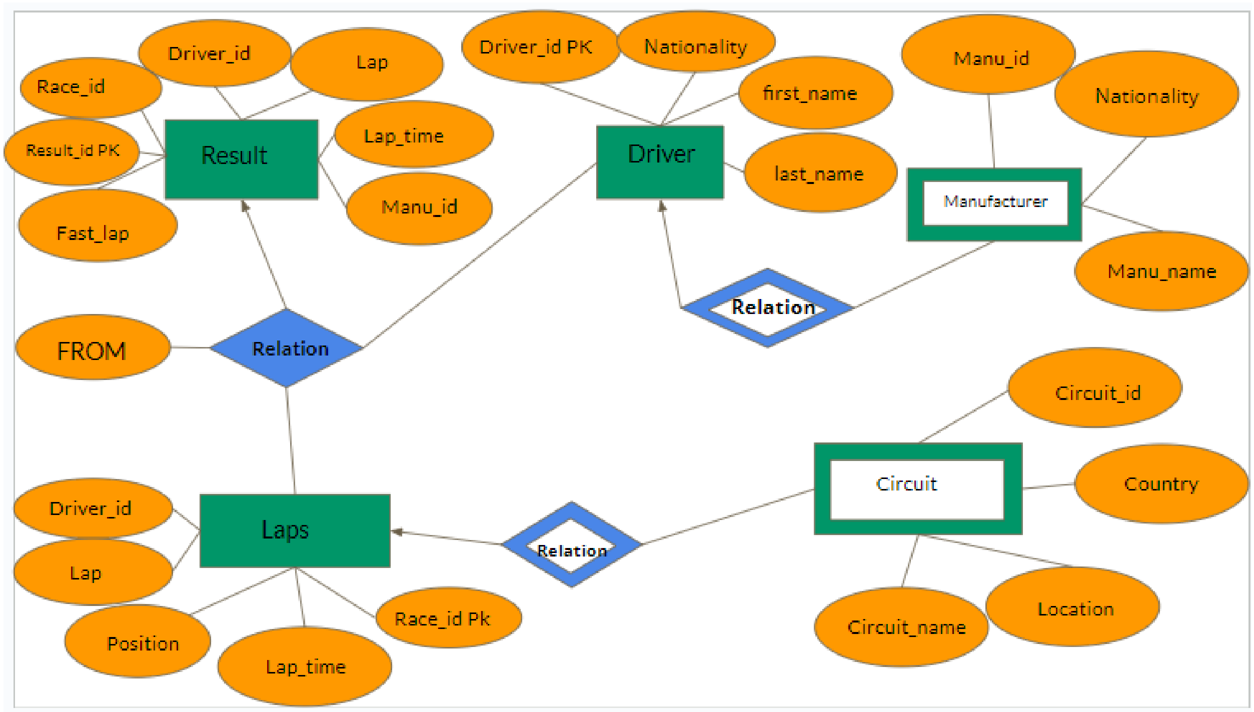
Database Project Report - Formula 1 racing

Introduction: Our project consists of collecting data for the Formula 1 racing series, or simply F1 racing. F1 racing is a racing series along with NASCAR, Red Bull racing, and NHRA, with events hosted all over the world. F1's uniqueness derives primarily from high speeds, fast laps, high performance, and track racing complexity. The main questions we wanted to be answered are:

1) Which car manufacturer produced the fastest laps?
2) Which circuit produced the fastest lap time?
3) What driver won the most races?

These questions are significant because we felt strongly that they greatly represent the reputation of F1 racing, how this type of racing appeals to the audience, and how F1 really stands out in competition with other racing events.

Methods/Approach: Our E/R diagram consists of 5 entity sets, each that have 3-7 attributes, as well as 3 relationships.



The Result entity set has the following attributes that are related to the driver's racing performance for a particular race:

- Fast_lap
- Result_id PK

- Race_id
- Driver_id
- Lap
- Lap_time

The Driver entity set has the following attributes that are related to the driver:

- Driver_id PK
- Nationality
- First_name
- last_name

The manufacturer entity set has the following attributes that are related to the car manufacturer and its origin:

- Manu_id
- Nationality
- Manu_name

The circuit entity set has the following attributes that are related to the racing location:

