Fakultet tehničkih nauka
Inženjerstvo informacionih sistema

Predmet: Projektovanje skladišta podataka

Projekat

Data Warehouse za analizu fudbalskih utakmica

Student: Vojin Ćetković IT21/2021

Sadržaj

1. Zadatak i ciljevi projekta 3
2. Opis postupka projektovanja *DW* sistema 4
3. Specifikacija zahteva korisnika 5
4. Specifikacija modela 6
4.1 Specifikacija izvora podataka 6
4.2 Specifikacija ciljanog *Data Warehouse* sistema 8
4.2.1 Specifikacija zahtevanih dimenzija 9
4.2.2 Specifikacija zahtevanih mera 9
5. Opis ETL procesa 12
5.1 Punjenje dimenzionih tabela 15
5.2 Punjenje činjenične tabele 19
6. Prikaz izveštaja 25
7. Zaključak 32

1. Zadatak i ciljevi projekta

Zadatak ovog projekta jeste da se razvije pouzdan i centralizovan izvor informacija o fudbalu, koji će služiti kao osnova za analitičke potrebe i izveštavanje. Polazna tačka su javno dostupni podaci sa sajta „*Transfermarkt*“, jednog od najpopularnijih i najpoznatijih izvora fudbalske statistike na svetu. „*Transfermarkt*“ redovno prati rezultate utakmica, učinak klubova i igrača, informacije o sudijama, menadžerima i stadionima, kao i brojne druge detalje iz fudbalskog sveta. Upravo zbog svoje sveobuhvatnosti i pouzdanosti, ovaj sajt predstavlja idealnu osnovu za transformaciju podataka u kvalitetno skladište namenjeno boljem donošenju zaključaka. Cilj projekta je da se sirovi i fragmentisani podaci iz CSV fajlova transformišu u *Data Warehouse* sistem, gde će oni biti standardizovani, očišćeni i organizovani na način koji omogućava brzu i efikasnu analizu. Na taj način korisnicima će biti omogućeno da kroz jasno definisane izveštaje dobiju sve relevantne informacije na jednom mestu. Sam *Data Mart* biće projektovan u obliku zvezdaste šeme, sa jasno definisanim dimenzijama i jednom činjeničnom tabelom. Granularnost činjenične tabele biće jedna utakmica, što znači da će svaki red u tabeli predstavljati jedinstveni meč sa svim njegovim karakteristikama i merama (golovi, kartoni, poseta...). Ovakav pristup obezbeđuje dovoljno detalja za sve analize, a istovremeno ostaje dovoljno jednostavan za implementaciju i korišćenje. Krajnji cilj projekta jeste da se, na osnovu izgrađenog skladišta podataka, kreiraju izveštaji u SSRS-u koji će odgovarati na unapred definisana pitanja korisnika. Time će se obezbediti projekat koji ne samo da skladišti podatke, već ih i pretvara u informacije korisne za donošenje zaključaka, prepoznavanje trendova i unapređenje razumevanja fudbalskih procesa.

