Teaching Case

Formula One – a database project from start to finish

Anthony Serapiglia
Anthony.Serapiglia@stvincent.edu
CIS Department
St. Vincent College
Latrobe, PA 15650

Abstract

The following is an applied database scenario based on a single season in the FIA Formula One (F1) World Championship of auto racing. This scenario builds database understanding and skills through data modeling, data acquisition, creation of a database schema through a database management system, query construction, and report creation. In the United States, Formula One falls squarely in the realm of "niche" sports, yet is regarded as one of the largest and most popular sports world wide. Various agencies have measured television viewership of F1 to be the largest for a seasonal sporting league in the world and only surpassed by quadrennial events such as the World Cup of Soccer and the Olympic Games. This dynamic allows for a great opportunity for undergraduate students. While most students will not be familiar with F1, they will have a basic understanding of racing in general. Also, while not in the consciousness of most Americans, the amount of sources of data for F1 abound worldwide. This creates a fertile environment abundant of data and the opportunity for easy entry to understanding of the environment by the novice. Utilizing this scenario will require that students learn about the world that the data is describing, enriching the experience with the personalities and energy of a world class sporting environment. The environment also provides multiple points of focus for modularization, but taken as a whole, allows for a full database creation experience, from "start" to "finish" - from modeling to reporting with all aspects in between.

Keywords: Data modeling, SQL, ER Diagram, Relational Database, data extraction, data manipulation

1. INTRODUCTION

One of the primary concerns when beginning database work is for students to appreciate the value of building an understanding of the subject matter the database is to be built around. There are many definitions for a database, but many hold the following in common - a database is a collection of inter-related data that describes objects or events that exist or occur in a defined world. One essential that is not often stated and assumed to be accepted, but is often neglected and forgotten is that a database must have a purpose. There must be some reason that it has come into being. Without a clearly defined purpose, those connections that define the interrelations of the collected data become weak, or nonexistent, allowing the structure to deteriorate and result in inefficient or unusable.

The purpose of this scenario is to build a database encompassing data from the current Formula One racing season. A user will be able to query this database to answer several basic questions of the data. Questions such as, "Who has won the most races?" or "What are the current team standings?" or "Who has qualified in the top ten the most times during this season?"

To create such a database, several steps are involved as evolutionary steps. First, an understanding of the world of F1 must be obtained so as to be able to identify what data is necessary, what is just useful, and what is not needed at all. To do this, sources for data and background context must be found. Second, a model needs developed to identify the entities that populate this world and their descriptors or attributes. Third, a schema needs to be developed within a database management system that will hold the F1 data. Fourth, a methodology must be developed to acquire data from sources outside of the database, that data transformed to fit the requirements of the database schema, and then loaded into the database. Fifth, queries will be necessary to select desired data from the database. Finally, sixth, reports can be developed to present the results of queries in a readily digestible and professional manner.

2. THE WORLD OF F1

Formula One (F1) is recognized by the Fédération Internationale de l'Automobile (FIA) as the highest level of open wheeled single seat auto racing in the world. Grand Prix racing can trace its roots to early road races in France, the earliest of which that utilized the title "Grand Prix" is recognized to be a 1901 event at the *Circuit du Sud-Ouest* around the streets of Pau (Rushby, 2011). Modern F1 history begins in 1950 with the first official World Championship for Drivers. Originally the championship season consisted of six events in Europe and the Indianapolis 500. The European events all had grown out of national championship races that had begun to appear prior to World War II. The first season saw Grand Prix races in Great Britain, Monaco, Switzerland, Belgium, France, and Italy. Other races were held as exhibitions, but did not count toward the championship. Non-points earning races would continue to be seen for many years until 1993. The number of races in the championship has varied from year to year, with the highest number being 21 events in 2016. While Europe remains the ancestral home of the series, races are now found across the globe including Asia, South America, North America, and the Far East. The 2017 season began in Australia in March and will end in Abu Dhabi in November.

Each season, two primary championships are contested, the World Constructors' Championship and the World Drivers' Championship. For the 2017 season, 10 teams will compete for the World Constructors' Championship. One team, Ferrari, has participated in every F1 season since 1950. The newest, Haas F1, enters its second season and was the first US based F1 team in over 20 years. Each team consists of two cars and two drivers. The drivers compete for the World Drivers' Championship. Both championships are amongst the most prestigious trophies in the sporting world and there are many times that the Constructors trophy has gone to a team other than the team for the championship wining driver.