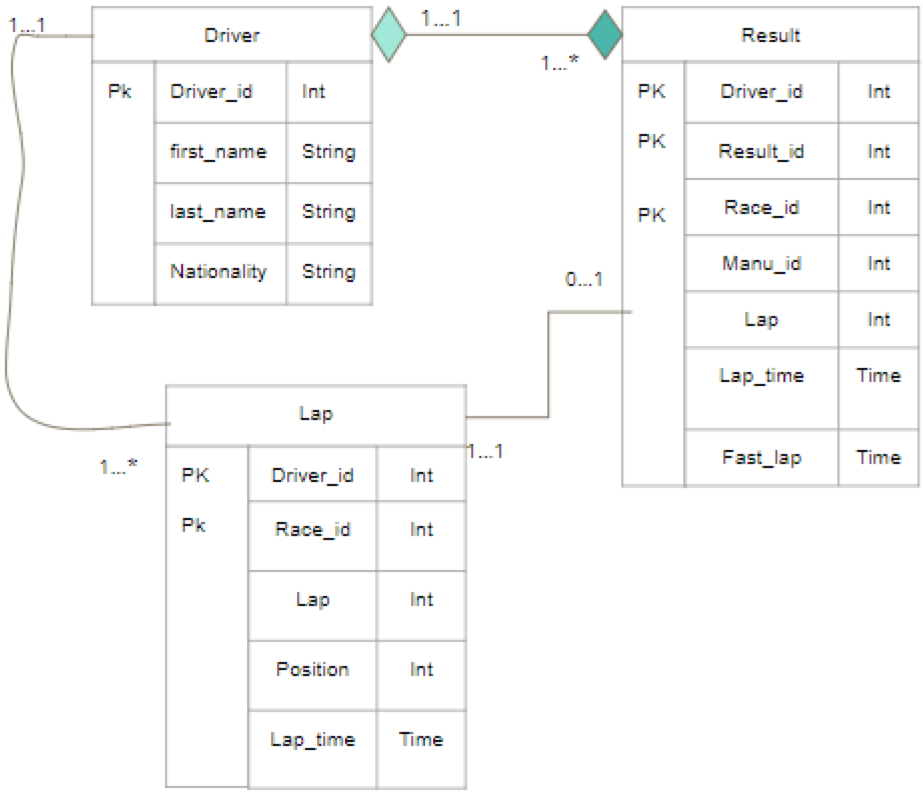
- Circuit_id
- Country
- Location
- Circuit_name

The Laps entity set has the following attributes that are related to the racing laps themselves with the drivers that were involved in a particular racing event:

- Driver_id
- Lap
- Position
- Lap_time
- Race_id PK

There is a relationship between Laps and Circuit because the Lap information couldn’t be possible without knowing the location that the laps took place at. A relationship exists between Laps and Result because there needs to be a way to present the fastest lap time performed by some driver under a car manufacturer, and also for a way to answer one of our key questions. A relationship exists between Driver and Manufacturer because we need to determine the manufacturer of a car that the driver used to participate in a race.

Database schema



Our database schema consists of 3 main tables: Driver, Result, and Lap. From the figure of our schema you can see that we have a lot of possible constraints and relationships going on here. The diamonds between the Driver and Result show aggregation and exclusive ownership taking place. The relationship between the driver and the lap shows that there can only be one driver to a set of lap records, where the lap can only have many drivers. It can also be said the same with lap and result relations. It shows that the result can only have one set of records of each lap but there can be many laps but only one record that is going to the result. Next up is the relation between Driver and Result. The figure shows that Result must have a Driver to show the driver results. The result can only have one set of records for the driver but the driver can have one or many results.

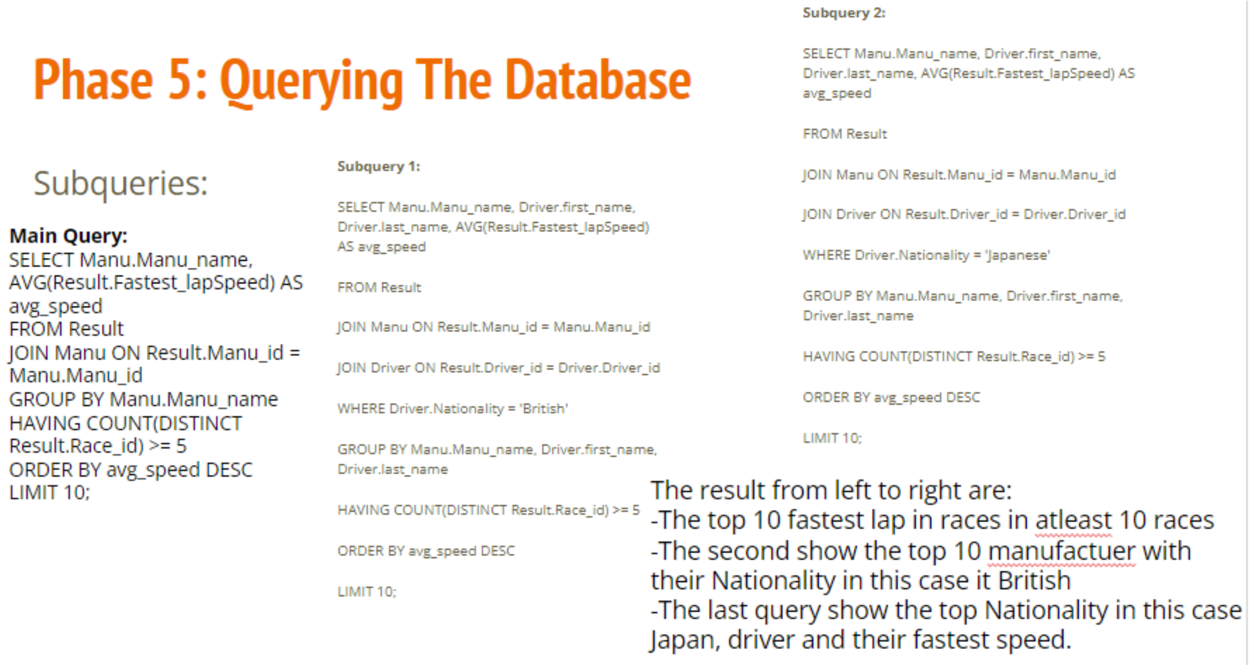
The data normalization consists of a functional dependency that exists between the Result_id and all others, as well as multivariable dependencies that exist between:

