Università degli Studi di Milano Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Informatica
Progetto di Datawarehouse
Anno accademico 2017/2018

Creazione di un datawarehouse per l'analisi delle quotazioni riferite ad eventi sportivi.

781229 Federico Giannini
781266 Mauro Manfredelli

Sommario

Indice delle figure

1. Contesto e obiettivi del progetto

In questa trattazione è descritta l'attività di creazione di un datawarehouse che permetta di effettuare un'analisi sulle quotazioni, riferite ad eventi sportivi, fornite da alcuni dei bookmaker presenti sul mercato.
Sono successivamente presentati i risultati delle analisi effettuate.
Sono state studiate, nello specifico, le quotazioni fornite dai seguenti bookmaker:

- Bet365
- Ladbrokes
- Pinnacle
- StanJames.

L'analisi delle quotazioni è stata effettuata in relazione a partite di calcio e match di tennis.

L'obiettivo del datawarehouse è quello di analizzare la relazione in essere tra le quotazioni dei bookmaker e il risultato effettivo di un evento sportivo.
Si vuole, successivamente, cercare di capire se le quotazioni, a competizione avviata, siano più coerenti con il risultato finale rispetto alle quotazioni relative ai primi match di una competizione.

Prima di definire, nel dettaglio, le analisi che il datawarehouse debba essere in grado di alimentare, diamo le seguenti definizioni:

- **Match_odd:** insieme delle quotazioni, fornite da un bookmaker e relative ad un match, per i seguenti risultati:
  - vittoria prima partecipante
  - vittoria secondo partecipante
  - pareggio.
- **Match_odd corretta:** match_odd per la quale la quotazione più bassa si riferisce al risultato corretto.
- **Prima/seconda parte di una stagione:**
  Per una partita di calcio la prima parte della stagione è rappresentata dalle partite giocate entro il mese di ottobre.
  La seconda parte è rappresenta dalle partite rimanenti.
  Per un match di tennis la prima parte della stagione è rappresentata dai seguenti round:
  - Round robin
  - Round 0
  - Round 1
  - Round 2
  - Round 3
  - Round 4.

  La seconda parte è rappresentata, invece, dai seguenti round:
  - Quarti di finale
  - Semifinale
  - Finale.
- **KO:** situazione in cui, in relazione ad un match, uno dei due partecipanti abbia vinto con un ampio scarto nel punteggio.
  Per una partita di calcio si considera un ampio scarto uno scarto di almeno 3 goal.
  Per una partita di tennis al meglio delle 3, si considera un ampio scarto uno scarto di 2 set.
  Per una partita di tennis al meglio delle 5, si considera un ampio scarto uno scarto di 3 set.
- **Errore grave in una match_odd:** situazione in cui, in caso di KO, la quotazione più bassa di una match_odd si riferisca alla squadra perdente.

Il datawarehouse deve permette di effettuare le seguenti analisi:

1. Percentuale di correttezza delle quote.
2. Confronto sulla correttezza delle quote nella prima parte e nella seconda parte della stagione di una competizione.
3. Quotazione media riferita alla squadra vincente in caso di KO.
4. Quotazione media riferita alla squadra perdente in caso di KO.
5. Percentuale di errori gravi nelle quote.

Ognuna delle cinque analisi deve essere effettuata:

- Su tutti i match presenti nel database e:
    - Per ogni singolo bookmaker (in questo caso trovare il bookmaker migliore)
    - Media dei diversi bookmaker
- Per ogni singola competizione (in questo caso trovare la competizione migliore) e:
    - Per ogni singolo bookmaker (in questo caso trovare il bookmaker migliore)
    - Media dei diversi bookmaker
- Per ogni singola stagione di ogni singola competizione (in questo caso trovare la stagione migliore)
    - Per ogni singolo bookmaker (in questo caso trovare il bookmaker migliore)
    - Media dei diversi bookmaker
- Per ogni sport (in questo caso trovare lo sport migliore) e:
    - Per ogni singolo bookmaker (in questo caso trovare il bookmaker migliore)
    - Media dei diversi bookmaker.
- Per ogni singolo partecipante (in questo caso trovare il partecipante migliore) e:
    - Per ogni singolo bookmaker (in questo caso trovare il bookmaker migliore)
    - Media dei diversi bookmaker.

2. Analisi delle sorgenti

Elenco delle sorgenti

Per la creazione del datawarehouse, sono state utilizzate le seguenti sorgenti dati:

- **European Soccer Database (ESDB)**
  Database, presente sulla piattaforma Kaggle, che contiene le informazioni relative alle partite di calcio, con le relative quote dei bookmaker.
  Il database è distribuito attraverso uno script in formato .sqlite.
- **Football-data (FD)**
  Dataset, presente sul sito football-data.co.uk, che contiene le informazioni relative alle partite di calcio, con le corrispondenti quote dei bookmaker.
  Il dataset è distribuito in formato csv. Esiste un file per ogni stagione di ogni campionato e un file che contiene le informazioni dei diversi campionati.
- **ATP Matches Dataset (ATPMDS)**
  Dataset, presente sulla piattaforma Kaggle, che contiene le informazioni relative alle partite di tennis.
  Il dataset è distribuito in formato csv. Esiste un unico file per tutti i match.
- **ATP Men's Tour (ATPMT)**
  Dataset, presente sulla piattaforma Kaggle, che contiene le informazioni relative alle partite di tennis, con le corrispondenti quote dei bookmaker.
  Il dataset è distribuito in formato csv. Esiste un file unico per tutti i match.
- **TD (Tennis-data)**
  Dataset, presente sul sito tennis-data.co.uk, che contiene le informazioni relative alle partite di tennis, con le corrispondenti quote dei bookmaker.
  Il dataset è distribuito in formato csv. Esiste un file per ogni anno solare.
- **Countries**
  Dataset contenente tutte le nazioni riconosciute
- **Cities**
  Dataset contenente le città più importanti di ogni nazione.

Schemi locali

ESDB

Schema E/R

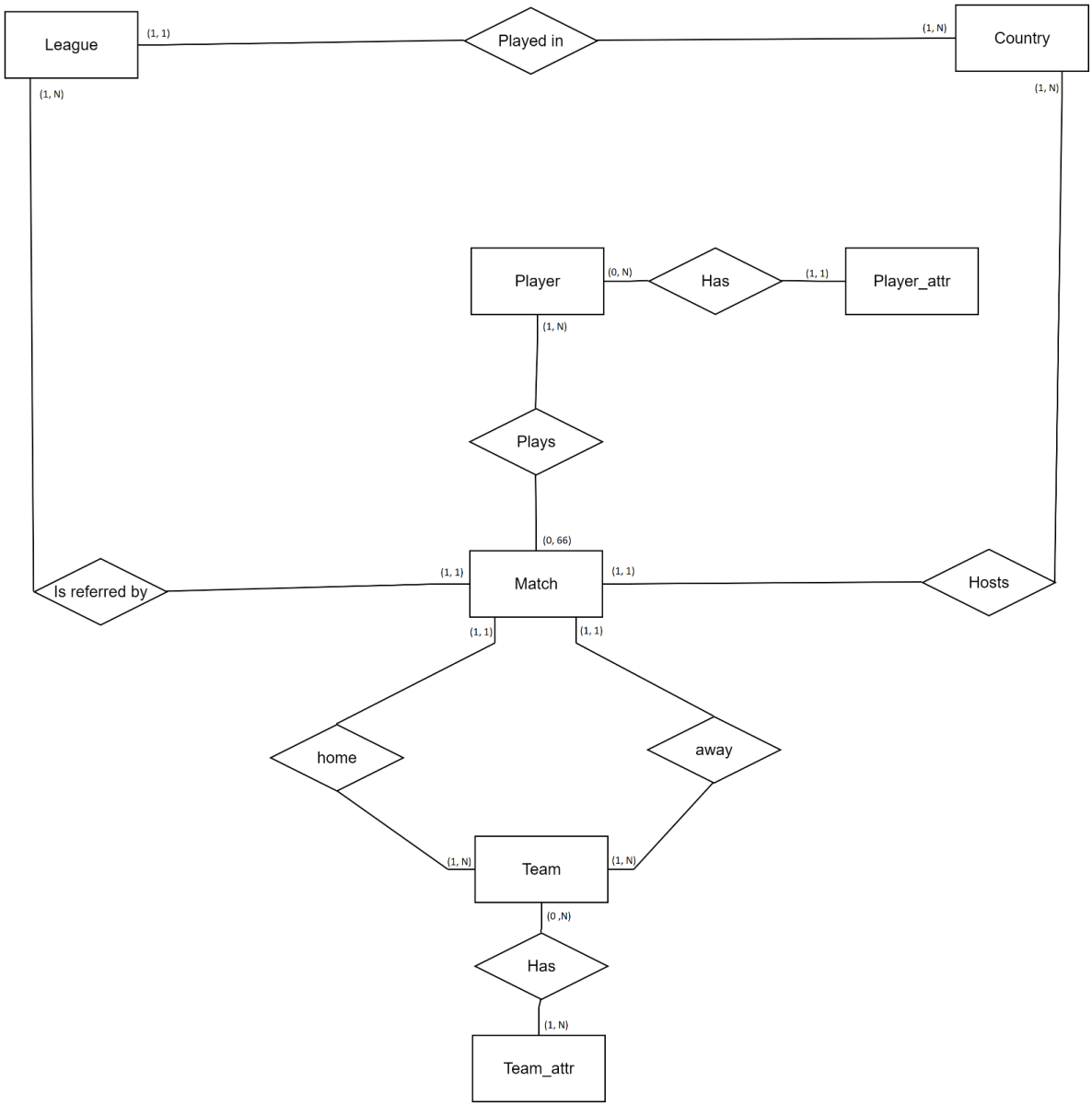

Figura 1: Schema E/R di ESDB.

Attributi

Team (
- **id**
- team_api_id
- team_fifa_api_id
- team_long_name
- team_short_name

);

Country (
- id
- name

);

Player (
- **id**
- player_api_id
- player_name
- player_fifa_api_id
- birthday
- height
- weight

);

Team_attr (
- **id**
- date
- buildUpPlaySpeed
- buildUpPlaySpeedClass
- buildUpPlayDribbling
- buildUpPlayDribblingClass
- buildUpPlayPassing
- buildUpPlayPassingClass
- buildUpPlayPositioningClass
- chanceCreationPassing
- chanceCreationPassingClass
- chanceCreationCrossing
- chanceCreationCrossingClass
- chanceCreationShooting
- chanceCreationShootingClass
- chanceCreationPositioningClass
- defencePressure
- defencePressureClass
- defenceAggression
- defenceAggressionClass
- defenceTeamWidth
- defenceTeamWidthClass
- defenceDefenderLineClass

);

Player_attr (
- **id**
- date
- overall_rating
- potential
- preferred_foot
- attacking_work_rate
- defensive_work_rate
- crossing
- finishing
- heading_accuracy
- short_passing
- volleys
- dribbling
- curve
- free_kick_accuracy
- long_passing
- ball_control
- acceleration
- sprint_speed
- agility
- reactions
- balance
- shot_power
- jumping
- stamina
- strength
- long_shots
- aggression
- interceptions
- positioning
- vision
- penalties
- marking
- standing_tackle
- sliding_tackle
- gk_diving
- gk_handling
- gk_kicking
- gk_positioning
- gk_reflexes

);

League (
- **id**
- name

);

Is_referred_by (
- stage
- season

);

Match (
- **id**
- date
- match_api_id
- goal
- shoton
- shotoff
- foulcommit
- card
- cross
- corner
- possession
- B365D
- BWD
- IWD
- LBD
- PSD
- WHD
- SJD
- VCD
- GBD
- BSD

);

Home (
- home_team_goal
- B365H
- BWH
- IWH
- LBH
- PSH
- WHH
- SJH
- VCH
- GBH
- BSH

)

Away (
- away_team_goal
- B365A
- BWA
- IWA
- LBA
- PSA
- WHA
- SJA
- VCA
- GBA
- BSA

);

FD

Schema E/R

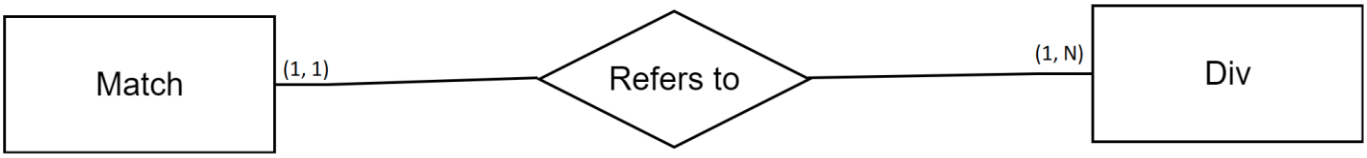

Figura 2: Schema E/R di FD.

Attributi

Match (

- Date *(dd/mm/yy)*
- HomeTeam *(Squadra di casa)*
- AwayTeam *(Squadra in trasferta)*
- FTHG *(Goal finali squadra in casa)*
- FTAG *(Goal finali squadra in trasferta)*
- FTR (*Risultato finale:* H=Vittoria squadra in casa, D=Pareggio, A=Vittoria squadra in trasferta)
- HTHG *(Goal squadra in casa alla fine del primo tempo)*
- HTAG *(Goal squadra in trasferta alla fine del primo tempo)*
- HTR *(Risultato alla fine del primo tempo)*
- Attendance *(Numero di spettatori)*
- Referee *(Nome dell'arbitro)*
- HS *(Tiri della squadra in casa)*
- AS *(Tiri della squadra in trasferta)*
- HST (*Tiri in porta della squadra in casa)*
- AST (*Tiri in porta della squadra in trasferta)*
- HHW (*Pali presi dalla squadra in casa)*
- AHW *(Pali presi dalla squadra in trasferta)*
- HC *(Calci d'angolo della squadra in casa)*
- AC *(Calci d'angolo della squadra in trasferta)*
- HF *(Falli commessi dalla squadra in casa)*
- AF *(Falli commessi dalla squadra in trasferta)*
- HO *(Offside della squadra in casa)*
- AO (*Offside della squadra in trasferta)*
- HY *(Cartellini gialli per la squadra in casa)*
- AY *(Cartellini gialli per la squadra in trasferta)*
- HR *(Cartellini rossi per la squadra in casa)*
- AR *(Cartellini rossi per la squadra in trasferta)*
- HBP *(Bookings Points squadra in casa: 10 = yellow, 25 = red)*
- ABP *(Bookings Points squadra in trasferta: 10 = yellow, 25 = red)*
- B365H
- B365D
- B365A
- BSH
- BSD
- BSA
- BWH
- BWD
- BWA

- GBH
- GBD
- GBA
- IWH
- IWD
- IWA
- LBH
- LBD
- LBA
- PSH
- PSD
- PSA
- SOH
- SOD
- SOA
- SBH
- SBD
- SBA
- SJH
- SJD
- SJA
- SYH
- SYD
- SYA
- VCH
- VCD
- VCA
- WHH
- WHD
- WHA
- PSH
- PSD
- PSA
- Bb1X2
- BbMxH
- BbAvH
- BbMxD
- BbAvD
- BbMxA
- BbAvA
- MaxH
- MaxD
- MaxA
- AvgH
- AvgD
- AvgA
- BbOU
- BbMx>2.5
- BbAv>2.5
- BbMx<2.5
- BbAv<2.5
- GB>2.5
- GB<2.5
- B365>2.5
- B365<2.5
- BbAH
- BbAHh
- BbMxAHH
- BbAvAHH
- BbMxAHA
- BbAvAHA
- GBAHH
- GBAHA
- GBAH
- LBAHH
- LBAHA
- LBAH
- B365AHH
- B365AHA
- B365AH

);

Div (
    id,
    name
);

ATPMDS

Match(
- tourney_id
- tourney_name
- surface
- draw_size
- tourney_level
- tourney_date
- match_num
- winner_id
- winner_seed
- Seed of winner
- winner_entry
- winner_name
- winner_hand
- winner_ht
- winner_ioc *(nazionalità vincitore: codice a 3 cifre)*
- winner_age
- winner_rank
- winner_rank_points
- loser_id
- loser_seed
- loser_entry
- loser_name
- loser_hand
- loser_ht
- loser_ioc *(nazionalità perdente: codice a 3 cifre)*
- loser_age
- loser_rank
- Loser_rank_points
- score
- best_of
- round
- minutes
- w_ace
- w_df
- w_svpt
- w_1stIn
- w_1stWon
- w_2ndWon
- w_SvGms
- w_bpSaved
- w_bpFaced
- l_ace
- l_df
- l_svpt
- l_1stIn
- l_1stWon
- l_2ndWon
- l_SvGms
- l_bpSaved
- l_bpFaced

);

Countries

Country (
    name
    appha-2
    alpha-3
    country-code
    iso_3166-2
    region
    sub-region
    region-code
    sub-region-code
);

ATPMT e TD

Match (

ATP	Loser	L4	EXW	SJW
WTA	WRank	W5	EXL	SJL
Location	LRank	L5	LBW	UBW
Tournament	WPts	Wsets	LBL	UBL
Data	LPts	Lsets	GBW	MaxW
Series	W1	Comment	GBL	MaxL
Tier	L1	B365W	IWW	AvgW
Court	W2	B365L	IWL	AvgL
Surface	L2	B&WW	PSW	
Round	W3	B&WL	PSL	
Best of	L3	CBW	SBW	
Winner	W4	CBL	SBL	

);

Cities

City (

city,
city_ascii, *(nome della città senza accenti)*
lat,
lng,
pop,
country, *(nome della nazione)*
iso2, *(codice iso2 della nazione)*
iso3, *(codice iso3 della nazione)*
province

);

Analisi delle informazioni contenute e valutazione della qualità

ESDB

Questo database contiene le informazioni relative ai match dei seguenti campionati:

- Jupiler League belga
- Premier League inglese
- Ligue 1 francese
- Bundesliga tedesca
- Serie A italiana
- Eredivisie olandese
- Ekstraklasa polacca
- Primeria Liga portghese
- Scotland Premier League scozzese
- Liga spagnola
- Super League svizzera.

Sono presenti inoltre le quotazioni di alcuni bookmaker (per i risultati 1, x o 2):

- Bet365 (B365H, B36D, B365A)
- Bet&Win (BWH, BWD, BWA)
- Interwetten (IWH, IWD, IWA)
- Ladbrokes (LBH, LBD, LBA)
- Pinnacle (PSH, PSD, PSA)
- William Hill (WHH, WHD, WHA)
- Stan James (SJH, SJD, SJA)
- VC Bet (VCH, VCD, VCA)
- Gamebookers (GBH, GBD, GBA)
- Blue square (BSH, BSD, BSA).

In alcuni casi non sono rispettati i vincoli sulla chiave esterna, in quanto, per alcuni match, le chiavi relative alla squadra in casa o alla squadra in trasferta, non si riferiscono a nessuna squadra.
Il dataset contiene i dati relativi alle stagioni comprese tra 2008/2009 e 2015/2016 (estremi inclusi).
Sono memorizzate le informazioni relative al numero di giornata di ogni match e alla stagione di riferimento.

FD

Questo dataset contiene le informazioni relative ai match dei seguenti campionati:

- Jupiler League belga
- Bundesliga tedesca
- Premier League inglese
- Ligue 1 francese
- Serie A italiana
- Eredivisie olandese
- Primeira Liga portoghese
- Scottish Premier League scozzese
- Liga spagnola.

