IE 7374 ST: BI Data Integration

Accelerating Insights:

A Cloud-Based Data Pipeline for F1 Racing Analytics

Group 8

Anurag Palanki

Rohith Kanakagiri

INTORDUCTION

Formula 1 (F1) is a world-renowned motor racing championship that has captivated millions of fans since its inception in 1950. Over the decades, the sport has evolved dramatically, with cutting-edge technology and innovation driving performance and shaping the industry's landscape. F1 racing teams and decision-makers rely on data-driven insights to optimize performance, increase efficiency, and sharpen their competitive edge.

PROBLEM DEFINITION

While the F1 racing dataset contains a wealth of historical and current information, the data remains largely untapped, and the full potential of data-driven decision-making in the F1 world remains unrealized. The primary challenge is to design and implement a robust data pipeline in Google Cloud Platform (GCP) that efficiently loads data into a data warehouse, enabling users to access and analyze data for strategic and operational insights.

DATASET LINK:

https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

Analysis Dimensions:

The following dimensions can be explored in the F1 racing dataset:

a. Driver Performance: Analyze individual drivers' performance over time, including wins, pole positions, and podium finishes. Identify key factors that contribute to a driver's success, such as team affiliation, car specifications, and driving styles.

b. Team Performance: Evaluate the success of F1 teams in terms of constructors' championship points, race wins, and podium finishes. Identify trends in team performance and potential areas of improvement.

c. Track Analysis: Investigate the impact of different tracks on race outcomes, examining factors such as track length, weather conditions, and average lap times.

d. Historical Trends: Identify patterns in race outcomes, driver and team performance, and technological advancements over time.

e. Race Strategy: Analyze pit stop strategies, tire choices, and fuel management to understand the impact on race results.

Strategic and Operational Insights

By building a data pipeline and data warehouse for the F1 racing dataset, several strategic and operational insights can be achieved:

a. Identify Success Factors: Uncover the key factors that contribute to successful drivers and teams, enabling stakeholders to optimize their resources and strategies for future races.

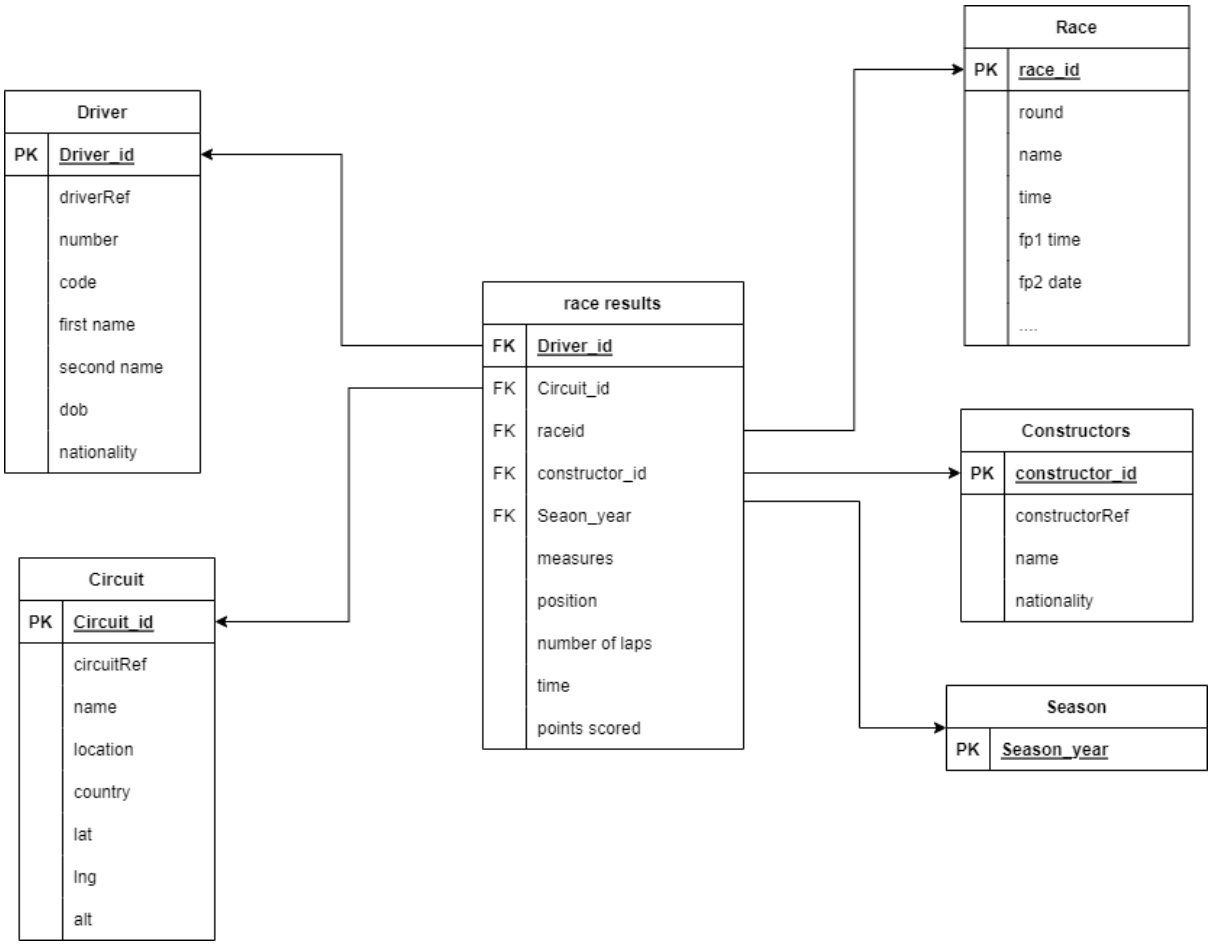
b. Optimize Race Strategies: Analyze historical data to determine the most effective race strategies for specific tracks, weather conditions, and driver-team combinations.

c. Benchmarking and Performance Tracking: Monitor the performance of drivers, teams, and cars against historical data, enabling stakeholders to set ambitious yet realistic goals and track progress over time.

d. Enhance Fan Engagement: Leverage data-driven insights to create engaging and interactive content for fans, deepening their connection to the sport and fostering a more informed and passionate fanbase.

Datawarehouse Schema

Here is the relational schema for our data warehouse with 5 dimensions and 1 fact table.



Implementation

Google Cloud Platform (GCP) is a leading suite of cloud computing services offered by Google, providing a range of powerful tools and infrastructure that enable businesses to build, deploy, and scale applications, websites, and services on the same robust foundation as Google itself. GCP is known for its cutting-edge technology, ease of use, and seamless integration with other Google services. Given its robust suite of tools and scalability, GCP is an ideal choice for implementing a data pipeline to load data into a data warehouse.

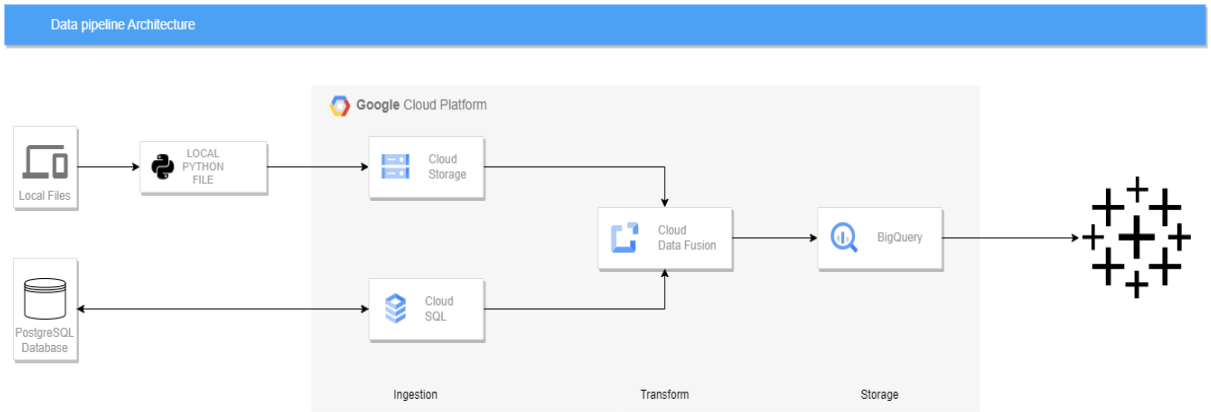
One of the primary reasons GCP is well-suited for this project is its ability to handle large-scale data processing tasks, which is essential given the extensive F1 racing dataset spanning over seven decades. GCP's suite of services, including Cloud Storage, Cloud SQL, Data Fusion, and BigQuery, enables seamless data ingestion, transformation, and storage, thereby streamlining the entire data pipeline process.

