INFORMATION INTEGRATION

Hello!

Team members for group project:

	●●●	D'Alessandro Giammarco	1753102
	●●●	Ionta Antonio	1469982

Abstract

▷ Realize a system for integrating data about football players coming from heterogeneous sources.

▷ Complete Information Integration workflow: GAV mapping with materialization and virtualization approaches for query answering.

3

1.
THE DATA

Soccer Players Performances

Source Datasets

FBref

It contains statistics about real soccer player performances for the Big 5 European leagues, for seasons from 2012/2013 to 2021/2022.

Data comes from FBref.com, famous international Web portal devoted to tracking statistics for football teams and players from around the world.

FM20

It contains fictional player attributes contained in the database of Football Manager 2020 (season 2019/20) videogame.

Football Manager simulation gaming has become particularly interesting also in real soccer scouting activities for its realistic recreation of professional world of football.

https://www.footballmanager.com/features/data-hub

The source datasets come from two distinct kaggle repositories.

They both deals with data about football players, but from two completely different perspectives.

Datasets are heterogeneous in schema and semantics, but not in data format, which is in both cases CSV.

13 CSV files

12 for FBref dataset + 1 for FM dataset

346,797 records

202,047 for FBref dataset + 144,750 for FM dataset

395 attributes

332 for FBref dataset + 63 for FM dataset

?

Is there a sort of relationship between hypothetic player attributes and actual performance on the field?

2.

IIS Formal Definition

Source Schema

Defense$_{/31}$	**Contains FBref data on a player's defensive performance.**
GCA$_{/24}$	**Contains FBref data on a player's goal and shot creation.**
Info$_{/13}$	**Contains FBref general information about a player.**
Keeper$_{/26}$	**Contains FBref data on goalkeeper performance.**
KeeperAdv$_{/33}$	**Contains FBref advanced data for goalkeeper performance.**
Misc$_{/24}$	**Contains FBref miscellaneous player performance data.**
Passing$_{/30}$	**Contains FBref passing data.**
PassingTypes$_{/33}$	**Contains FBref data about players' pass types.**
PlayingTime$_{/29}$	**Contains FBref data about a player's playing time.**
Possession$_{/32}$	**Contains FBref possession data.**
Shooting$_{/25}$	**Contains FBref data about a player's shooting performance.**
Standard$_{/32}$	**Contains an overview of FBref player performance data.**
FM20$_{/64}$	**Contains Football Manager 2020 videogame fictional player attributes.**

Global Schema

Player$_{/13}$	**Contains general information about the players.**
DefenseStat$_{/18}$	**Contains data on a player's defensive performance focusing on most relevant statistics.**
GoalkeeperStat$_{/10}$	**Contains data on goalkeeper performance focusing on most relevant statistics.**
GoalStat$_{/18}$	**Contains data on a player's attacking performance in relation to his ability to score and/or to create opportunities.**
PassingStat$_{/18}$	**Contains passing data focusing on most relevant statistics.**
PlaymakingStat$_{/20}$	**Contains data on a player's performance in relation to his ability to playmake: creating opportunities for the team mates, being at the center of the action.**
PresenceStat$_{/16}$	**Contains data on a player's presence on the field and the corresponding results of his team.**
GoalkeeperAbility$_{/5}$	**Contains fictional player attributes relative to goalkeeper.**
MentalAbility$_{/10}$	**Contains fictional player attributes relative to mental abilities.**
PhysicalAbility$_{/8}$	**Contains fictional player attributes relative to physical abilities.**
TechnicalAbility$_{/14}$	**Contains fictional player attributes relative to technical abilities.**

GAV Mapping – Snippets (1.1/3)

$\forall$

$name, position, foot, height, weight, dob, age, cityob, countryob, squad, value, wage, ca, pa.$

$\exists$

$id,$

$\exists$

i_2, i_{11}, i_{12}

$\exists$

$s_5, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32}$

$\exists$

$f_1, f_2, f_3, f_4, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35},$
$f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43}, f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{49}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64}.$

$Info$

$(id, i2, name, position, foot, height, weight, dob, cityob, countryob, i11, i12, age)$

$\wedge Standard$

$(id,$ **"2019-2020", "ITA", "1. Serie A",**

$s_5, squad, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32})$

$\wedge FM20$

$(f_1, name, f_3, f_4,$ **"Italian Serie A"**, $f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, value, wage, ca, pa,$
$f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35}, f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43},$
$f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{49}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64})$

$\rightarrow Player$

$(name, position, foot, height, weight, dob, age, cityob, countryob, value, wage, ca, pa)$

GAV Mapping – Snippets (1.2/3)

$\forall$

$name, position, foot, height, weight, dob, age, cityob, countryob, squad, value, wage, ca, pa.$