Sono presenti i dati relativi alle stagioni comprese tra 1993/1994 e 2016/2017 (estremi inclusi).
Ad ogni match sono associate le quote (per i risultati 1, x o 2) dei seguenti bookmaker:

- Bet365 (B365H, B36D, B365A)
- Bet&Win (BWH, BWD, BWA)
- Interwetten (IWH, IWD, IWA)
- Ladbrokes (LBH, LBD, LBA)
- Pinnacle (PSH, PSD, PSA)
- William Hill (WHH, WHD, WHA)

- Stan James (SJH, SJD, SJA)
- VC Bet (VCH, VCD, VCA)
- Gamebookers (GBH, GBD, GBA)
- Blue square (BSH, BSD, BSA)
- Sporting Odds (SOH, SOD, SOA)
- Sporting Bet (SBH, SBD, SBA)
- Stanleybet (SYH, SYD, SYA).

Il dataset è denormalizzato e non esistono identificativi per le squadre ad esclusione del nome.
Non sono direttamente memorizzate le informazioni relative al numero di giornata di ogni match e alla stagione di riferimento.

ATPMDS

Questo dataset contiene le informazioni relative ai match di diversi tornei dal 2008 al 2016.
L'importanza del dataset è rappresentata dalla presenza delle nazionalità dei giocatori espresse tramite codice iso3.
ATPMDS non contiene però gli stessi match e gli stessi giocatori di ATPMT e TD.
Alcuni tennisti rimarranno, quindi, senza nazionalità.

ATPMT e TD

Questi dataset contengono le informazioni relative ai match di diversi tornei dal 2000 al 2016.
Per ogni match oltre alla location, al nome del torneo, alla data, al round del torneo, al nome del vincente, al nome del perdente e ad altre informazioni, sono memorizzate le quote associate alla vittoria del tennista che successivamente ha vinto il match (W) e che successivamente ha perso il match (L).
Tra i bookmaker troviamo:

- Bet365 (B365W, B365L)
- Ladbrokes (LBW, LBL)
- Pinnacle (PSW, PSL)
- StanJames (SJW, SJL).

I dataset sono denormalizzati e non esistono identificativi per i tornei.
Il campo ATP non è univoco, match riferiti allo stesso torneo possono avere ATP diverso e match riferiti a tornei diversi possono avere ATP uguale.
Lo stesso torneo può essere indicato quindi con nomi diversi, ATP diversi e location diverse.
I due dataset condividono gli stessi nomi dei tennisti (se presenti), i quali identificano univocamente i tennisti.

Countries

Contiene le informazioni relative a tutte le nazioni riconosciute:

- name
- iso2
- iso3
- identificativo
- iso_3166-2
- continente
- sotto-continente
- identificativo del continente
- identificativo del sotto-continente.

Cities

Contiene le informazioni relative alle principali città dei diversi stati del mondo:

- nome,
- nome della città senza accenti
- latitudine
- longitudine
- popolazione
- nome della nazione
- codice iso2 della nazione
- codice iso3 della nazione
- provincia.

I nomi delle città non sono univoci, alcuni nomi sono presenti in più stati (esempio: London o Valencia).

Selezione delle informazioni utili e modifica degli schemi locali

È stato scelto di effettuare l'analisi sui match dal 2008 al 2016 in modo da avere dati più affidabili.
È stato selezionato il sottoinsieme dei bookmaker presenti in entrambi gli sport, il quale contiene:

- Bet365
- Ladbrokes
- Stan James
- Pinnacle.

Su ogni schema locale sono state effettuate le modifiche riportate di seguito.

ESDB

Schema E/R

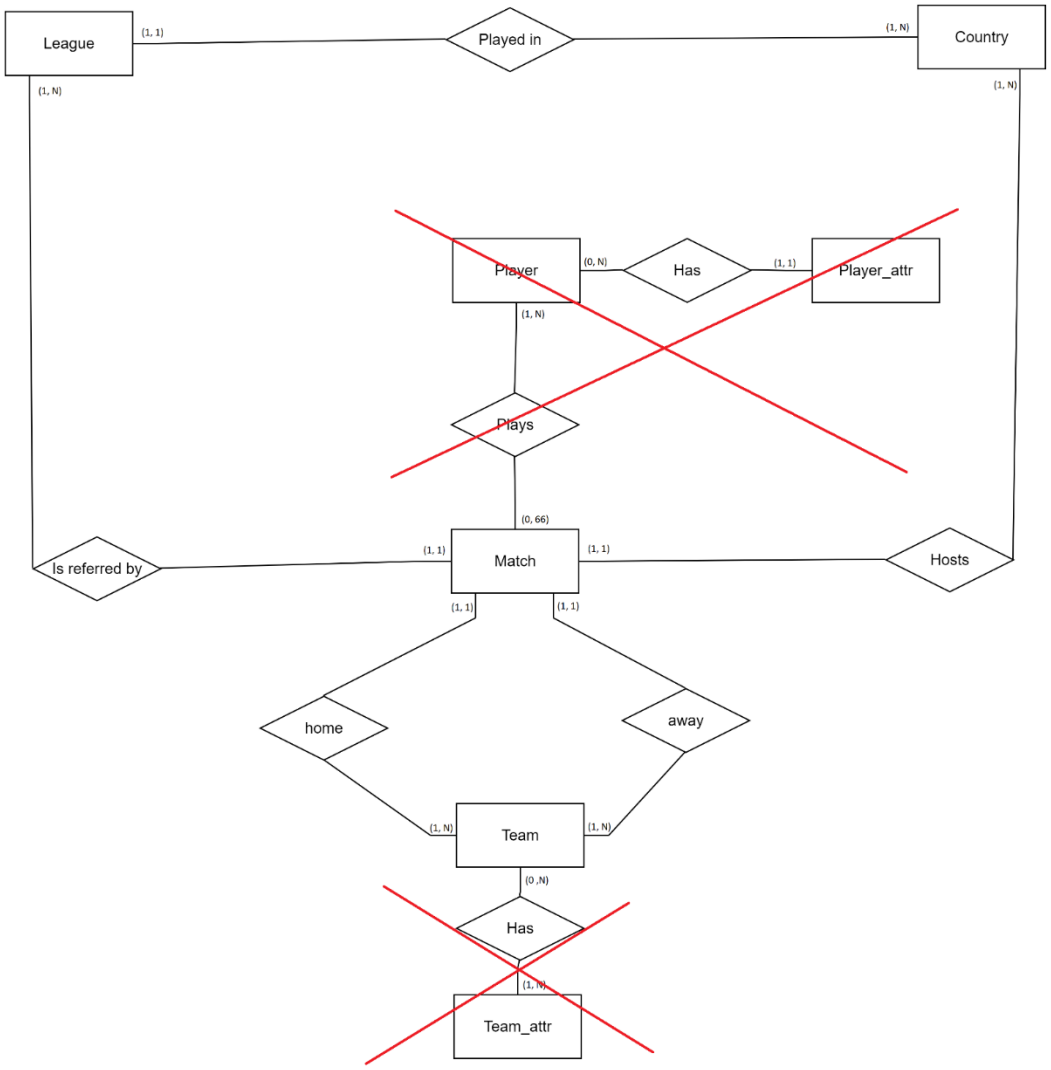

Figura 3: Selezione delle informazioni rilevanti di ESDB.

Attributi

Team (
    id
    team_api_id
    team_long_name
);

Country (
    id
    name
);

League (
    id
    name
);

Match (
    id
    date
    B365D
    LBD
    PSD
    SJD
);

Is_referred_by (
    season,
)

Home (
    home_team_goal
    B365H
    LBH
    PSH
    SJH
)

Away (
    away_team_goal
    B365A
    LBA
    PSA
    SJA
)

FD

Match (
    Date
    HomeTeam
    AwayTeam
    FTHG
    FTAG
    B365H
    B365D
    B365A
    LBH
    LBD
    LBA
    PSH
    PSD
    PSA
    SJH
    SJD
    SJA
)

Div (
    id,
    name
);

ATP Matches Dataset

Match (
    winner_name
    winner_ioc
    loser_name
    loser_ioc
)

Countries

Country (
    name
    alpha-3
    region
    sub-region
)

ATP Men's Tour e Tennis-data

Match (
    ATP
    Tournament
    Location
    Data
    Round
    Best of
    Winner
    Loser
    Wsets
    Lsets
    B365W
    B365L
    LBW
    LBL
    PSW
    PSL
    SJW
    SJL
)

Cities

City (
    city_ascii, *(nome della città senza accenti)*
    country, *(nome della nazione)*
    iso2, *(codice iso2 della nazione)*
    iso3, *(codice iso3 della nazione)*
);

Normalizzazione degli schemi

I diversi schemi locali, dopo una ricognizione, sono stati normalizzati come segue.

ESDB

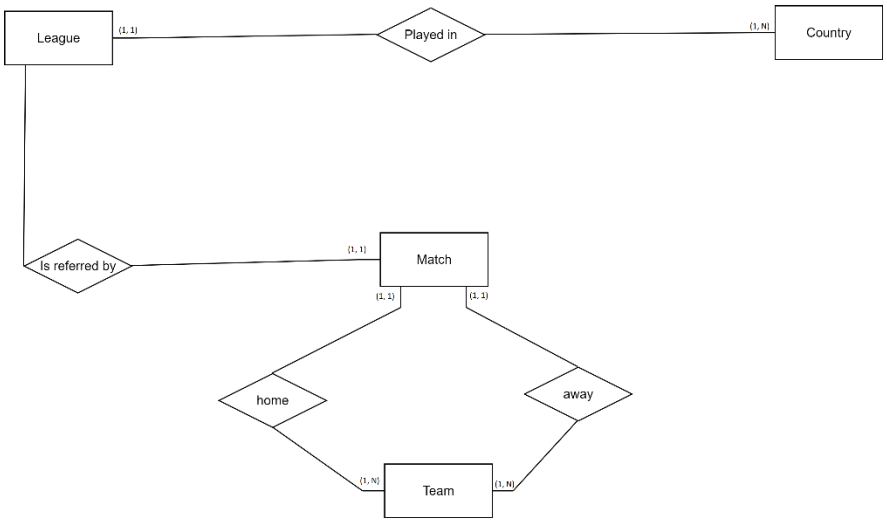

Figura 4: Normalizzazione dello schema E/R di ESDB.

Gli attributi non vengono modificati.

FD

Schema E/R

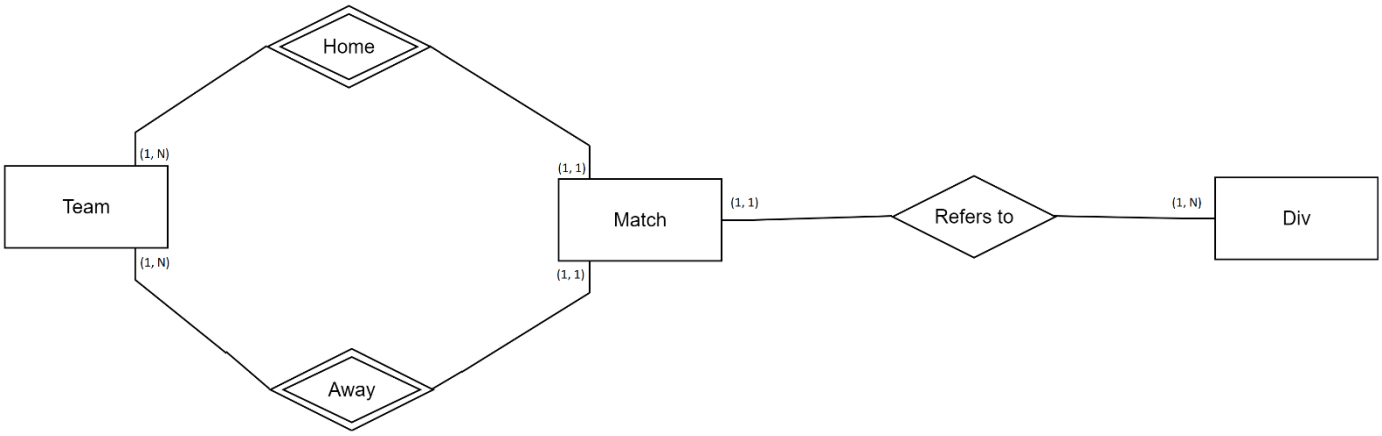

Figura 5: Normalizzazione dello schema E/R di FD.

Attributi

Match (
    Date
    B365D
    LBD
    PSD
    SJD
)

Div (
    id
    name
)

Home (
    FTHG
    B365H
    LBH
    PSH
    SJH
)

Away (
    FTAG
    B365A
    LBA
    PSA
    SJA
)

Le associazioni Home e Away sono identificative.

ATPMDS

Schema E/R

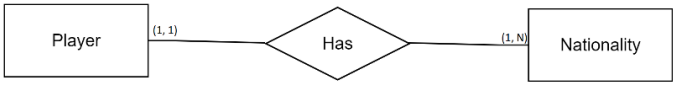

Figura 6: Normalizzazione dello schema E/R di ATPMDS.

Attributi

Player (
    name
)

Nationality (
    code_3
)

Il nome rappresenta univocamente il giocatore.

ATPMT e TD

Schema E/R

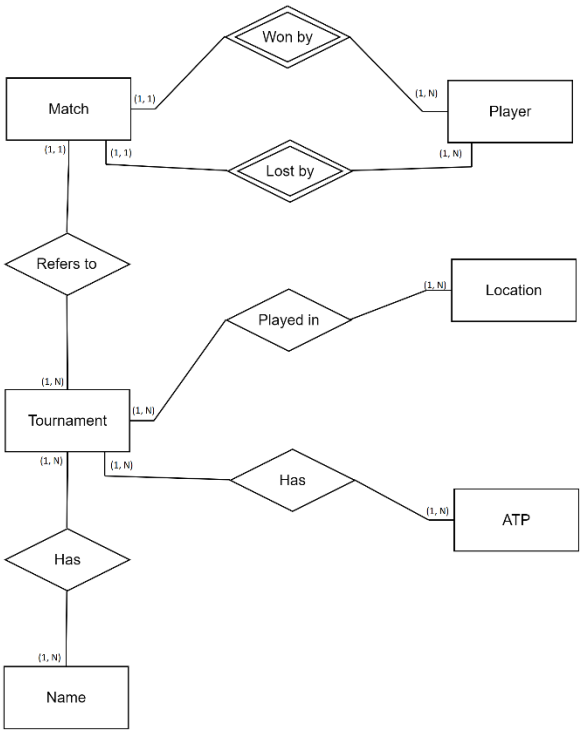

Figura 7: Normalizzazione degli schemi E/R di ATPMT e TD.

Attributi

Tournament (
    name
)

Name (
    name
)

Location (
    Location

Refers_to (
    Round
)

Player (
    name
)

Won_by (
    Wsets
    B365W
    LBW
    PSW
    SJW
)

Lost_by (
    Lsets

)

Match (
Date
)

B365L
LBL
PSL
SJL
)

Le associazioni Won By e Lost by sono identificative. Insieme a queste associazioni, la data identifica il match.

Cities

Schema E/R

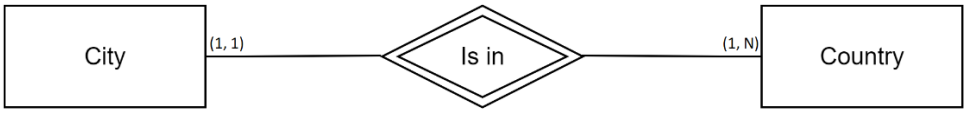

Figura 8: Normalizzazione dello schema E/R di Cities.

Attributi

City (
city_ascii
);

Country (
country,
iso2,
iso3
);

3. Analisi dei requisiti

Per quanto riguarda l'architettura del datawarehouse è stato scelto, per semplicità, di adottare un'architettura a datawarehouse centralizzato che contenga il livello dei dati riconciliati.
La dimensione del problema non necessita infatti l'introduzione di singoli datamart.

Glossario dei requisiti

Per evitare di complicare la struttura del datawarehouse, è stato scelto di utilizzare un unico fatto che rappresenti tutti gli eventi sportivi, siano essi associato al calcio o al tennis.

Fatto: Quotazioni, associate ad un evento sportivo, fornite da un bookmaker.
Misure: Valori delle quotazioni in relazione ai possibili risultati:
- Vittoria del primo partecipante
- Vittoria del secondo partecipante
- Pareggio.

Dimensioni:
- Match
- Data del match
- Competizioni
- Nazioni di riferimento delle competizioni
- Stagioni delle competizioni
- Porzioni delle stagioni (prima o seconda parte)
- Partecipanti
- Nazionalità dei partecipanti
- Sport
- Risultato del match
- KO
- Bookmaker.

Carico di lavoro preliminare

Fatto: Quotazioni, associate ad un evento sportivo, fornite da un bookmaker.
Interrogazioni:
- Analisi 1
    a) Considerando tutti i match contenuti nel datawarehouse, per ogni bookmaker:
        - Calcolo del numero di volte in cui la quotazione minima, di ogni match, si riferisca al risultato corretto.
        - Calcolo del numero di match_odd
        - Calcolo della percentuale.
    b) Calcolo della media delle percentuali ottenute, per tutti i bookmaker, in precedenza.
    c) Ri-esecuzione dei punti a e b considerando solo i match relativi:
        - alle singole competizioni (trovare successivamente la competizione con una percentuale maggiore).
        - alle singole stagioni di ogni competizione (trovare successivamente la stagione con una percentuale maggiore).
        - ai singoli sport (trovare successivamente lo sport con una percentuale maggiore)
        - ai singoli partecipanti (trovare successivamente il partecipante con una percentuale maggiore).
- Analisi 2
    a) Considerando tutti i match contenuti nel datawarehouse, per ogni bookmaker e per ogni porzione di stagione (prima parte o seconda parte):

- Calcolo del numero di volte in cui la quotazione minima, di ogni match, si riferisca al risultato corretto.
- Calcolo del numero di match_odd
- Calcolo della percentuale.

b) Calcolo della media delle percentuali ottenute, per tutti i bookmaker, in precedenza.
c) Ri-esecuzione dei punti a e b considerando solo i match relativi:
   - alle singole competizioni (trovare successivamente la competizione con una percentuale maggiore).
   - alle singole stagioni di ogni competizione (trovare successivamente la stagione con una percentuale maggiore).
   - ai singoli sport (trovare successivamente lo sport con una percentuale maggiore)
   - ai singoli partecipanti (trovare successivamente il partecipante con una percentuale maggiore).

- Analisi 3 e 4
  a) Considerando tutti i match contenuti nel datawarehouse, per ogni bookmaker:
     - Selezione dei match in cui ci sia stato un KO
     - Calcolo della quotazione media riferita alla squadra vincente
     - Calcolo della quotazione media riferita alla squadra perdente.
  b) Calcolo delle media delle quotazioni medie ottenute, per tutti i bookmaker, in precedenza.
  c) Ri-esecuzione dei punti a e b considerando solo i match relativi:
     - alle singole competizioni (trovare successivamente la competizione con una percentuale maggiore).
     - alle singole stagioni di ogni competizione (trovare successivamente la stagione con una percentuale maggiore).
     - ai singoli sport (trovare successivamente lo sport con una percentuale maggiore)
     - ai singoli partecipanti (trovare successivamente il partecipante con una percentuale maggiore).
- Analisi 5
  a) Considerando tutti i match contenuti nel datawarehouse, per ogni bookmaker:
     - Selezione dei match in cui ci sia stato un KO
     - Calcolo del numero di volte in cui, per ogni match, la quotazione della squadra perdente sia la più bassa
     - Calcolo del numero di match_odd
     - Calcolo della percentuale
  b) Calcolo delle media delle percentuali ottenute, per tutti i bookmaker, in precedenza.
  c) Ri-esecuzione dei punti a e b considerando solo i match relativi:
     - alle singole competizioni (trovare successivamente la competizione con una percentuale maggiore).
     - alle singole stagioni di ogni competizione (trovare successivamente la stagione con una percentuale maggiore).
     - ai singoli sport (trovare successivamente lo sport con una percentuale maggiore)
     - ai singoli partecipanti (trovare successivamente il partecipante con una percentuale maggiore).

