F1 EDA - Big Data Final Term Project

Enrique Ulises Báez Gómez Tagle
Mauricio Iván Ascencio Martínez
Sara Rocío Miranda Mateos

May 26, 2024

Abstract

Contents

1 Introduction

2 Infrastructure / Architecture

- **Hadoop Cluster Configuration:** Central to our project is a Hadoop cluster, designed for scalability and resilience. We configured it with one master node to manage the cluster and multiple worker nodes to process the data. This setup allows for efficient data processing and analysis.
- **Hive Database Integration:** For querying capabilities, we integrated Hive with our Hadoop cluster. Hive facilitates querying and managing large datasets residing in distributed storage using SQL-like syntax. It is particularly advantageous for EDA, allowing us to query the data stored in the Hadoop Distributed File System (HDFS) with ease.
- **Deployment Process:**
  1. We initiated our deployment by setting up the Google Cloud SDK, enabling us to interact with Google Cloud services seamlessly.
  2. A Cloud Storage bucket was created to host our Hive data, carefully chosen to be in proximity to our compute resources to minimize latency.

This infrastructure serves as the backbone for our project, enabling us to leverage big data technologies effectively to extract meaningful insights from the Formula 1 dataset. The deployment emphasizes scalability, efficiency, and cost-effectiveness, tailored to meet the demands of processing and analyzing large-scale data.

Figure 1: **Architecture Diagram**

3 Data Analysis with F1 Dataset

3.1 Query Analysis and Insights

3.1.1 Average Number of Laps per Grand Prix

Figure 2: Number of Laps GCP Querie Result

Figure 3: Bar Chart Average Number of Laps per Grand Prix

Description:

Interpretation: The bar graph visually represents the average number of laps for each Grand Prix, sorted in descending order. Longer races may indicate a need for different tire strategies or fuel management plans. Conversely, shorter races might lead to more aggressive racing tactics. We can say that Monaco GP is the longest and Belgian GP is the shortest with a difference of around 16 laps.

4 Challenges Encountered

During the development and analysis phases of our project, we encountered several challenges that fell into two main categories: infrastructure-related and query/code-related. Below, we detail these issues and how we addressed them.

4.1 Query and Code-Related Challenges

5 Conclusions

Through the utilization of big data technologies such as Hadoop and Hive on the Google Cloud Platform, this project has provided significant insights into the 2023 Formula 1 season. Our comprehensive analysis spanned multiple aspects of the sport, from individual driver performance and tire strategies to the influence of weather conditions on race dynamics.

Insights and Impact: The queries conducted revealed:

- The average number of laps per Grand Prix can inform race strategies and pit stop planning.
- Tire compound usage rates can help teams optimize tire strategies for different tracks and conditions.

Future Directions:

Final Thoughts:

6 Project Repository

`https://github.com/enriquegomeztagle/BigData/tree/main/FinalTerm/F1-GridGuru-Project`

7 References