Round	Grand Prix	Circuit	Date
1	Australian Grand Prix	Melbourne Grand Prix Circuit, Melbourne	26 March
2	Chinese Grand Prix	Shanghai International Circuit, Shanghai	9 April
3	Bahrain Grand Prix	Bahrain International Circuit, Sakhir	16 April
4	Russian Grand Prix	Sochi Autodrom, Sochi	30 April
5	Spanish Grand Prix	Circuit de Barcelona-Catalunya, Barcelona	14 May
6	Monaco Grand Prix	Circuit de Monaco, Monte Carlo	28 May
7	Canadian Grand Prix	Circuit Gilles Villeneuve, Montreal	11 June
8	Azerbaijan Grand Prix	Baku City Circuit, Baku	25 June
9	Austrian Grand Prix	Red Bull Ring, Spielberg	9 July
10	British Grand Prix	Silverstone Circuit, Silverstone	16 July
11	Hungarian Grand Prix	Hungaroring, Budapest	30 July
12	Belgian Grand Prix	Circuit de Spa-Francorchamps, Stavelot	27 August
13	Italian Grand Prix	Autodromo Nazionale Monza, Monza	3 September
14	Singapore Grand Prix	Marina Bay Street Circuit, Singapore	17 September
15	Malaysian Grand Prix	Sepang International Circuit, Kuala Lumpur	1 October
16	Japanese Grand Prix	Suzuka International Racing Course, Suzuka	8 October
17	United States Grand Prix	Circuit of the Americas, Austin, Texas	22 October
18	Mexican Grand Prix	Autódromo Hermanos Rodríguez, Mexico City	29 October
19	Brazilian Grand Prix	Autódromo José Carlos Pace, São Paulo	12 November
20	Abu Dhabi Grand Prix	Yas Marina Circuit, Abu Dhabi	26 November

Figure 1 – the 2017 Formula One Race Calendar.

Points are awarded at each race. For many years, 1960 through 2002, F1 was one of the stingiest points allotters of any sport. Only six places earned points. This often led to dominant teams quickly creating insurmountable leads and settling the championship well before the end of the season. In an effort to combat this situation, more points were allotted to more finishers with less gap between. The current point system was put in place for the 2010 season and allows for 10 finishers to earn point on a scale of 25 for the winner, 18 – 2nd, 15 – 3rd, 12 – 4th, 10 – 5th, 8 – 6th, 6 - 7th, 4 – 8th, 2 – 9th, and 1 for tenth.

A race weekend involves three days. Each event will allow for two "free" practice session on Fridays. (The exception being Monaco which begins on Thursday to allow for the traditional Friday market day in the principality…). Saturday begins with a third "free" practice followed by qualifying. Sunday is race day. "Free" practices earned the name due to the fact that although timing and scoring will display results of the sessions, they do not count towards anything. The teams are "freed" of the normal race regulations and allowed to test new parts, as well as occasionally test new drivers. It is not uncommon to see a Friday test driver take part the first session before giving way to the normal race driver for the rest of the weekend.

Qualifying formats in F1 have also evolved over the years in an effort to both reduce costs and to produce a better show for fans. The current system of Qualifying consists of three separate sessions, Q1, Q2, and Q3. All cars have an opportunity to participate in Q1. At the end of 18 minutes, the top 15 move onto Q2. After a short break, Q2 lasts 15 minutes where only the fastest 10 drivers move onto Q3. The fastest driver in Q3 earns the pole position, or the right to start from the first position in Row 1 at the start of the race.

Each race has a different lap length and overall length of the race. Most races are around 300KM and last approximately 90 minutes. At the end of each race, the top three drivers make up the podium ceremony. Embracing a rich nationalistic history, during the podium ceremony the national anthem for the winning driver is played followed by the national anthem for the winning car constructor. In 2017, the 10 teams represent 8 different countries of home base, while amongst the 20 drivers 15 nations are represented.

Section 2 Assignment: Write a 1 to 2 page summary explaining the world of Formula 1. Include: What defines a season. Who are the major participants (Teams, Drivers)? How does a race weekend run? What are the two major championships being fought for? How are results of practice session, qualifying, races reported? Include at least three websites as sources for data for results as well as background information about F1.

3. DATA MODELING

Following the work done in section one, a basic understanding of the actors/dynamics of the world of F1 should have been achieved. From this background, a model of what data describes this world can be determined.

Before starting in to build out a database in the database management system (DBMS), it is essential to organize and plan ahead of time. A data model helps to visualize the data and to aid in identifying the relationships that exist between the different groups. Data models can be very simple, or more complex depending on the needs of the developer, the data involved, and the eventual end "product" that is to be strived for.