4. Progettazione concettuale

Sulla base del glossario dei requisiti è stato prodotto lo schema di fatto DFM riportato in figura 9.
Per ottimizzare le analisi ed evitare di complicare troppo la struttura, gli attributi dimensionali relativi a risultato e partecipanti sono stati collegati direttamente al fatto.
Tra il fatto e l'attributo dimensionale dei partecipanti vi è un arco multiplo che rappresenta i due partecipanti che prendono parte al match.
I due partecipanti sono stati distinti attraverso le notazioni "Home" e "Away", che hanno una semantica precisa nel calcio, e, per il tennis, rappresentano dei nomi convenzionali.
Lo schema presenta inoltre una gerarchia condivisa tra i partecipanti e i match, le quali rappresentano lo sport e la nazione di riferimento.
La misura odd_x assumerà un valore solo per i fatti riferiti a partite di calcio, in quanto una partita di tennis non può finire in pareggio.
Per i match di tennis, il partecipante in casa rappresenta il vincitore, quello in trasferta il perdente.

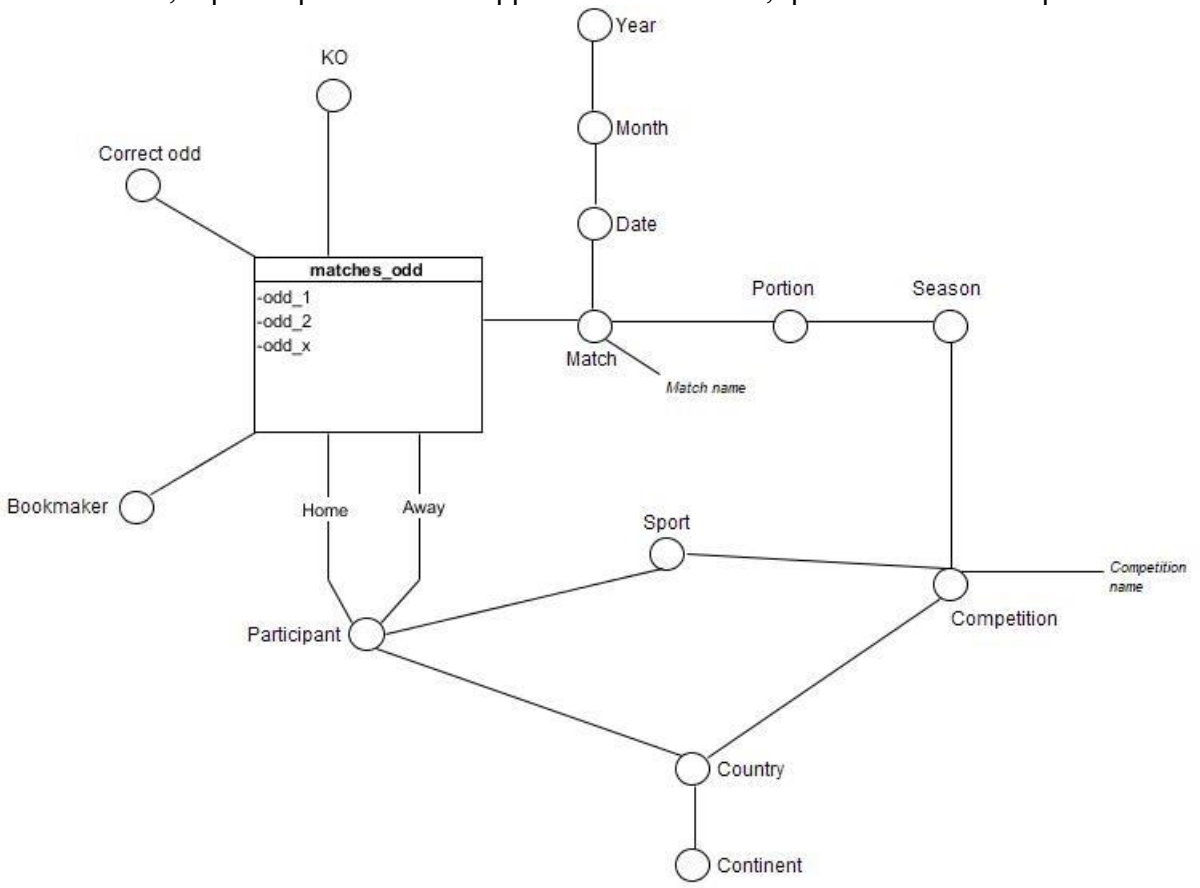

Figura 9: Progettazione concettuale, schema di fatto DFM.

5. Livello dei dati riconciliati (ODS)

L'architettura contiene un livello dei dati riconciliati.
Per questo livello è stata scelta una struttura transitoria in quanto si memorizzano informazioni relative a eventi che non mutano nel tempo.
Le uniche informazioni che possono cambiare sono quelle relative ai nomi delle competizioni e ai nomi delle squadre di calcio. Per quanto riguarda queste informazioni non è necessario mantenere la storicizzazione: una competizione o una squadra verrà sempre rappresentata col nome attuale.

Integrazione degli schemi locali

Lo schema riconciliato è il risultato dell'integrazione effettuata sugli schemi locali.
Questa operazione è stata suddivisa in quattro fasi:

- Preintegrazione: scelta della strategia di integrazione da attuare.
- Comparazione degli schemi: identificare correlazioni e conflitti tra gli schemi da integrare
- Allineamento degli schemi: risoluzione dei conflitti applicando primitive di trasformazione
- Fusione degli schemi: costruzione dello schema riconciliato.

Preintegrazione

La figura 10 mostra la strategia di integrazione scelta.
Inizialmente vengono integrati i due schemi relativi ai match di calcio.
Successivamente viene integrato TD con lo schema Cities. Questo viene fatto perché TD contiene le città in cui vengono giocati i match, le quali sono memorizzate in Cities.
La terza integrazione prevede l'integrazione tra il risultato della seconda integrazione con ATPMDS, in modo da poter collegare i tennisti dei match con le relative nazionalità.
Gli schemi relativi ai match di tennis e alle partite di calcio vengono integrati nel passo 5.
L'ultima integrazione aggiunge lo schema relativo alle nazioni.

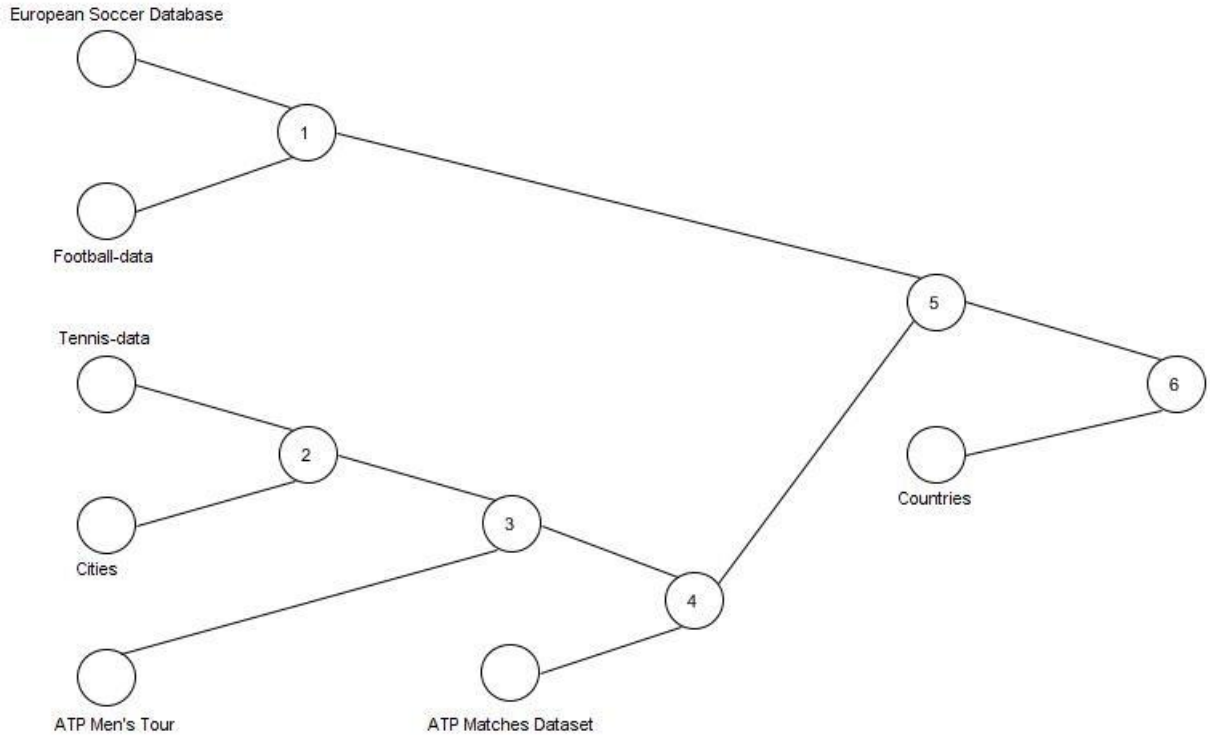

Figura 10: Preintegrazione.

Di seguito sono riportati i dettagli delle diverse integrazioni.

Integrazione 1

Comparazione

1. Sinonimia tra Div e League, entrambi rappresentano un campionato.
2. Conflitto strutturale: in FD le associazioni tra Match e Team sono identificative, in ESDB Match ha un proprio identificativo
3. Sinonimia tra home_team_goal e FTHG
4. Sinonimia tra away_team_goal e FTAG

Allineamento

1. Adozione del nome League
2. Adozione della soluzione proposta da ESDB e, nel caso di record di FD, inserimento di una chiave surrogata.
3. Adozione del nome home_team_goal
4. Adozione del nome away_team_goal.

Fusione

Schema E/R

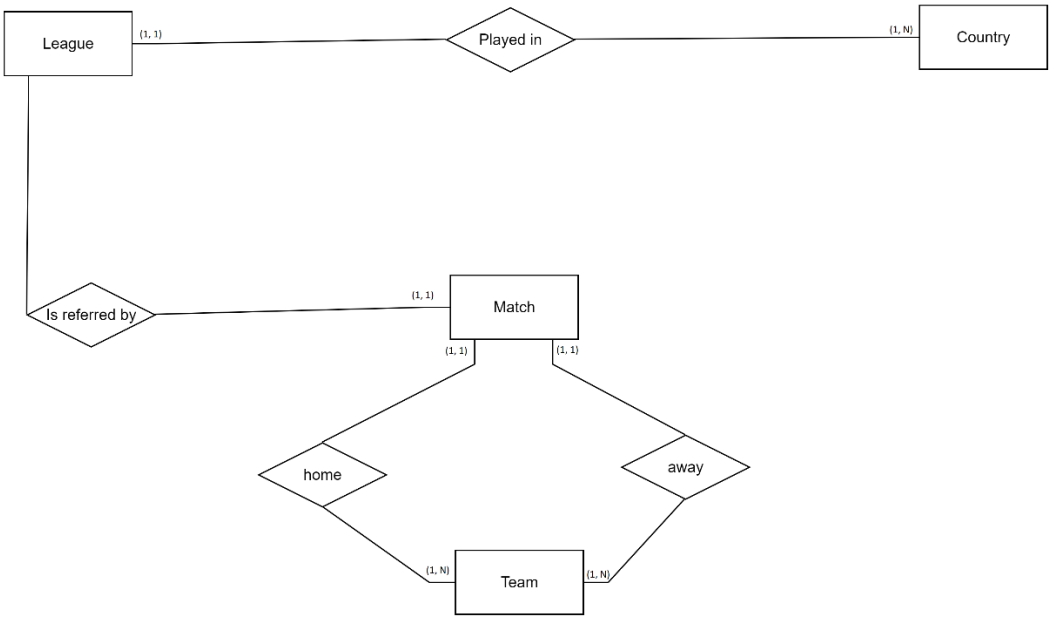

Figura 11: Schema E/R risultante dall'integrazione 1.

Attributi

Team (
    id
    team_api_id
    team_long_name
);

Country (
    id
    name
);

League (
    id
    name

Match (
    id
    date
    B365D
    LBD
    PSD
    SJD
);

Is_referred_by (
    season,
    stage
)

Home (
    home_team_goal
    B365H
    LBH
    PSH
    SJH
)

Away (
    away_team_goal
    B365A
    LBA
    PSA
    SJA

);

)

Integrazione 2

Comparazione

1. Sinonimia tra le entità City e Location.

Allineamento

1. Utilizzo del nome Location.

Fusione

Schema E/R

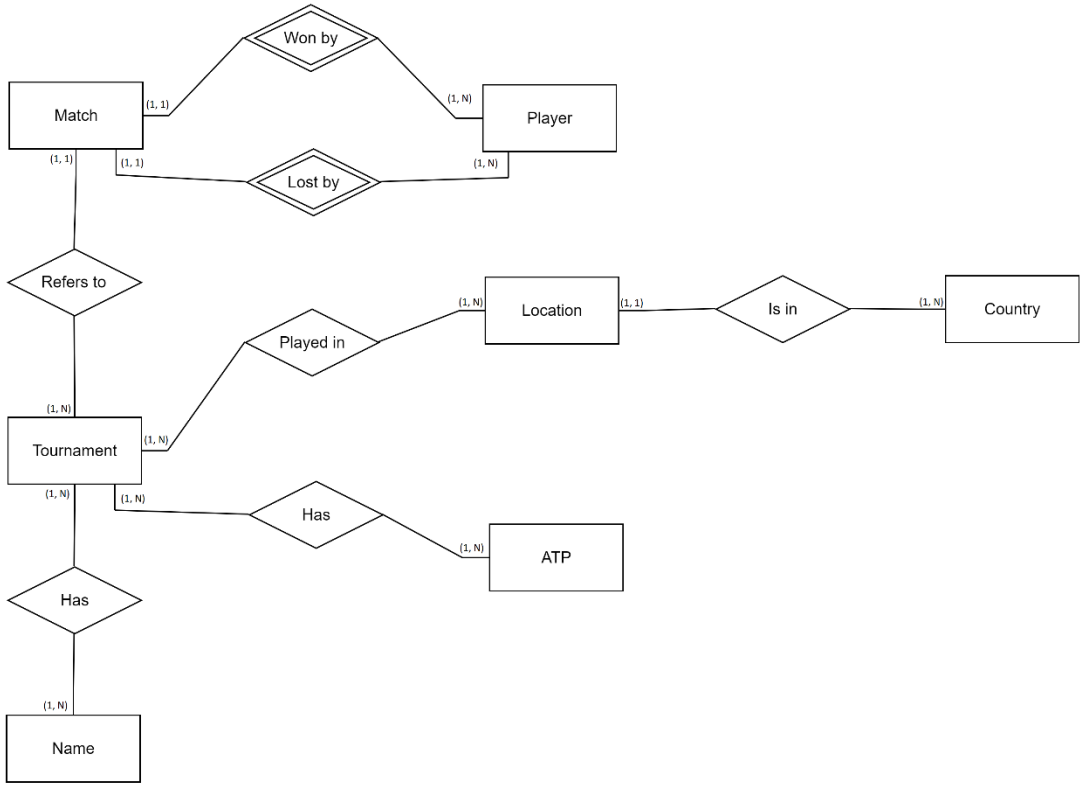

Figura 12: Schema E/R risultante dall'integrazione 2.

Attributi:

Tournament (
name
)

Name (
name
)

Location (
Location
)

Match (
Date
)

Refers_to (
Round
)

Player (
name
)

Country (
Iso3
)

Won_by (
Wsets
B365W
LBW
PSW
SJW
)

Lost_by (
Lsets
B365L
LBL
PSL
SJL
)

Integrazione 3

Non ci sono conflitti tra i due schemi.
Il risultato è lo schema risultante dall'integrazione 2.

Integrazione 4

Comparazione

1. Sinonimia tra Country e Nationality.
2. Sinonimia tra iso3 e code_3.

Allineamento

1. Utilizzo del nome Country e modifica del nome dell'associazione "Has" in "Nationality".

Fusione

Schema E/R

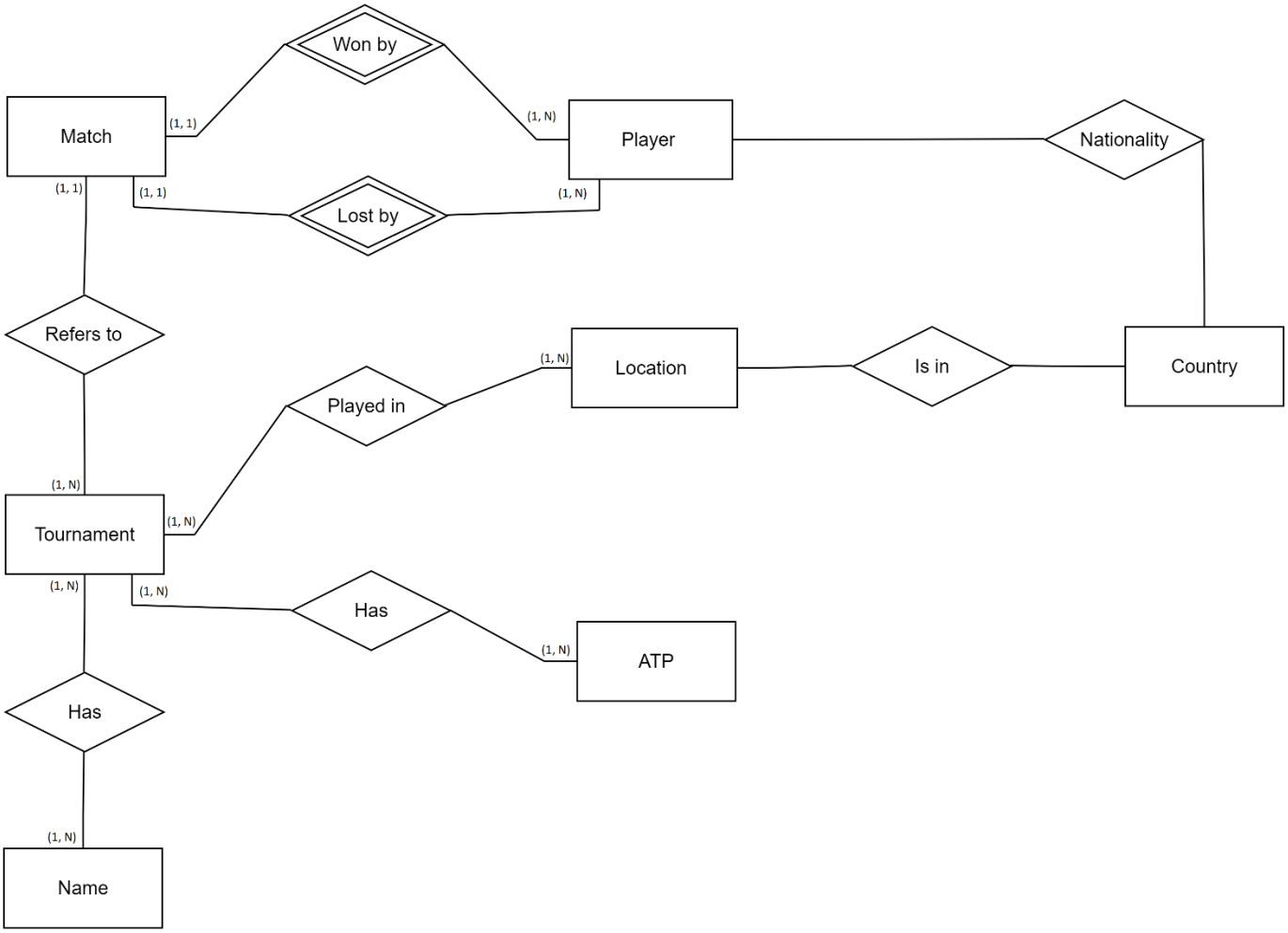

Figura 13: Schema E/R risultante dall'integrazione 4.

