F1 Data Loading

The data loading process starts with Main\_Task.R. This file runs Libraries.R which loads in the required libraries. It creates the following functions: dbKillConnections() which terminates any MySQL connections, schema\_nuke() which drops all the season schemas & the formula\_1 schema, schema\_nuke\_years(), truncate\_current() which truncates all the tables in a specified season schema, & truncate\_all() which truncates every table in the season schemas & the formula\_1 schema. The formula\_1 schema is created if the schema doesn’t exist with the seasons table which contains years & number of rounds for that year, & it creates the table\_insert\_tracker which is used to track the data insertion process & used to figure out where the process failed. The years listed in the API (<http://ergast.com/mrd/>) are gathered, so only valid values are recalled. The table names are also listed. The max year completed from the table\_insert\_tracker is retrieved to determine the start\_year for creating the seasons schema. If a f1\_[year] schema doesn’t exists & the year value is in the valid API years, that schema & its tables are created.

Schema\_Creator.R is run to create the f1\_[year] schemas. It creates the following tables: circuits, driver\_standings, drivers, lap\_times, races, results, constructors, qualifying, pit\_stops, constructor\_standings, constructor\_results, seasons, & status. The values in these tables reflect what is returned by the API. They are not normalized. These tables are created for each valid year.

All the connections strings found in the R scripts in the Connections folder are schema specific. These are run at various points of the program. No more than 16 connections can be open at a time.

Then, the data is loaded into the f1\_[year] tables using Data\_Load.R. The start\_year is generated using the method above. The data cleaning functions found in Data\_Cleaning\_Tools.R are loaded in & stop\_loop is set to FALSE which is used if the data loading process fails. The data loading process is as follows:

* Runs a for loop from start\_year to length(years)
  + Run dbKillConnections()
  + Get the number of rounds for the season/year
  + Insert the year & number of rounds into the seasons table in the formula\_1 & f1\_[year] schemas
  + Insert the year, number of rounds, & start time into the table\_insert\_tracker table
    - This table is updated after data is inserted into each table in the f1\_[year] schema
  + Run a for loop from 1 to number of rounds
    - Everything is inside a tryCatch in case the process fails
    - Get all the values from the API for the season/year & each round
    - Insert the data into the tables created by Schema\_Creator.R
    - If the process fails/throws an error
      * Print the error message
      * Set stop\_loop = TRUE
      * Truncate all the tables in the current f1\_[year] schema that just failed
      * Retrieve all the values from the table\_insert\_tracker table for the year that failed & show the results
      * Delete the rows in the table\_insert\_tracker table for the year that failed
      * Delete the failed year value from the formula\_1.seasons table
  + Update the table\_inset\_tracker table to say that year has completed insertion
* Run dbKillConnections() & remove the dataframes that contain the API values to save RAM

Then, the data in the f1\_[year] schema is normalized by running Main\_Task\_pt2.R. The following functions are created: truncated\_master() which truncates the tables in the f1\_master schema & schema\_nuke\_master() which drops the f1\_master schema. If the f1\_master schema does not exist, it is created by running Master/Schema\_Creator\_master.R. This script also creates the necessary tables which have the same names as the tables in the f1\_[year] schema, but the tables have slightly different values & data types since the inserted data will be normalized. Then, it checks if multiple f1\_[year] schemas exists to determine if which data load path to take. One is for updating to f1\_master schema tables while the other is for populating those tables with all the data. Only Data\_load\_master.R will be explained. The two scripts are very similar.

Data\_load\_master.R works by compiling the data from all the f1\_[year] schemas together. Then, it assigns the drivers, constructors, races, status, & circuits unique numerical IDs for each unique value. Columns from the f1\_[year] schemas are renamed to their proper name if needed. Unneeded columns are dropped. The data is organized in chronological order before being inserted. The new ID values are matched to be associated with the proper values in the other tables.

Finally, schema\_nuke\_years() is run to drop all the f1\_[year] schemas because they are no longer needed & take up space.