- Result_id, Driver_id to all others
- Race_id, Driver_id to Laps
- Driver_id, Nationality to Manufacturer

There were normal form violations for our Result, Driver, and Lap tables that needed to be addressed, but among all the tables, the most common form had no non-prime key through BCNF. Our Result table's

3NF had no transitive dependency, but the 4NF contained a multivariable dependency, which was Driver_id, Race_id, Result_id to Lap and Lap_time. Our driver table also had no transitive dependency for 3NF and had no multivariable dependency for 4NF. Our Lap table's 4NF has no multivariable dependency, but the 3NF has a transitive dependency which is Race_id, Lap, Position, and Lap_time is dependent on Driver_id.

Implementation: Our implementation was performed using Kaggle for data collection, MySQL workbench for SQL coding, and Microsoft Excel for data formatting. We also utilized a website where we had to convert CSV files into Insert SQL. The querying process involves 3 main queries: top 10 fastest laps, top 10 manufacturers with their respective nationality, and the top nationality with the fastest speed.



Phase 5: Querying The Database

Subqueries:

Main Query:

```
SELECT Manu.Manu_name,
AVG(Result.Fastest_lapSpeed) AS
avg_speed
FROM Result
JOIN Manu ON Result.Manu_id =
Manu.Manu_id
GROUP BY Manu.Manu_name
HAVING COUNT(DISTINCT
Result.Race_id) >= 5
ORDER BY avg_speed DESC
LIMIT 10;
```

Subquery 1:

```
SELECT Manu.Manu_name, Driver.first_name,
Driver.last_name, AVG(Result.Fastest_lapSpeed)
AS avg_speed

FROM Result

JOIN Manu ON Result.Manu_id = Manu.Manu_id

JOIN Driver ON Result.Driver_id = Driver.Driver_id

WHERE Driver.Nationality = 'British'

GROUP BY Manu.Manu_name, Driver.first_name,
Driver.last_name

HAVING COUNT(DISTINCT Result.Race_id) >= 5

ORDER BY avg_speed DESC

LIMIT 10;
```

Subquery 2:

```
SELECT Manu.Manu_name, Driver.first_name,
Driver.last_name, AVG(Result.Fastest_lapSpeed) AS
avg_speed

FROM Result

JOIN Manu ON Result.Manu_id = Manu.Manu_id

JOIN Driver ON Result.Driver_id = Driver.Driver_id

WHERE Driver.Nationality = 'Japanese'

GROUP BY Manu.Manu_name, Driver.first_name,
Driver.last_name

HAVING COUNT(DISTINCT Result.Race_id) >= 5

ORDER BY avg_speed DESC

LIMIT 10;
```

The result from left to right are:
-The top 10 fastest lap in races in atleast 10 races
-The second show the top 10 manufactuer with their Nationality in this case it British
-The last query show the top Nationality in this case Japan, driver and their fastest speed.

For demonstration, we will insert values into our Circuit table with the SQL commands:

INSERT INTO Circuit (Circuit_id, Circuit_name, Location, Country)

VALUES (6, 'Monza', 'Monza, Italy', 'Italy');

These commands will insert the following values in order of our Circuit() function: circuit ID, circuit name, location, and country of origin.

