University of Manitoba

Ergast F1 Dataset

Includes: Part 3 Report/Write Up

Owen Ostermann

Martin Popper

Christian Trites

Fillip Karamanov

COMP 3380 A01

Group 27

Anifat Olawoyin

February 28, 2023

We decided as a group to use the ErgastF1-Dataset since we all have a shared interest in F1 and we want to acknowledge Relational Dataset Repository - ErgastF1 for the data set that we used. The data consists of:

- 544,056 rows
- 98 columns
- 31,313 instances
- 60.4 MB file size

This database keeps track of Formula 1 races from the 1950s to current and includes data such as drivers and their lap times, qualifications, constructors/teams and their standings, pit stops, circuit locations, and more. In the tables, they separated the teams, drivers, and their standings. They did this so they could keep track of them all separately since a driver may change teams midway through the season or a driver may be in position 1 but the team as a whole may not be. They also created tables for pit-stop, circuit, results, qualifying, and seasons to try and use qualifying data and other information prior to a race to predict the outcome of a race. A difficulty we faced while designing the EER model was figuring out the participation and cardinality of the 3-way relationships, specifically, the cardinality for drivers stopping at a pit stop during a race. This can be modeled in multiple ways, but we decided to model it as a particular driver stopping for multiple pits during a single race.

The data set that we used did fit quite cleanly into a relational model since the creators designed it for use within a relational model. This made it easy to follow through from our design in part one to our implementation in part two. We, unfortunately, did not receive any feedback on our part one submission which made it difficult to

decide if there was anything we needed to change/fix for part two. A regret we have about part one that would have made part two a bit easier is since the data set is quite large, it did contain some data that was a bit unnecessary like the q1, q2, and q3 in the ‘Qualify’ table. It would have made it easy not to include this data since it did need to be converted due to impedance mismatching and it didn’t have a huge impact on the rest of the data. We decided not to include the ‘status’ table from the original data set for this same reason and we do wish we applied this logic to some other areas.

The data could likely be modeled in other ways like a graph database but, since it was collected and designed with the idea that it would be used within a relational database, this made it relatively easy to implement and would likely have caused more work to be converted for use in a different model. A relational database isn’t needed but, it would certainly be more work to rework the design to fit another model. I do believe this data set could be a good teaching tool for later courses but, a caution to be aware of is that the row count and the number of tables are both well above the recommended amount, which means there is a lot of data that may need to be ‘cleaned’ and sorted through depending on the system being used and if there are any changes students would want to make. As well, many people are unfamiliar with F1, and how the scoring system works which can be confusing especially since the driver’s and constructor’s points are scored differently, making it a bit of an unfriendly teaching example.

The queries that can be run are as follows:

1. Select the number of wins for each driver and order them by the number of wins. This will return the names of the drivers and the number of wins they have in descending order
2. Select the results of the race with id 200 and order them by the position of the drivers. This will return the results of the race in order of the positions with names, IDs, and nationality of the drivers and the name of the constructors.
3. Select the number of races held in each country and order them by the number of races. This will return a list of countries along with the number of races that have taken place within that country, displayed in descending order.
4. Get the number of wins and total points from each driver. This will return a list of driver ids, names, a count of each driver’s wins, and a sum of each driver's total standing points.
5. Get the average duration of pitstops for each race. This will return a list of each race with the average time for a pit stop (in milliseconds)
6. Get the constructor standings for raceid 100. This will return the name, the number of standing points, and the number of wins for each constructor for race id = 100, displayed in descending order.
7. The fastest lap speed from each driver during the 2022 Montreal grand prix to see which drivers had the fastest laps during the race. This will return the names of each driver and their fastest lap speed, displayed in descending order.

The EER diagram is as follows: