Project report 1 for Data Engineering course

Members: Albert Unn (33%), Karen Roht(33%), Jasper Luik(33%)

Business brief

Objective: Investigate how weather conditions influence football match outcomes.

Stakeholders: Football coaches and players (to adapt strategies for different weather conditions), sports betting sites (to refine odds and risk assessment, improve predictive models), fans and viewers (to understand how weather affects game experience, attendance).

Key Metrics (KPIs): Goals scored per game, winrate, ball possession percentage.

Business Questions: Is the variability of match outcomes higher in extreme weather conditions (e.g, very hot vs. very cold, strong wind, heavy rain)? Are certain teams more resilient to difficult weather conditions? How much does attendance depend on the weather conditions? Are home teams more/less affected by extreme weather? Does the number of fouls committed depend on weather conditions? Does bad weather impact penalty shootouts?

Datasets

1. ESPN Soccer data
   - contains 30 000+ soccer match data for the 2024-2025 season.
   - updated daily
   - source: https://www.kaggle.com/datasets/excel4soccer/espn-soccer-data/data.

2. Weather data
   - contains historical and real-time weather parameters. Data can be queried based on match date and location.
   - updated hourly
   - source: open-meteo.com.

Tooling choice: which tools you would use per lifecycle stage (no coding needed yet). (Karen)

1. Data ingestion: our ESPN data is updated daily so a batch-based data ingestion is used.
   - Airflow to orchestrate ingestion workflows and schedule daily batch jobs.
   - Python to implement the actual ingestion tasks: downloading daily CSVs from Kaggle and fetching weather data from Open Meteo API. Python scripts are also used to perform initial cleaning (if needed) as the data is loaded into the staging schema.

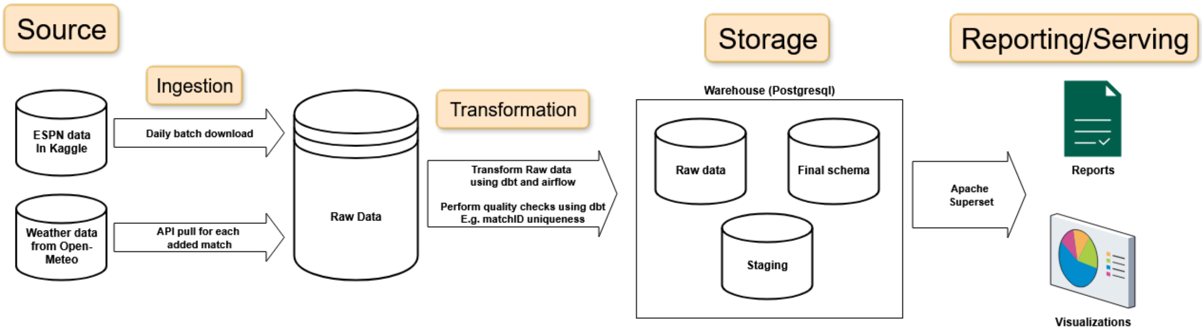
2. Data transformation: after ingestion, raw data is cleaned and transformed.
    - dbt to perform transformations on the data: reads data from the staging area and executes SQL code to build and update the final dimension and fact tables. This also includes running dbt tests for quality checks.
    - Airflow to orchestrate the pipeline to ensure transformations occur after ingestion and before data is loaded into the warehouse.

3. Data storage: data is stored across multiple layers of the data architecture to separate raw and transformed data.
    - PostgreSQL acts as the primary data warehouse. The storage will be partitioned into schemas: raw (source data), staging (cleaned, standardised data), and the final schema containing the dimensional model.

