UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

SoccerSense⚽

A Unified Framework for Football Data Collection,Storage,and Analysis

Supervisor: **Sergi Nadal, Achraf Hmimou**
Student: Jintao Ma

Content

- Context
- Data Sources and Ingestion
- Landing Zone
- Trust Zone
- Exploitation Zone
- Consumption Zone

Motivation

In modern professional football, **data-driven** decision-making has become a key to success.

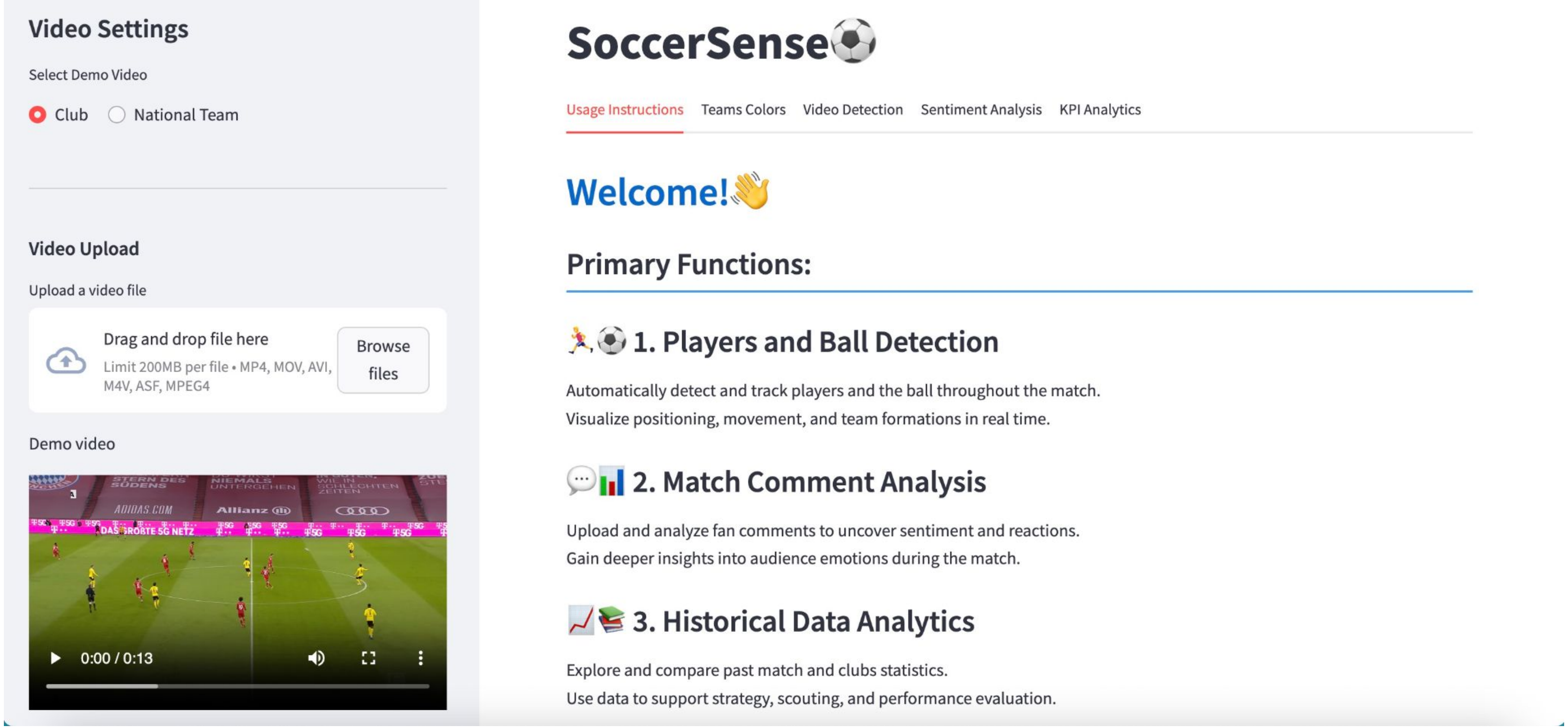
- Explosive growth in **data volume**
- Widespread adoption of game **video analytics**
- Rising influence of **social media**
- **Cross-disciplinary** integration

Objectives

SoccerSense emerges in response to these trends, aiming to build a **future-oriented**, **end-to-end** football analytics platform with the following goals:

- Integrate heterogeneous **multi-source** data
- Build a big **data processing** architecture
- Enable **automated analysis** and real-time visualization
- Support coaches and **analysts**

My Product / Demo



Video Settings

Select Demo Video

● Club ○ National Team

Video Upload

Upload a video file

Drag and drop file here

Limit 200MB per file • MP4, MOV, AVI, M4V, ASF, MPEG4

Demo video

SoccerSense⚽

Welcome!👋

Primary Functions:

🏃⚽ 1. Players and Ball Detection

Automatically detect and track players and the ball throughout the match.
Visualize positioning, movement, and team formations in real time.