$\exists$

$id,$

$\exists$

i_2, i_{11}, i_{12}

$\exists$

$s_5, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32}$

$\exists$

$f_1, f_2, f_3, f_4, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35},$
$f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43}, f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{49}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64}.$

$Info$

$(id, i2, name, position, foot, height, weight, dob, cityob, countryob, i11, i12, age)$

$\wedge Standard$

$(id, \text{“2019-2020”}, \text{“GER”}, \text{“1. Bundesliga”},$

$s_5, squad, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32})$

$\wedge FM20$

$(f_1, name, f_3, f_4, \text{“Bundesliga”}, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, value, wage, ca, pa,$

$f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35}, f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43},$

$f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{49}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64})$

$\rightarrow Player$

$(name, position, foot, height, weight, dob, age, cityob, countryob, value, wage, ca, pa)$

GAV Mapping – Snippets (1.3/3)

$\forall$

$name, position, foot, height, weight, dob, age, cityob, countryob, squad, value, wage, ca, pa.$

$\exists$

$id,$

$\exists$

i_2, i_{11}, i_{12}

$\exists$

$s_5, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32}$

$\exists$

$f_1, f_2, f_3, f_4, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35},$
$f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43}, f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{49}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64}.$

$Info$

$(id, i2, name, position, foot, height, weight, dob, cityob, countryob, i11, i12, age)$

$\wedge Standard$

$(id, \text{``2019-2020''}, \text{``ESP''}, \text{``1. La Liga''},$
$s_5, squad, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32})$

$\wedge FM20$

$(f_1, name, f_3, f_4, \text{``Spanish First Division''}, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, value, wage, ca, pa,$
$f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35}, f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43},$
$f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{49}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64})$

$\rightarrow Player$

$(name, position, foot, height, weight, dob, age, cityob, countryob, value, wage, ca, pa)$

GAV Mapping – Snippets (1.4/3)

$\forall$

$name, position, foot, height, weight, dob, age, cityob, countryob, squad, value, wage, ca, pa.$

$\exists$

$id,$

$\exists$

i_2, i_{11}, i_{12}

$\exists$

$s_5, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32}$

$\exists$

$f_1, f_2, f_3, f_4, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35},$
$f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43}, f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{49}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64}.$

$Info$

$(id, i2, name, position, foot, height, weight, dob, cityob, countryob, i11, i12, age)$

$\wedge Standard$

$(id, \text{"2019-2020"}, \text{"ENG"}, \text{"1. Premier League"},$

$s_5, squad, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{20}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32})$

$\wedge FM20$

$(f_1, name, f_3, f_4, \text{"English Premier Division"}, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, value, wage, ca, pa,$
$f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35}, f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43},$
$f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{49}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64})$

$\rightarrow Player$

$(name, position, foot, height, weight, dob, age, cityob, countryob, value, wage, ca, pa)$

14

GAV Mapping – Snippets (2/3)

$\forall$

$player\_name, player\_squad, sca, sca\_per90, gca, gca\_per90, assists, touches, touches\_def\_3rd, touches\_mid\_3rd, touches\_att\_3rd,$
$touches\_att\_pen\_area, dribbles\_completed, dribbles, dribbles\_completed\_pct, miscontrols, dispossessed, pass\_targets, assists\_per90, fouled.$

$\exists$

$id, g_5, g_7, g_8, g_{11}, g_{12}, g_{13}, g_{14}, g_{15}, g_{16}, g_{19}, g_{20}, g_{21}, g_2, g_{23}, g_{24}$

$\exists$

$pa_5, pa_7, pa_8, pa_9, pa_{10}, pa_{11}, pa_{12}, pa_{13}, pa_{14}, pa_{15}, pa_{16}, pa_{17}, pa_{18}, pa_{19}, pa_{20}, pa_{21}, pa_{22}, pa_{24}, pa_{25}, pa_{26}, pa_{27}, pa_{28}, pa_{29}, pa_{30}$

$\exists$

$po_5, po_7, po_8, po_{10}, po_{15}, po_{19}, po_{20}, po_{21}, po_{22}, po_{23}, po_{24}, po_{25}, po_{26}, po_{30}, po_{31}, po_{32}$

$\exists$

$s_5, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32}$

$\exists$

$m_5, m_7, m_8, m_9, m_{10}, m_{11}, m_{12}, m_{14}, m_{15}, m_{16}, m_{17}, m_{18}, m_{19}, m_{20}, m_{21}, m_{22}, m_{23}, m_{24}$