Attributi

Tournament (
name
)

Name (
name
)

Refers_to (
Series
Round
)

Player (
name
)

Won_by (
Wsets
B365W
LBW
PSW
SJW
)

Location (
Location
)

Match (
Date
)

Country (
code_3
)

Lost_by (
Lsets
B365L
LBL
PSL
SJL
)

Integrazione 5

Comparazione

1. Sinonimia tra League e Tournament
2. Sinonimia tra Team e Player
3. Sinonimia tra team_long_name e Player.name
4. Conflitto strutturale: in 4 le associazioni tra Match e Team sono identificative, in 1 Match ha un proprio identificativo.
5. Conflitto strutturale: in 1 League è direttamente associata a Country, in 4, tra le due entità, c'è l'entità Location.
6. Sinonimia tra home e Won_by
7. Sinonimia tra away e Lost_by
8. Sinonimia tra home_team_goals e Wsets
9. Sinonimia tra away_team_goals e Lsets
10. Sinonimia tra B365H e B365W
11. Sinonimia tra LBH e LBW
12. Sinonimia tra PSH e PSW
13. Sinonimia tra SJH e SJW
14. Sinonimia tra B365A e B365L
15. Sinonimia tra LBA e LBL
16. Sinonimia tra PSA e PSL
17. Sinonimia tra SJA e SJL

Allineamento

1. Adozione del termine Competition
2. Adozione del termine Participant
3. Utilizzo del termine name.
4. Adozione della struttura utilizzata in 1 e, nel caso di record di 3, inserimento di una chiave surrogata.
5. Eliminazione dell'entità Location.

6-17. Adozione della struttura di 1 rinominando home_team_goals in home_points e away_team_goals in away_points.

- Memorizzazione in uno schema esterno dei diversi nomi e dei diversi ATP associati ad un torneo di tennis.
  Utilizzo del primo nome utilizzato per l'attributo name e di una chiave surrogata per l'attributo id.
- Aggiunta di un attributo portion che indichi la porzione della stagione.
- Rimozione degli attributi stage e round.

Fusione

Schema E/R

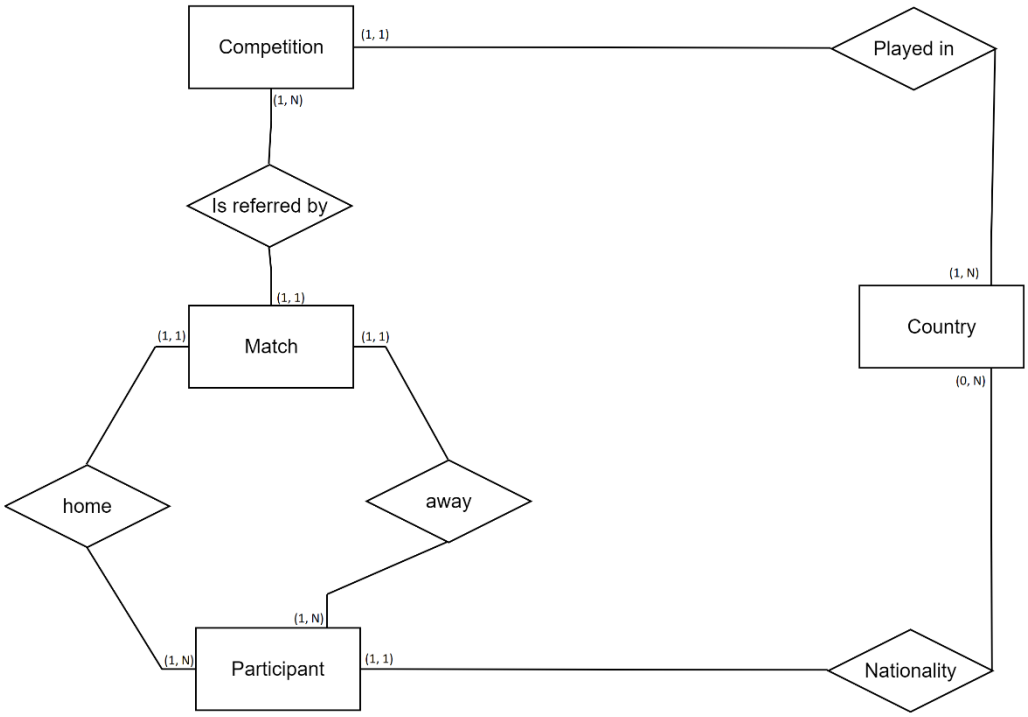

Figura 14: Schema E/R risultante dall'integrazione 5.

Attributi

Participant (
    id
    name
);

Country (
    id
    name
);

Competition (
    id
    name
    sport
);

Match (
    id
    date
    B365D
    LBD
    PSD
    SJD
);

Is_referred_by (
    season
    portion
);

Home (
    home_points
    B365H
    LBH
    PSH
    SJH
);

Away (
    away_points
    B365A
    LBA
    PSA
    SJA
);

Integrazione 6

Comparazione e allineamento

I due schemi non hanno conflitti.

Fusione

Schema E/R

Lo schema E/R è lo stesso dell'integrazione 5.

Attributi

Participant (
- **id**
- name

);

Country (
- **id**
- name
- code-3
- continent
- sub-continent

);

Competition (
- **id**
- name
- sport

);

Is_referred_by (
- season
- portion

);

Match (
- **id**
- date
- B365D
- LBD
- PSD
- SJD

);

Home (
- home_points
- B365H
- LBH
- PSH
- SJH

);

Away (
- away_points
- B365A
- LBA
- PSA
- SJA

);

Schema logico dell'ODS

La figura 15 mostra lo schema logica dell'ODS, risultato dell'integrazione degli schemi locali.
Per comodità tutti gli attributi (escluse le chiavi surrogate), sono stati rappresentati attraverso stringhe.

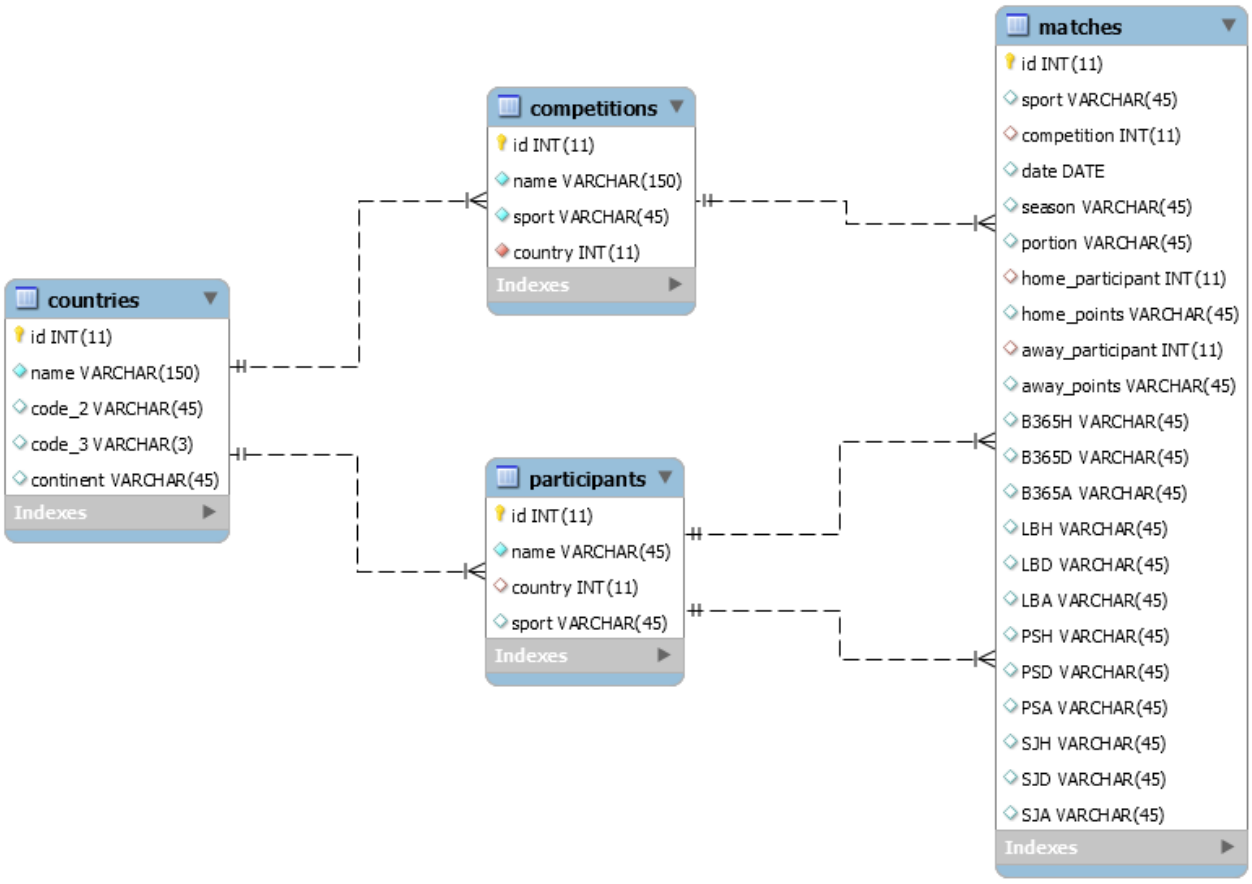

Figura 15: Modello logico-relazionale dell'ODS.

Processo di ETL

Questo processo utilizza una Staging Area in cui alcune tabelle sono memorizzate in maniera permanente (tabelle di lookup che permettono di effettuare un mapping tra le chiavi surrogate delle tabelle dell'ods e le chiavi dei dataset originali). Altre tabelle, invece, sono memorizzate in maniera temporanea.
Per la realizzazione del processo, oltre a script SQL, è stato utilizzato il linguaggio Java.
Sono state utilizzate 4 tabelle di lookup:

- Competitions_football_lookup: associa all'identificativo di ogni competizione nell'ods, gli identificativi del dataset FD e del dataset ESDB.

La prima fase di questo processo prevede l'esportazione del database ESDB in un file csv per ogni tabella.
I file relativi ai campionati e alle nazioni vengono puliti attraverso la classe FileCleaner, i cui metodi eliminano i caratteri speciali e/o indesiderati.
Per quanto riguarda il file dei campionati di FD, i nomi delle nazioni "England" e "Scotland" vengono sostituiti col nome ufficiale del Regno Unito di Gran Bretagna.

Alimentazione della tabella Countries (Creazione del file countries_cleaned)

Trascurando i possibili cambiamenti nell'ordinamento delle nazioni, questa tabella viene riempita una sola volta durante tutta la vita del Datawarehouse.
Le righe presenti nel dataset Countries vengono estratte staticamente.
Su di esse viene effettuata una pulizia volta alla rimozione di caratteri speciali (e/o indesiderati) e alla proiezione delle colonne rilevanti:

- Name

- Code_2
- Code_3
- Continent
- Sub_continent.

L'operazione di pulitura viene effettuata tramite codice Java, attraverso metodo clean_countries della classe FileCleaner.
Questo metodo lavora direttamente sul file csv del dataset e produce in output un nuovo file csv (countries_cleaned).
Al dataset originale è stata aggiunta una nazione "_None" che verrà utilizzata per rappresentare la nazionalità dei partecipanti per i quali non è presente una nazione associata.
Il file risultante dal metodo Java viene caricato, tramite uno script sql, in una tabella temporanea della staging area, la quale verrà successivamente utilizzata per riempire la tabella countries dell'ODS.

Procedure di estrazione e pulitura necessarie per l'alimentazione delle tabelle Competitions, Participants e Matches.

Per poter alimentare le tabelle rimanenti dell'ODS è stato necessario effettuare delle operazioni pregresse comuni a tutti le tabelle.
I match dei diversi dataset vengono estratti in maniera incrementale utilizzando come parametro la data del match. Il risultato dell'estrazione viene salvato in un file csv matches_<nome_dataset>cleaned.
Prima di eseguire l'estrazione viene memorizzata in un file, tramite uno script SQL, la data massima presente nella tabella Matches dell'ODS (max_date).
Nel caso in cui questa data sia nulla viene memorizzata la data 31/12/2007, in modo che vengano estratti match la cui data sia dal 2008 in poi.
Questo file viene utilizzato dai metodi della classe FileCleaner per selezionare le righe che si riferiscano ad un match con una data successiva a quella presente nel file.
Una mappa che associa, ad ogni posssibile round di un match di tennis, la portion corrispondente è stata costruita manualmente e memorizzata in un file csv (portion_tennis).
Questa mappa viene tradotta, dalla classe FileCleaner, in una HashMap Java.

Estrazione dei match di ESDB (Creazione di matches_esdb_cleaned, countries_esdb_cleaned, leagues_esdb_cleaned)

L'estrazione viene effettuata dal metodo clean_esdb().
Il metodo inizialmente scrive nel file di output (matches_esdb_cleaned) l'intestazione delle colonne rilevanti.
Si effettua poi una scansione lineare del file csv relativo ai match del dataset ESDB.
Per ogni riga si trasforma la data in formato DD/MM/YYYY e si confronta questa data con max_date.
Nel caso in cui la data sia successiva a max_date, si inseriscono le colonne rilevanti nel file di output, rimuovendo i caratteri speciali e/o indesiderati.
Oltre all'inserimento delle colonne rilevanti, viene anche calcolata la porzione di stagione del match in base al mese in cui il match è stato giocato:

- Se il mese è compreso tra luglio e ottobre il match si riferisce alla prima parte della stagione (portion=1).
- Altrimenti il match si riferisce alla seconda parte della stagione (portion=2).

Matches_esdb_cleaned, insieme ai file puliti relativi ai campionati e alle nazioni (countries_esdb_cleaned, leagues_esdb_cleaned) vengono importati in un database SQL (european_soccer).
Dopo l'importazione vengono modificati, per tutte le tabelle, i riferimenti alle nazioni "England" e "Scotland", in modo che si riferiscano alla nazione che rappresenta il Regno Unito.

Estrazione dei match di FD (Creazione di matches_fd_cleaned)

I match del dataset FD sono memorizzati in più file csv, uno per ogni stagione di ogni campionato.
I nomi dei file è seguono il pattern:

<id_campionato><stagione in formato YY-YY>. Gli id dei campionati contenenti solo due caratteri sono seguiti dal carattere '_'.
I diversi file non hanno la stessa struttura e non contengono le quote degli stessi bookmaker.
Per questo motivo è stata creata manualmente una mappa che, ad ogni campionato (indicato dall'id eventualmente seguito da '_') associa la colonna di inizio delle informazioni relative alle quotazioni (considerando 0 come prima colonna).
Questa mappa è stata memorizzata in un file csv (odds_map).
Il metodo Java clean_fd() utilizza il path della cartella contenente tutti i file dei match del dataset FD.
Inizialmente trasforma il file odds_map in una HashMap Java.
Successivamente, per ogni file della directory, legge la prima riga (la quale rappresenta l'intestazione).
Solo per il primo file, inserisce, nel file di output (matches_fd_cleaned), l'intestazione del file contenente l'intestazione delle colonne di interesse a cui vengono aggiunte id, portion, season e le intestazioni relative alle colonne delle quotazioni dei bookmaker di riferimento.
Per ogni file, estrae l'id del campionato dal nome del file e cerca la colonna di inizio delle quote nell'HashMap corrispondente.
Sulla base dei nomi delle colonne, contenuti nella prima riga del file, costruisce poi una nuova mappa contenente, per i bookmaker di interesse, la colonna relativa alla quotazione per la vittoria della squadra in casa.
Per le righe successive trasforma la data del match in formato DD/MM/YYYY e, se quest'ultima è maggiore di max_date:

1. Scrive, sul file di output, il numero della riga, memorizzato in un contatore. Questo sarà l'identificativo univoco del match.
2. Scrive le prime 6 colonne sul file di output, eliminando i caratteri speciali e/o indesiderati.
3. Calcola la season in base alla data del match e la scrive sul file di output.
4. Calcola la portion in base al mese del match e la scrive sul file di output.
5. Cicla su tutte le chiavi della seconda mappa per cercare le colonne relative alle quotazioni di riferimento, scrivendo i valori nel file di output.
   Nel caso in cui non ci siano le quotazioni di uno dei bookmaker scrive le colonne vuote.

Estrazione dei match di ATPMT (Creazione di matches_atpmt_cleaned)
Il metodo clean_atpmt() scrive inizialmente nel file di output (matches_atpmt_cleaned) l'intestazione delle colonne rilevanti (comprese id, portion e season).
Successivamente, per ogni riga del file csv dei match:

1. Trasforma la data in formato DD/MM/YYYY
2. Controlla che la data sia maggiore di max_date (in caso contrario passa alla riga successiva senza eseguire il punto 3)
3. Scrive sul file di output il numero della riga che sta leggendo (memorizzato in un contatore) che rappresenta l'id del match.
4. Scrive sul file di output le colonne di interesse, la season (anno della data del match) e la portion (calcolata attraverso la HashMap portion_tennis).
   Le colonne di riferimento vengono pulite cancellando i caratteri speciali e/o indesiderati.

Estrazione dei match di TD (Creazione di matches_td_cleaned)
I diversi file del dataset sono stati inizialmente fusi in unico file cercando unificando la struttura delle colonne.
Il metodo clean_td() effettua operazioni simili al metodo clean_atpmt(), ad esclusione della conversione della data, in quanto nel dataset la data è già memorizzata nel formato corretto.

Alimentazione della tabella Competitions

Competizioni calcistiche
L'estrazione delle competizioni calcistiche viene effettuata in maniera statica su entrambi i dataset.

Inizialmente viene eseguito uno script SQL che esporta, in un file csv (competitions_esdb), il risultato dell'operazione di join tra i campionati e le nazioni di ESDB.
Successivamente, tramite il metodo integrate della classe CompetitionsMaker, viene effettuata un'integrazione dei due file.
Il file di ouput (competitions_football) contiene i seguenti campi:

- Id_esdb: identificativo della competizione riferito ad ESDB
- Id_fd: identificativo della competizione riferito a FD
- name: nome della competizione di ESDB
- country: nome della nazione
- country_id_esdb: identificativo della nazione riferito a ESDB.

Viene considerato il nome di ESDB perché per le squadre dovrà essere memorizzato il nome di ESDB, in modo da poter rilevare i cambiamenti nei nomi.
Due competizioni vengono considerate uguali se hanno la stessa nazione di riferimento e il nome di una è contenuto nel nome dell'altra.
Il metodo integrate esegue un nested loop sulle righe di competitions_esdb (file1) e sulle righe del file csv che contiene le informazioni dei campionati di FD (file2).
Per ogni iterazione del ciclo interno, se le nazioni rappresentate hanno lo stesso nome, si effettua un confronto dei nomi delle competizioni: nel caso in cui, l'insieme delle parole che compongono uno dei due nomi sia contenuto nell'insieme delle parole che compongono l'altro, le due competizioni combaciano.
In questo caso si scrivono le informazioni relative su competitions_football, si memorizza il numero della riga corrente del file2, e si procede con una nuova iterazione del ciclo esterno.
Se un'iterazione del ciclo esterno finisce senza una corrispondenza tra competizioni, vengono scritte, su competitions_football, le informazioni della competizione relativa alla riga corrente del file1, lasciando vuote le colonne relative alle informazioni del file2.
Una volta terminato il nested loop, si esegue un ciclo sul file2 che, per ogni riga che non abbia dato una corrispondenza con una competizione del file1, scrive su competitions_football le informazioni relative.
Per confrontare i nomi delle competizioni presenti nei due file, viene eliminata la prima parola relativa ai nomi delle competizioni di ESDB, in quanto questa rappresenta sempre il nome della nazione.