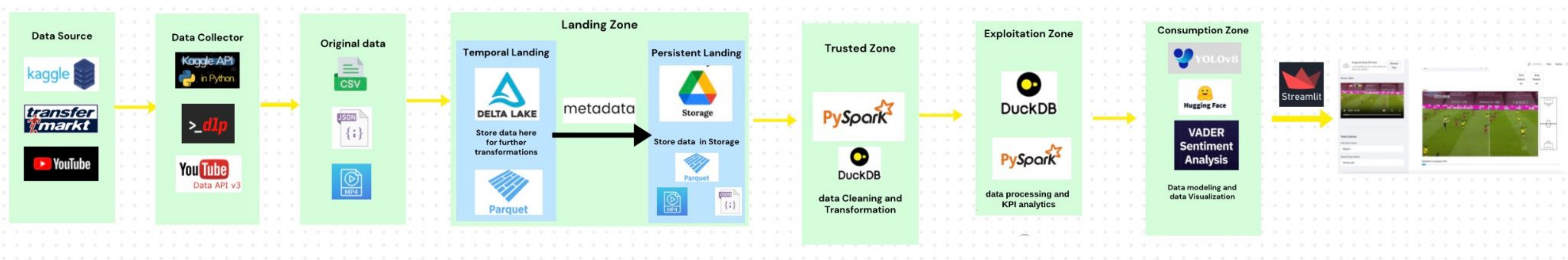
💬📊 2. Match Comment Analysis

Upload and analyze fan comments to uncover sentiment and reactions.
Gain deeper insights into audience emotions during the match.

📈📚 3. Historical Data Analytics

Explore and compare past match and clubs statistics.
Use data to support strategy, scouting, and performance evaluation.

Big Data Architecture



Data Ingestion: Collecting and importing raw data from diverse sources into a pipeline

Landing Zone: Storing raw ingested data in its original format without modification.

Trusted Zone: General data cleaning and transformation to ensure consistency and reliability.

Exploitation Zone: Adding computed as sets like KPIs —to support efficient and flexible analysis.

Consumption Zone: Utilizing data by connecting it to downstream systems—such as models, dashboards, or simulations.

Data Sources

- **Structured data（CSV）**

- Kaggle / Transfermarkt

- **Semi-structured（JSON Comment）**

- YouTube Comment Section

- **Unstructured: （MP4）**

- YouTube videos

Data Source

Data Ingestions

- **Structured data（CSV）**

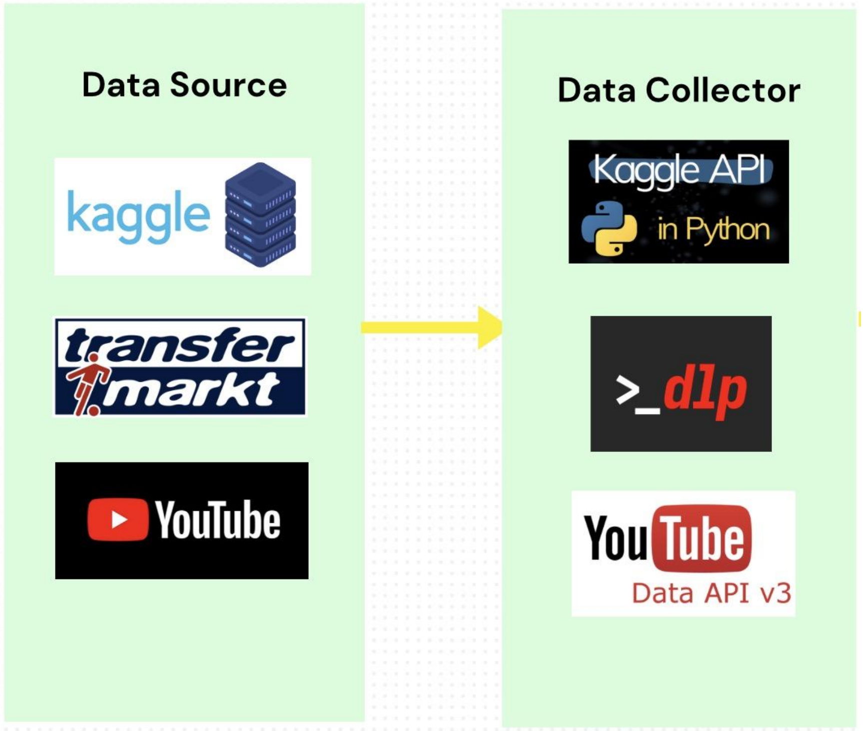
  - Use the **Kaggle API** and **web scraper** to obtain structured CSV data

- **Semi-structured（JSON Comment）**

  - Use the YouTube **Data API v3** to collect comments and save them as JSON files