Entities should be identified, the general categories that will contain multiple entries and their descriptors or attributes. In the case of F1, the race calendar can be a starting point. "Events" or "Races" can be used as a label for this entity. A listing can then be developed that includes the pieces of data that describe each event. What number the event is on the calendar, what country the race is located in, the dates the event takes place, the track name, the city name, the "official" sponsored title of the event, etc… can all be included as attributes.

Following this identification, Entity Relationship (ER) Diagrams should be developed to visually display the connections between these collections of data. Does a driver appear in a collection of drivers? Teams? Race Results? Practice Results? Qualifying results?

Section 3 Assignment:
Begin to identify data points in the world of F1. In step 1, create a document that identifies entities that exist. Under each entity, bullet out the attributes that serve as descriptors for each item that would belong in that collection. Begin to include a data type to describe what form the data will be in for each of the attributes. In step 2, develop a preliminary ER diagram of your entities. Identify attributes that the entities have in common.

4. BUILDING OUT THE DATABASE

Different database management systems can refer to the software structures that make up the tables and other objects within the database in different ways. For this scenario, MySQL will be used as a reference for general vocabulary describing the structures within the DBMS. Thus "schema" will refer to the collection of tables, saved queries, and other objects that function as the "F1 Database" for this scenario.

Most DBMS packages include a visual design component. It is important to note that while these GUI (Graphical User Interface) interfaces can be extremely convenient, in almost all cases there will be certain advanced tasks that will still need to be completed through issuing standard SQL commands. It is always a good idea to become familiar with the specific SQL commands to issue that will satisfy even the most fundamental tasks such as table creation.

A common best practice is to save SQL commands in a plain text file for later reference. It is always a good idea to keep a running log of each command committed to the database in the case that it may need recreation from scratch. Having these commands saved off will allow anyone to be able to retrace footsteps and ensure that no steps are missed in the worst case recovery scenario.

This is also the time for verifying several hard choices that are made in the data modeling stage. When transitioning from "paper" to "practice" the absolute need to adhere to several database rules comes into play. The F1 world allows for several options when choosing primary keys for different tables. It also presents several "red herrings" that can be tempting to use, but could lead to duplication in several instances. Can the car number be used as a primary key within the Drivers table?

Section 4 Assignment:
Build out a shell of a database structure within a DBMS that fulfills your data model. For each step, save the SQL commands that were utilized in creation of tables and fields, the setting of datatypes, and identification of primary keys. Important constraints should also be identified, such as no-nulls. These SQL commands should be collected on a text file and clearly identified with comments defining what each section of commands is for.

5. ETL

With a shell of a database built, it must be filled with data. Just like many other sports, F1 has a wealth of available statistics available over the Internet. One of the best sources is the official F1 web site - https://www.formula1.com/. In the results section of the website, a report is available with the summary statistics for each of the five sessions that cars are on track.
One of the most common and most important tasks of database work is managing an ETL process – Extract – Transform – Load.

In the case of a basic F1 database structure, a handful of common tables will prove to be essentially static, that is once data is loaded there will be little change or updates during the season. A table for events, the table containing teams, and the table for drivers are all relatively static. They are all also relatively small. For each it is manageable that the data they contain could very well be directly keyed in. While this is possible, it is not good practice. Anyone who takes this route may well be tempted to continue in this fashion. They will soon find it will be much to their advantage to develop a system to prepare datasets of session results and to import them through SQL scripts. While a team table may have only ten records, through the course of a season, twenty drivers per session with five sessions per event and twenty events per season will result in at least 2,000 records of results. While this may not be considered "Big Data" by some, it does demand that a more efficient method than direct keying of data be found.

In deciding a method of extraction and transformation there is no singular "best" way, although there are some "better" ways. Much of this depends on the source and shape of the data source that is providing the raw material. For most, an approachable method can consist of a copy and paste of data from the http://formulaone.com website into a spreadsheet. The spreadsheet can be arranged to transform the data, performing replacement of event, driver, and team names with corresponding foreign key identifiers for example. These transformed data sets can then be saved off as comma delimited (CSV) files which in turn can be imported into corresponding tables within the database. More advanced methods of ETL can include data scraping utilities, or custom programs in Java or C++ to collect data from a target and produce an arranged CSV file. Loading of data can also be broken into several steps. It is not always a good idea to load data straight into the final destination table. Holding tables can be

created as an initial import destination where data can then be validated and further moved to its final resting location.

