F1 EDA - Big Data Final Term Project

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

June 1, 2024

Abstract

In this project, we explore the use of big data technologies to analyze the Formula 1 dataset. Our objective is to derive meaningful insights that can inform strategic decisions in the sport.

Contents

1 Introduction

This project focuses on leveraging big data technologies to analyze the Formula 1 dataset. By implementing a robust data lakehouse architecture, we aim to perform detailed data analysis and extract insights that can drive data-driven decision-making in the context of Formula 1 racing.

2 Infrastructure / Architecture

- **Data Lakehouse Configuration:** Central to our project is a Data Lakehouse architecture designed for scalability and resilience within an AWS environment. This setup includes multiple layers of data storage and processing capabilities, ensuring efficient data management and analysis.
- **Storage Layers:**
  - **Raw Data Storage (RAW Layer):** Utilizes Amazon S3 to store unprocessed, raw data ingested from various sources.
  - **Cleansed Data Storage (Cleansed Layer):** Stores cleaned and transformed data in Amazon S3, ready for further analysis.
  - **Data Warehouse Storage (Data Warehouse Layer):** Contains structured and optimized data in Amazon S3, prepared for querying and visualization.
- **Data Ingestion:**
  - **Real-Time Data Ingestion (AWS Lambda):** Handles the real-time ingestion of data, storing it in the RAW Layer.
  - **Batch Data Ingestion (Static Dataset):** Manages the batch ingestion of static datasets, such as CSV files, into the RAW Layer.
- **Data Transformation:**
  - **ETL Processing Jobs (AWS Glue):** Performs ETL (Extract, Transform, Load) operations to clean and transform data from the RAW Layer to the Cleansed Layer.
- **Data Analysis:**
  - **Data Analytics and Processing (AWS EMR):** Utilizes Amazon EMR to run advanced data processing and analytical tasks on data from the Cleansed Layer, producing structured results stored in the Data Warehouse Layer.
- **Data Cataloging:**
  - **Metadata Management (AWS Glue Data Catalog):** Catalogs the metadata of datasets in the Cleansed Layer, facilitating efficient data discovery and access.

- **Data Visualization:**
  - **Visualization and Analysis (RStudio):** Processes the curated datasets in the Data Warehouse Layer locally to create insightful visualizations and analysis, supporting decision-making.
- **Deployment Process:**
  1. We initiated our deployment by setting up the AWS CLI, enabling us to interact with AWS services seamlessly.
  2. Amazon S3 buckets were created to host our data layers, carefully chosen to be in proximity to our compute resources to minimize latency.
  3. For infrastructure provisioning, we utilized both AWS CloudFormation and Terraform. AWS CloudFormation was used for managing AWS-specific resources, ensuring consistent and repeatable deployments within the AWS ecosystem. Terraform was employed for managing multi-cloud resources and ensuring infrastructure as code principles across various environments.
  4. AWS Glue was configured to automate our ETL processes, facilitating efficient data transformation and loading into our cleansed and data warehouse layers.
  5. Amazon EMR clusters were set up for big data processing, with configurations tailored to optimize performance and cost.
  6. Continuous Integration and Continuous Deployment (CI/CD) pipelines were implemented using AWS CodePipeline and Jenkins, automating the deployment process and ensuring that updates to our infrastructure and applications could be rolled out smoothly.
  7. Monitoring and logging were established using Amazon CloudWatch and AWS CloudTrail, providing real-time insights and audit trails for all activities within our environment.
  8. Security best practices were followed, including the use of IAM roles and policies to control access, as well as encryption of data at rest and in transit to protect our sensitive information.

This infrastructure is pivotal to our project, allowing us to effectively utilize big data technologies to derive valuable insights from the Formula 1 dataset. The design prioritizes scalability, efficiency, and cost-effectiveness, ensuring it meets the high demands of large-scale data processing and analysis. Ultimately, this setup empowers us to make data-driven decisions that enhance our understanding and strategic planning within the context of Formula 1.

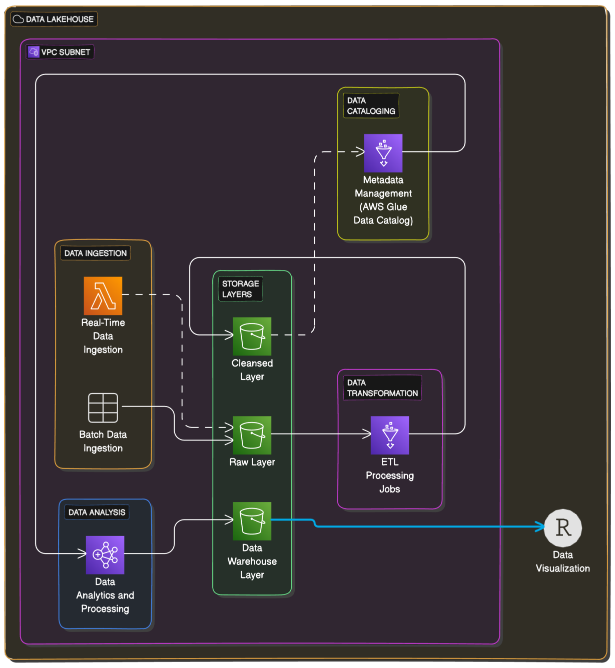

Figure 1: **Architecture Diagram**

3 Data Analysis with F1 Dataset

```
# Placeholder for Python code
```

3.1 Query Analysis and Insights

3.1.1 Average Number of Laps per Grand Prix

Description: The average number of laps per Grand Prix provides insights into race length and potential strategy implications for teams.

Figure 2: Number of Laps GCP Query Result

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

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 Infrastructure-Related Challenges

Description: Some of the challenges faced were related to the setup and configuration of the AWS services, ensuring efficient data flow between different layers, and maintaining security and compliance standards.

4.2 Query and Code-Related Challenges

Description: Challenges included optimizing queries to handle large datasets efficiently and debugging code to ensure accurate analysis results.

5 Conclusions

Through the utilization of big data technologies such as AWS and the implementation of a robust data lakehouse architecture, this project has provided significant insights into the Formula 1 dataset. 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:

- Insights into race strategies based on the number of laps per Grand Prix.
- Data-driven decisions for tire management and fuel strategies.

Future Directions: Future work could involve incorporating additional data sources, such as weather data, to enhance the analysis and insights.

Final Thoughts: This project demonstrates the power of big data technologies in sports analytics, providing a framework for future analysis and strategic planning in Formula 1.

6 Project Repository

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

7 References

References

[1] Author, *Title of the paper*, Journal Name, Volume, Year.