- **Unstructured:（MP4）**

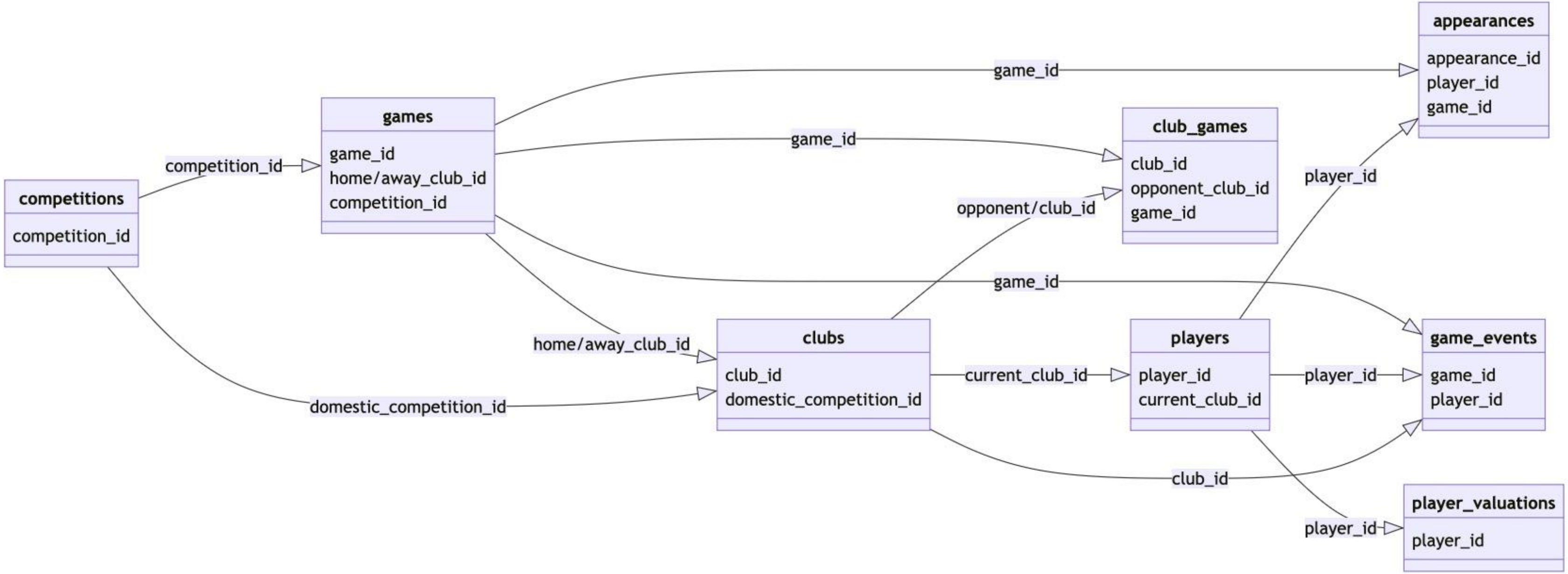
  - Use **t-dlp** to download YouTube MP4 videos