$\exists$

$i_2, i_4, i_5, i_6, i_7, i_8, i_9, i_{10}, i_{11}, i_{12}, i_{13}.$

GCA

$(id, \text{“2019-2020”}, \text{“ITA”}, \text{“1. Serie A”}, g_5, player\_squad, g_7, g_8, sca, sca\_per90, g_{11}, g_{12}, g_{13}, g_{14}, g_{15}, g_{16}, gca, gca\_per90, g_{19}, g_{20}, g_{21}, g_2, g_{23}, g_{24})$

$\wedge Passing$

$(id, \text{“2019-2020”}, \text{“ITA”}, \text{“1. Serie A”}, pa_5, player\_squad, pa_7, pa_8, pa_9, pa_{10}, pa_{11}, pa_{12}, pa_{13}, pa_{14}, pa_{15},$
$pa_{16}, pa_{17}, pa_{18}, pa_{19}, pa_{20}, pa_{21}, pa_{22}, assists, pa_{24}, pa_{25}, pa_{26}, pa_{27}, pa_{28}, pa_{29}, pa_{30})$

$\wedge Possession$

$(id, \text{“2019-2020”}, \text{“ITA”}, \text{“1. Serie A”}, po_5, player\_squad, po_7, po_8,$
$touches, po_{10}, touches\_def\_3rd, touches\_mid\_3rd, touches\_att\_3rd, touches\_att\_pen\_area,$
$po_{15}, dribbles\_completed, dribbles, dribbles\_completed\_pct, po_{19}, po_{20}, po_{21}, po_{22},$
$po_{23}, po_{24}, po_{25}, po_{26}, miscontrols, dispossessed, pass\_targets, po_{30}, po_{31}, po_{32})$

$\wedge Standard$

$(id, \text{“2019-2020”}, \text{“ITA”}, \text{“1. Serie A”}, s_5, player\_squad, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}, s_{13},$
$s_{14}, s_{15}, s_{16}, s_{17}, s_{18}, s_{19}, assists\_per90, s_{21}, s_{22}, s_{23}, s_{24}, s_{25}, s_{26}, s_{27}, s_{28}, s_{29}, s_{30}, s_{31}, s_{32})$

$\wedge Misc$

$(id, \text{“2019-2020”}, \text{“ITA”}, \text{“1. Serie A”}, m_5, player\_squad, m_7, m_8, m_9, m_{10}, m_{11}, m_{12}, fouled, m_14, crosses, m_{16}, m_{17}, m_{18}, m_{19}, m_{20}, m_{21}, m_{22}, m_{23}, m_{24})$

$\wedge Info$

$(id, i_2, player\_name, i_4, i_5, i_6, i_7, i_8, i_9, i_{10}, i_{11}, i_{12}, i_{13})$

$\rightarrow PlaymakingStat$

$(player\_name, player\_squad, sca, sca\_per90, gca, gca\_per90, assists, touches, touches\_def\_3rd,$
$touches\_mid\_3rd, touches\_att\_3rd, touches\_att\_pen\_area, dribbles\_completed, dribbles,$
$dribbles\_completed\_pct, miscontrols, dispossessed, pass\_targets, assists\_per90, fouled)$

GAV Mapping – Snippets (3/3)

$\forall$

$player\_name, tec, tck, pen, pas, mar, lth, lon, hea, fir, fin, dri, cro, cor.$

$\exists$

$f_1, f_2, f_3, f_4, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}, f_{15}, f_{16}, f_{17}, f_{18}, f_{19}, f_{20}, f_{22}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29},$
$f_{32}, f_{33}, f_{34}, f_{35}, f_{39}, f_{40}, f_{41}, f_{43}, f_{44}, f_{45}, f_{48}, f_{50}, f_{51}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64}.$

$FM20$

$(f_1, player\_name, f_3, f_4, \text{“Italian Serie A”}, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}, f_{15}, f_{16}, f_{17}, f_{18}, f_{19}, f_{20},$
$tec, f_{22}, tck, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, pen, pas, f_{32}, f_{33}, f_{34}, f_{35}, mar, lth, lon, f_{39}, f_{40}, f_{41}, hea, f_{43}, f_{44},$
$f_{45}, fir, fin, f_{48}, dri, f_{50}, f_{51}, cro, cor, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64})$

$\rightarrow TechnicalAbility$

$(player\_name, tec, tck, pen, pas, mar, lth, lon, hea, fir, fin, dri, cro, cor)$

3.

IIS Implementation

Data Pre-processing

Workflow

Input	APACHE DRILL	Output

Input

▷ **13 CSV files**
12 for FBref +
1 for FM

▷ **346,797 records**
202,047 for FBref +
144,750 for FM

▷ **395 attributes**
332 for FBref +
63 for FM

SQL engine designed to enable data exploration and analytics on non-relational datastores.