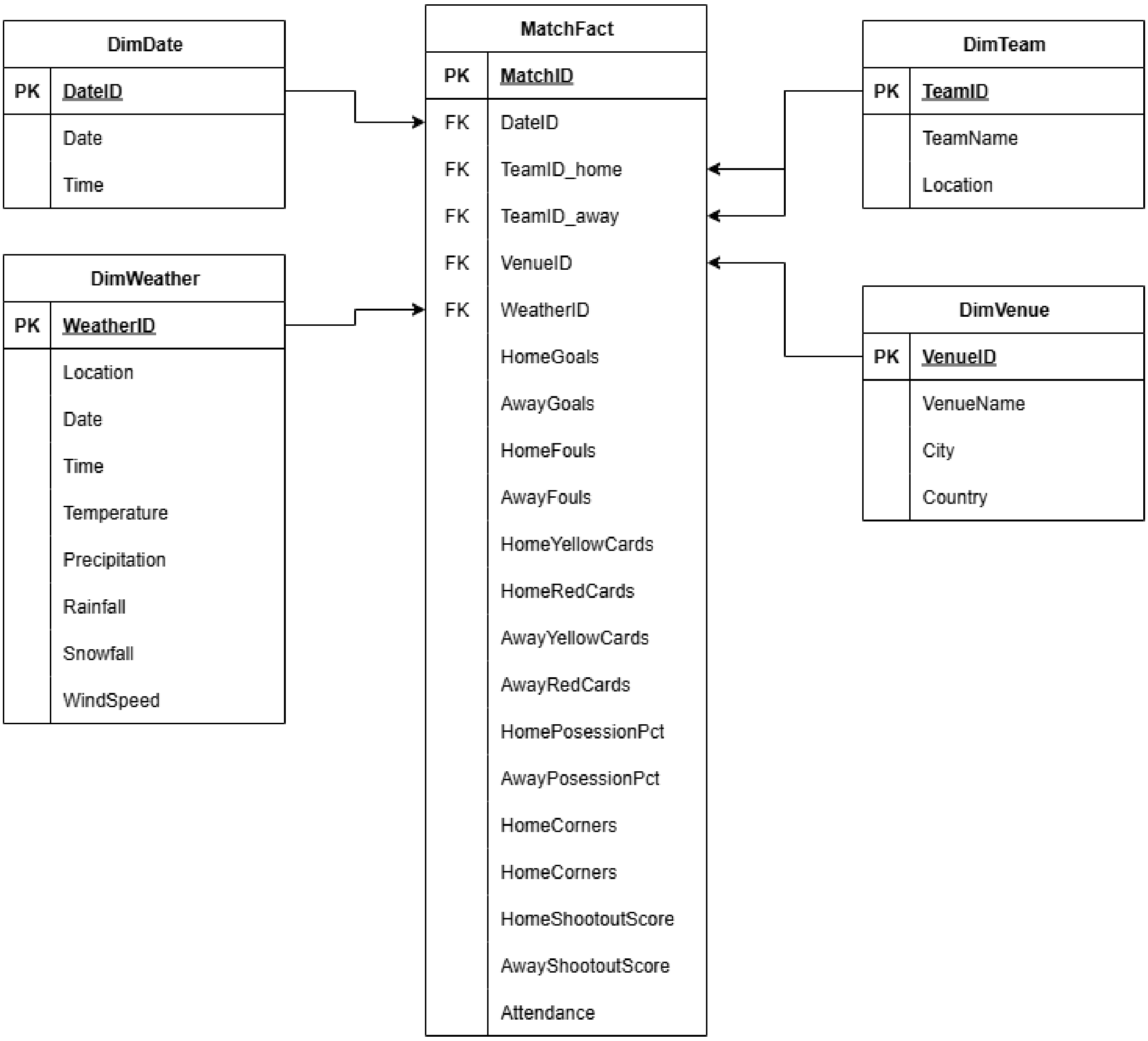
4. Data serving:
    - SQL to run custom queries.
    - Apache Superset connects to the PostgreSQL dimensional schema, providing an interactive platform for users to explore the data.

5. Development: these tools manage the development environment:
    - Docker to create isolated environments for all components.
    - Github to manage team collaboration and track changes.