Data Source

Data Collector

CSV Data



competitions
competition_id

games
game_id
home/away_club_id
competition_id

clubs
club_id
domestic_competition_id

club_games
club_id
opponent_club_id
game_id

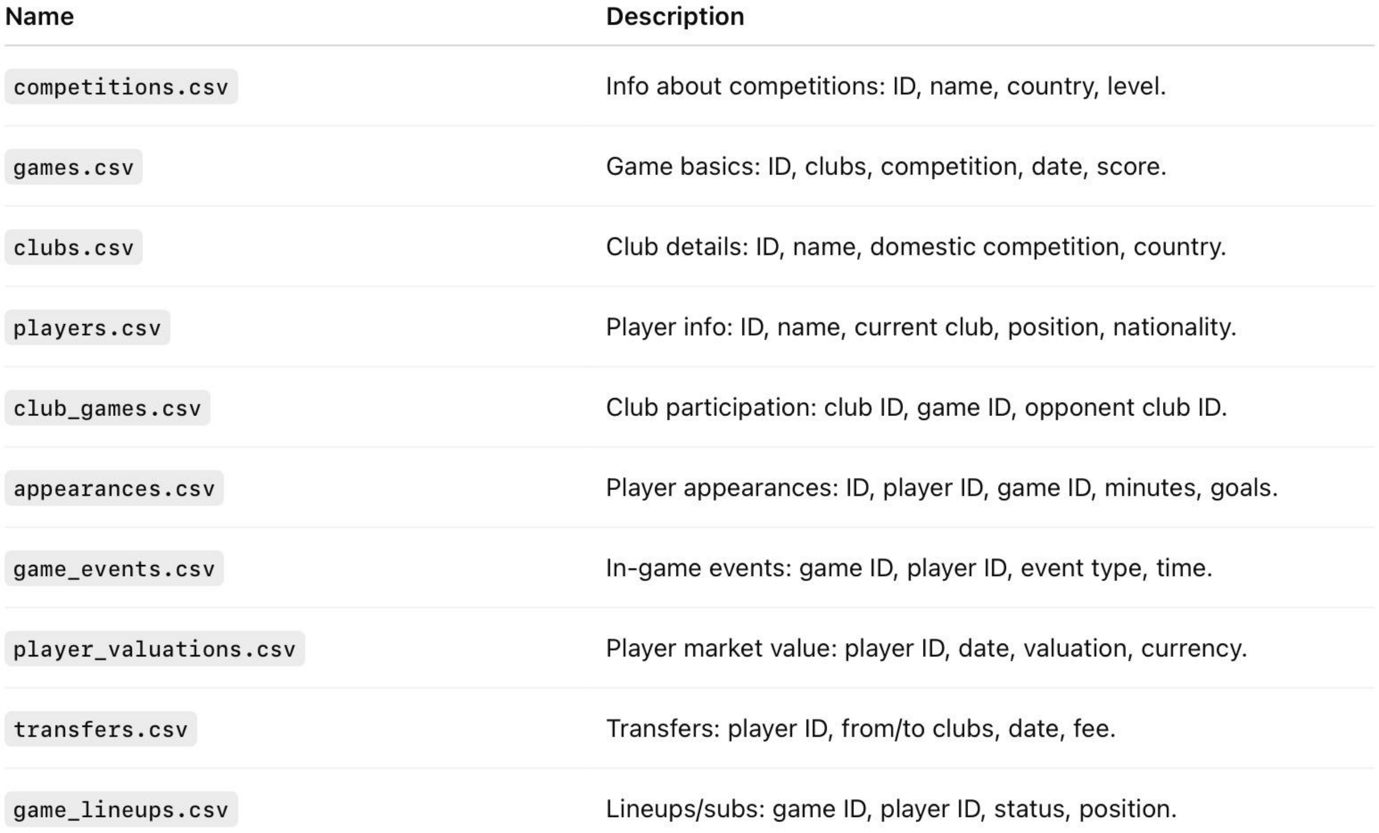
players
player_id
current_club_id

appearances
appearance_id
player_id
game_id

game_events
game_id
player_id

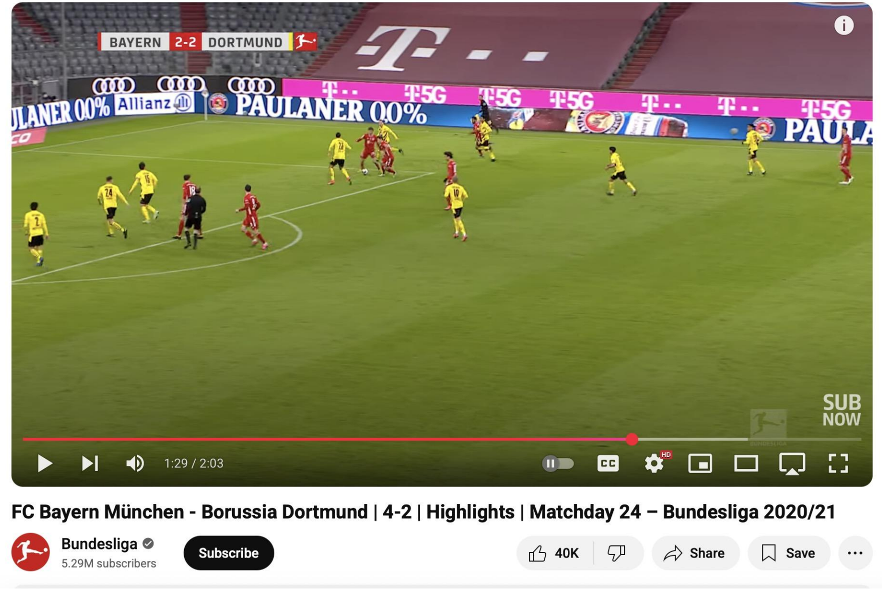
player_valuations
player_id

CSV Data



Name	Description
`competitions.csv`	Info about competitions: ID, name, country, level.
`games.csv`	Game basics: ID, clubs, competition, date, score.
`clubs.csv`	Club details: ID, name, domestic competition, country.
`players.csv`	Player info: ID, name, current club, position, nationality.
`club_games.csv`	Club participation: club ID, game ID, opponent club ID.
`appearances.csv`	Player appearances: ID, player ID, game ID, minutes, goals.
`game_events.csv`	In-game events: game ID, player ID, event type, time.
`player_valuations.csv`	Player market value: player ID, date, valuation, currency.
`transfers.csv`	Transfers: player ID, from/to clubs, date, fee.
`game_lineups.csv`	Lineups/subs: game ID, player ID, status, position.

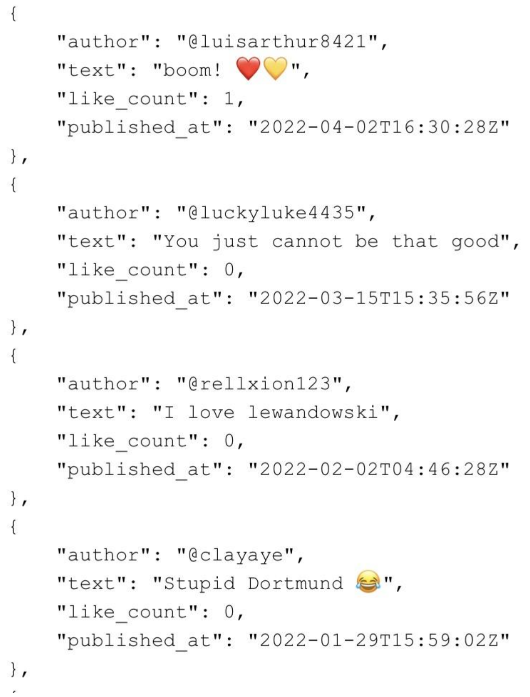
Data Ingestions



FC Bayern München - Borussia Dortmund | 4-2 | Highlights | Matchday 24 – Bundesliga 2020/21