Depending on the time of year, several races will have already been completed, with several more to take place. The ETL procedure can thus be tested with exiting data, and the process validated with new data as it arrives with the results of new races.

Section 4 Assignment:
Locate a good source of data that includes all necessary items to fill your database. Develop a method of capturing data, transforming it into a format that can then be loaded into your database. Detail a set of instructions for your procedure. The instructions must be specific and clear enough that another person unfamiliar with your database and the data will be able to follow them to a positive result. Load all possible data into your database.

6. QUERY THE DATA

Data has potential. If allowed to simply sit in its static state within a database, that potential value will never be realized. To unleash the value of the data, it must be processed into information. Processing of data comes in many forms, from simply separating it from other pieces of data to advanced analytics and mathematical formulas. The purpose of the data processing is to answer a question – information informs. Data has potential, information is kinetic. Ask questions of your data.

The relational data model separates several key components that are normally found together. A normalized F1 database design would not have a table that contained a record that included a team name with the first and last names of their two drivers for example. Once data is loaded into the database, verify and validate the structure of the data model by performing several basic queries.

Simple queries to start with will look to combine data in just two tables. The example of a query providing a list of all teams with their drivers is a good example. To perform this query properly a basic SQL JOIN statement can be used to link the two tables together.

Once the basic structure has been confirmed for two table combinations, further test structure by producing a query that ties together four tables to display a race or practice session result. This should require a join between tables for events, teams, drivers, and session or race result table.

Section 6 Assignment:
Produce SQL queries that will pull data to answer the following questions: Who are the drivers for each team? What were the results from the second practice session of the fifth race of the season? What is the current standings for the Constructor's Championship? How many drivers have started on pole position and how many times for each? What races have been run and who were the winners to this point in the season? Who are the top ten drivers in terms of laps completed during practice sessions?

7. OPTIONAL - REPORTING

(This section optional dependent on access to reporting software)

One of the biggest differences between enterprise level DBMS packages and a personal database program is the inclusion of a reporting component. Users of programs such as Microsoft Access will find relatively robust and user friendly modules for the production of reports generated from the data within the database. Traditionally, enterprise level database systems such as Oracle, Microsoft SQL (MSSQL), and MySQL have not included such features. Third party applications such as Crystal Reports and open sourced coding through PHP web pages have been utilized dependent on the need and the end audience.

In more recent releases of MSSQL, Microsoft has developed a report server that includes a free report builder component that can be used independently of MSSQL and the server component. A leading consumer application that is a competitor to Crystal Reports is DBxtra.

Reporting out of a database is a critical component to realizing the ultimate potential value of the data it contains. Whether the data is pushed out to recipients on a scheduled basis, pulled on demand by power users, or distilled for easy digestion through dashboards – reporting is often the final step in turning data into information by making it consumable by decision makers.

Section 7 assignment:
Based in the queries completed in the previous section, produce reports that will display results of each practice session, qualifying session, and race. Follow with preset reports that will display current driver and constructor standings.

8. OPTIONAL – FINAL ASSIGNMENT

With the recent takeover of the F1 commercial rights by Liberty media have opened up a opportunity for a special 21st race to be added to the 2017 season! Held in (insert exotic location somewhere in the world here) the race went off in secret until now. F1snoops.local (a local webserver to your location) has received race reports for the "lost weekend".
Based on all of the experience gathered through the past several assignments:

- Import the results of the event weekend into your F1 Database.
- Update any new drivers participating in Friday Free practice 1 in the Drivers table
- Produce reports for each of the five sessions of the weekend
- Produce updated reports with the final driver and constructors championship standings
- Provide a copy of the procedure followed in your ETL process.

9. REFERENCES

A brief history of Formula One. (n.d.). Retrieved May 3, 2017, from http://en.espn.co.uk/f1/motorsport/story/3831.html

Elmasri, R., & Navathe, S. (2011). Fundamentals of database systems. Boston: Addison-Wesley.

Official Formula One Website. (n.d.). Retrieved May 12, 2017, from https://www.formula1.com/

Rushby, K. (2011, April 29). Speeed date: motor racing returns to Pau. Retrieved May 10, 2017, from https://www.theguardian.com/travel/2011/apr/30/pau-france-race-circuit

Spurgeon, B. (2016, March 18). Is Formula One Still on Top? Retrieved April 21, 2017, from https://www.nytimes.com/2016/03/19/sports/autoracing/is-formula-one-still-on-top.html

Williamson, Martin "A Brief History of formula One" http://en.espn.co.uk/f1/motorsport/story/3831.html