Additionally, GCP's serverless architecture ensures that resources are allocated efficiently and automatically scaled based on demand, allowing for cost-effective data processing and storage. This is particularly useful when dealing with large datasets, as it minimizes the need for manual intervention and ensures optimal performance.

Furthermore, GCP's security features and compliance certifications provide an additional layer of protection for the sensitive data being processed and stored, making it a reliable and secure choice for this project.

Building the Data Pipeline

The data pipeline for the F1 racing dataset was built using several GCP services, following these steps:



Loading the Dimensions and Fact Data

First, the dimension data was loaded into Google Cloud Storage buckets using a local Python file. This process involved reading the data from the source files, pre-processing it, and uploading it to the Cloud Storage buckets. Cloud Storage provides a scalable, durable, and cost-effective solution for storing and retrieving large amounts of data.

The fact data, which contains race results, was loaded into a Cloud SQL connected PostgreSQL server using pgAdmin. Cloud SQL is a fully-managed relational database service that simplifies database setup, maintenance, and administration. By using Cloud SQL, we can ensure the fact data is efficiently stored, managed, and retrieved for further processing.

Transforming Data with Google Data Fusion

Google Data Fusion was used to transform the data from the Cloud Storage buckets and the Cloud SQL PostgreSQL server. Data Fusion is a cloud-native data integration service that helps users build and manage ETL (extract, transform, load) pipelines with a visual interface, making it easier to clean, prepare, and blend data from various sources.

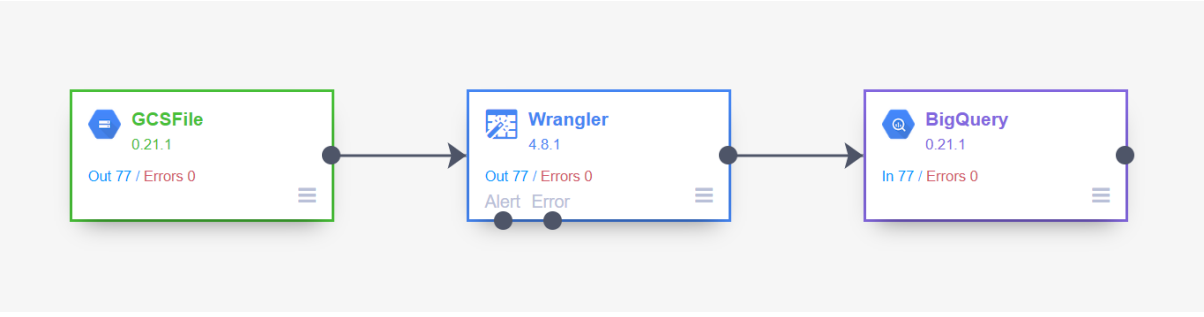
In this step, Data Fusion read the dimension data from Cloud Storage and fact data from Cloud SQL. Custom transformations were applied to clean and enrich the data, ensuring that it was in the appropriate format for further analysis. The transformed data was then prepared for loading into the data warehouse.