Output

▷ **11 CSV files**
10 for FBref +
1 for FM

▷ **16,868 records**
12,628 for FBref +
4240 for FM

▷ **229 attributes**
181 for FBref +
48 for FM

Apache Drill
Query Snippets (1/2)

```
alter session set `store.format`='csv';

create table export.fbrefStandardExtract as
select
id,season,country,comp_level,lg_finish,squad,age,games,games_starts,mnts,m
inutes_90s,goals,assists,goals_pens,pens_made,pens_att,cards_yellow,cards_
red,goals_per90,assists_per90,goals_assists_per90,goals_pens_per90
from input.`fbref-standard.csv`
where
    season = '2019-2020'
    and country in ('ITA', 'GER', 'ESP', 'ENG')
    and comp_level in ('1. Serie A', '1. Bundesliga', '1. La Liga', '1.
Premier League')
;
```

Apache Drill
Query Snippets (2/2)

```
alter session set `store.format`='csv';

create table export.fm20Extract as
select Name, Position, Club, Division, Nation, Height, Weight,
Age, Preferred_Foot, Best_Pos, Best_Role, Value, Wage, CA, PA,
Wor, Vis, Tec, Tea, Tck, Str, Sta, TRO, Ref, Pen, Pas, Pac, 1v1,
OtB, Mar, L_Th, Lon, Ldr, Jum, Hea, Han, Fla, Fir, Fin, Dri, Cro,
Cor, Cnt, Cmd, Agi, Agg, Aer, Acc
from input.`fm20.csv`
where Division in ('Spanish First Division', 'Italian Serie A',
'Bundesliga', 'English Premier Division')
;
```

4.

IIS Implementation

Data Integration Main Process

Technological Stack

Apache Spark™

Multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

Slick

Modern database query and access library for Scala. It features an extensible query compiler which can generate code for different backends.

ElephantSQL

Automates every part of setup and running of PostgreSQL clusters. Available on all major cloud and application platforms all over the world.

Workflow

Load and query the pre-processed data with **Apache Spark**.

Define the Global Schema through **Slick**.

Materialize the Integration result on an **ElaphantSQL** remote instance.

Spark SQL Query Snippet

```
select
    info.name as name,
    info.position as position,
    info.foot as foot,
    cast(info.height as double) as height,
    cast(info.weight as double) as weight,
    cast(fm.Age as integer) as age,
    info.dob as dob,
    info.cityob as cityOb,
    info.countryob as countryOb,
    cast(fm.Value as double) as value,
    cast(fm.Wage as double) as wage,
    cast(fm.CA as double) as ca,
    cast(fm.PA as double) as pa
from Fm20 fm, FbrefInfo info
where (fm.Name = info.name)
```

Scala-Slick Code Snippets
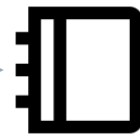
```scala
case class Player(
    name       : Option[String],
    position   : Option[String],
    foot       : Option[String],
    height     : Option[Double],
    weight     : Option[Double],
    age        : Option[Int],
    dob        : Option[String],
    cityOb     : Option[String],
    countryOb  : Option[String]
    value      : Option[Double],
    wage       : Option[Double],
    ca         : Option[Double],
    pa         : Option[Double]
)
```

```scala
class PlayerEntity(tag: Tag) extends Table[Player](tag, "Player")
{
    def name        = column[String]("name");
    def position    = column[Option[String]]("position");
    def foot        = column[Option[String]]("foot");
    def height      = column[Option[Double]]("height");
    def weight      = column[Option[Double]]("weight");
    def age         = column[Option[Int]]("age");
    def dob         = column[Option[String]]("dob");
    def cityOb      = column[Option[String]]("city_ob");
    def countryOb   = column[Option[String]]("country_ob");
    def value       = column[Option[Double]]("value");
    def wage        = column[Option[Double]]("wage");
    def ca          = column[Option[Double]]("ca");
    def pa          = column[Option[Double]]("pa");
    def * = (name.?, position, foot, height, weight, age, dob,
cityOb, countryOb, value, wage, ca, pa) <> (Player.tupled,
Player.unapply);
}
```

5.
Querying

Retrived Global DB and Unfolding

Query (1/6)

Retrieve the number of goals scored by the players who have the same finalization ability (FM fictional attribute) of "Ciro Immobile".