Bundesliga

5.29M subscribers



```
{
    "author": "@luisarthur8421",
    "text": "boom! ❤️💛",
    "like_count": 1,
    "published_at": "2022-04-02T16:30:28Z"
},
{
    "author": "@luckyluke4435",
    "text": "You just cannot be that good",
    "like_count": 0,
    "published_at": "2022-03-15T15:35:56Z"
},
{
    "author": "@rellxion123",
    "text": "I love lewandowski",
    "like_count": 0,
    "published_at": "2022-02-02T04:46:28Z"
},
{
    "author": "@clayaye",
    "text": "Stupid Dortmund 😂",
    "like_count": 0,
    "published_at": "2022-01-29T15:59:02Z"
},
```

- **Video Data**

- **JSON Data**

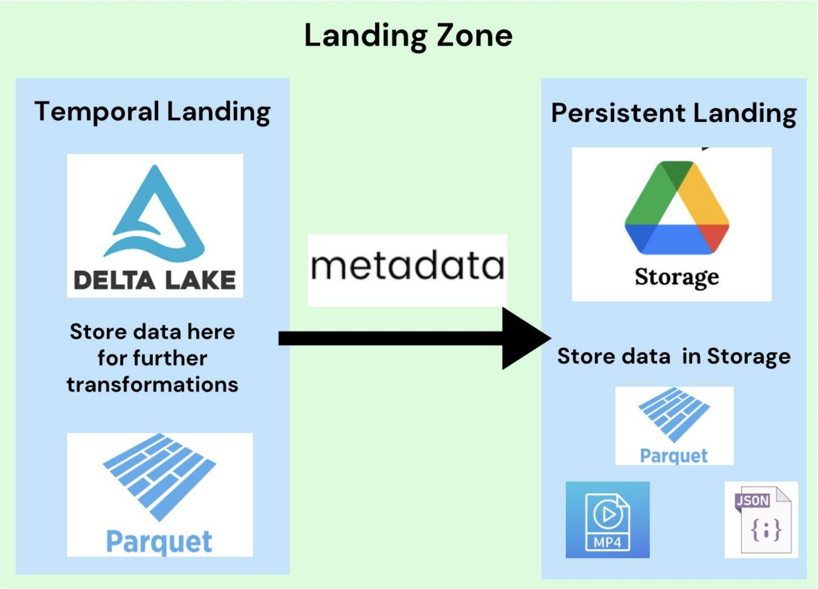
Landing Zone

Temporal Landing Zone

- **Purpose**: Temporary storage & quick preprocessing of raw data
- **Tools**: `Delta Lake`
- **Advantages**:
  - Fast access with minimal impact on the source system
  - ACID transactions & schema enforcement
  - Data saved in Parquet format for efficient querying

Persistent Landing Zone+Meta Data Management

- **Purpose**: Long-term storage of cleaned data, ensuring version cont and traceability
- **Tools**: Delta Lake metadata management + JSON tracking file
- ✅ **Metadata Fields**:
  - File name, column names, row count
  - Ingestion/last modified timestamps
  - Delta storage path
- 📂 **Metadata Format**: `metadata.json` stores detailed information for each ingested CSV file



Landing Zone

Temporal Landing

Store data here for further transformations

metadata

Persistent Landing

Store data in Storage

Trusted Zone

- ## For Structured data(CSV)

- Loaded datasets: `players`, `games`, `clubs`, etc. via **PySpark**
- Removed nulls & duplicates
- Validated **key relationships** (e.g., player IDs across tables)
- Saved as **Parquet**, imported into **DuckDB** for SQL queries

- ## For Unstructured data:(Youtube MP4)

- Extracted **duration**, **resolution**, and **read status** with **OpenCV + PySpark**
- Created Spark RDDs → DataFrames
- Integrated video metadata for future tasks (e.g., object detection)

- ## For Semi-structured data(YouTube Comments JSON)

- Parsed with **PySpark**, explored schema
- Key fields: `author`, `text`, `likes`
- Ready for **sentiment** and **trend analysis**

Trusted Zone

PySpark

DuckDB

Exploitation Zone

- ## KPI Generation

- Read cleaned data from **Trusted Zone** (appearances, games, clubs)
- Generated new KPI tables:

- Player **Performance Scores**
- Club **Win Rates**
- Player **Contribution Rates**

- Saved as **Parquet files** for efficient access

- ## Visualization

- Used **Matplotlib** for visual insights:
  - ○ 🔝 Top Goal Scorers
  - ○ 🏆 Clubs with Highest Win Rates
  - ○ 💪 Most Valuable Players (MVPs)
  - ○ Conducted Focused analysis (e.g., Bayern, Bundesliga)

