Real-Time Soccer Analytics Pipeline - Project Documentation

# 1. Introduction

Objective: Explain the goal of the project—to create a real-time analytics pipeline tracking player movements and match events.

Technologies Used: Apache Kafka, Apache Spark Streaming, Google BigQuery/AWS Redshift, dbt, Google Data Studio/Tableau Public.

# 2. Data Ingestion

Source of Data: Describe where you get the soccer data (e.g., Football API or historical datasets).

Kafka Setup: Explain how Kafka is configured to stream real-time match events and player movements.

Simulating Data: Describe the Python script used to simulate real-time data streams of player movements and match events.

Screenshots/Code: Include relevant code snippets and screenshots of Kafka setup.

# 3. Data Processing

Spark Streaming Setup: Explain how Apache Spark Streaming is used to process and clean the data in real time.

Aggregation: Describe any data aggregations or transformations (e.g., calculating player distances, match events).

Screenshots/Code: Include relevant code and screenshots.

# 4. Data Storage & Transformation

Data Storage: Explain how the processed data is stored in Google BigQuery or AWS Redshift.

Partitioning Strategy: Describe how data is partitioned (by match, team, player) for optimized queries.

dbt for Transformation: Explain how dbt is used to transform the stored data for specific metrics.

Screenshots/Code: Include code for SQL queries and dbt transformations.

# 5. Data Visualization

Google Data Studio/Tableau: Describe the dashboards created, such as player heatmaps, match events, and real-time stats.

Screenshots: Include visual examples of dashboards.

# 6. Challenges & Solutions

Challenges: Discuss any technical challenges (e.g., managing large data volumes, real-time latency).

Solutions: Explain how you solved them, for example, by tuning Kafka or optimizing Spark jobs.

# 7. Conclusion

Results: Summarize the key insights generated from the data.

Lessons Learned: Reflect on what you learned from building the pipeline.

Future Improvements: Suggest future enhancements (e.g., additional metrics, integrating player tracking data).