2. Opis postupka projektovanja DW sistema

Proces projektovanja skladišta podataka počeo je analizom izvornog seta podataka. Kao polazni materijal korišćen je *dataset* sajta „*Transfermarkt*“, koji u sebi sadrži informacije o odigranim fudbalskim utakmicama. Prvi korak bio je da se detaljno pregleda sadržaj fajla, da se razume koji atributi postoje, kakva je njihova struktura i šta oni predstavljaju. U podacima su identifikovani osnovni elementi svake utakmice – datum odigravanja, takmičenje i sezona, klubovi koji učestvuju, menadžeri i sudije, stadion na kojem je utakmica igrana, kao i osnovne statistike meča poput broja golova, kartona i posećenosti. Već na ovom koraku postalo je jasno da se radi o bogatom skupu podataka, ali i da je potrebno dodatno čišćenje i standardizacija da bi oni bili spremni za analitičku upotrebu. Nakon upoznavanja sa sadržajem, bilo je neophodno doneti odluku o granularnosti skladišta. Posle razmatranja različitih opcija, zaključeno je da je najprirodniji i najlogičniji izbor da granularnost bude jedna utakmica. Na taj način svi ključni podaci ostaju vezani za jedinstven događaj, a skladište zadržava preglednost i jednostavnost. Sledeći korak bilo je modelovanje skladišta podataka. Odlučeno je da se koristi zvezdasta šema, jer je ona najpogodnija za ovakav tip analitičkih sistema. U centru šeme postavljena je činjenična tabela koja sadrži mere (golovi, kartoni, posećenost, indikatori ishoda), dok su oko nje definisane dimenzije – vreme, takmičenje, sezona, stadion, sudija, klub i menadžer. Kada je model bio definisan, sledeća faza bila je implementacija ETL procesa u SSIS-u. Podaci iz CSV fajla su uvezeni u staging deo, gde su zatim prošli kroz niz transformacija: čišćenje nepotpunih zapisa, zamena nedostajućih vrednosti sa oznakama tipa „Nepoznato“, formatiranje kolona... Konačni korak u postupku bio je razvoj izveštaja u SSRS-u. Na osnovu definisanih zahteva korisnika kreirani su izveštaji koji korisnicima omogućavaju uvid u podatke.

3. Specifikacija zahteva korisnika

Da bi sve imalo smisla, tj. da bi skladište imalo potrebnu praktičnu vrednost, potrebno je definisati neka pitanja na koja sistem treba da pruži odgovore. Skladišta ne postoje samo da bi se podaci čuvali, već i da pruže korisnicima olakšanu analizu i donošenje odluka. Pre implementaciju određeno je šest ključnih pitanja koja pomažu korisnicma da lakše sagledaju situaciju. Prvo pitanje odnosi se na prosečan broj golova po takmičenju i sezoni. Menadžment, sportski analitičari pa i navijači žele da imaju uvid u atraktivnost određenih liga i sezona, da vide da li su mečevi takmičenja efikasni i koliko se trend golova menja tokom vremena. Drugi zahtev se tiče sudija i broja kartona. Potrebno je utvrditi koje sudije dodeljuju najviše žutih i crvenih kartona u proseku po utakmici. Ovakva analiza pomaže sudijskoj komisiji da prepozna stroge i tolerantne sudije, ali i da se prati njihovo suđenje. Treći zahtev odnosi se na posećenost stadiona. Klubovi, lige i mediji žele da znaju koji stadioni imaju najveću prosečnu posetu i kako se posećenost menja tokom sezona. Takvi podaci su važni za planiranje organizacije utakmica, dodelu važnijih mečeva, ali i za komercijalne i marketinške aktivnosti. Takođe taj podatak često povećava zainteresovanost kod turista. Četvrti zahtev definiše analizu učinka klubova kod kuće i na strani. Vrh kluba, trener i navijači žele jasnu sliku o tome kako se timovi ponašaju u različitim uslovima, kolika im je stopa pobeda, koliko često igraju nerešeno i kakva je gol-razlika. Na osnovu tih podataka može se proceniti koliko je klub zaista dominantan kod kuće i koliko je ranjiv na gostovanjima. Peto pitanje usmereno je na performanse menadžera. Ovde se posmatra koliko su menadžeri uspešni u svom radu, kroz win rate, prosečnu gol-razliku... I na kraju analiziraćemo top 5 timova po sezoni koji imaju najbolji „*fair-play index*“ kao statistika koja je izuzetno bitna prilikom odlučivanja osvajača lige ako više ekipa isti broj bodova. Definisanjem ovih šest pitanja obezbeđeno je da skladište podataka ne bude samo tehnički projekat, već sistem koji donosi realnu vrednost i odgovara na konkretne analitičke potrebe korisnika.

4. Specifikacija modela

U ovom delu ćemo proći kroz analizu izvora podataka i OLTP šeme. Takođe u drugom delu ove sekcije ćemo opisati i OLAP šemu kao i svaku dimenziju posebno, na kraju biće opisana činjenična tabela.