Data architecture diagram: flow, ingestion method, update frequency, ≥1 data-quality check. (Albert)



Dimensional model: star schema with grain, ≥1 fact + ≥3 dims; justify SCD type per dim.(Jasper)



Grain = one row per match

SCD types

Dimension Table	SCD Type	Justification
DimDate	Static	The date dimension table contains fixed calendar data about match dates and times, which do not change over time.
DimWeather	Static	Each weather record corresponds to a specific date and time at a fixed location. Since historical weather conditions do not change after being recorded, this dimension remains static and requires no modifications.
DimTeam	Type 2	Team attributes like name or location can change over

Dimension Table	SCD Type	Justification
		time (e.g., due to rebranding or relocation). To ensure that past matches remain linked to the team attributes at the time of the match, a Type 2 dimension is used.
DimVenue	Type 1	Although venue names may occasionally change, historical venue details are not critical for match analysis. Therefore, overwriting existing records is an appropriate approach.

Simple data dictionary: tables, columns, data types. (Karen)

Table 1. MatchFact. The central fact table storing football events. Each row represents a single match, linked to weather, data, teams, and venue dimensions.

Variable	Type	Description
MatchID	SERIAL	Unique identifier for each match.
DateID	INT	Foreign key referencing the ‘DimDate’ table.
TeamID_home	INT	Foreign key referencing the ‘DimTeam’ table. Identifies the home team - the team hosting the match at their home venue.
TeamID_away	INT	Foreign key referencing the ‘DimTeam’ table. Identifies the away team - the team visiting the home team’s venue.
VenuneID	INT	Foreign key referencing the ‘DimVenue’ table.
WeatherID	INT	Foreign key referencing the ‘DimWeather’ table.
HomeGoals	INT	Number of goals scored by the home team.
AwayGoals	INT	Number of goals scored by the away team.
HomeFouls	INT	Number of fouls committed by the home team.
AwayFouls	INT	Number of fouls committed by the away team.
HomeYellowCards	INT	Number of yellow cards issued to home team players.
HomeRedCards	INT	Number of red cards issued to home team players.
AwayYellowCards	INT	Number of yellow cards issued to away team players.
AwayRedCards	INT	Number of red cards issued to away team players.

Variable	Type	Description
HomePossessionPct	DECIMAL (5,2)	Percentage of ball possession controlled by the home team.
AwayPossessionPct	DECIMAL (5,2)	Percentage of ball possession controlled by the away team.
HomeCorners	INT	Number of corner kicks awarded to the home team.
AwayCorners	INT	Number of corner kicks awarded to the away team.
HomeShootoutScore	INT	Number of goals scored by the home team during a penalty shootout.
AwayShootoutScore	INT	Number of goals scored by the away team during a penalty shootout.
Attendance	INT	Number of spectators at the match.

Table 2. DimDate. A dimension table that stores calendar information about match dates and times.

Variable	Type	Description
DateID	SERIAL	Unique identifier for each date.
Date	DATE	Calendar date.
Time	TIME	Match start time.

Table 3. DimWeather. Holds historical weather information for watch match location and time.

Variable	Type	Description
WeatherID	SERIAL	Unique identifier for each weather record.
Location	VARCHAR	City or venue location where the match was played.
Date	DATE	Calendar date of the weather observation.
Time	TIME	Exact time of the weather observation.
Temperature	DECIMAL(5,2)	Air temperature (°C) measured at the given date and time.
Precipitation	DECIMAL(5,2)	Precipitation amount (mm) recorded during the observation period.
Rainfall	DECIMAL(5,2)	Rainfall amount (mm) measured at the location and

Variable	Type	Description
		time.
Snowfall	DECIMAL(5,2)	Snowfall amount (mm) measured at the location and time.
WindSpeed	DECIMAL(5,2)	Wind speed (m/s) at the location and time.

Table 4. DimTeam. Stores detailed information about all soccer teams.

Variable	Type	Description
TeamHomeID	SERIAL	Unique identifier for the team.
TeamName	VARCHAR	Full name of the home team.
Location	VARCHAR	City where the team is based.

Table 5. DimVenue. A dimension table that holds information about stadiums and locations where matches are played.

Variable	Type	Description
VenueID	SERIAL	Unique identifier for each venue.
VenueName	VARCHAR	Name of the venue.
City	VARCHAR	City of the venue.
Country	VARCHAR	Country of the venue.

Demo SQL: queries that answer the business questions using your star schema. (Albert)
Queries located in repository in “SQL queries folder”.