L'alimentazione della tabella competitions dell'ods si basa su una tabella di lookup (competitions_football_lookup) la quale, per ogni competizione inserita in ods.competitions, contiene:

- id_ods: chiave surrogata dell'ods
- id_fd: identificativo relativo a FD
- id_esdb: identificativo relativo a ESDB.

Il file delle competizioni di fd viene importato in una tabella temporanea della staging area (divisions_fd).
Per selezionare le competizioni, contenute solo in FD e che abbiano subito una modifica nel nome, si effettuano due join:

- La prima tra divisions_fd e competitions_football_lookup sull'identificativo della competizione relativo a FD, selezionando le righe in cui sia nullo l'identificativo di esdb.
- La seconda tra il risultato della prima e ods.competitions sull'identificativo della competizione relativo all'ODS, selezionando le righe in cui il nome in divisions_fd sia diverso dal nome in ods.competitions.

Si effettua successivamente un'update di ods.competitions .
In maniera analoga vengono aggiornati i nomi delle competizioni presenti in esdb, in questo caso, nella prima join, non sarà necessario controllare che l'identificativo relativo a FD sia nullo, in quanto, in caso di corrispondenza nelle competizioni tra i due dataset, è stato selezionato il nome di ESDB.
In seguito competitions_football viene importato in una tabella della staging area.
Si vanno poi a memorizzare in una tabella temporanea della staging area (competitions_to_insert) le competizioni presenti in competitions_football e non presenti in ods.competitions, effettuando una left outer join tra competitions_football e competitions_football_lookup.
Successivamente, effettuando una join sul nome della nazione, vengono aggiunti a competitions_football i riferimenti alle nazioni presenti in ods.countries.

Le righe presenti in competitions_football vengono inserite in ods.competitions creando una chiave surrogata incrementale per ciascuna di esse.
Vengono poi inserite le righe di competitions_football all'interno della tabella di lookup competitions_football_lookup.
Prima di effettuare l'inserimento devono essere recuperate le chiavi surrogate dell'ods effettuando una join tra competitions_football e ods.competitions sul nome della competizione e sull'identificativo della nazione.
Come ultima operazione viene aggiornata una nuova tabella di lookup (countries_football_lookup), utile per l'inserimento dei partecipanti.
Questa tabella associa gli identificativi di ESDB, relativi alle nazioni, e gli identificativi di FD, relativi ai campionati, con la chiave surrogata della nazione corrispondente dell'ods.

Competizioni di tennis
L'estrazione delle competizioni di tennis viene effettuata dai match estratti nelle operazioni pregresse.
Le informazioni che le riguardano sono presenti nei dataset TD e ATPMT, ognuno dei quali è composto da un singolo file denormalizzato.
Non esiste una colonna che identifichi in maniera univoca una competizione in quanto:

- Una competizione può essere giocata in più location
- Una competizione può avere valori della colonna ATP diversi
- Due competizioni diverse possono avere valori della colonna ATP uguali
- Una competizione può avere nomi diversi.

E' stato quindi scelto di identificare una competizione in base alla nazione in cui viene giocata.
Se una nazione ospita più competizioni, le diverse competizioni verranno rappresentate come un'unica competizione.
Il riferimento alla competizione non è direttamente presente nei dataset, per questo motivo è stato utilizzato il dataset cities.
I file di ATPMT e TD, dopo la pulitura, sono stati importati in due tabelle temporanee (matches_atpmt e matches_td) della staging area.
Ognuna delle due tabelle è stata aggiornata inserendo la nazione di riferimento, effettuando una join con la tabella cities (avendo precedentemente pulito il file relativo e avendolo importato nella staging area) sul nome della location (per le tabelle dei match) e sul nome della città (per la tabella cities).
Il risultato di queste join ha evidenziato il fatto che i nomi delle città nel dataset cities non siano univoci.
Per questo motivo sono state cancellate manualmente le righe relative agli omonimi delle location dei match.
Le location non presenti in cities sono, invece, state aggiunte manualmente.
Inizialmente vengono trovate le competizioni in comune ai due dataset effettuando una join tra le nazioni (indicate tramite codice iso3) presenti in matches_atpmt e le nazioni presenti in matches_td e memorizzando i risultati in una tabella competitions_tennis.
Successivamente vengono effettuate due join:

- Una join semplice tra competitons_tennis e ods.countries in modo da estrarre il nome e la chiave surrogata della nazione
- Una left outer join tra il risultato della prima join e ods.competitions in modo da selezionare solo le competizione che non siano già presenti in ods.

Le righe risultanti dai due join vengono inserite in ods.competitions, selezionando come nome della competizione la stringa ottenuta concatenando il codice iso della nazione con la stringa " Open".
In maniera analoga verranno inseriti, per ognuno dei due dataset, sono stati inserite le competizioni presenti solo nel dataset.
Per trovare i match presenti in un dataset è stata effettuata una outer join sulla nazione (rappresentata in codice iso3).
Dopo ogni inserimento nell'ods, si inseriscono le competizioni anche in una tabella di lookup (competitions_tennis_lookup). Prima di inserire i dati nella tabella di lookup bisogna recuperare la chiave surrogata dell'ods, per fare questo si effettua una join tra la tabella contenente le competizioni da inserire e ods.competions, indicando come condizione del join che il codice iso_3 della nazione, memorizzato nella prima tabella e concatenato alla stringa " Open", sia uguale al nome della competizione.

Competitions_tennis_lookup memorizza, per ogni competizione, la chiave surrogate dell'ods (id_ods) e il codice iso_3 della nazione relativa (country).

Partecipanti di calcio

L'estrazione dei partecipanti viene effettuata utilizzando i match estratti nelle operazioni pregresse.
Per ESDB non è possibile estrarre i partecipanti solo dalla tabella delle squadre perché, nello schema locale, i partecipanti non sono collegati direttamente alla nazione.
Per recuperare questo collegamento è necessario effettuare una doppia join tra partecipanti e match (una per la squadra in casa e una per la squadra in trasferta).
Viene esportato il risultato della join in un file csv (participants_esdb) sul quale sono memorizzati, per ogni singolo partecipante, il nome (team_long_name), il team_api_id (identificativo della squadra utilizzato come chiave esterna in ESDB) e l'identificativo della nazione.
Successivamente, per ogni riga di matches_fd_cleaned, viene letto il nome della squadra in casa.
Se questo nome non è già presente in una HashMap di nomi:

- si scrive il nome e il codice del campionato sul file csv participants_esdb
- si aggiunge il nome alla mappa.

In caso contrario non si effettua alcuna operazione.
La stessa operazione viene effettuata relativamente al nome della squadra in trasferta.
In questo modo si inserisce ogni una sola volta.
I file participants_esdb e participants_fd vengono integrati effettuando un nested loop.
Ad ogni iterazione del ciclo interno, i nomi delle due squadre vengono confrontati: se l'insieme delle parole che costituisce un nome è contenuto nell'insieme delle parole che costituisce l'altro, le due squadre sono considerate combacianti.
In questo caso viene scritta sul file participants_football una riga contenente:

- Nome di esdb (adottato come nome della squadra)
- Nome di esdb
- Nome di fd
- Team_api_id di esdb
- Esdb (dataset di riferimento).

Viene poi memorizzato, in una HahsMap m, il numero della riga del file relativo al ciclo interno.
Se un'iterazione del ciclo esterno si conclude senza una corrispondenza tra due squadre, le informazioni della squadra vengono scritte su participants_football, lasciando vuote le colonne relative all'altro file e inserendo, come nome adottato. In questo caso il nome adottato sarà, ovviamente, il nome del file relativo al ciclo esterno.
Una volta concluso il nested loop, si effettua un ciclo sul file relativo al ciclo interno del nested loop.
Per tutte le righe non presenti nella HashMap m, si scrivono su participants_football le informazioni relative alla squadra (lasciando vuote le colonne relative all'altro file).
Il file participants_football viene importato in una tabella della staging area.
Dopo una prima esecuzione dell'algoritmo ci si è accorti che esso non considerasse i diminutivi.
A questo proposito sono stati creati manualmente due file che sistemassero le situazioni non prese in considerazione dall'algoritmo:

- Sinonimi: indica le parole da considerare uguali (ad esempio Man e Manchester)
- Squadre: indica i nomi delle squadre dei due dataset che corrispondono alla stessa squadra (esempio: Reggio Calabria e Reggina).

L'algoritmo prende in input questi due file e li utilizza per effettuare il controllo sui nomi.
Per tutte le squadre presenti nel database european_soccer, si controlla che queste non abbiano subito variazioni nel nome.
Questo controllo può essere fatto solo su ESDB poiché FD non è un dataset denomarlizzato e l'unico identificativo disponibile per una squadra è il nome stesso.
Le squadre che hanno cambiato nome vengono recuperate effettuando una doppia join:

- Tra la tabella delle squadre di ESDB e participants_football_lookup sul team_api_id della squadra
- Tra participants_football_lookup e ods.participants sulla chiave surrogata

e selezionando le righe in cui il nome della tabella delle squadre di european_soccer è diverso dal nome del partecipante nell'ODS.
Si esegue successivamente un update in modo da aggiornare i nomi ai valori contenuti in esdb.
In seguito, per selezionare i partecipanti di participants_football non presenti nell'ods, si effettua una left outer join che controlli che il team_api_id di ESDB o il nome di FD non siano presenti in participants_football_lookup.
Con l'utilizzo di countries_football_lookup si traducono i codici delle nazioni di esdb e dei campionati di fd in nazionalità dei partecipanti e si aggiorna la tabella contenente i partecipanti da inserire.
I partecipanti vengono inseriti nell'ods e si aggiorna la tabella di lookup participants_football_lookup. Per recuperare la chiave surrogata dell'ods si effettua una join tra la tabella utilizzata per inserire i partecipanti nell'ods e ods.participants, imponendo come condizione di join che il nome e la nazionalità dei partecipanti della prima tabella siano uguali a quelli della seconda.

Partecipanti di tennis
L'estrazione viene effettuata utilizzando i match estratti nelle operazioni pregresse.
Il metodo Java della classe ParticipantsMaker scansione il file matches_td_cleaned e, per ogni riga, estrae i nomi del vincitore e del perdente.
Se ogni nome non è presente in una mappa di nomi, scrive il nome in un file di output (participants_tennis_1) e lo si aggiunge alla mappa
Lo steso metodo esegue la stessa cosa, usando la stessa mappa, sul file matches_atpmt_cleaned.
In questo modo tutti i nomi dei tennisti presenti nei due dataset vengono memorizzati una volta sola.
L'algoritmo si basa sull'assunzione che i nomi utilizzati nei due dataset siano gli stessi, se presenti.

Per l'estrazione delle nazionalità dei tennisti, viene utilizzato ATPMDS.
In maniera analoga a quanto svolto per ATPMT e TD, per ogni riga, tramite il metodo make_nationalities della classe ParticipantsMaker, si estraggono i nomi e le nazionalità (espresse in iso3) del vincitore e del perdente.
Se ogni nome non è presente in una mappa di nomi, si aggiunge il nome alla mappa (memorizzando come valore la nazionalità) e si scrive la coppia nome, nazionalità nel file participants_nationalities.

I file participants_tennis_1 e participants_nationalities vengono successivamente integrati con un nested loop simile a quello utilizzati per i calciatori.
I nomi presenti in ATPMDS sono memorizzati. nel formato (f2) <Nome Cognome>, i nomi presenti in TD e ATPMT sono memorizzati nel formato (f1) <Cognome I.> dove I. è l'iniziale del nome (o le iniziali dei nomi).
I due nomi vengono quindi confrontati:

- Si prendono inizialmente tutti i cognomi presenti nel formato f1 e si controlla che siano presenti nel formato f2. I cognomi saranno contenuti dalla prima alla penultima parola.
- Successivamente si controlla che le iniziali contenute nell'ultima parola del nome nel formato f1, siano le iniziali dei nomi nel formato f2.

Nel caso in cui i due nomi combacino viene scritto sul file di output (participants_tennis) il nome nel formato f1 e la nazionalità presente in participants_nat.
Alla fine di ogni iterazione, che non abbia rilevato una corrispondenza tra nomi, viene scritto su participants_tennis il nome nel formato f1 seguito da una colonna vuota.
Non viene eseguito il ciclo finale su participants_nat in quanto ci interessano solo i partecipanti presenti in ATPMT e TD.

In maniera analoga a quanto effettuato per i calciatori, il file participants_tennis viene importato in una tabella temporanea della staging area.
Viene effettuata una left outer join sul nome con ods.participants, in modo da selezionare i nomi non presenti nell'ods.
I codici iso3 vengono tradotti nelle chiavi surrogate di ods.countries, effettuando una left join sull'attributo code_3 di quest'ultima.

Viene effettuata una left join perché alcuni tennisti non hanno una nazionalità nota.
Le chiavi esterne nulle che rappresentano la nazionalità vengono tradotte in un riferimento ad una nazione fittizia "_None" inserita appositamente in ods.countries.
I tennisti vengono inseriti successivamente in ods.participants.

Match di calcio

Il file matches_fd_cleaned viene importato in una tabella della staging area (matches_fd).
L'estrazione incrementale è già stata effettuata durante la pulizi, i match presenti nei due file saranno tutti da inserire all'interno dell'ods.
Successivamente si estraggono i match in comune tra FD e ESDB effettuando 5 join:

- Tra matches_fd e european_soccer.matches sulla data del match
- Tra matches_fd e participants_football_lookup sul nome della squadra in casa
- Tra matches_fd e participants_football_lookup sul nome della squadra in trasferta
- Tra european_soccer.matches e participants_football_lookup sul team_api_id della squadra in casa
- Tra european_soccer.matches e participants_football_lookup sul team_api_id della squadra in trasferta.

Vengono selezionate le righe in cui le chiavi surrogate dell'ods presenti nelle tabelle di lookup siano uguali per le due squadre in casa e per le due squadre in trasferta dei due dataset.
Come fonte più attendibile viene considerata FD, vengono quindi prese le informazioni in comune presenti in FD.
La tabella risultante dalle join (matches_football) viene aggiornata inserendo gli identificativi delle competizioni, effettuando una join con competitions_football_lookup.
I match vengono successivamente inseriti nell'ODS.

In seguito vengono inseriti i match presenti solo in uno dei due dataset.
Per selezionare i match presenti solo in uno dei due dataset di effettua una left outer join con matches_football, la quale contiene gli identificativi per i match dei due dataset.
L'operazione viene effettuata prima per FD e, successivamente per ESDB.
Le tabelle risultanti dalle left outer join, vengono aggiornate con gli identificativi della competizione e delle squadre partecipanti (effettuando join con competitions_football_lookup e participants_football_lookup).
I match vengono successivamente inseriti.

Match di tennis

Per questa operazione si utilizzano le tabelle matches_atmpt e matches_td create durante l'inserimento delle competizioni di tennis.
Si estraggono inizialmente i match in comune tra i due dataset, effettuando una join tra i due dataset che imponga come condizione di join l'uguaglianza di:

- Data del match
- Nome del vincitore
- Nome del perdente.

Anche questa operazione si basa sulla constatazione che i nomi dei tennisti vengono salvati nello stesso modo nei due dataset.
La tabella risultante dal join (matches_tennis) viene aggiornata inserendo gli identificativi della competizione e dei due partecipanti (effettuando una join su competitions_tennis_lookup e due join su ods.participants).
I match vengono poi inseriti in ods.matches.

Analogamente a quanto fatto per le partite di calcio, si estraggono successivamente i match presenti solamente in uno dei due dataset, effettuando una left outer join tra la tabella dei match del dataset e matches_tennis, la quale contiene gli identificativi dei match dei due dataset.
Le due tabelle risultanti dalle left outer join vengono aggiornate inserendo gli identificativi delle competizioni e dei partecipanti.

I match vengono poi inseriti in ods.matches.

6. Progettazione logica

Il DFM è stato tradotto in uno star schema. Ogni gerarchia è quindi rappresentata da una dimension table.
Le dimensioni degeneri BookMaker, Correct odd, Result e KO sono state importate direttamente nella fact table.
La porzione di gerarchia condivisa relativa alla nazionalità è stata tradotta effettuando uno snowflake e rappresentando le nazioni in una tabella distinta.
La porzione di gerarchia condivisa relativa allo sport è stata tradotta inserendo l'attributo dimensionale sport sia nella dimension table relativa ai partecipanti, sia nella dimension table relativa ai match.
E' stato effettuato un ulteriore snowflake per quanto riguarda la porzione di gerarchia dei match relativa alla data. Le date sono state memorizzate quindi in una tabella distinta.
La chiave primaria della tabella dates è rappresentata dalla data stessa, in quanto questo è un valore che rappresenta univocamente la data e che non verrà mai modificato. Non è quindi necessario l'inserimento di una chiave surrogata.
Sono state generate delle chiavi surrogate per le seguenti tabelle:

- Countries
- Participants
- Matches.

Per comodità, per le tabelle Participants e Countries, sono state memorizzate le chiavi primarie relative alle tabelle corrispondenti nell'ODS.
L'attributo dimensionale match è rappresentato dagli attributi:

- Match: identificativo del match nell'ODS.
- Match_name nome composto dalla concatenazione dei nomi dei partecipanti.

L'attributo dimensionale competition è rappresentato dagli attributi:

- Competition: identificativo della competizione nell'ODS.
- Competition_name: nome della competizione.

Per quanto riguarda la gestione delle slowly changing dimension, per tutte le gerarchie, è stato deciso di adottare una gerarchia dinamica di tipo 1 che consente la gestione di uno scenario temporale "Oggi per ieri". Si considera quindi solo l'attuale configurazione della gerarchia, non tenendo traccia del passato.
Se un partecipante o una competizione dovessero cambiare nome, il nuovo verrà sovrascritto al nome precedente.
Il cambiamento di nome potrà avvenire solo per calciatori o campionati di calcio, in quanto i nomi dei tennisti non subiscono modifiche (a meno di rarissimi cambi anagrafici) e le competizioni di tennis sono state memorizzate con un nome fittizio che contiene il nome dello stato in cui la competizione viene giocata.
Come detto nei capitoli precedenti, i cambiamenti dei nomi delle squadre potranno essere rilevati solo per il dataset ESDB, in quanto FD non contiene identificativi univoci.
Questi sono gli casi in cui il datawarehouse richieda di gestire cambiamenti, in quanto non è stato gestito il cambiamento di ordinamento degli stati.