4.1 Specifikacija izvora podataka

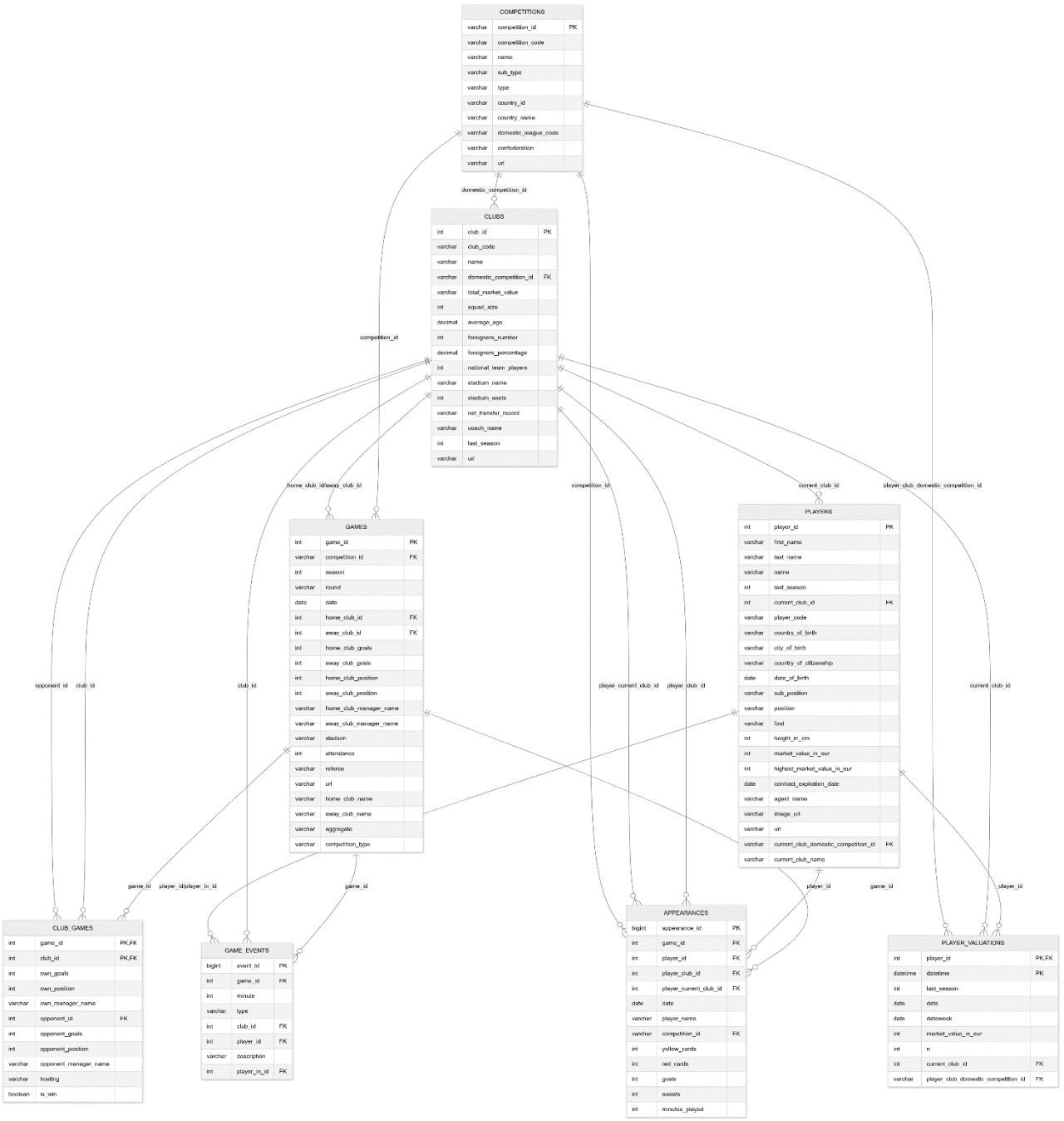
Izvor podataka je javni skup sa *Keggle* u CSV formatu. Iako fizički ne postoji OLTP baza, ovi fajlovi predstavljaju operativni sloj sistema: svaki fajl odgovara jednom poslovnom entitetu , sadrži ključeve i polja koja se kasnije transformišu i standardizuju tokom ETL procesa. U nastavku su fajlovi *dataset*-a i njihova uloga:

- *competitions.csv* - Sadrži metapodatke o ligama/kupovima: *competition_id / competition_code (NK), competition_name, competition_type/sub_type (league/cup), country_name, confederation, domestic_league_code, url*…Ovo je tabela na koju se naslanjaju klubovi i mečevi (veze *Competition–Club*, *Competition–Game*).
- *clubs.csv* - Osnovni opis klubova: *club_id (NK), club_code, name, country, domestic_competition_id, foundation_year, stadium_name/seat(s), last_season, market_value*…Vezuje se sa *competitions* (*domestic* liga), sa *games* (*home/away* klubovi) i posredno sa *managers/players*.
- *games.csv* - Utakmice (centralni operativni događaj). Atributi: *game_id, competition_id (FK), season/round, date, stadium, referee_id, home_club_id, away_club_id, home/away_goals, home/away_positions, attendance*…Veze: N–1 ka većini entiteta (*Competition, Season, Stadium, Referee, Club, Manager*).
- *players.csv* - Identitet i biografija: *player_id (NK), first_name, last_name, date_of_birth, citizenship, height, position, foot, agent, current_club_id, contract_until, image_url, last_season*… Koristi se primarno za analize na nivou igrača (nije u fokusu našeg *mart*-a, ali je važan za kasnija proširenja).
- *appearances.csv* – Atributi: *appearance_id (NK), game_id (FK), player_id (FK), player_club_id, minutes, goals, assists, yellow/red cards, position, rating*…ž
- *game_events.csv* - Detaljni log: *event_id (NK), game_id (FK), minute, event_type (goal, yellow, red…), player_id (FK), club_id (FK), description*…Pogodan je za

dodatni mart (*FactEvent*) u slučaju proširenja.

- *player_valuations.csv* - Serija tržišnih vrednosti: *player_id (FK), date, market_value, current_club_id…*
- *club_games.csv* – Statistika kluba po meču. Atributi: *game_id (FK), club_id (FK), goals_for/against, xG, possession…*

Izvorni OLTP sistem je napravljen na osnovu CSV fajlova i sadrži više međusobno povezanih tabela. U centru se nalazi tabela *Games,* koja opisuje svaku odigranu utakmicu kroz osnovne atribute: identifikatore takmičenja i sezone, domaći i gostujući klub, menadžere, postignute golove, pozicije, prisutne gledaoce i osnovne pokazatelje ishoda. Tabela *Competitions* čuva podatke o svim takmičenjima (liga ili kup), uključujući naziv, kod i zemlju, i povezuje se sa tabelama *Games* i *Clubs*. Tabela *Clubs* sadrži osnovne informacije o klubovima (naziv, država, stadion, vrednost tima) i povezuje se sa *Games*, kao domaći ili gostujući klub. Tabela *Players* opisuje osnovne podatke o igračima (ime, datum rođenja, pozicija, državljanstvo, trenutni klub). Njihovi nastupi na utakmicama evidentirani su u tabeli *Appearances*, koja povezuje igrače i mečeve i čuva detalje o minutima, golovima, asistencijama i kartonima. Tabela *Game Events* beleži detaljne događaje sa utakmica (golovi, žuti i crveni kartoni) i povezuje se sa *Games* i *Players*. Dodatno, tabela *Player Valuations* čuva istorijske vrednosti igrača kroz vreme, dok tabela *Club Games* daje sažetu statistiku kluba po meču. Na ovaj način OLTP model obuhvata sve ključne entitete – takmičenja, klubove, igrače i utakmice – i njihove međusobne veze, što čini dobru osnovu za dalje razvijanje skladišta podataka.