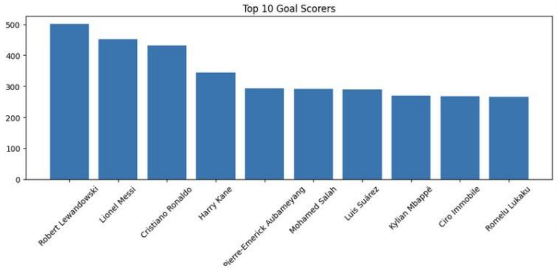
Exploitation Zone

DuckDB

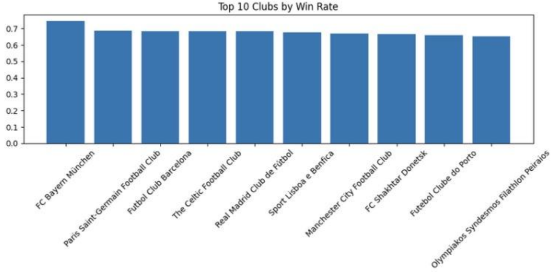
PySpark

data processing and KPI analytics

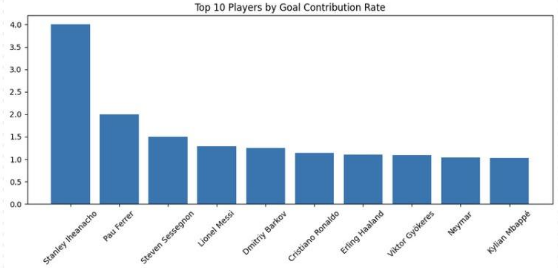
KPI Analytics



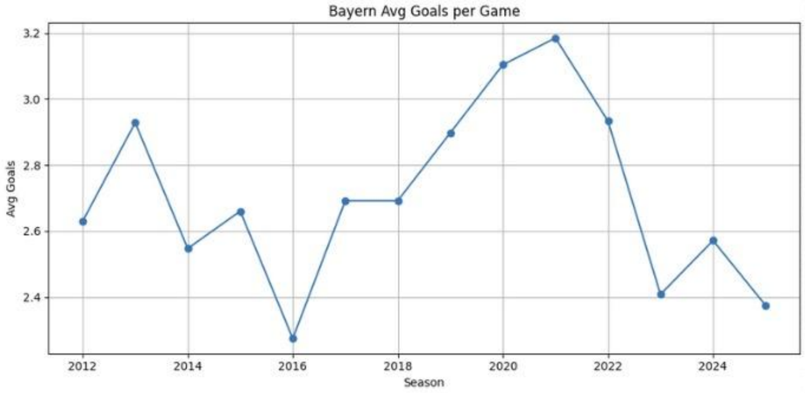
Top 10 Goal Scorers



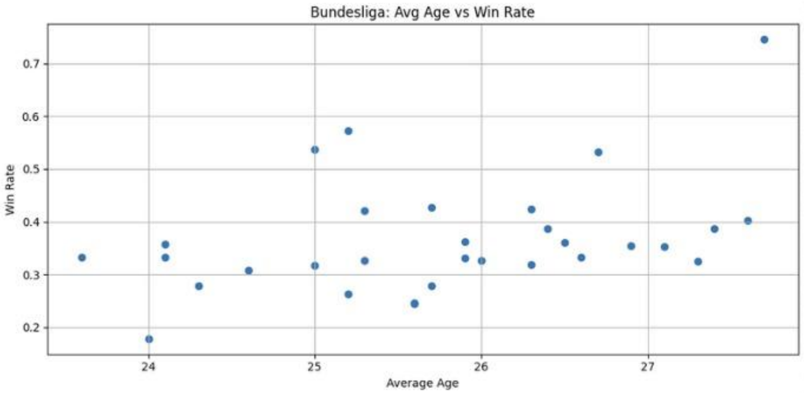
Top 10 Clubs by Win Rate



Top 10 Players by Goal Contribution Rate



Bayern Avg Goals per Game



Bundesliga: Avg Age vs Win Rate

Comsuption Zone

Real-time Video Detection(YouTube Videos)

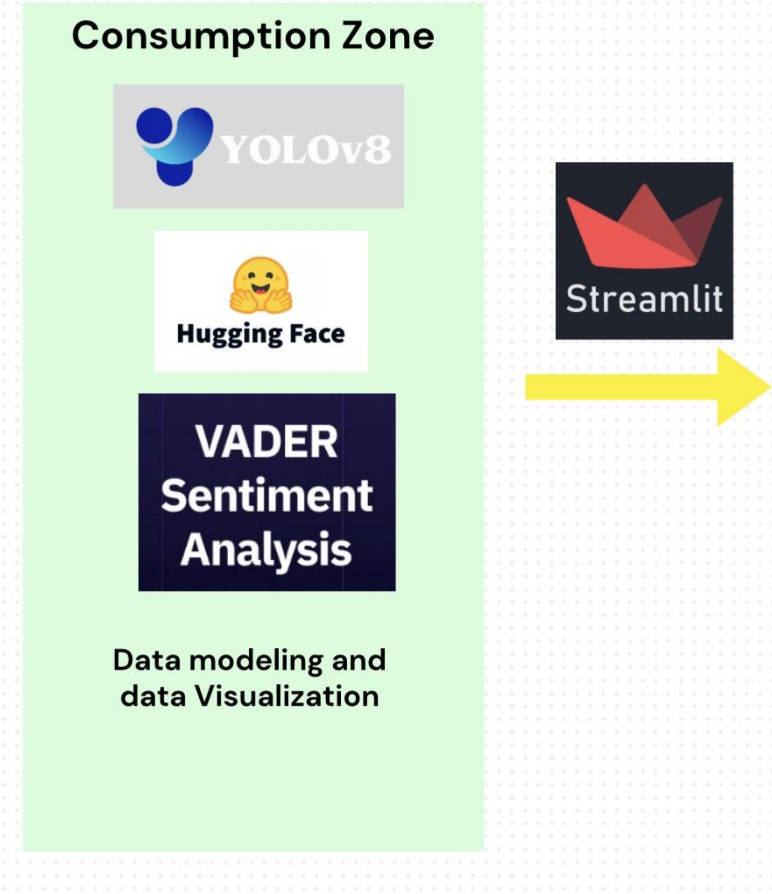
- 📌 Implemented using **YOLOv8** (players + field keypoints)
- 📂 Supports **video upload** or use of default clips
- 🎯 Outputs: Player bounding boxes, ball trajectories
- 🧠 Enables **tactical understanding** and **positioning analysis**