As for updating, here is an example of how we work with the syntax:

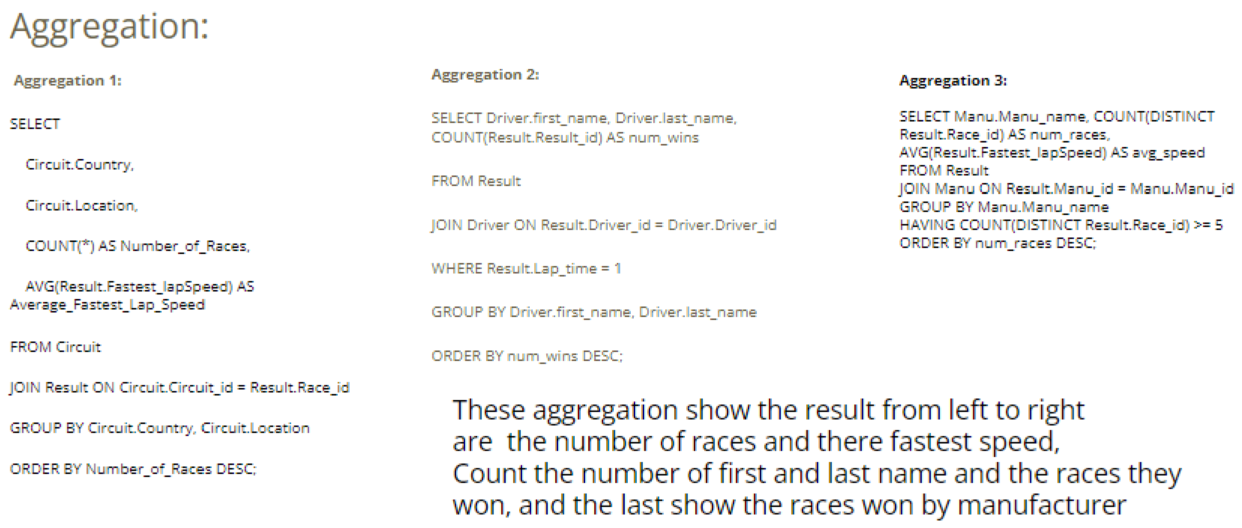
UPDATE Circuit

SET Location = 'Los Angeles', Country = 'United State'

WHERE Circuit_id = 300;

These commands will go to the circuit table and update the location set to “Los Angeles”, country to “United States”, for a racing circuit with the ID of 300.

Our aggregation process involves gathering the number of races with the fastest speed recorded, next we obtain the first and last names of the drivers and take the number of races won, and finally we display the respective manufacturer for those races won.



Aggregation:

Aggregation 1:

```
SELECT

   Circuit.Country,

   Circuit.Location,

   COUNT(*) AS Number_of_Races,

   AVG(Result.Fastest_lapSpeed) AS
Average_Fastest_Lap_Speed

FROM Circuit

JOIN Result ON Circuit.Circuit_id = Result.Race_id

GROUP BY Circuit.Country, Circuit.Location

ORDER BY Number_of_Races DESC;
```

Aggregation 2:

```
SELECT Driver.first_name, Driver.last_name,
COUNT(Result.Result_id) AS num_wins

FROM Result

JOIN Driver ON Result.Driver_id = Driver.Driver_id

WHERE Result.Lap_time = 1

GROUP BY Driver.first_name, Driver.last_name

ORDER BY num_wins DESC;
```

Aggregation 3:

```
SELECT Manu.Manu_name, COUNT(DISTINCT
Result.Race_id) AS num_races,
AVG(Result.Fastest_lapSpeed) AS avg_speed
FROM Result
JOIN Manu ON Result.Manu_id = Manu.Manu_id
GROUP BY Manu.Manu_name
HAVING COUNT(DISTINCT Result.Race_id) >= 5
ORDER BY num_races DESC;
```

These aggregation show the result from left to right are the number of races and there fastest speed, Count the number of first and last name and the races they won, and the last show the races won by manufacturer

Results: Based on the CSV files that were fed into our database and the results that we’ve obtained, we can proudly say that all of our questions that we wanted from the very beginning were successfully answered. The manufacturer with the fastest lap speed was BMW Sauber with an average speed of 202.08 mph. The driver winning the most races was Sergio Perez with 47 wins. The circuit that produced the fastest lap time was Ontario, Canada with 22 races held and having the fastest average lap speed recorded at 235.62 mph.

Conclusion: In summary, our end goal was to discover some basic insight about the racing statistics of the Formula 1 racing series and were successful in utilizing various database tools to find our answers. Although we have accomplished our end goal, this project could be expanded outside of the scope of F1 racing. To make this project more interesting, there are some aspects to add that could emphasize the mission of determining which racing event series appeals to the race car community the most. For instance, we could do something along the lines of obtaining ratings from the racing audience among F1, NASCAR, NHRA, and Red Bull racing. We could also apply similar ideas that we had from the very beginning to other racing event competitors that exist.