1- OLTP šema

4.2 Specifikacija ciljanog *Data Warehouse* sistema

Ciljani sistem je projektovan u obliku zvezdaste šeme u kojoj centralno mesto zauzima činjenična tabela *Game*, dok su oko nje raspoređene dimenzije koje obezbeđuju deskriptivne podatke. Granularnost je jasno definisana – jedan red u činjeničnoj tabeli

predstavlja jednu utakmicu.

4.2.1 Specifikacija zahtevanih dimenzija

Dimenzije predstavljaju deskriptivne tabele koje sadrže atribute pomoću kojih se činjenični podaci mogu posmatrati iz različitih perspektiva. One su ključne za analitičke procese jer omogućavaju filtriranje, grupisanje i poređenje podataka. Uloga dimenzija je da obogate mere kontekstom – na primer, da se golovi posmatraju po sezoni, po takmičenju ili po stadionu. U ovom modelu imamo sledeće dimenzije:

- *Date* - sadrži vremenske atribute (dan, mesec, kvartal, godina, indikator vikenda) i omogućava vremenske analize.
- *Competition* - beleži podatke o takmičenjima (šifra, naziv, tip, zemlja, konfederacija).
- *Season* - čuva podatke o sezoni i rundi, što omogućava poređenja performansi kroz različite cikluse.
- *Stadium* - sadrži naziv stadiona i koristi se za analize posećenosti.
- *Referee* - beleži identitet sudija, što omogućava analize strogoće i broja kartona.
- *Club* - opisuje klubove (šifra, ime) i u tabeli činjenica se koristi u dve uloge – domaći i gostujući tim.
- *Manager* - čuva podatke o menadžerima, što omogućava analize trenerskih performansi.

Na ovaj način dimenzije obezbeđuju različite uglove posmatranja i daju smisao merama iz činjenične tabele..

4.2.2 Specifikacija zahtevanih mera

Mere predstavljaju kvantitativne vrednosti koje se nalaze u činjeničnoj tabeli i koje se analiziraju pomoću dimenzija. One su uvek vezane za tačno definisanu granularnost – u ovom slučaju za jednu utakmicu. Na osnovu mera korisnici dobijaju odgovore na pitanja na primer o rezultatima, disciplinama, posećenosti...

U činjeničnoj tabeli nalaze se sledeće mere koje omogućavaju analizu utakmica:

- Broj golova domaće i gostujuće ekipe – osnovni pokazatelj uspešnosti timova na

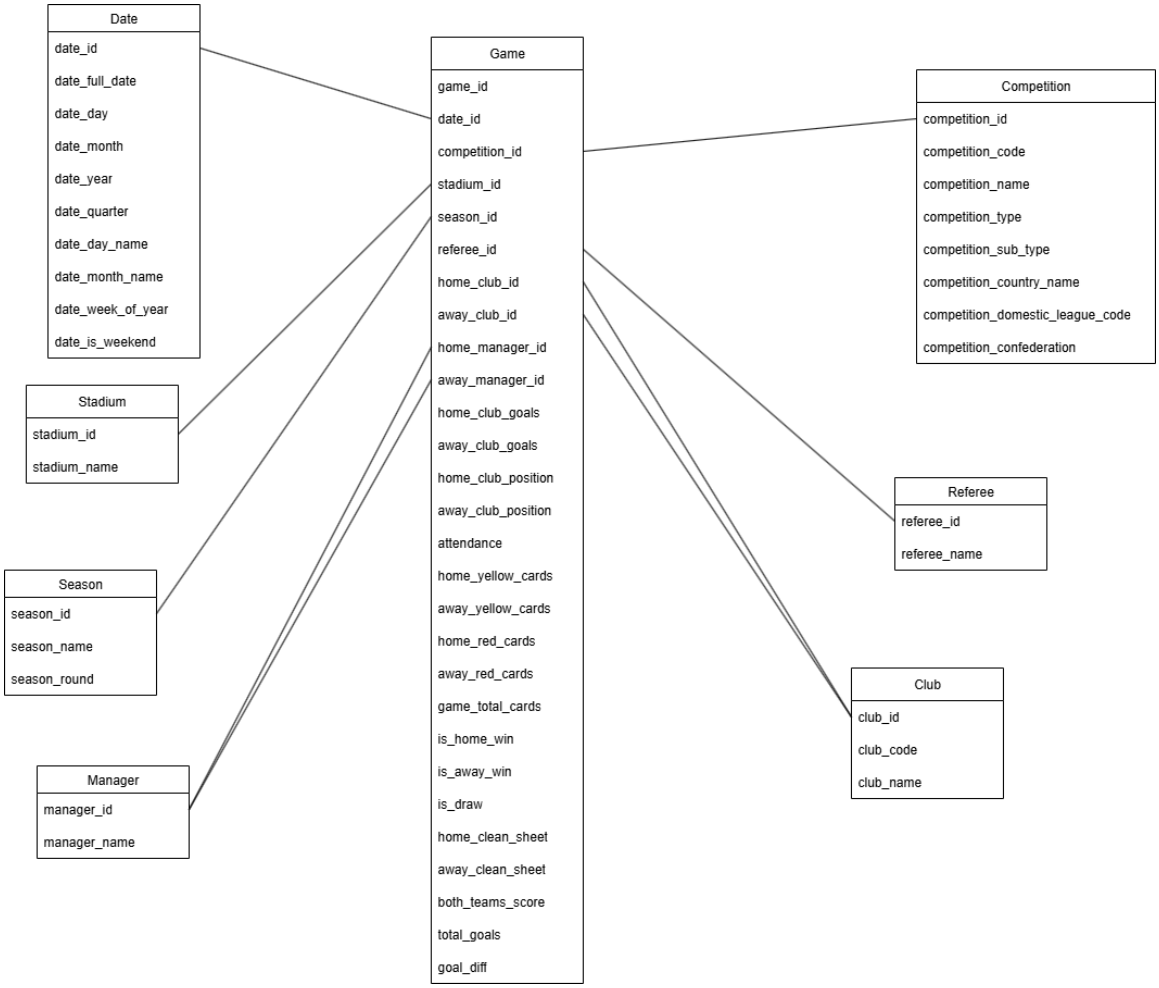
pojedinačnoj utakmici.

- Ukupan broj golova – zbir golova oba tima, koji se koristi za analizu efikasnosti mečeva.
- Gol razlika – razlika između postignutih i primljenih golova, što je važan pokazatelj dominacije jednog tima.
- Broj žutih i crvenih kartona po timu – pokazatelji disciplinske statistike, koji omogućavaju analize fer-pleja i strogoće suđenja.
- Ukupan broj kartona na meču – zbir svih kartona, koristan za poređenje utakmica i takmičenja po disciplinskim kriterijumima.
- Posećenost – broj gledalaca prisutnih na stadionu, što omogućava analize popularnosti timova i kapaciteta stadiona.

Pored numeričkih mera imamo i neke logične indikatore, koji nam takodje služe u analizi, lako se sabiraju i daju pokazatelje na primer broja pobeda domaćina, pobeda u gostima... To su:

- *is_home_win* – označava da li je domaći tim pobedio.
- *is_away_win* – označava da li je gostujući tim pobedio.
- *is_draw* – označava da li je utakmica završena nerešeno.
- *home_clean_sheet* – indikator da li domaćin nije primio gol.
- *away_clean_sheet* – indikator da li gost nije primio gol.
- *both_teams_score* – indikator da li su oba tima postigla gol.

Ove mere i indikatori čine osnovu svih analiza. Kada se povežu sa dimenzijama (takmičenje, sezona, klub, stadion, sudija), korisnicima se omogućava da rezultate sagledaju iz različitih perspektiva i da izvuku praktične uvide o fudbalskim mečevima.