FOL

$q_1 = \{(\textbf{player\_name}, \textbf{goals}, \textbf{fin}) |$

$\exists g_3, g_4, g_5, g_6, g_7, g_8, g_9, g_{10}, g_{11}, g_{12}, g_{13}, g_{14}, g_{15}, g_{16}, g_{17}$
$\exists t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{12}, t_{13}, t_{14}.$
$GoalStat(\textbf{player\_name}, \textbf{goals}, g_3, g_4, g_5, g_6, g_7, g_8, g_9, g_{10}, g_{11}, g_{12}, g_{13}, g_{14}, g_{15}, g_{16}, g_{17})$
$\wedge\, TechnicalAbility(\textbf{player\_name}, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, \textbf{fin}, t_{12}, t_{13}, t_{14})$
$\wedge\, TechnicalAbility(\text{``\textbf{Ciro Immobile}''}, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, \textbf{fin}, t_{12}, t_{13}, t_{14})\}$

SQL

```sql
select gs.player_name, goals, ta1.fin
from GoalStat gs, TechnicalAbility ta1, TechnicalAbility ta2
where
    gs.player_name = ta1.player_name
    and ta1.fin = ta2.fin
    and ta2.player_name = 'Ciro Immobile';
```

Query (2/6)

Retrieve the number of yellow cards received by the players who have 17 as value of the aggression ability (FM fictional attribute).

FOL

$$q_2 = \{(\boldsymbol{player\_name}, \boldsymbol{player\_squad}, \boldsymbol{cards\_yellow}, \boldsymbol{games}, \boldsymbol{agg})$$

$$\mid$$

$$\exists d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15}, d_{16}, d_{18}$$

$$\exists p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16}$$

$$\exists m_2, m_3, m_4, m_5, m_6, m_7, m_8, m_9.$$

$$DefenseStat(\boldsymbol{player\_name}, \boldsymbol{player\_squad}, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15}, d_{16}, \boldsymbol{cards\_yellow}, d_{18})$$

$$\wedge PresenceStat(\boldsymbol{player\_name}, \boldsymbol{player\_squad}, \boldsymbol{games}, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{16})$$

$$\wedge MentalAbility(\boldsymbol{player\_name}, m_2, m_3, m_4, m_5, m_6, m_7, m_8, m_9, \boldsymbol{agg})$$

$$\wedge \boldsymbol{agg} = \boldsymbol{17}\}$$

SQL

```sql
select ds.player_name, ds.player_squad, cards_yellow, games, agg
from DefenseStat ds, MentalAbility ma, PresenceStat ps
where
    (ds.player_name = ps.player_name and ds.player_squad = ps.player_squad)
    and ps.player_name = ma.player_name
    and ma.agg = 17;
```

Query (3.1/6)

Retrieve the number of assists realized by the players who have a value of the vision ability (FM fictional attribute) not equal to 20 nor 19 nor 18 nor 17 nor 16.

FOL

$q_3 = \{(\boldsymbol{player\_name, player\_squad, position, assists, vis})$

$\quad \mid$

$\quad \exists pm_3, pm_4, pm_5, pm_6, pm_8, pm_9, pm_{10}, pm_{11}, pm_{12}, pm_{13}, pm_{14}, pm_{15}, pm_{16}, pm_{17}, pm_{18}, pm_{19}, pm_{20}$

$\quad \exists p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}$

$\quad \exists m_2, m_4, m_5, m_6, m_7, m_8, m_9, m_{10}.$

$\quad PlaymakingStat(\begin{matrix} \boldsymbol{player\_name, player\_squad}, pm_3, pm_4, pm_5, pm_6, \boldsymbol{assists}, \\ pm_8, pm_9, pm_{10}, pm_{11}, pm_{12}, pm_{13}, pm_{14}, pm_{15}, pm_{16}, pm_{17}, pm_{18}, pm_{19}, pm_{20} \end{matrix})$

$\quad \wedge Player(\boldsymbol{player\_name, position}, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13})$

$\quad \wedge MentalAbility(\boldsymbol{player\_name}, m_2, \boldsymbol{vis}, m_4, m_5, m_6, m_7, m_8, m_9, m_{10})$

$\quad \wedge \neg(\boldsymbol{vis = 20}) \wedge \neg(\boldsymbol{vis = 19}) \wedge \neg(\boldsymbol{vis = 18}) \wedge \neg(\boldsymbol{vis = 17}) \wedge \neg(\boldsymbol{vis = 16})\}$

Query (3.2/6)

Retrieve the number of assists realized by the players who have a value of the vision ability (FM fictional attribute) not equal to 20 nor 19 nor 18 nor 17 nor 16.

```sql
SQL
select pm.player_name, pm.player_squad, position, assists, vis
from PlaymakingStat pm, Player p, MentalAbility m
where
    pm.player_name = p.name
    and pm.player_name = m.player_name
    and vis != 20 and vis != 19 and vis != 18 and vis != 17 and vis != 16;
```

Query (4/6)

Retrieve the points per match gained by Juventus with respect to the presence of its players. Retrieve also the wage and the strength of the players (FM fictional attributes).