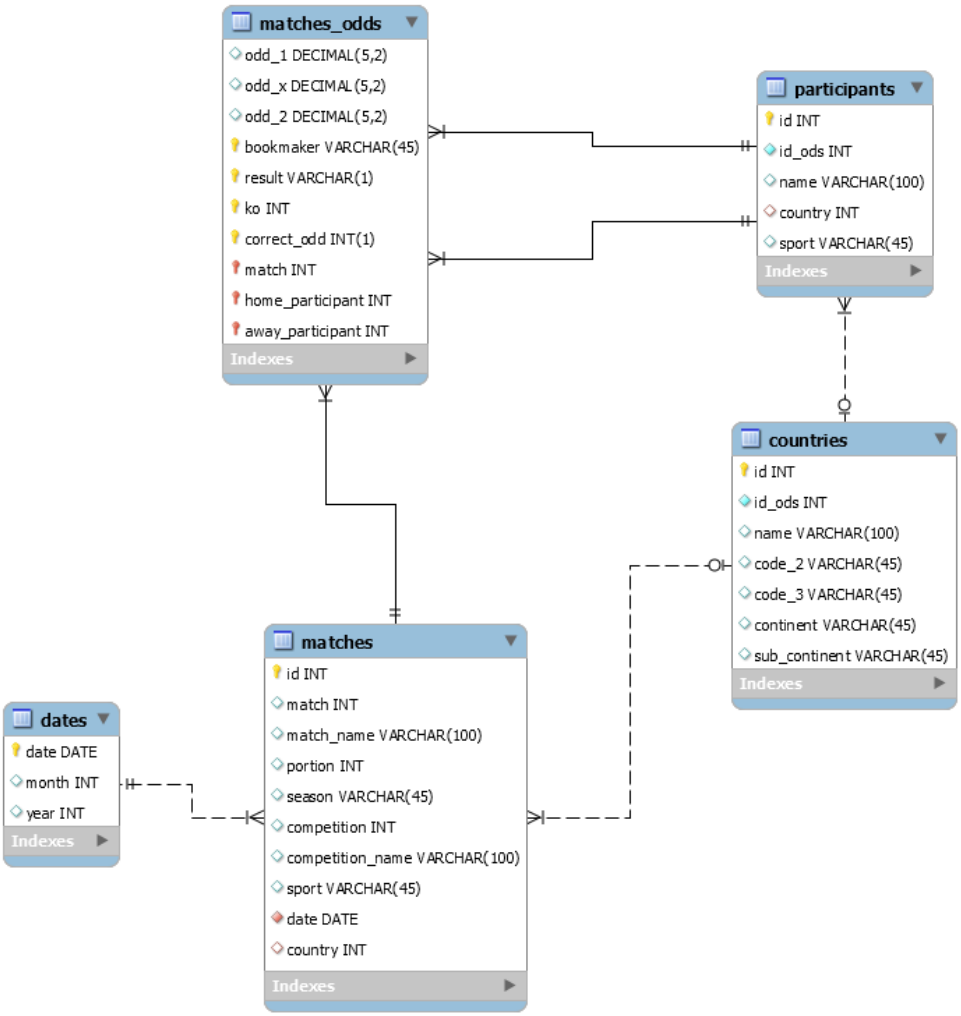

Figura 16: Modello logico-relazionale del datawarehouse.

Processo di ETL

Avendo a disposizione il livello dei dati riconciliati, il processo di ETL viene eseguito direttamente tramite codice SQL.
Il datawarehouse è stato memorizzato in uno schema ss.
Il processo utilizza una staging area temporanea.

Alimentazione e aggiornamento della tabella countries

Anche in questo caso l'alimentazione della tabella viene effettuata una sola volta durante la vita del datawarehouse.
Tutte le righe della tabella ods.countries vengono inserite nella tabella ss.countries.
Per ogni riga viene generata una chiave surrogata incrementale e viene memorizzata la chiave surrogata dell'ods.

Alimentazione e aggiornamento della tabella participants

Questa alimentazione viene effettuata tramite un'estrazione statica dalla tabella ods.participants.
Prima di tutto si controlla se ci sono partecipanti per i quali il nome è da aggiornare.
La tabella ss.participants contiene la chiave surrogata utilizzata in ods.participants.
E' quindi possibile effettuare una join tra ods.participants e ss.participants sulla chiave surrogata riferita all'ods.
Si selezionano poi le righe in cui i due nomi sono diversi e si effettua un'update che aggiorni i valori di ss in base ai valori di ods.
In seguito si selezionano i partecipanti da inserire effettuando una left outer join tra ods.participants e ss.participants, sempre sulla chiave surrogata riferita all'ods.

Le righe della tabella risultante vengono inserite nella tabella ss.participants, generando una chiave surrogata per ognuna di esse e memorizzando la chiave surrogata riferita all'ods.

Aggiornamento delle competizioni

In maniera analoga a quanto effettuato per i partecipanti, si cercano le competizioni che hanno cambiato nome rispetto all'ultimo aggiornamento del datawarehouse.
I nomi delle competizioni sono memorizzati nella tabella ss.matches, la quale contiene anche la chiave surrogata, riferita all'ods, della competizione.
Si esegue quindi una join tra la tabella ods.competitions e ss.matches sulla chiave surrogata, riferita all'ods, della competizione.
Si selezionano poi le righe in cui i nomi delle competizioni siano diversi e, la tabella risultante, viene utilizzata per effettuare un'update della colonna competition_name della tabella ss.matches.

Alimentazione delle tabelle matches e dates

L'alimentazione viene effettuata eseguendo un'estrazione incrementale sulle righe della tabella ods.matches.
Si recupera inizialmente la data maggiore presente in ss.matches (max_date).
Max_date rappresenta il parametro utilie per scartare i match, memorizzati in ods.matches, già presenti in ss.matches.
I match di ods.matches, con una data successiva a max_date, vengono quindi estratti e memorizzati nella tabella matches_to_insert della staging area.
Le date dei match vengono inserite nella tabella ss.dates. Per ogni riga viene calcolato automaticamente il mese e l'anno della data.
Sulle righe della tabella matches_to_insert vengono aggiornate le chiavi esterne relative ai partecipanti e alle nazioni, in modo che esse si riferiscano alle chiavi surrogate del datawarehouse.
Per fare questo, si effettuano 4 join:

- tra matches_to_insert e ss.participants sulla chiave surrogata riferita all'ods, per il participante home
- tra matches_to_insert e ss.participants sulla chiave surrogata riferita all'ods, per il participante away
- tra matches_to_insert e ss.countries sulla chiave surrogate riferita all'ods
- tra matches_to_insert e ods.competitions sulla chiave surrogata dell'ods. Questa join è utile per recuperare il nome della competizione.

I match vengono quindi inseriti nella tabella ods.matches, per ognuno di essi viene generata una chiave surrogata incrementale.
I riferimenti ai partecipanti verranno memorizzati direttamente nella fact table.

Alimentazione della fact table

Questa alimentazione si basa sulla tabella matches_to_insert creata durante la fase di alimentazione delle tabelle matches e dates.
La tabella matches_to_insert viene aggiornata inserendo una colonna che contiene la chiave surrogata del match relativa alla tabella ss.matches.
L'aggiornamento viene effettuato tramite una join tra matches_to_insert e matches sulla chiave surrogata relativa all'ods (attributo match della tabella ss.matches).
Per ogni bookmaker viene effettuata una insert nella tabella ods.matches.
Si selezionano solamente i match per i quali il bookmaker ha fornito le quote (le quote sono diverse da null).
Per ogni inserimento vengono calcolati i valori delle dimensioni degeneri:

- Result: in base al punteggio della squadra in casa e della squadra in trasferta si stabilisce se il risultato sia 1, x o 2.
- KO: si considera KO nei casi in cui:
    - lo sport sia 'football' e il valore assoluto della differenza tra i punti della squadra in casa e i punti della squadra in trasferta sia almeno di 3.
    - Lo sport sia tennis, il partecipante in casa (e quindi il vincente, in quanto il partecipante di casa per il tennis è il vincente) abbia totalizzata 2 o 3 punti e il partecipante in trasferta (il perdente) abbia totalizzato zero punti.
- Correct odd: la quotazione del bookmaker è corretta nei casi in cui:

- Odd_1 sia la quota minore e result='1'
- Odd_x sia la quota minore e result='x'
- Odd_2 sia la quota minore e result='2'.

Nel codice sql si distinguono i casi in cui odd_x sia uguale o diverso da null (per i match di tennis odd_x è null), in modo da evitare di confrontare dei valori numerici con null.

7. Analisi dei risultati

In questa sezione vengono descritte le interrogazioni che sono state effettuate per permettere l'analisi delle quotazioni e dei risultati relativi ai match.
Le interrogazioni sono state realizzate in parte tramite un ambiente di business analytics OLAP definito con Pentaho e in parte con script SQL. In entrambi i casi i risultati sono stati esportati in csv e organizzati in un foglio di lavoro Excel.
Di seguito vengono riportate alcune interrogazioni e parte dei risultati.

Percentuale di correttezze delle quote

Per questa interrogazione ci si è basati sulla dimensione correct_odd. Questa dimensione assume un valore booleano 0/1 che ci permette di capire quali sono i match che sono stati quotati correttamente.
In una fase iniziale i dati sono stati raggruppati per bookmaker.

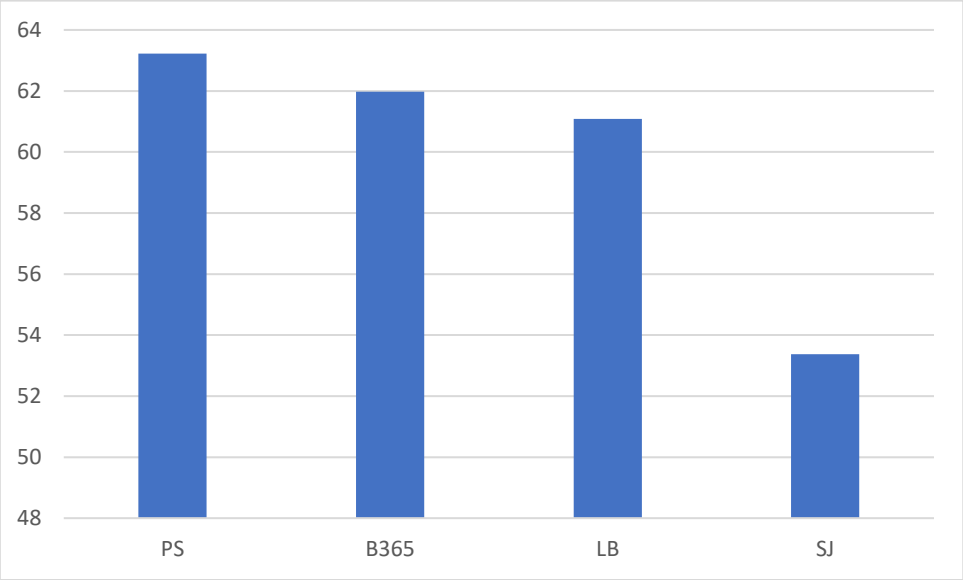

Figura 17: Percentuale di correttezza delle quote per bookmaker.

Da questa prima interrogazione si evince che il miglior bookmaker è Pinnacle.
Nelle interrogazioni successive, i dati sono stati filtrati anche secondo altre dimensioni.

Percentuale di correttezza delle quote per competizione

Per facilitare l'interpretazione dei dati si è scelto di dividere l'analisi per sport per effettuare successivamente un'analisi totale.
In queste analisi si considera la media delle percentuali ottenute dai quattro bookmaker.

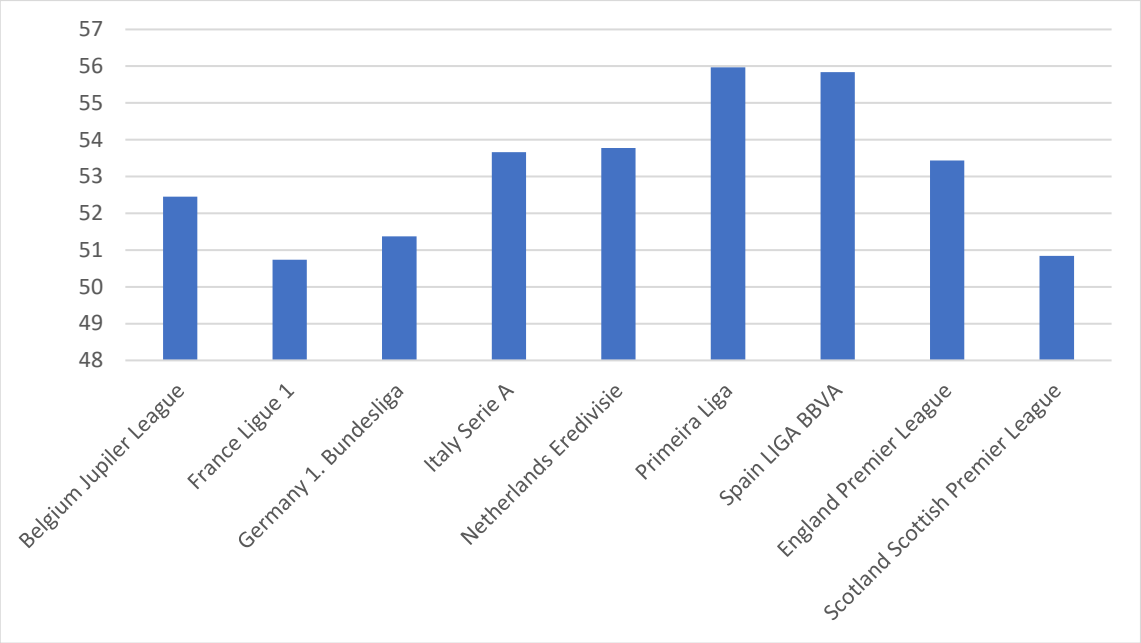

Figura 18: Percentuale di correttezza delle quote per i campionati di calcio.

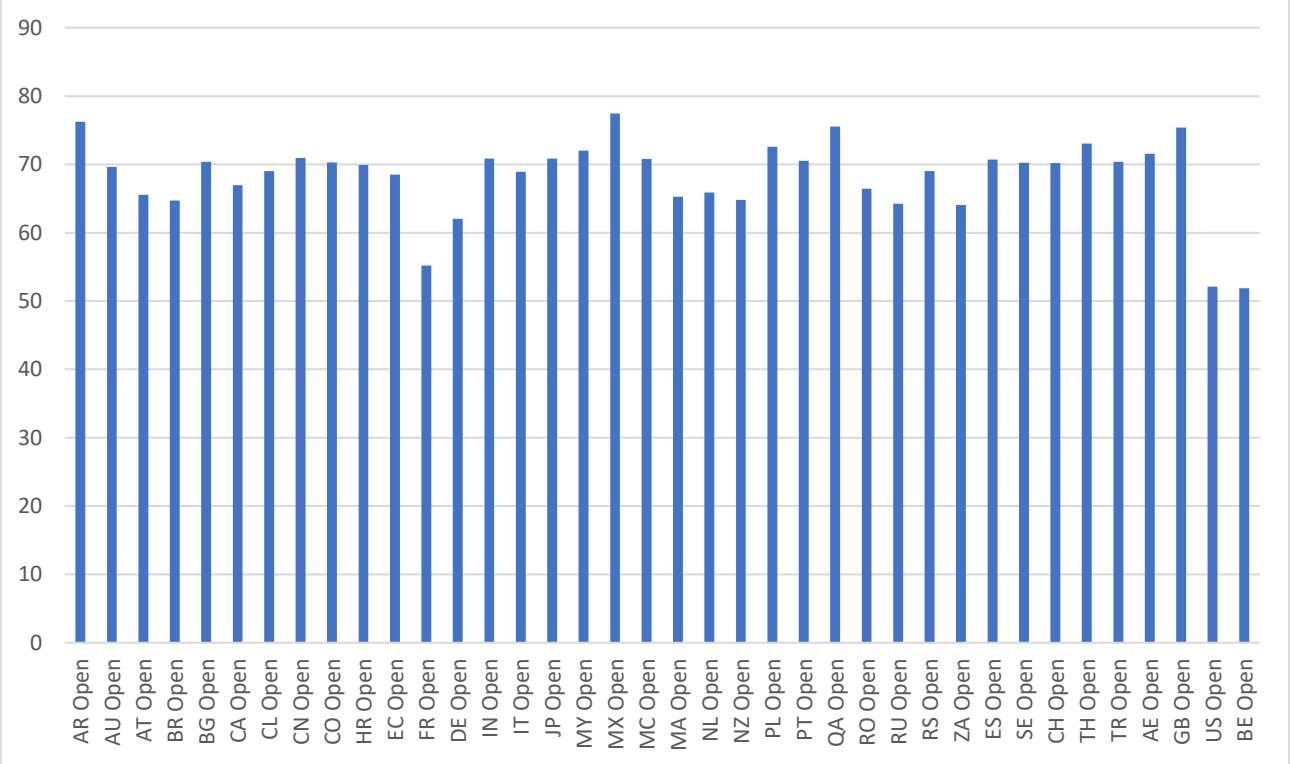

Figura 19: Percentuale di correttezza delle quote per i tornei di tennis.

Per quanto riguarda i campionati di calcio, la lega più indovinata è risultata quella portoghese, mentre, per il tennis, i tornei giocati in messico.
In media i tornei di tennis presentano quote più corrette rispetto a quelle dei campionati di calcio, infatti sono solo tre i paesi in cui la media di quotazioni corrette per i tornei di tennis è sotto il 60% (Francia, Stati Uniti e Belgio), mentre, per i campionati di calcio, neanche un campionato supera questa percentuale.

Questo può essere dovuto al numero di risultati possibili: per il tennis è presente solo l'1 o il 2, mentre per il calcio è presente anche il pareggio 'x'.

Percentuale di correttezza delle quote per partecipante

Per questo tipo di analisi è stato scelto di stilare una classifica dei migliori e peggiori partecipanti, secondo la percentuale di correttezza delle quote.
Analogamente a quanto effettuato per le competizioni, i partecipanti sono stati divisi per sport.
Anche in questo caso si considera la media delle percentuali ottenute dai quattro bookmaker.

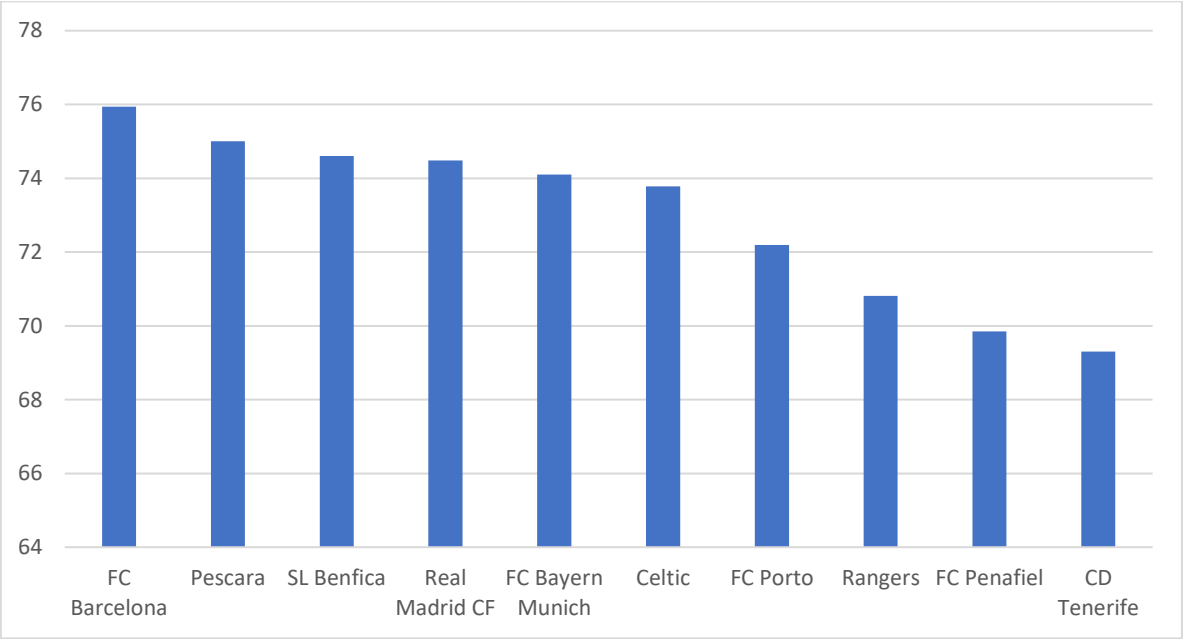

Figura 20: Migliori squadre di calcio per correttezza delle quote.

