CIRCUIT_NAME	varchar
LOCATION	varchar
COUNTRY	varchar
LAT	float
LNG	float
ALT	float



dim_drivers	
DRIVER_ID	int
DRIVER_REF	varchar
DRIVER_NUMBER	int
DRIVER_CODE	varchar
DRIVER_FORENAME	varchar
DRIVER_SURNAME	varchar
DRIVER_NATIONALITY	varchar

fct_driver_lap_analysis	
RACE_ID	int
DRIVER_ID	int
LAP	int
LAP_MILLISECONDS	int
STOP_MILLISECONDS	int
PIT_STOP_NUMBER	int
SEASON	int
DRIVER_NAME	varchar

dim_race_circuit	
RACE_ID	int
SEASON	int
RACE_ROUND	int
CIRCUIT_ID	int
RACE_NAME	varchar
RACE_DATE	date
RACE_TIME	time
CIRCUIT_REF	varchar
CIRCUIT_NAME	varchar
LOCATION	varchar
COUNTRY	varchar
LAT	float
LNG	float
ALT	float

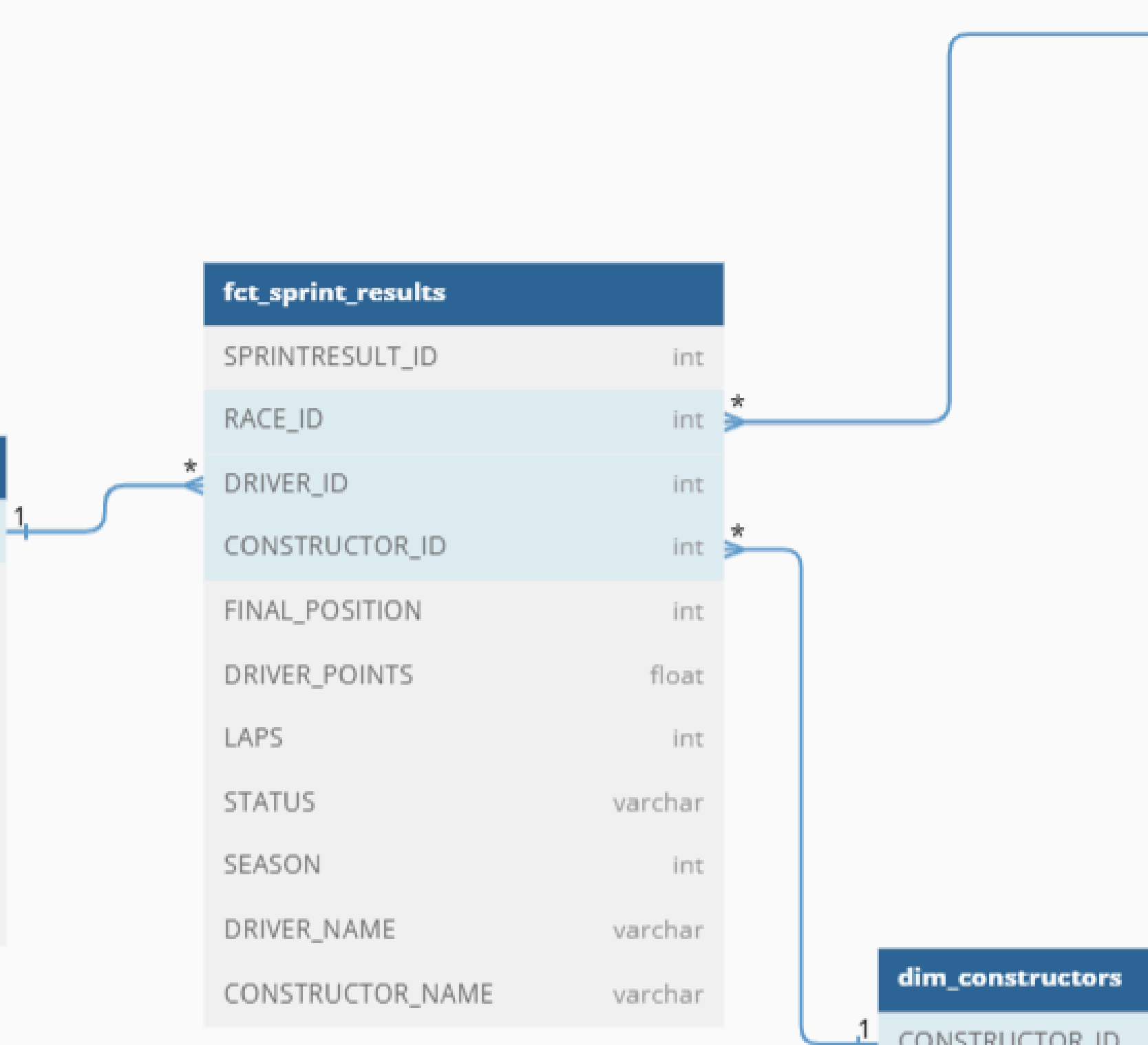


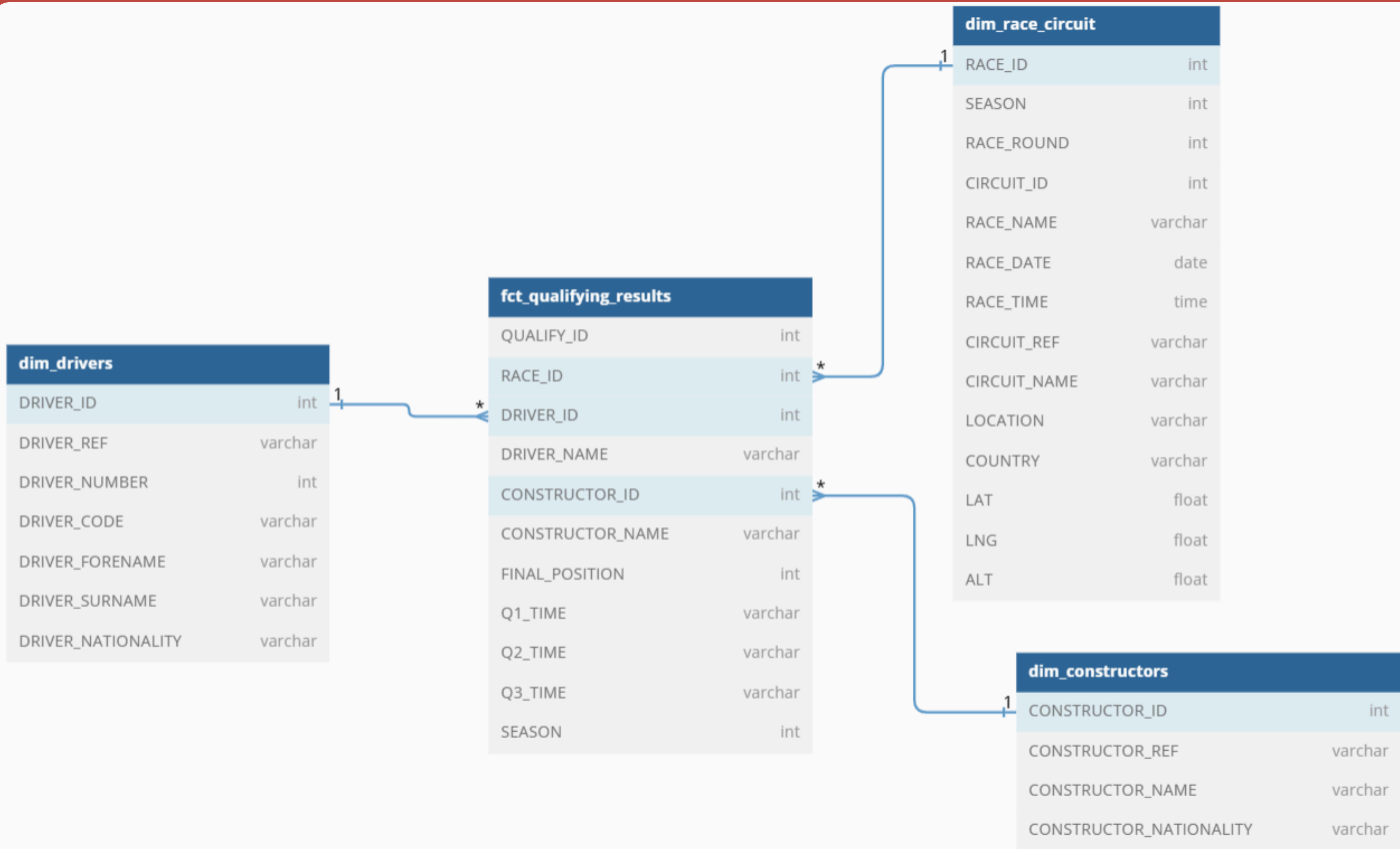
dim_drivers	
DRIVER_ID	int
DRIVER_REF	varchar
DRIVER_NUMBER	int
DRIVER_CODE	varchar
DRIVER_FORENAME	varchar
DRIVER_SURNAME	varchar
DRIVER_NATIONALITY	varchar

fct_sprint_results	
SPRINTRESULT_ID	int
RACE_ID	int
DRIVER_ID	int
CONSTRUCTOR_ID	int
FINAL_POSITION	int
DRIVER_POINTS	float
LAPS	int
STATUS	varchar
SEASON	int
DRIVER_NAME	varchar
CONSTRUCTOR_NAME	varchar

dim_constructors	
CONSTRUCTOR_ID	int
CONSTRUCTOR_REF	varchar