Sentiment Analysis (YouTube Comments)

- 🧪 Used **VADER** for emotion classification
- 👍 Categorized as **positive**, **negative**, or **neutral**
- 📊 Visualized using **pie charts**
- 📈 Understand **fan reactions**, brand perception, and emotional trends

Interactive Dashboard

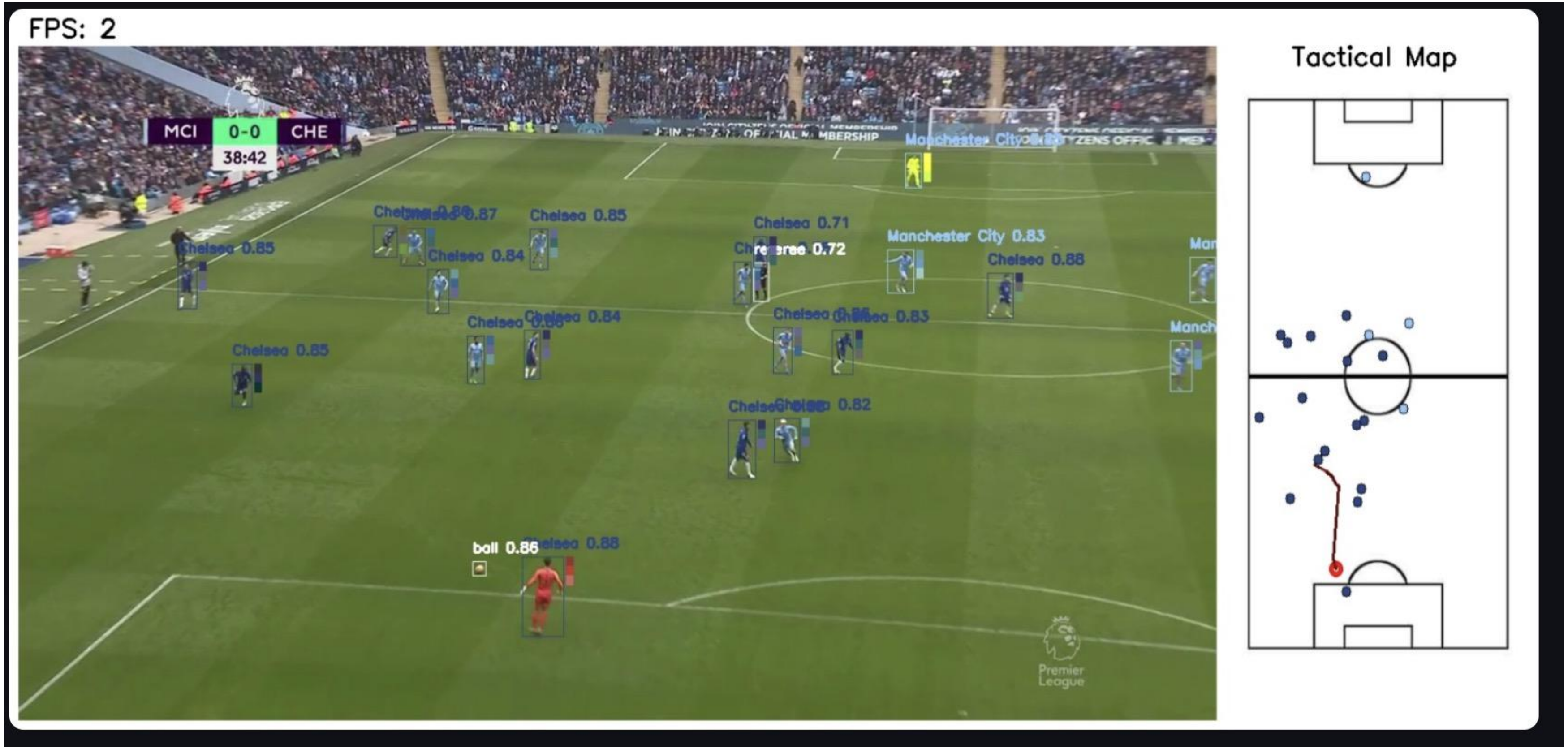
- ⚙️ Built using **Streamlit** + **DuckDB**
- 🧩 Query and visualize **KPI data** (e.g., player performance, club stats)
- 📉 Provides **actionable insights** with a user-friendly frontend
- 🔍 Useful for **coaches, analysts, and scouts**