Loading Data into BigQuery Dataset

After the transformation process, the data was loaded into a BigQuery dataset. BigQuery is Google's fully-managed, serverless data warehouse solution designed for large-scale data analytics. It enables super-fast SQL queries using the processing power of Google's infrastructure.

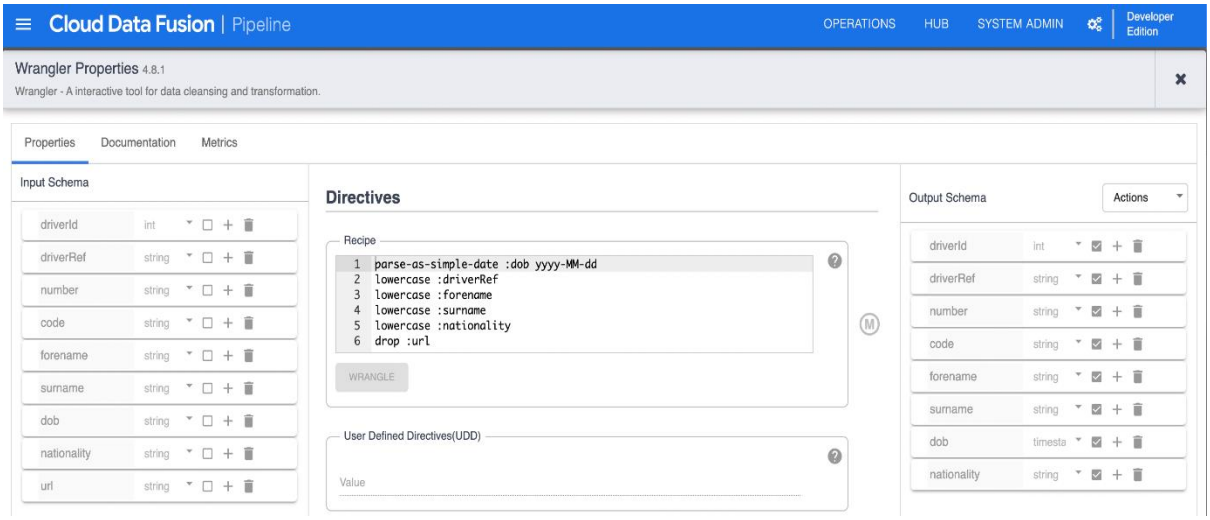
By loading the F1 racing dataset into BigQuery, we can perform complex queries and analysis on the data, uncovering insights into driver performance, team performance, track analysis, historical trends, and race strategies. This enables stakeholders to make data-driven decisions and drive improvements in the F1 racing world.

These is the data pipeline process image for loading the dimensions into the big query table.

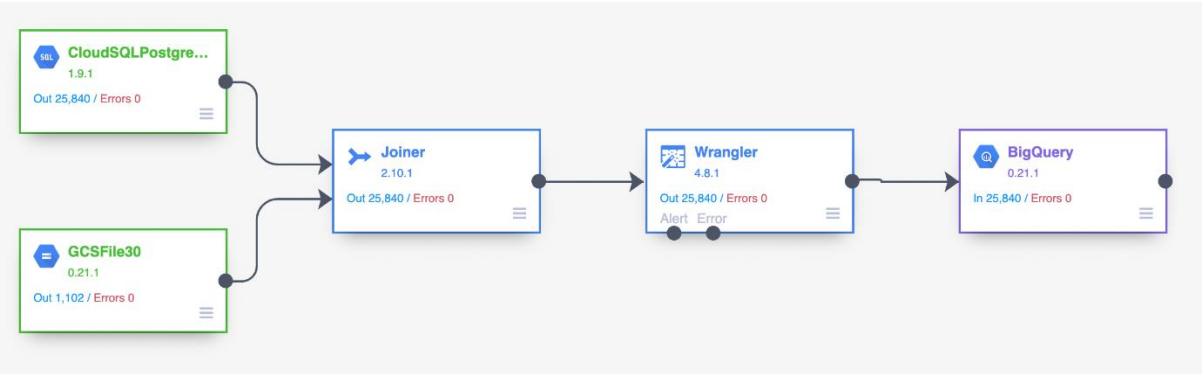


The process is similar for the other dimensions as well.

