1. Project Statement, Why you have chosen this project (Motivation, Importance, etc), and for whom you are building this II system (set of close-to-real applications, use-cases, stakeholders, and purpose), and focus on innovative application(s)/use-case(s)

The "Data Integration for Sports Analytics" project aims to enhance sports decision-making by integrating diverse data sources like player stats and match results. It provides actionable insights for improving player performance, aiding coaches in strategy decisions, and uncovering hidden trends. Key stakeholders include sports teams and organizations, performance tracking and tactical analysis. The goal is to revolutionize sports analytics through a cutting-edge, comprehensive system.

2. Defining the underlying data sources (Type of sources - relational/non-relation, etc, Schemas and/or APIs (input and output). Populating the data in these sources, what query execution support they have, how do they communicate with the Data Integration system (i.e. communication protocol)
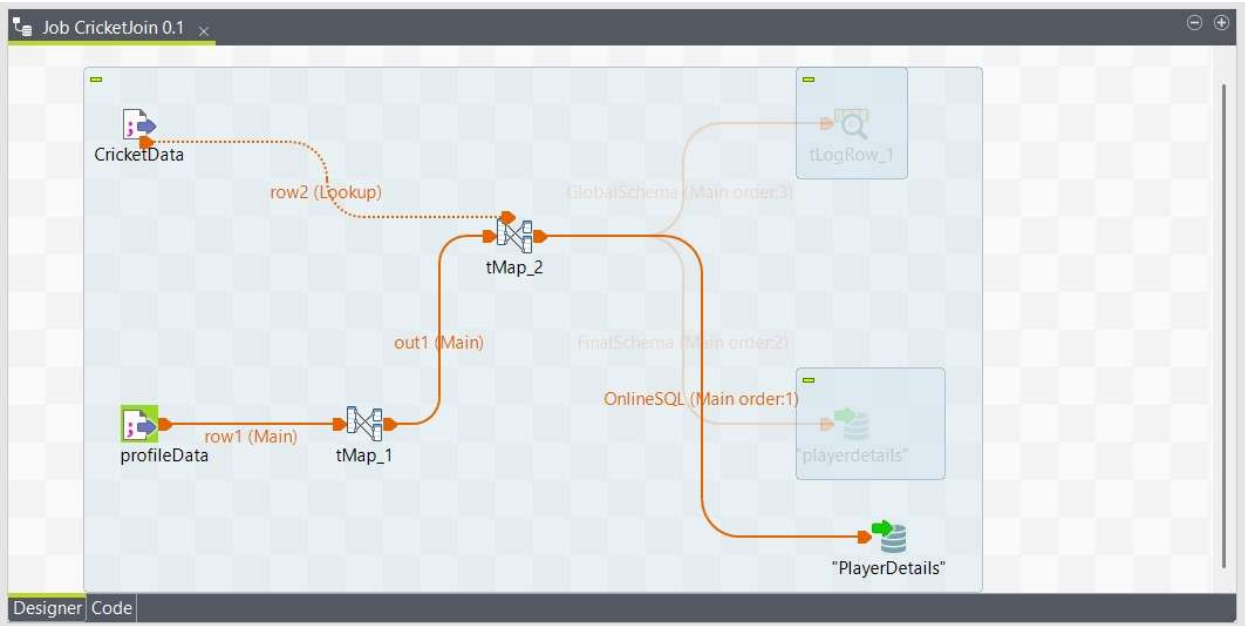
cricket.csv file schema

fullname	battingstyle	bowlingstyle	position

Profile_data.csv file schema

id	firstname	lastname	Image_path	dateofbirth	gender	Continent_id	Continent_name	Country_id	Country_name	Country_image_path

We have integrated both the tables using cricket.fullname = profile_data.firstname +" "+profile.lastname

Relational Database: CREATE TABLE cricket_data.PlayerDetails (

    PlayerName VARCHAR(100) NOT NULL,

    image_path VARCHAR(255),

    dateofbirth DATE,

    gender CHAR(1),

    country_name VARCHAR(100),

    country_image_path VARCHAR(255),

    battingstyle VARCHAR(100),

    bowlingstyle VARCHAR(100),

    position VARCHAR(50)

);