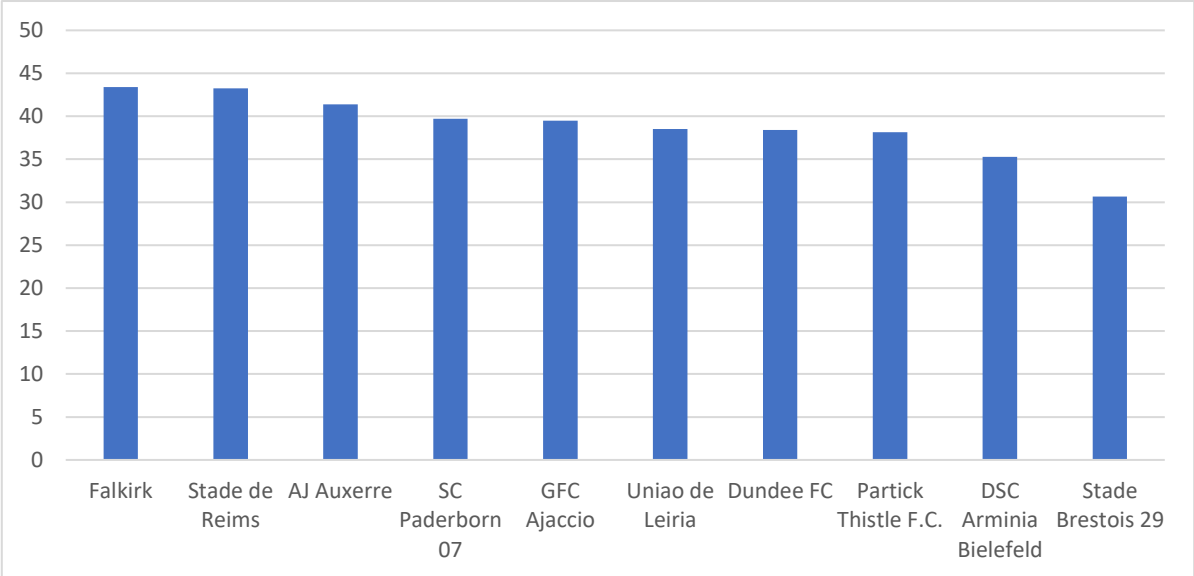

Figura 21: Peggiori squadre di calcio per percentuale di correttezza delle quote.

In media le squadre che si trovano nella top 10, sono quelle che dominano nel loro campionato, fatta eccezione per il Pescara.
Rispetto alle altre squadre, il Pescara ha una quota molto bassa di sconfitta. Questo è testimoniato dal fatto che le quote corrette che vedono il Pescara coinvolto, la diano per perdente.
Nella flop 10 troviamo La Stade Brestois 29, squadra francese, che ha una percentuale di correttezza delle quote (30,67%) inferiore alla metà della percentuale di correttezza delle quote del FC Barcelona (75,94%).

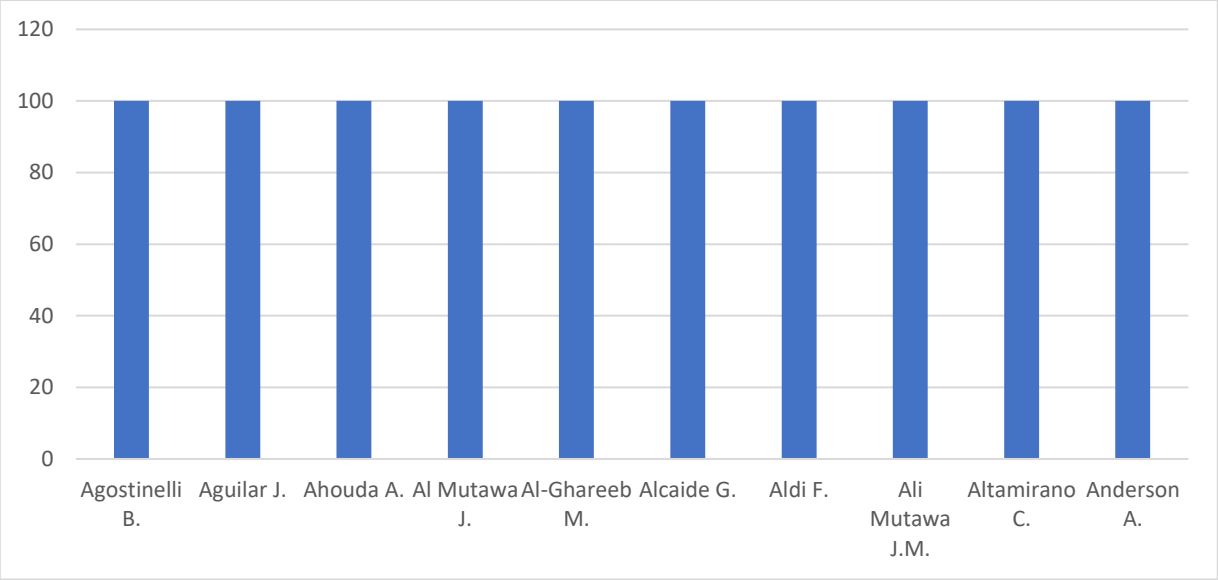

Figura 22: Migliori dieci tennisti per percentuale di correttezza delle quote.

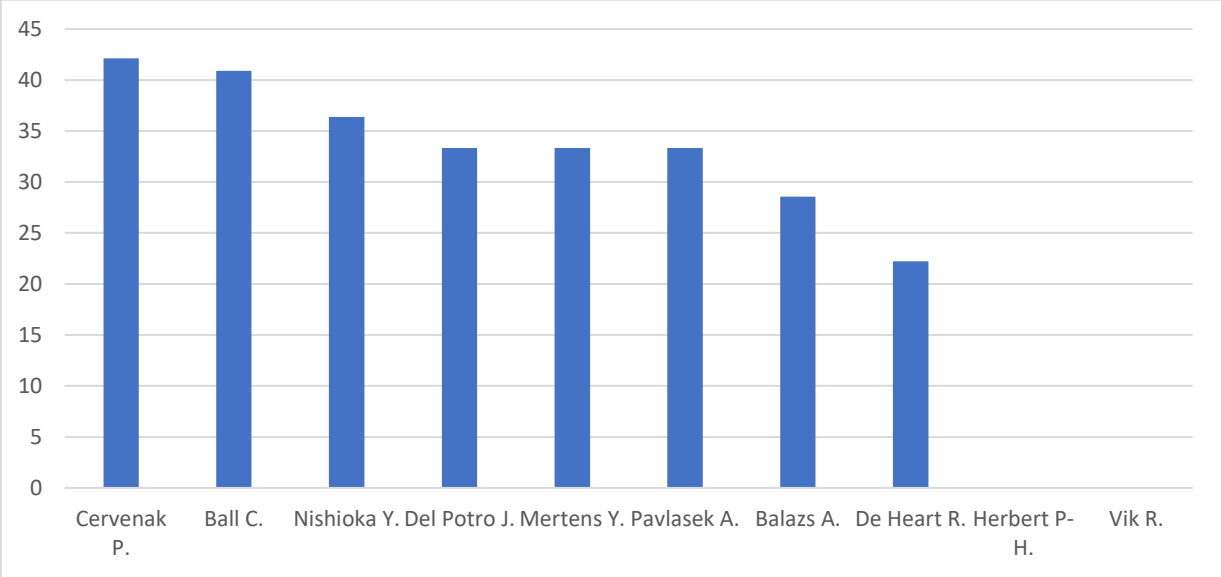

Figura 23: Peggiori 10 tennisti per percentuale di correttezza delle quote.

I tennisti presenti nella top 10 hanno una percentuale di correttezza delle quote del 100%. Anche le posizioni successive presentano giocatori con una percentuale del 100%.
Questo perché molti giocatori hanno un numero di match giocati molto basso.
Più interessante è l'analisi dei peggiori 10 tennisti, in cui compaiono due giocatori per cui la percentuale di correttezza delle quote è dello 0%.
Analizzando più a fondo le partite di questi due giocatori scopriamo che Vik R. ha giocato un solo match ed è stato dato, erroneamente, vincente dall'unico bookmaker che l'ha quotato (B365), stesso discorso vale anche per Herbert P-H.

Percentuale di correttezza delle quote per sport

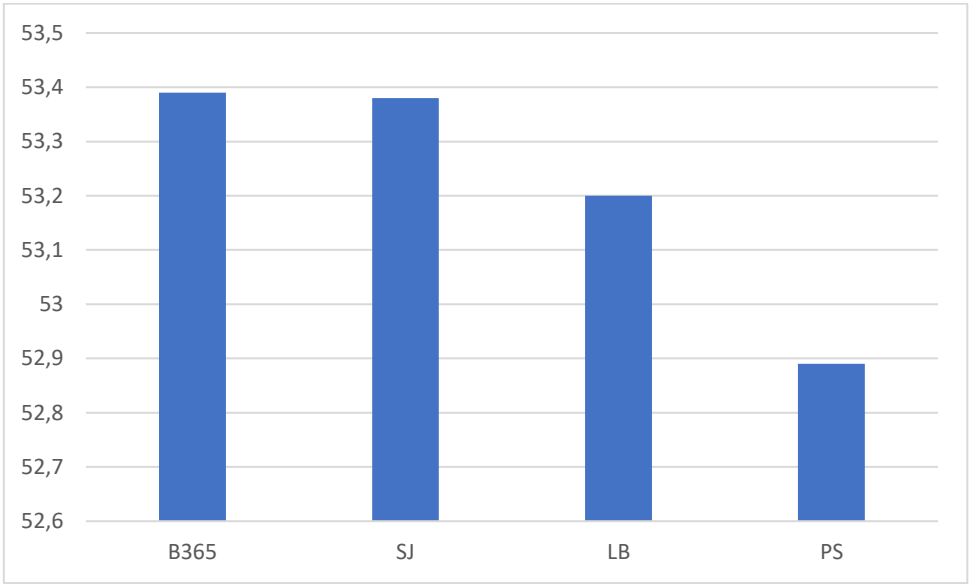

Figura 24: Percentuale di correttezza delle quote relative a partite di calcio, per i diversi bookmaker.

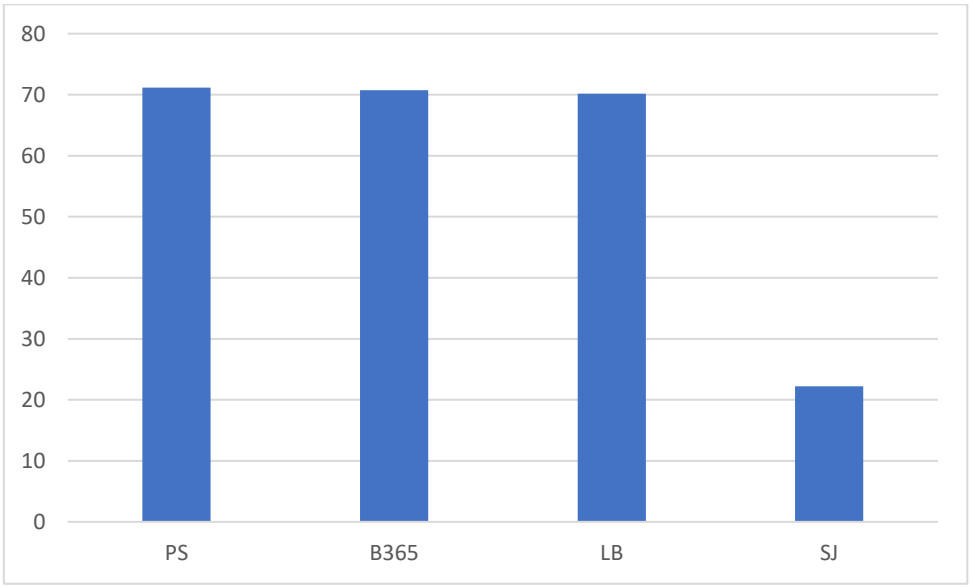

Figura 25: Percentuale di correttezza delle quote relative a partite di tennis, per i diversi bookmaker.

Per quanto riguarda le partite di calcio, tutti i bookmaker hanno ottenuto una percentuale simile: lo scarto tra il migliore e il peggiore è di solo mezzo punto.
Per quanto riguarda i match di tennis PS, B365 e LB hanno ottenuto risultati molto simili, con una percentuale poco al di sopra del 70%.
SJ ha ottenuto invece risultati molto più bassi (22,22%).

Percentuale di correttezza delle quote divise per anno e bookmaker

L'ultima separazione è data dagli anni in cui i bookmaker hanno fornito delle quotazioni.
Per ogni bookmaker si è analizzato l'andamento della correttezza delle quote fornite.

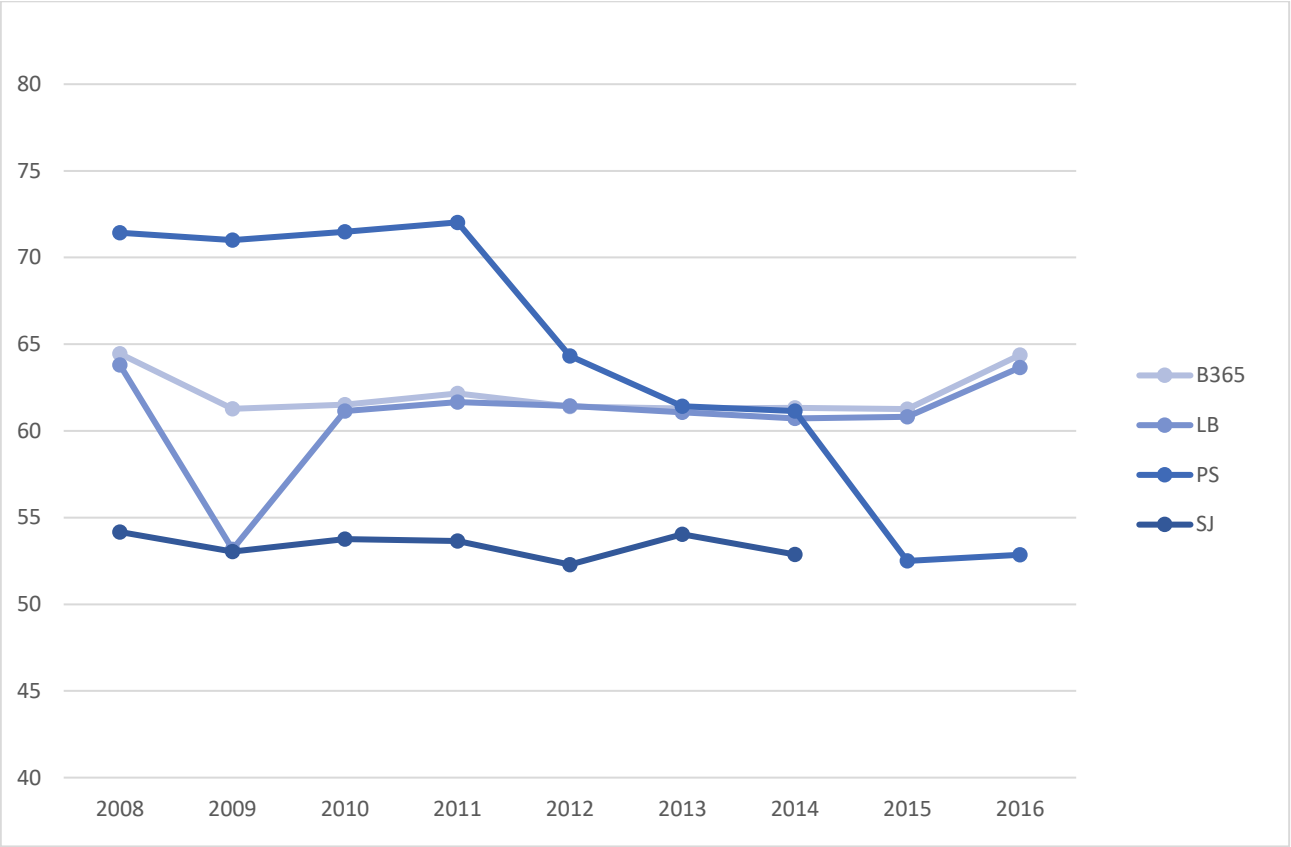

Figura 26: Andamento dei bookmaker per anno (percentuale di correttezza delle quote).

Dal 2008 al 2011 PS è stato il bookmaker con il maggior numero di quote corrette, toccando un massimo del 72,03%. Negli anni successivi ha avuto un drastico calo arrivando, come punto più basso, al 52,5% nel 2015.
B365 ha avuto un andamento abbastanza stabile che si è aggirato tra il 61,27% e il 63,66%.
Lo stesso può essere detto per LB, il quale ha avuto un andamento simile a B365.
Nel 2009 LB ha avuto un crollo che lo ha portato al 53% di quote corrette.
SJ è il bookmaker con andamento peggiore ma costante: dal 52,29% al 53,03%.

Percentuale di errori gravi nelle quote

Successivamente è stata analizzata la situazione in cui, in caso di KO, la quotazione più bassa di una match_odd si riferisca alla squadra perdente. Le separazioni che sono state scelte sono le stesse dell'interrogazione precedente.

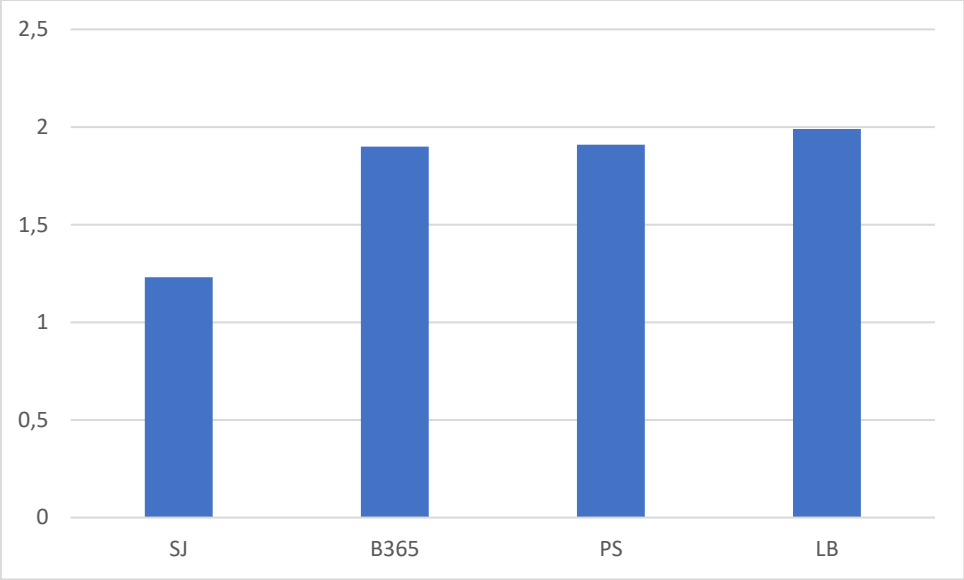

Figura 27: percentuali di errori gravi per i diversi bookmaker.

Il bookmaker che ha commesso più errori gravi è risultato LB.
Nonostante SJ abbia sbagliato il maggior numero di quote per i match di tennis, non ha commesso errori gravi e, sotto questo aspetto, è risultato il migliore bookmaker.

Percentuale di errori gravi sulle quote per competizione

Per facilitare l'interpretazione dei dati, si è scelto di dividere l'analisi per sport, per poi fare un'analisi totale. Per queste analisi è stata considerata la media delle percentuali ottenute dai 4 bookmaker.

