**Formula1 Racing**

**Overview**

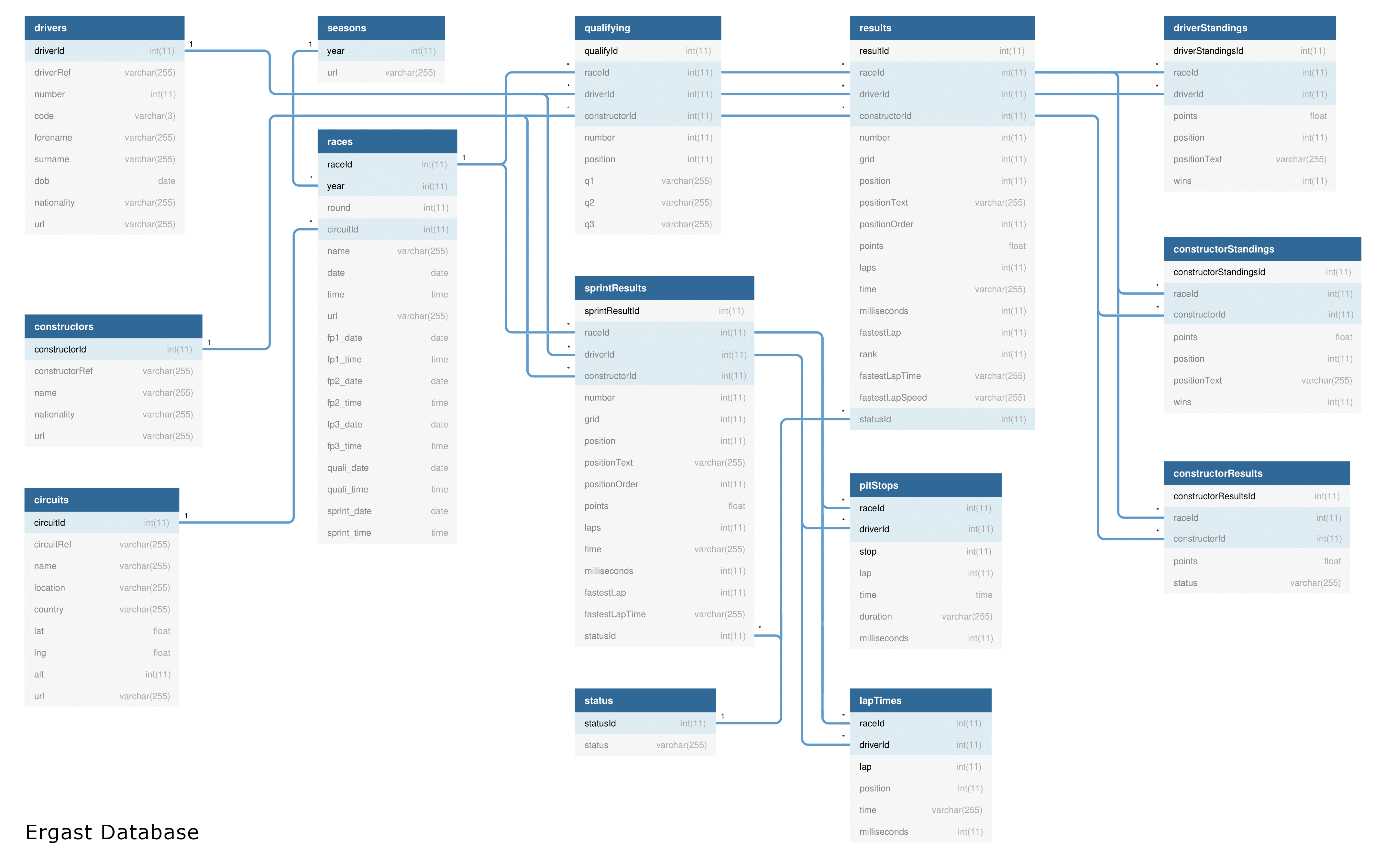
Formula One is the highest class of international racing for open-wheel single-seater formula racing cars. Every season happened once a year, each race happened over weekends (Friday to Sunday). Each race is conducted in individual circuits. 10 Teams/Constructors will be participated. 2 Drivers will be assigned in a team. Saturday will be a qualifying round for the Sundays match. 50-70 Laps will be there on each race. Pitstop will be available to change tire or damages. Race results included driver standing and constructure standing.

**Source Date Files**

We are referring open-source data from website [Ergast Developer API](http://ergast.com/mrd/). Data available from 1950 till 2022.

|  |  |
| --- | --- |
| Circuits | CSV |
| Races | CSV |
| Constructors | Single Line JSON |
| Drivers | Single Line Nested JSON |
| Results | Single Line JSON |
| PitStops | Multi Line JSON |
| LapTimes | Split CSV Files |
| Qualifying | Split Multi Line JSON Files |

Data Model (http://ergast.com/images/ergast\_db.png)



**Data Ingestion Requirement**

* Ingest All 8 Files into the data lake
* Ingested data must have the schema applied
* Ingested data must have audit columns
* Ingested data must be stored in a columnar format
* Must be able to analyze the ingest data
* Ingestion logic must be able to handle incremental load

**Data Transformation Requirements**

* Join the key information required for reporting to create a new table
* Join the key information required for analysis to create a new table
* Transformed tables must have audit columns
* Must be able to analyze the transformed data via SQL

**Reporting Requirements**

* Driver Standing
* Constructure Standing

**Analysis Requirements**

* Dominant Drivers
* Dominant Teams
* Visualize the output
* Create Databricks Dashboard

**Scheduling Requirements**

* Scheduling to run every Sunday 10PM
* Ability to monitor pipelines
* Ability to re-run failed pipelines
* Ability to set-up alerts on failures

**Technologies/Tools Used:**

* Pyspark
* Spark SQL
* Delta Lake
* Azure Databricks
* Azure Data Factory
* Azure Date Lake Storage Gen2
* Azure Key Fault
* Power BI