FOL

$q_4 = \{(\boldsymbol{player\_name, player\_squad, points\_per\_match, mnts, str, wage})$

$\quad\quad |$

$\quad\quad \exists pr_3, pr_5, pr_6, pr_7, pr_8, pr_9, pr_{10}, pr_{11}, pr_{13}, pr_{14}, pr_{15}, pr_{16}$

$\quad\quad \exists p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{12}, p_{13}$

$\quad\quad \exists ph_3, ph_4, ph_5, ph_6, ph_7, ph_8.$

$\quad\quad PresenceStat(\boldsymbol{player\_name, player\_squad}, pr_3, \boldsymbol{mnts}, pr_5, pr_6, pr_7, pr_8, pr_9, pr_{10}, pr_{11}, \boldsymbol{points\_per\_match}, pr_{13}, pr_{14}, pr_{15}, pr_{16})$

$\quad\quad \wedge Player(\boldsymbol{player\_name}, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, \boldsymbol{wage}, p_{12}, p_{13})$

$\quad\quad \wedge PhysicalAbility(\boldsymbol{player\_name, str}, ph_3, ph_4, ph_5, ph_6, ph_7, ph_8)$

$\quad\quad \wedge \boldsymbol{player\_squad} = \text{"Juventus"}\}$

SQL

```sql
select  pr.player_name, pr.player_squad, points_per_match, mnts, str, wage
from PresenceStat pr, Player p, PhysicalAbility ph
where
    pr.player_name = p.name
    and pr.player_name = ph.player_name
    and pr.player_squad = 'Juventus';
```

Query (5.1/6)

Retrieve the potential ability (FM fictional attribute) and the save percentage of the goalkeepers whose age is 20 or 21 or 22.

FOL

$q_5 = \{(\boldsymbol{player\_name}, \boldsymbol{player\_squad}, \boldsymbol{age}, \boldsymbol{pa}, \boldsymbol{save\_pct})$

$\mid$

$(\exists p_3, p_4, p_5, p_6, p_8, p_9, p_{10}, p_{11}, p_{12}$

$\exists g_3, g_4, g_5, g_6, g_8, g_9, g_{10}.$

$Player(\boldsymbol{player\_name}, \text{``GK''}, p_3, p_4, p_5, p_6, \boldsymbol{age}, p_8, p_9, p_{10}, p_{11}, p_{12}, \boldsymbol{pa})$

$\wedge GoalkeeperStat(\boldsymbol{player\_name}, \boldsymbol{player\_squad}, g_3, g_4, g_5, g_6, \boldsymbol{save\_pct}, g_8, g_9, g_{10})$

$\wedge \boldsymbol{age = 20}))$

$\vee$

$(\exists p_3', p_4', p_5', p_6', p_8', p_9', p_{10}', p_{11}', p_{12}'$

$\exists g_3', g_4', g_5', g_6', g_8', g_9', g_{10}'.$

$(Player(\boldsymbol{player\_name}, \text{``GK''}, p_3', p_4', p_5', p_6', \boldsymbol{age} p_8', p_9', p_{10}', p_{11}', p_{12}', \boldsymbol{pa})$

$\wedge GoalkeeperStat(\boldsymbol{player\_name}, \boldsymbol{player\_squad}, g_3', g_4', g_5', g_6', \boldsymbol{save\_pct}, g_8', g_9', g_{10}')$

$\wedge \boldsymbol{age = 21}))$

$\vee$

$(\exists p_3'', p_4'', p_5'', p_6'', p_8'', p_9'', p_{10}'', p_{11}'', p_{12}''$

$\exists g_3'', g_4'', g_5'', g_6'', g_8'', g_9'', g_{10}''.$

$(Player(player\_name, \text{``GK''}, p_3'', p_4'', p_5'', p_6'', \boldsymbol{age}, p_8'', p_9'', p_{10}'', p_{11}'', p_{12}'', \boldsymbol{pa})$

$\wedge GoalkeeperStat(player\_name, player\_squad, g_3'', g_4'', g_5'', g_6'', \boldsymbol{save\_pct}, g_8'', g_9'', g_{10}'')$

$\wedge \boldsymbol{age = 22}))\}$

Query (5.2/6)

Retrieve the potential ability (FM fictional attribute) and the save percentage of the goalkeepers whose age is 20 or 21 or 22.

```sql
SQL
select player_name, player_squad, age, pa, save_pct
from Player p, GoalkeeperStat g
where
    p.name = g.player_name and p.position = 'GK'
    and age = 20
union
select player_name, player_squad, age, pa, save_pct
from Player p, GoalkeeperStat g
where
    p.name = g.player_name and p.position = 'GK'
    and age = 21
union
select player_name, player_squad, age, pa, save_pct
from Player p, GoalkeeperStat g
where
    p.name = g.player_name and p.position = 'GK'
    and age = 22;
```

Query (6/6)

Retrieve the players whose dribbling ability (FM fictional attribute) is equal to 18.