PlayerName	image_path	dateofbirth	gender	country_name	country_image_path	battingstyle	bowlingstyle	position
Ahmed Shehzad	https://cdn.sportmonks.com/images/cricket/players/...	1991-11-22	m	Pakistan	https://cdn.sportmonks.com/images/countries/png/sh...	right-hand-bat	legbreak	Batsman
Anwar Ali	https://cdn.sportmonks.com/images/cricket/players/...	1987-11-24	m	Pakistan	https://cdn.sportmonks.com/images/countries/png/sh...	right-hand-bat	right-arm-fast-medium	Bowler
Sarfraz Ahmed	https://cdn.sportmonks.com/images/cricket/players/...	1987-05-21	m	Pakistan	https://cdn.sportmonks.com/images/countries/png/sh...			Allrounder
Azhar Ali	https://cdn.sportmonks.com/images/cricket/players/...	1985-02-18	m	Pakistan	https://cdn.sportmonks.com/images/countries/png/sh...	right-hand-bat	legbreak	Batsman
Fakhar Zaman	https://cdn.sportmonks.com/images/cricket/players/...	1990-04-09	m	Pakistan	https://cdn.sportmonks.com/images/countries/png/sh...			Allrounder
Imam ul Haq	https://cdn.sportmonks.com/images/cricket/players/...	1995-12-11	m	Pakistan	https://cdn.sportmonks.com/images/countries/png/sh...	left-hand-bat		Batsman
Babar Azam	https://cdn.sportmonks.com/images/cricket/players/...	1994-10-14	m	Pakistan	https://cdn.sportmonks.com/images/countries/png/sh...	right-hand-bat	right-arm-offbreak	Batsman
Asad Shafiq	https://cdn.sportmonks.com/images/cricket/players/...	1986-01-27	m	Pakistan	https://cdn.sportmonks.com/images/countries/png/sh...	right-hand-bat	right-arm-offbreak	Batsman

Non-relational database:

Structure

{"_id":{"$oid":"6745e8307256773c9634ede4"},

"playerName":"Deepak Hooda",

"batting":[{"6s":{"$numberInt":"4"},"4s":{"$numberInt":"1"},"runs":{"$numberInt":"41"},"sr":{"$numberDouble":"178.26"},"matchid":"T20I # 1984","matchdate":"Jan 3 2023"},

    {"6s":{"$numberInt":"0"},"4s":{"$numberInt":"0"},"runs":{"$numberInt":"9"},"sr":{"$numberInt":"75"},"matchid":"T20I # 1985","matchdate":"Jan 5 2023"},],

"bowling":[{"overs":"2","wickets":{"$numberInt":"0"},"economy":"7.00","runs":"14","matchid":"T20I # 1990","matchdate":"Jan 27 2023"},

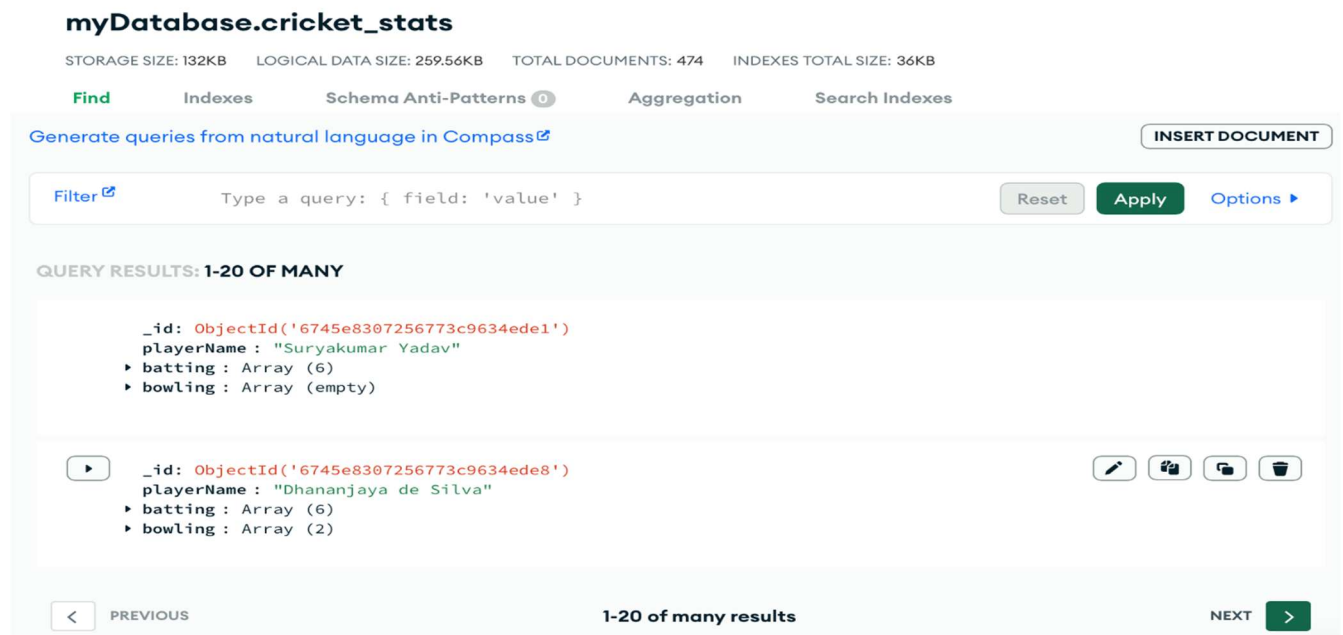   {"overs":"4","wickets":{"$numberInt":"1"},"economy":"4.25","runs":"17","matchid":"T20I # 1991","matchdate":"Jan 29 2023"}]

}

To populate the database we had 2 csv file which we integrated using python script and outputed it in the form of json format

We are using respective queries in databases to fetch the required data.