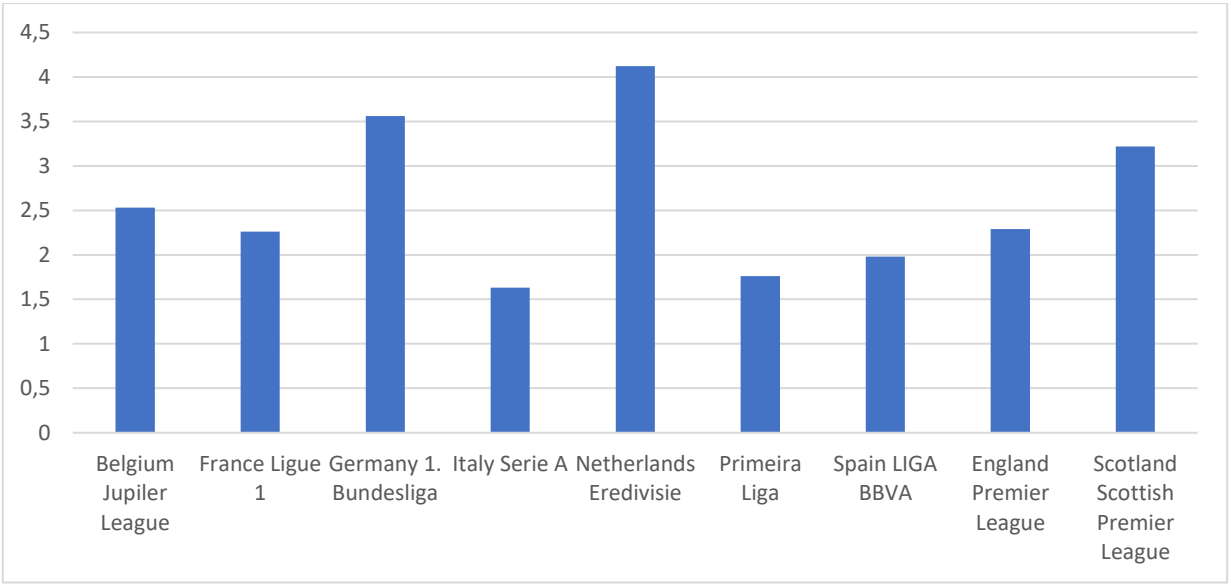

Figura 28: Percentuale di errori gravi per il campionato di calcio.

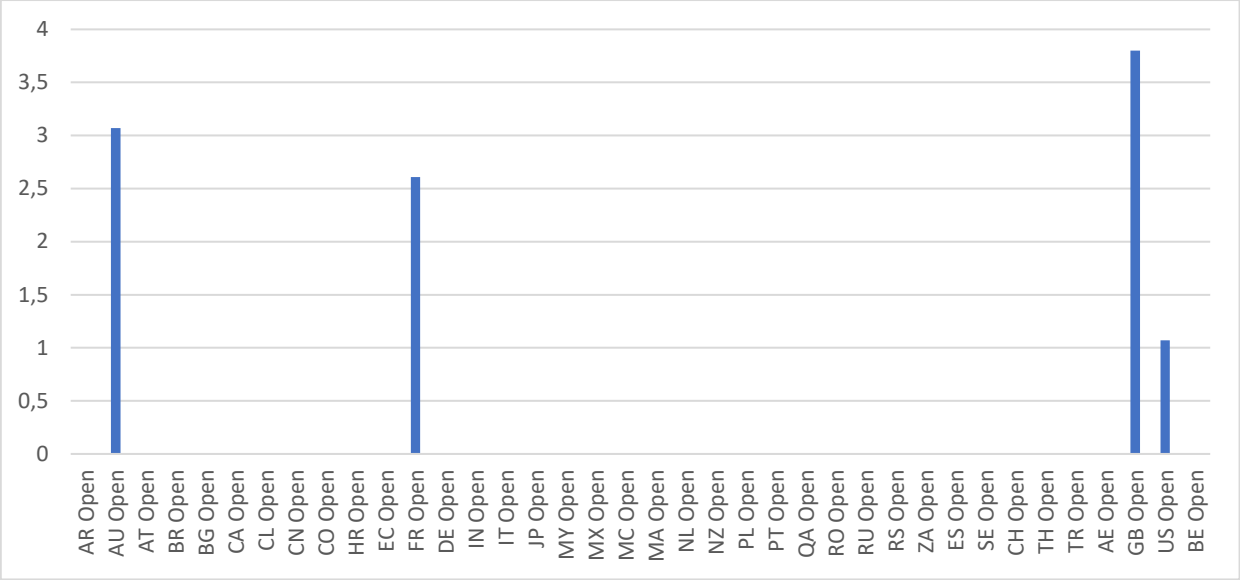

Figura 29: Percentuale di errori gravi per i tornei di tennis.

Eredivise per il calcio è risultato il campionato con la maggior percentuale di errori gravi, mentre la Serie A è risultata il campionato con la percentuale minore.
Nel tennis sono i tornei della Gran Bretagna ad avere la percentuale di errori maggiore.
Sono solo quattro i tornei di tennis in cui sono stati commessi degli errori gravi.
I motivi sono potrebbero essere simili a quelli dell'interrogazione precedente.
In più è possibile pensare che il KO, come definito dalla nostra analisi, è molto più probabile nel calcio che nel tennis.

Percentuale di errori gravi sulle quote per partecipante

L'analisi si è concentrata sui dieci partecipanti dei due sport per cui sono stati commessi il maggior numero di errori gravi.
Anche per queste analisi è stata considerata la media delle percentuali ottenute dai 4 bookmaker.

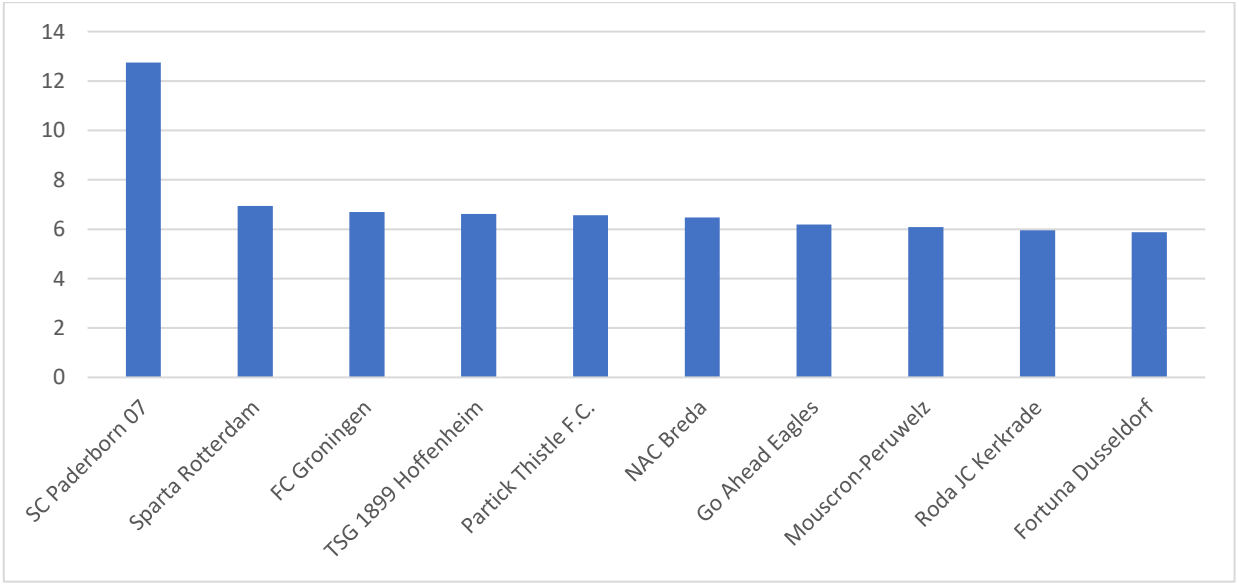

Figura 30: Peggiori 10 squadre di calcio per percentuale di errori gravi.

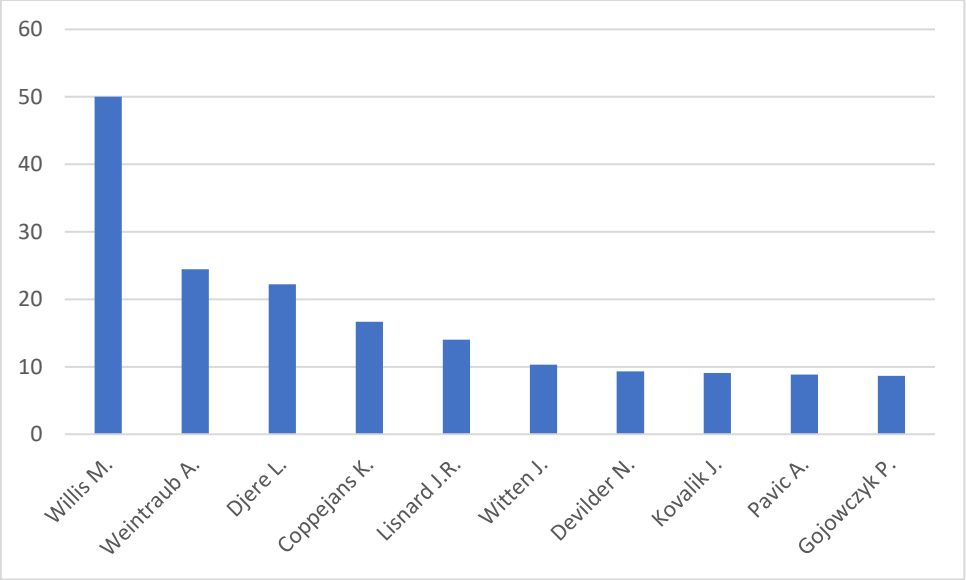

Figura 31: Peggiori 10 tennisti per percentuale di errori gravi.

Per quanto riguarda entrambi gli sport, la squadra per con una percentuale maggiore ha un valore che è circa il doppio rispetto a quella di tutti gli altri.
SC Paderborn ha giocato una sola stagione in Bundesliga (2014/2015) dove ha subito un gran numero di KO.
Willis M. è presente in solo due partite, finite entrambe con punteggio di KO.
In una delle due partite ha vinto, ma era dato per perdente dai bookmaker.
Paragonando le competizioni dei due sport, l'errore grave è maggiormente presente nel tennis.

Percentuale di correttezza delle quote per sport, relativamente ai diversi bookmaker

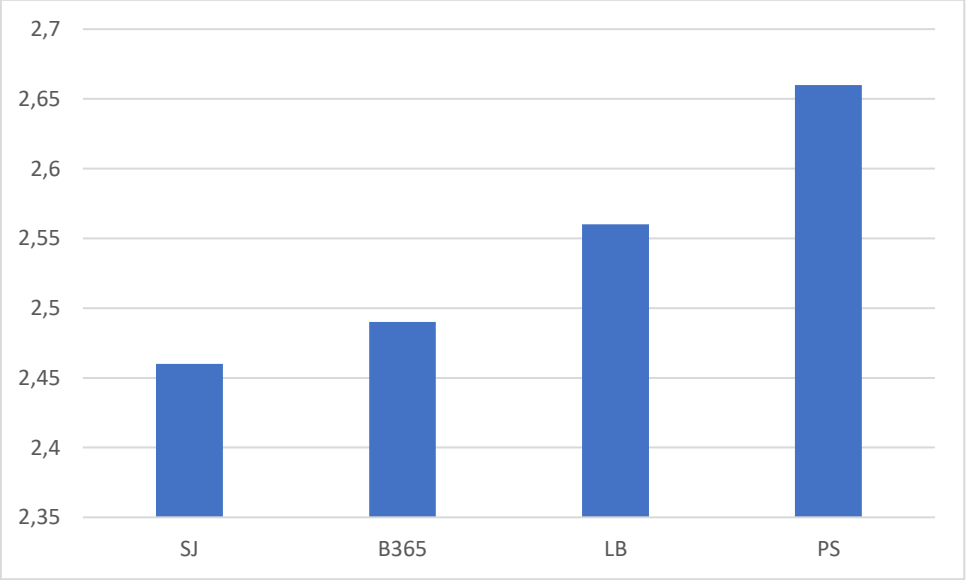

Figura 32: percentuale di errori gravi per le partite di calcio, relativamente ai diversi bookmaker.

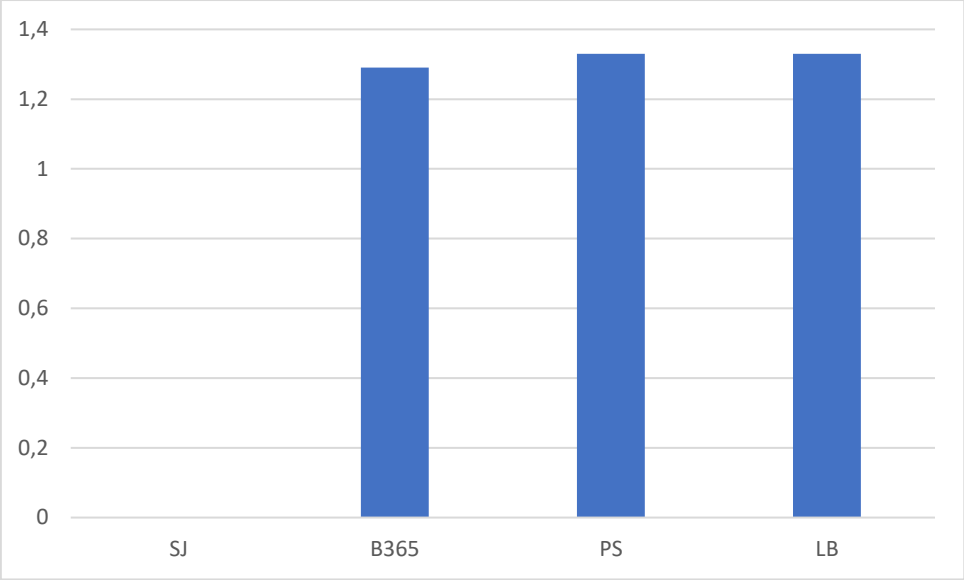

Figura 33: percentuale di errori gravi per le partite di tennis, relativamente ai diversi bookmaker.

Il bookmaker SJ non ha mai commesso errori gravi nella quotazione dei match di tennis.
PS che dalla prima separazione risulta il peggiore, lo rimane anche in quest'ultima dove è il peggiore nel calcio e il secondo peggiore nel tennis.

Quotazione media riferita alla squadra vincente in caso di KO

Per l'esecuzione di questa analisi, i risultati saranno forniti in forma tabellare.
Con squadra vincente si intende la squadra che nel KO ha vinto il match, sia che il risultato fosse 1 o 2.

Media quota KO W bookmaker
1.86

bookmaker	Media W KO
B365	1.81
PS	1.81
LB	1.82
SJ	1.99

Nome competizione	Media KO W bookmaker
Belgium Jupiler League	2.02
France Ligue 1	2.15
Germany 1. Bundesliga	2.12
Italy Serie A	2.03
Netherlands Eredivisie	1.97
Primeira Liga	1.79
Spain LIGA BBVA	1.84
England Premier League	2.01
Scotland Scottish Premier League	2.17
AU Open	1.44
FR Open	1.53
GB Open	1.56
US Open	1.52

Nome competizione	Bookmaker	Media KO W
Belgium Jupiler League	LB	1.95
Belgium Jupiler League	B365	2.01
Belgium Jupiler League	SJ	2.02
Belgium Jupiler League	PS	2.11
England Premier League	SJ	1.95
England Premier League	LB	1.98
England Premier League	B365	2.01
England Premier League	PS	2.11
France Ligue 1	LB	2.12
France Ligue 1	SJ	2.16
France Ligue 1	B365	2.16
France Ligue 1	PS	2.17
Germany 1. Bundesliga	LB	2.08
Germany 1. Bundesliga	B365	2.1
Germany 1. Bundesliga	SJ	2.12
Germany 1. Bundesliga	PS	2.19
Italy Serie A	LB	2
Italy Serie A	B365	2.02
Italy Serie A	PS	2.03

Italy Serie A	SJ	2.05
Netherlands Eredivisie	LB	1.91
Netherlands Eredivisie	SJ	1.95
Netherlands Eredivisie	B365	1.96
Netherlands Eredivisie	PS	2.05
Primeira Liga	LB	1.75
Primeira Liga	B365	1.78
Primeira Liga	SJ	1.8
Primeira Liga	PS	1.82
Scotland Scottish Premier League	SJ	2.07
Scotland Scottish Premier League	B365	2.15
Scotland Scottish Premier League	LB	2.16
Scotland Scottish Premier League	PS	2.28
Spain LIGA BBVA	LB	1.83
Spain LIGA BBVA	SJ	1.83
Spain LIGA BBVA	B365	1.84
Spain LIGA BBVA	PS	1.86
AU Open	PS	1.41
AU Open	B365	1.42
AU Open	LB	1.48
FR Open	B365	1.51
FR Open	PS	1.51
FR Open	LB	1.56
GB Open	PS	1.52
GB Open	B365	1.54
GB Open	LB	1.63
US Open	B365	1.5
US Open	PS	1.52
US Open	LB	1.54

Sport	Media KO W Bookmaker
football	2
tennis	1.51

Sport	Bookmaker	Media W KO
football	LB	1.96
football	SJ	1.99
football	B365	2
football	PS	2.06
tennis	PS	1.49
tennis	B365	1.49
tennis	LB	1.55

Figura 34: Quotazione media riferita alla squadra vincente in caso di KO

.

Quotazione media riferita alla squadra perdente in caso di KO

Per l'effettuazione di questa analisi, i risultati saranno forniti in forma tabellare.
Con squadra perdente si intende la squadra che nel KO ha perso il match, sia che il risultato fosse 1 o 2.

Media quota KO L bookamker
7.06

Bookmaker	Media KO L
SJ	6.58
PS	7.03
B365	7.13
LB	7.51

Nome competizione	Media KO L bookmaler
Belgium Jupiler League	5.49
France Ligue 1	5.61
Germany 1. Bundesliga	6.09
Italy Serie A	6.08
Netherlands Eredivisie	6.39
Primeira Liga	8.22
Spain LIGA BBVA	9.93
England Premier League	6.68
Scotland Scottish Premier League	5.81
AU Open	8.32
FR Open	8.04
GB Open	7.31
US Open	7.93

Nome competizione	Bookmaker	Media KO L
Belgium Jupiler League	LB	5.16
Belgium Jupiler League	SJ	5.44
Belgium Jupiler League	B365	5.6
Belgium Jupiler League	PS	5.74
England Premier League	LB	6.36
England Premier League	PS	6.37
England Premier League	SJ	6.98
England Premier League	B365	6.99
France Ligue 1	LB	5.13
France Ligue 1	SJ	5.31
France Ligue 1	B365	5.57
France Ligue 1	PS	6.42
Germany 1. Bundesliga	SJ	5.46
Germany 1. Bundesliga	LB	5.67
Germany 1. Bundesliga	B365	5.87
Germany 1. Bundesliga	PS	7.37
Italy Serie A	LB	5.62
Italy Serie A	SJ	5.62
Italy Serie A	B365	5.96
Italy Serie A	PS	7.11
Netherlands Eredivisie	LB	6.01
Netherlands Eredivisie	PS	6.14

Netherlands Eredivisie	SJ	6.68
Netherlands Eredivisie	B365	6.73
Primeira Liga	SJ	7.57
Primeira Liga	LB	7.61
Primeira Liga	B365	8.13
Primeira Liga	PS	9.57
Scotland Scottish Premier League	SJ	5.69
Scotland Scottish Premier League	LB	5.77
Scotland Scottish Premier League	PS	5.86
Scotland Scottish Premier League	B365	5.93
Spain LIGA BBVA	LB	8.69
Spain LIGA BBVA	SJ	9.27
Spain LIGA BBVA	B365	9.52
Spain LIGA BBVA	PS	12.23
AU Open	PS	6.83
AU Open	B365	7.83
AU Open	LB	10.29
FR Open	PS	6.29
FR Open	B365	7.78
FR Open	LB	10.04
GB Open	PS	5.64
GB Open	B365	7.15
GB Open	LB	9.13
US Open	PS	6.39
US Open	B365	7.78
US Open	LB	9.62

Sport	Media KO L bookmaker
football	6.83
tennis	7.91

Sport	Bookmaker	Media KO L
football	LB	6.32
football	SJ	6.58
football	B365	6.83
football	PS	7.6
tennis	PS	6.3
tennis	B365	7.64
tennis	LB	9.79

Figura 35: Quotazione media riferita alla squadra perdente in caso di KO.