Date
date_id
date_full_date
date_day
date_month
date_year
date_quarter
date_day_name
date_month_name
date_week_of_year
date_is_weekend
Game
game_id
date_id
competition_id
stadium_id
season_id
referee_id
home_club_id
away_club_id
home_manager_id
away_manager_id
home_club_goals
away_club_goals
home_club_position
away_club_position
attendance
home_yellow_cards
away_yellow_cards
home_red_cards
away_red_cards
game_total_cards
is_home_win
is_away_win
is_draw
home_clean_sheet
away_clean_sheet
both_teams_score
total_goals
goal_diff
Competition
competition_id
competition_code
competition_name
competition_type
competition_sub_type
competition_country_name
competition_domestic_league_code
competition_confederation
Stadium
stadium_id
stadium_name
Referee
referee_id
referee_name
Season
season_id
season_name
season_round
Club
club_id
club_code
club_name
Manager
manager_id
manager_name

2- OLAP Šema

5. Opis ETL procesa

Izvorni podaci sadržali su veliki broj duplikata i neusaglašenih formata, iz tog razloga je tokom ETL proces izvršen veliki broj transformacija nad podacima. Čitav proces započeo je kreiranjem uslovnog *drop*-a činjenične tabele i dimenzionih tabela, koje će biti kreiranje u drugom koraku.

```
IF OBJECT_ID('ProjectDW.FactGame','U') IS NOT NULL
DROP TABLE  ProjectDW.FactGame

IF OBJECT_ID('ProjectDW.DimDate','U') IS NOT NULL
DROP TABLE  ProjectDW.DimDate

IF OBJECT_ID('ProjectDW.DimStadium','U') IS NOT NULL
DROP TABLE  ProjectDW.DimStadium

IF OBJECT_ID('ProjectDW.DimSeason','U') IS NOT NULL
DROP TABLE  ProjectDW.DimSeason

IF OBJECT_ID('ProjectDW.DimManager','U') IS NOT NULL
DROP TABLE  ProjectDW.DimManager

IF OBJECT_ID('ProjectDW.DimCompetition','U') IS NOT NULL
DROP TABLE  ProjectDW.DimCompetition

IF OBJECT_ID('ProjectDW.DimReferee','U') IS NOT NULL
DROP TABLE  ProjectDW.DimReferee

IF OBJECT_ID('ProjectDW.DimClub','U') IS NOT NULL
DROP TABLE  ProjectDW.DimClub

IF SCHEMA_ID('ProjectDW') IS NOT NULL
DROP SCHEMA ProjectDW
```

3 – Drop tables

Drugi SQL Script u flow-u je za kreiranje dimenzionih tabela i činjenične tabele. Prvo je kreirana šema „*ProjectDw*“ u okviru koje su kreirane potrebne tabele. Korišćen je lokalni server.

```
GO
CREATE SCHEMA ProjectDW
GO

CREATE TABLE ProjectDW.DimDate (
  date_id int not null identity,
  date_full_date date not null,
  date_day int not null,
  date_month int not null,
  date_year int not null,
  date_quarter int not null,
  date_day_name varchar(15) not null,
  date_month_name varchar(15) not null,
  date_week_of_year int not null,
  date_is_weekend bit not null,

  CONSTRAINT PK_DimDate PRIMARY KEY(date_id)
)
GO
CREATE TABLE ProjectDW.DimStadium (
  stadium_id int not null identity,
  stadium_name nvarchar(400) not null

  CONSTRAINT PK_DimStadium PRIMARY KEY(stadium_id)
)
GO
CREATE TABLE ProjectDW.DimSeason (
  season_id int not null identity,
  season_name nvarchar(50) not null,
  season_round nvarchar(50) not null,

  CONSTRAINT PK_DimSeason PRIMARY KEY(season_id)
)
GO
CREATE TABLE ProjectDW.DimManager (
  manager_id int not null identity,
  manager_name nvarchar(50) not null,

  CONSTRAINT PK_DimManager PRIMARY KEY(manager_id)
)
GO
```