The wrangler has transformation steps in it as show below.



For loading the race results fact we built the following pipeline which gets the circuit id and season year foreign keys from the race table. The race results in a Cloud SQL instance and the race dimension fact is in Google cloud store so we performed a Join before transforming the data.

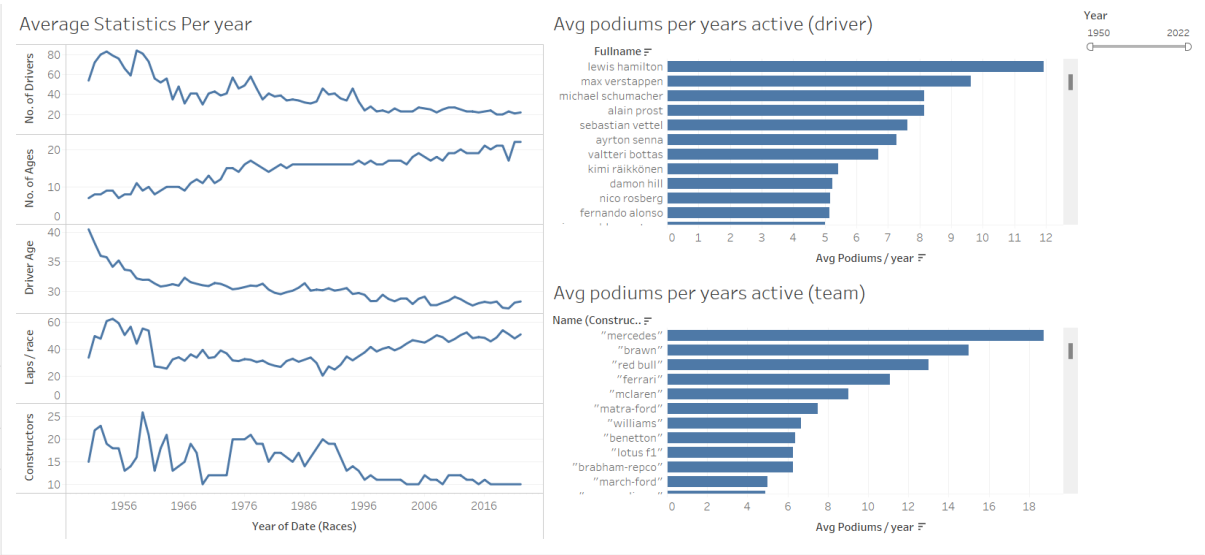


In the wrangler we dropped the redundant columns and pushed the data into big query

The data pipeline built using GCP services allows for efficient ingestion, transformation, and storage of the F1 racing dataset. By leveraging the power of GCP, the data pipeline ensures seamless data processing, scalability, and cost-effective performance, enabling stakeholders to unlock valuable insights and revolutionize the F1 racing world.