FOL

$q_6 = \{(\boldsymbol{player\_name, dri})|$

$\exists t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{11}, t_{13}, t_{14}$

$Technical Ability(\boldsymbol{player\_name}, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{11}, \boldsymbol{dri}, t_{13}, t_{14})$

$\wedge \boldsymbol{dri} = \boldsymbol{18}\}$

SQL

```sql
select player_name, dri
from TechnicalAbility
where dri = 18;
```

Query (6/6) – Unfolding (1/2)

Retrieve the players whose dribbling ability (FM fictional attribute) is equal to 18.

FOL

$q_6 = \{(\boldsymbol{player\_name}, \boldsymbol{dri}) \mid$

$(\exists f_1, f_3, f_4, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}, f_{15}, f_{16}, f_{17}, f_{18}, f_{19}, f_{20}, f_{21}, f_{22}, f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33},$
$f_{34}, f_{35}, f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43}, f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64}$

$f_1, \boldsymbol{player\_name}, f_3, f_4, \text{``Italian Serie A''}, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}, f_{15}, f_{16}, f_{17}, f_{18}, f_{19}, f_{20}, f_{21}, f_{22},$
$FM20(f_{23}, f_{24}, f_{25}, f_{26}, f_{27}, f_{28}, f_{29}, f_{30}, f_{31}, f_{32}, f_{33}, f_{34}, f_{35}, f_{36}, f_{37}, f_{38}, f_{39}, f_{40}, f_{41}, f_{42}, f_{43}, f_{44}, f_{45}, f_{46}, f_{47}, f_{48}, \boldsymbol{dri}, \quad)$
$f_{50}, f_{51}, f_{52}, f_{53}, f_{54}, f_{55}, f_{56}, f_{57}, f_{58}, f_{59}, f_{60}, f_{61}, f_{62}, f_{63}, f_{64}$

$\wedge \, \boldsymbol{dri} = 18)$
$\vee$

$(\exists f_1', f_3', f_4', f_6', f_7', f_8', f_9', f_{10}', f_{11}', f_{12}', f_{13}', f_{14}', f_{15}', f_{16}', f_{17}', f_{18}', f_{19}', f_{20}', f_{21}', f_{22}', f_{23}', f_{24}', f_{25}', f_{26}', f_{27}', f_{28}', f_{29}', f_{30}', f_{31}', f_{32}', f_{33}',$
$f_{34}', f_{35}', f_{36}', f_{37}', f_{38}', f_{39}', f_{40}', f_{41}', f_{42}', f_{43}', f_{44}', f_{45}', f_{46}', f_{47}', f_{48}', f_{50}', f_{51}', f_{52}', f_{53}', f_{54}', f_{55}', f_{56}', f_{57}', f_{58}', f_{59}', f_{60}', f_{61}', f_{62}', f_{63}', f_{64}'$

$f_1', \boldsymbol{player\_name}, f_3', f_4', \text{``Bundesliga''}, f_6', f_7', f_8', f_9', f_{10}', f_{11}', f_{12}', f_{13}', f_{14}', f_{15}', f_{16}', f_{17}', f_{18}', f_{19}', f_{20}', f_{21}', f_{22}',$
$FM20(f_{23}', f_{24}', f_{25}', f_{26}', f_{27}', f_{28}', f_{29}', f_{30}', f_{31}', f_{32}', f_{33}', f_{34}', f_{35}', f_{36}', f_{37}', f_{38}', f_{39}', f_{40}', f_{41}', f_{42}', f_{43}', f_{44}', f_{45}', f_{46}', f_{47}', f_{48}', \boldsymbol{dri},)$
$f_{50}', f_{51}', f_{52}', f_{53}', f_{54}', f_{55}', f_{56}', f_{57}', f_{58}', f_{59}', f_{60}', f_{61}', f_{62}', f_{63}', f_{64}'$

$\wedge \, \boldsymbol{dri} = 18)$
$\vee$

$(\exists f_1'', f_3'', f_4'', f_6'', f_7'', f_8'', f_9'', f_{10}'', f_{11}'', f_{12}'', f_{13}'', f_{14}'', f_{15}'', f_{16}'', f_{17}'', f_{18}'', f_{19}'', f_{20}'', f_{21}'', f_{22}'', f_{23}'', f_{24}'', f_{25}'', f_{26}'', f_{27}'', f_{28}'', f_{29}'', f_{30}'', f_{31}'', f_{32}'',$
$f_{33}'', f_{34}'', f_{35}'', f_{36}'', f_{37}'', f_{38}'', f_{39}'', f_{40}'', f_{41}'', f_{42}'', f_{43}'', f_{44}'', f_{45}'', f_{46}'', f_{47}'', f_{48}'', f_{50}'', f_{51}'', f_{52}'', f_{53}'', f_{54}'', f_{55}'', f_{56}'', f_{57}'', f_{58}'', f_{59}'', f_{60}'', f_{61}'', f_{62}'', f_{63}'', f_{64}''$