We assume that they only have fetch support for the querries.

3. Defining the Data Integration System (Target) (Virtual or Materialized or Hybrid or Rule based or ...., data Schemas, ETLing/Ingesting the data from sources (for materialized approach)how it communicates with the underlying data sources)

A Hybrid Approach would be ideal. Materialize critical player details (batting, bowling, career details) in a data warehouse for fast queries and use a virtual approach for frequently changed data like match statistics.

For player details ,since it won't be changing that fast may be once a year so we have added reload function to our backend which we can use whenever needed to fetch all details again.

For player statistics we use non -relational database and it fetches only when required /when the query happens at frontend side.

4. Schema matching and Mapping between Target and Sources

We are mapping on the basis of PlayerDetails.PlayerName=cricket_stats.playerName

```
searchQuery: Ishan Kishan
Connected to MongoDB database! and querying :  Ishan Kishan
Combined data:
  {
  PlayerName: 'Ishan Kishan',
  image_path: 'https://cdn.sportmonks.com/images/cricket/players/18/3122.png',
  dateofbirth: 1998-07-16T18:30:00.000Z,
  gender: 'm',
  country_name: 'India',
  country_image_path: 'https://cdn.sportmonks.com/images/countries/png/short/in.pn
g',
  battingstyle: 'left-hand-bat',
  bowlingstyle: '',
  position: 'Wicketkeeper',
  stats: {
    batting: [ [Object], [Object], [Object], [Object], [Object], [Object] ],
    bowling: []
  }
}
```

5. Design of Query Interface
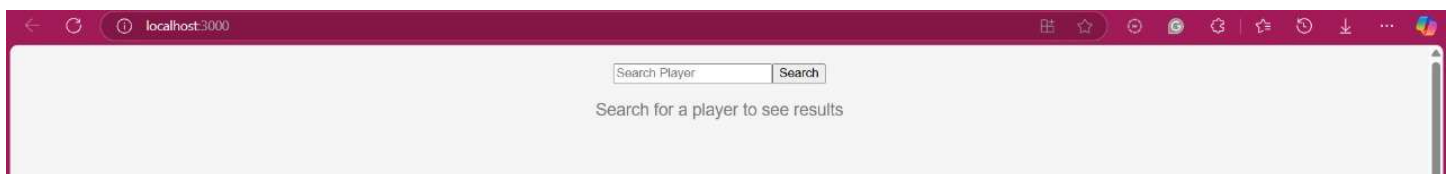
We can search required player details and their stats from here , if its there in database it will be shown .