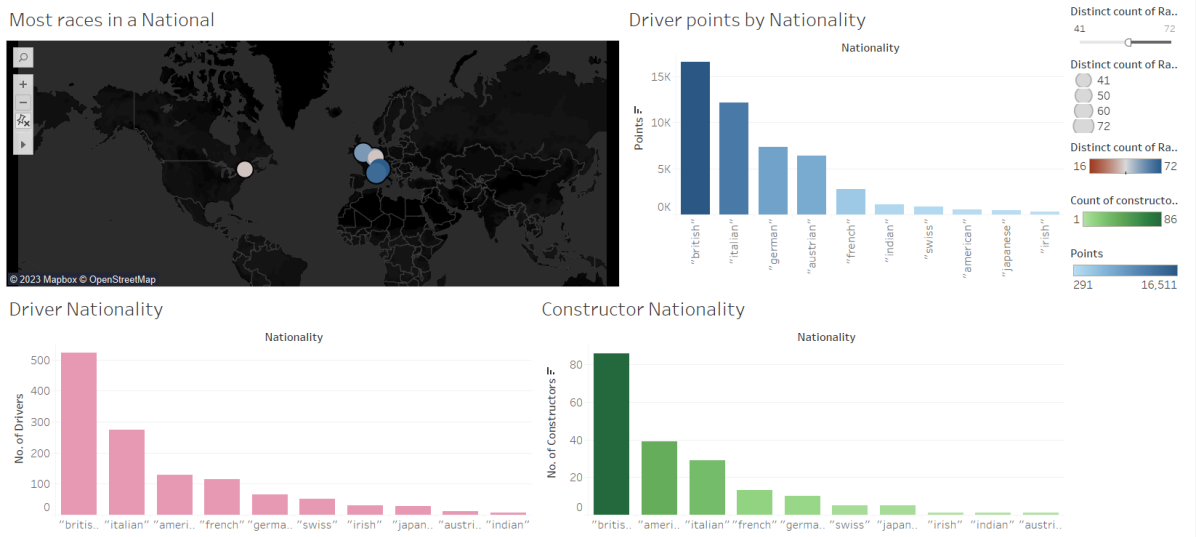
Analysis

The big query data warehouse was connected to a tableau notebook to analyze the data and gain the insights that we wanted to gain. We first analyzed the basic statistics.

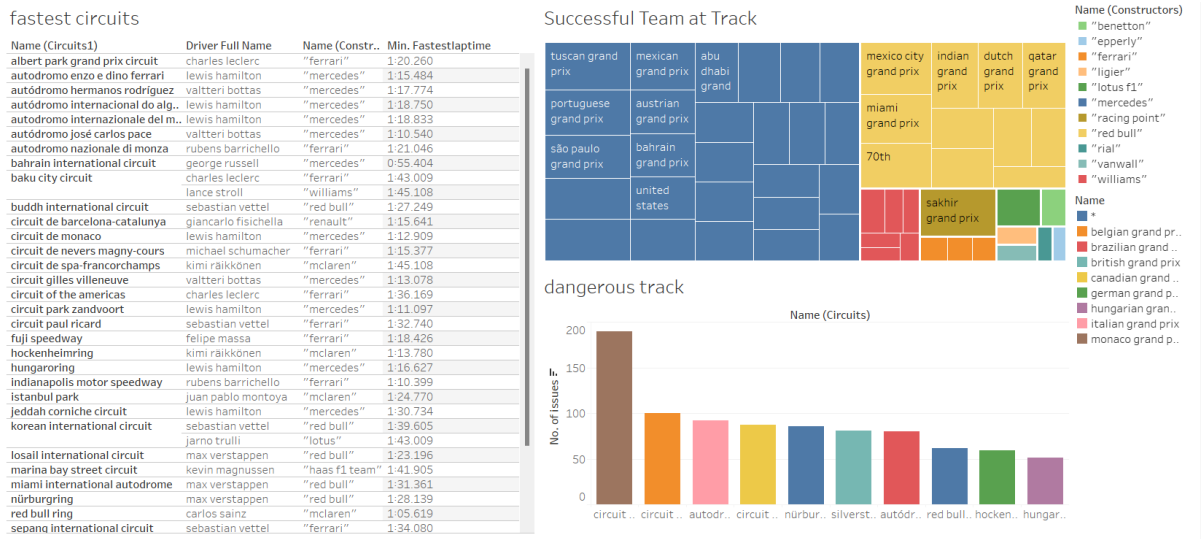


The number of races and the average age of the drivers increases as the years go on but the number of drivers, the number of issues the cars have decreases over time.