CONSTRUCTOR_NAME	varchar
CONSTRUCTOR_NATIONALITY	varchar

dim_race_circuit	
RACE_ID	int
SEASON	int
RACE_ROUND	int
CIRCUIT_ID	int
RACE_NAME	varchar
RACE_DATE	date
RACE_TIME	time
CIRCUIT_REF	varchar
CIRCUIT_NAME	varchar
LOCATION	varchar
COUNTRY	varchar
LAT	float
LNG	float
ALT	float





dim_constructors	
CONSTRUCTOR_ID	int
CONSTRUCTOR_REF	varchar
CONSTRUCTOR_NAME	varchar
CONSTRUCTOR_NATIONALITY	varchar

fct_results	
RESULT_ID	int
RACE_ID	int
DRIVER_ID	int
DRIVER_NAME	varchar
CONSTRUCTOR_ID	int
CONSTRUCTOR_NAME	varchar
GRID_POSITION	int
FINAL_POSITION	int
DRIVER_POINTS	float
CONSTRUCTOR_POINTS	float
MILLISECONDS	float
FASTESTLAP	float
FASTESTLAP_RANK	float
FASTESTLAPSPEED	float
FASTESTLAPTIME	float
STATUS	varchar
SEASON	int

dim_race_circuit	
RACE_ID	int
SEASON	int
RACE_ROUND	int
CIRCUIT_ID	int
RACE_NAME	varchar
RACE_DATE	date
RACE_TIME	time
CIRCUIT_REF	varchar
CIRCUIT_NAME	varchar
LOCATION	varchar
COUNTRY	varchar
LAT	float
LNG	float
ALT	float

dim_drivers	
DRIVER_ID	int
DRIVER_REF	varchar
DRIVER_NUMBER	int
DRIVER_CODE	varchar
DRIVER_FORENAME	varchar
DRIVER_SURNAME	varchar
DRIVER_NATIONALITY	varchar