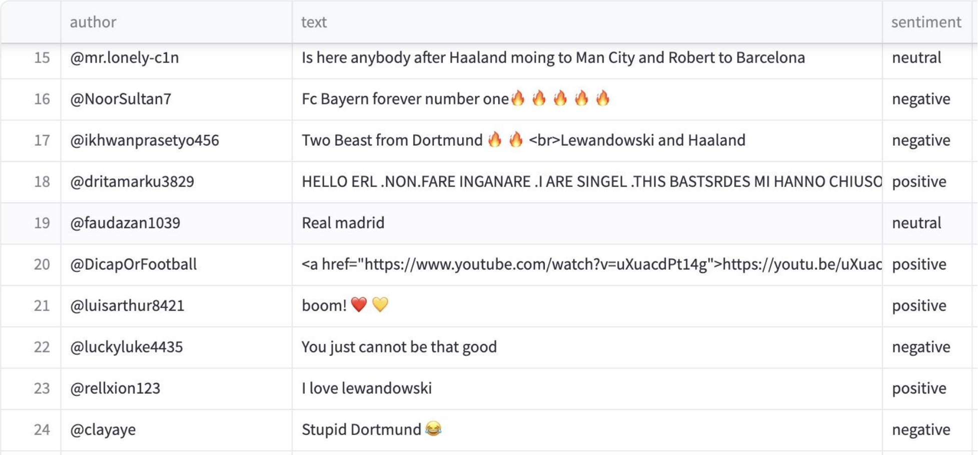
Consumption Zone

Data modeling and data Visualization

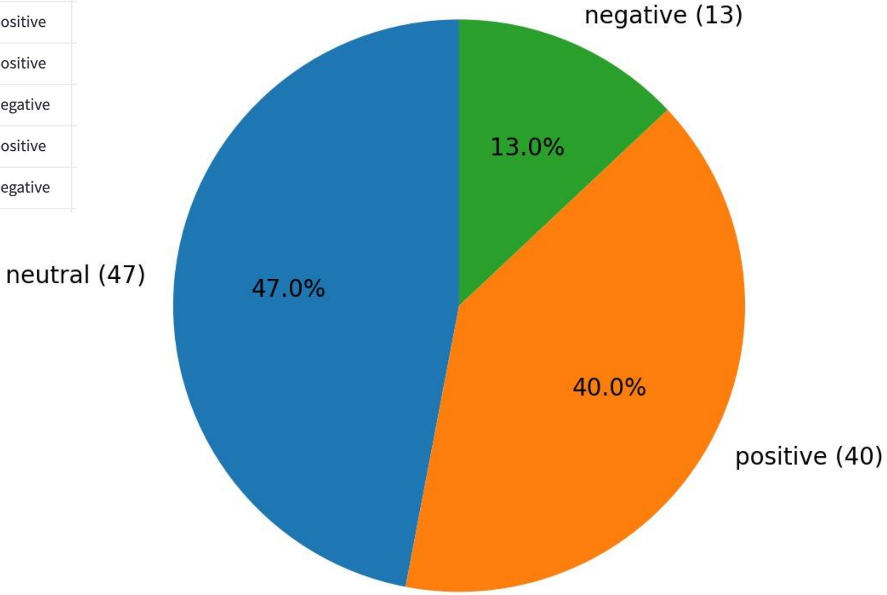
Video Detection



Sentiment Analysis



	author	text	sentiment
15	@mr.lonely-c1n	Is here anybody after Haaland moing to Man City and Robert to Barcelona	neutral
16	@NoorSultan7	Fc Bayern forever number one🔥 🔥 🔥 🔥 🔥	negative
17	@ikhwanprasetyo456	Two Beast from Dortmund 🔥 🔥 Lewandowski and Haaland	negative
18	@dritamarku3829	HELLO ERL .NON.FARE INGANARE .I ARE SINGEL .THIS BASTSRDES MI HANNO CHIUSO	positive
19	@faudazan1039	Real madrid	neutral
20	@DicapOrFootball	<a href="https://www.youtube.com/watch?v=uXuacdPt14g">https://youtu.be/uXuac	positive
21	@luisarthur8421	boom! ❤️ 💛	positive
22	@luckyluke4435	You just cannot be that good	negative
23	@rellxion123	I love lewandowski	positive
24	@clayaye	Stupid Dortmund 😂	negative



Justification of 3 Important Tools

Delta Lake

- **ACID Transactions**: Guarantees reliable insert/update/delete
- **Data Versioning**: Supports **time travel** & rollback via Parquet

PySpark

- **Scalable Processing**: Distributed handling of large football datasets
- **Data Transformation**: Enables joins & aggregations for **KPI extraction**

DuckDB

- **In-Memory SQL Engine**: Lightweight, no server needed
- **Dashboard Integration**: Seamless with **Python + Streamlit** for visual insights

UPC

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Thanks!