$f_1'', \boldsymbol{player\_name}, f_3'', f_4'', \text{``Spanish First Division''}, f_6'', f_7'', f_8'', f_9'', f_{10}'', f_{11}'', f_{12}'', f_{13}'', f_{14}'', f_{15}'', f_{16}'', f_{17}'', f_{18}'', f_{19}'',$
$FM20(f_{20}'', f_{21}'', f_{22}'', f_{23}'', f_{24}'', f_{25}'', f_{26}'', f_{27}'', f_{28}'', f_{29}'', f_{30}'', f_{31}'', f_{32}'', f_{33}'', f_{34}'', f_{35}'', f_{36}'', f_{37}'', f_{38}'', f_{39}'', f_{40}'', f_{41}'', f_{42}'', f_{43}'', f_{44}'', f_{45}'', \quad)$
$f_{46}'', f_{47}'', f_{48}'', \boldsymbol{dri}, f_{50}'', f_{51}'', f_{52}'', f_{53}'', f_{54}'', f_{55}'', f_{56}'', f_{57}'', f_{58}'', f_{59}'', f_{60}'', f_{61}'', f_{62}'', f_{63}'', f_{64}''$

$\wedge \, \boldsymbol{dri} = 18)$
$\vee$

$(\exists f_1''', f_3''', f_4''', f_6''', f_7''', f_8''', f_9''', f_{10}''', f_{11}''', f_{12}''', f_{13}''', f_{14}''', f_{15}''', f_{16}''', f_{17}''', f_{18}''', f_{19}''', f_{20}''', f_{21}''', f_{22}''', f_{23}''', f_{24}''', f_{25}''', f_{26}''', f_{27}''', f_{28}''', f_{29}''', f_{30}''', f_{31}''', f_{32}''', f_{33}''',$
$f_{34}''', f_{35}''', f_{36}''', f_{37}''', f_{38}''', f_{39}''', f_{40}''', f_{41}''', f_{42}''', f_{43}''', f_{44}''', f_{45}''', f_{46}''', f_{47}''', f_{48}''', f_{50}''', f_{51}''', f_{52}''', f_{53}''', f_{54}''', f_{55}''', f_{56}''', f_{57}''', f_{58}''', f_{59}''', f_{60}''', f_{61}''', f_{62}''', f_{63}''', f_{64}'''$

$f_1''', \boldsymbol{player\_name}, f_3''', f_4''', \text{``English Premier Division''}, f_6''', f_7''', f_8''', f_9''', f_{10}''', f_{11}''', f_{12}''', f_{13}''', f_{14}''', f_{15}''', f_{16}''', f_{17}''', f_{18}''', f_{19}''',$
$FM20(f_{20}''', f_{21}''', f_{22}''', f_{23}''', f_{24}''', f_{25}''', f_{26}''', f_{27}''', f_{28}''', f_{29}''', f_{30}''', f_{31}''', f_{32}''', f_{33}''', f_{34}''', f_{35}''', f_{36}''', f_{37}''', f_{38}''', f_{39}''', f_{40}''', f_{41}''', f_{42}''', f_{43}''', f_{44}''', \quad)$
$f_{45}''', f_{46}''', f_{47}''', f_{48}''', \boldsymbol{dri}, f_{50}''', f_{51}''', f_{52}''', f_{53}''', f_{54}''', f_{55}''', f_{56}''', f_{57}''', f_{58}''', f_{59}''', f_{60}''', f_{61}''', f_{62}''', f_{63}''', f_{64}'''$

$\wedge \, \boldsymbol{dri} = 18)\}$

Query (6/6) – Unfolding (2/2)

Retrieve the players whose dribbling ability (FM fictional attribute) is equal to 18.

```sql
SQL
select name, dri
from FM20
where
    division = 'Italian Serie A' and dri = 18
union
select name, dri
from FM20
where
    division = 'Bundesliga' and dri = 18
union
select name, dri
from FM20
    where division = 'Spanish First Division' and dri = 18
union
select name, dri
from FM20
    where division = 'English Premier Division' and dri = 18;
```

Thanks!

Share ideas and get involved!

Visit the GitHub page of the project