6. Query Federation, Execution and Results Integration

We make a query using a players name if there is a exact match in database table ,then it shows that persons details otherwise it uses levenshtein Distance to get names closest to it and gives nearest 5 names as suggestions.
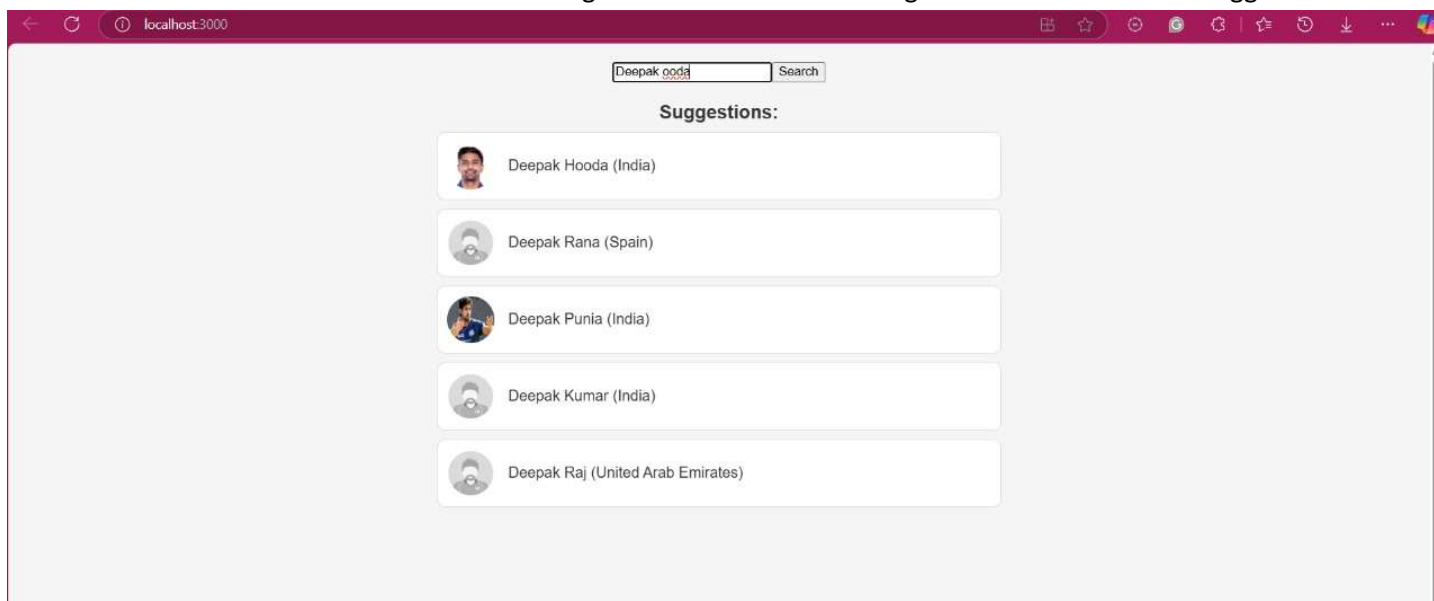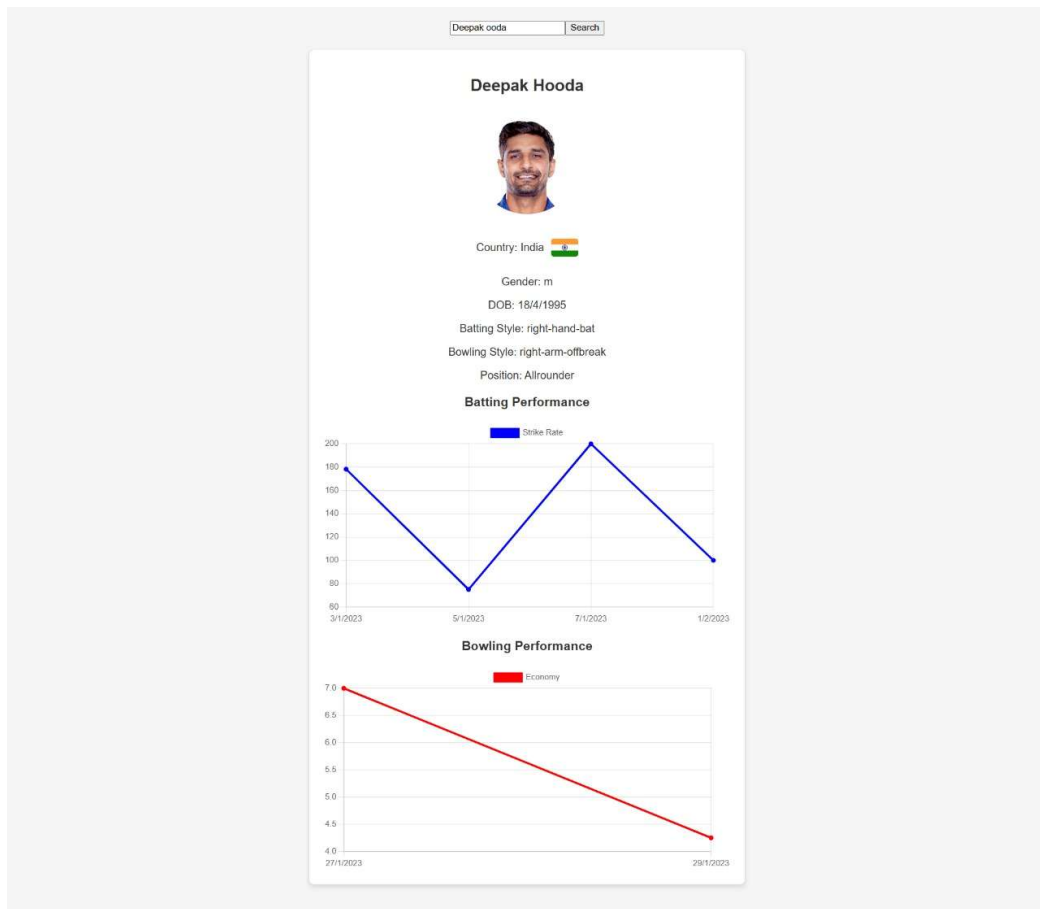
7. Innovation and Scalability

Since streaming platforms and old companies related to this industry have their data and make informed decisions but someone who is outside industry and want to enter it for example they want to make their own team ,they can make informed decisions based on past data .We scaleup our application by adding more data sources not only cricket but also different sports it can also be integrated with ML techniques which may also give us some new insights .