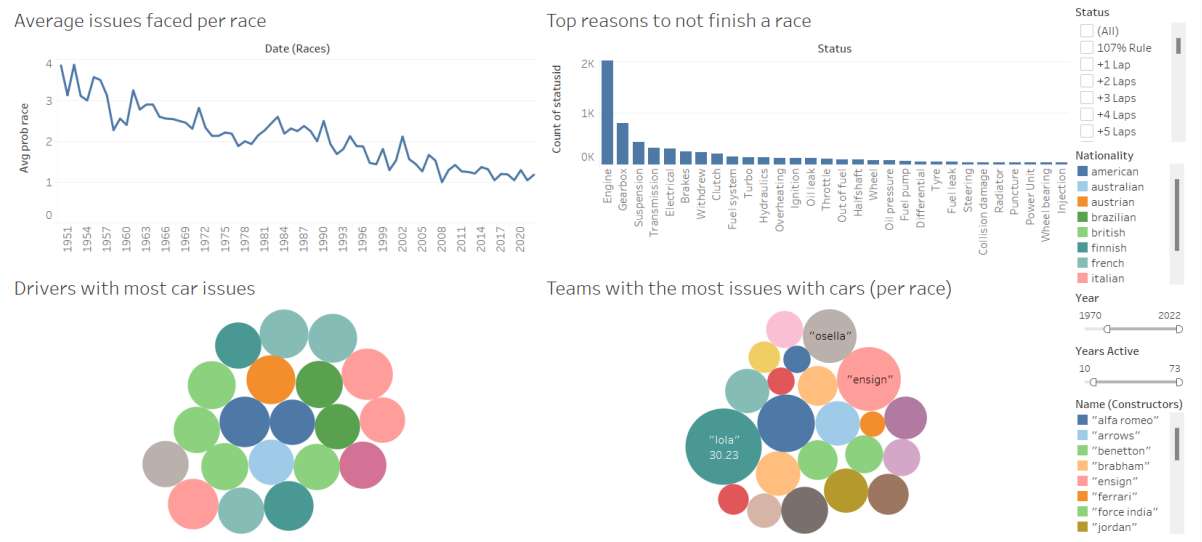
Sir Lewis Hamilton and Mercedes remain the most successful constructors and drivers by a huge margin over taking even Ferrari and red bull if we count the number of years they've been active in the sport.



We then turned to analyse the nationality and nation hierarchies of the teams, drivers and circuit dimensions revealing the domination the European natives. The top 5 locations where races are held is in Italy, Monaco and Britain while subsequently they also dominate the number of points drivers win, the number of drivers from those nations and the nationality from where most constructors are from.



This is to analyze who has the fastest lap ever at each track. It mentions the constructor, driver name and the time he put up. It also shows the drivers and teams who have the highest average points at each circuit. The last graph shows the track at which most issues occur. Monaco is notorious for cause a lot of crashing and straining the car more.



This is a dashboard is to analyse the average issues per race over the years. Then the reason for most issues being the engine. Then a bubble plot showing the drivers and constructors having the most issues.

Conclusion

In summary, our team has utilized a data warehouse to gain valuable insights into the world of Formula 1 racing. By connecting a Big Query data warehouse to a Tableau notebook, we were able to analyse data from the five dimensions (circuit, driver, constructor, race, status and season) and one fact table (race results) to gain a deeper understanding of the sport.

Through our analysis, we were able to identify trends and patterns in driver and team performance, historical data, race locations, and car issues. We found that the number of races and average driver age increases over time, but the number of drivers and car issues decreases. We also discovered that Sir Lewis Hamilton and Mercedes remain the most successful drivers and constructors in the sport by a significant margin.

By analysing the nationality and nation hierarchies of teams, drivers, and circuit dimensions, we were able to identify the dominance of European natives in the sport. We also found that Italy, Monaco, and Britain are the top locations for races and the nationality from where most constructors are from.

Our analysis of the fastest lap times and drivers and teams with the highest average points at each circuit, as well as the dashboard that analyses the average issues per race over the years, will help teams to optimize their race strategy and identify potential areas for improvement in their car design and performance.

Overall, utilizing a data warehouse and Google Cloud Platform has been crucial in gaining these valuable insights, as it allowed us to easily store, manage, and analyse large amounts of data from various sources. These insights can help teams to make informed decisions and improve their performance in future races.