4 – Create (1)

```
GO
CREATE TABLE ProjectDW.DimReferee (
  referee_id int not null identity,
  referee_name nvarchar(50) not null,

  CONSTRAINT PK_DimReferee PRIMARY KEY(referee_id)
)
GO
CREATE TABLE ProjectDW.DimCompetition (
  competition_id int not null identity,
  competition_natural_key nvarchar(50) not null,
  competition_code nvarchar(50) not null,
  competition_name nvarchar(50) not null,
  competition_type nvarchar(50) not null,
  competition_sub_type  nvarchar(50) not null,
  competition_country_name nvarchar(50) not null,
  competition_domestic_league_code nvarchar(50) not null,
  competition_confederation nvarchar(50) not null,

  CONSTRAINT PK_DimCompetition PRIMARY KEY(competition_id)
)
GO
CREATE TABLE ProjectDW.DimClub (
  club_id int not null identity,
  club_natural_key int not null,
  club_code nvarchar(50) not null,
  club_name nvarchar(60) not null

  CONSTRAINT PK_DimClub PRIMARY KEY(club_id)
)
GO
```

5 – Create (2)

```
CREATE TABLE ProjectDW.FactGame (
  game_id int not null identity,
  date_id int,
  competition_id int,
  stadium_id int,
  season_id int,
  referee_id int,
  home_club_id int,
  away_club_id int,
  home_manager_id int,
  away_manager_id int,
  home_club_goals int not null,
  away_club_goals int not null,
  home_club_position int not null,
  away_club_position int not null,
  attendance nvarchar(50) not null,
  home_yellow_cards int not null,
  away_yellow_cards int not null,
  home_red_cards int not null,
  away_red_cards int not null,
  game_total_cards int not null,
  is_home_win bit not null,
  is_away_win bit not null,
  is_draw bit not null,
  home_clean_sheet bit not null,
  away_clean_sheet bit not null,
  both_teams_score bit not null,
  total_goals int not null,
  goal_diff int not null,

  CONSTRAINT PK_FactGame PRIMARY KEY(game_id),
  CONSTRAINT FK_FactGame_DimDate FOREIGN KEY(date_id) REFERENCES ProjectDW.DimDate(date_id),
  CONSTRAINT FK_FactGame_DimCompetition FOREIGN KEY(competition_id) REFERENCES ProjectDW.DimCompetition(competition_id),
  CONSTRAINT FK_FactGame_DimStadium FOREIGN KEY(stadium_id) REFERENCES ProjectDW.DimStadium(stadium_id),
  CONSTRAINT FK_FactGame_DimSeason FOREIGN KEY(season_id) REFERENCES ProjectDW.DimSeason(season_id),
  CONSTRAINT FK_FactGame_DimReferee FOREIGN KEY(referee_id) REFERENCES ProjectDW.DimReferee(referee_id),
  CONSTRAINT FK_FactGame_DimHomeClub FOREIGN KEY(home_club_id) REFERENCES ProjectDW.DimClub(club_id),
  CONSTRAINT FK_FactGame_DimAwayClub FOREIGN KEY(away_club_id) REFERENCES ProjectDW.DimClub(club_id),
  CONSTRAINT FK_FactGame_DimHomeManager FOREIGN KEY(home_manager_id) REFERENCES ProjectDW.DimManager(manager_id),
  CONSTRAINT FK_FactGame_DimAwayManager FOREIGN KEY(away_manager_id) REFERENCES ProjectDW.DimManager(manager_id),
)
```

6 – Create (3)

5.1 Punjenje dimenzionih tabela

U ovoj sekciji će biti prikazan način na koji su punjene dimenzione tabele, transforamcija izvornih podataka. Neke od komponenti koje su korišćene u ovim transformacijama:

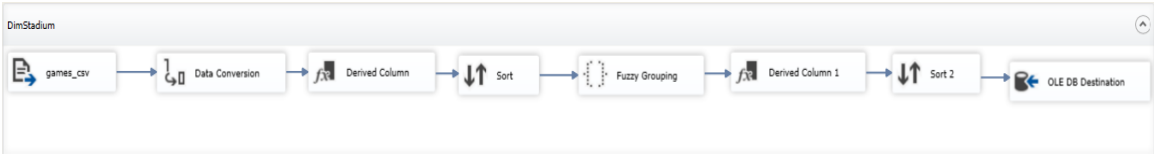
- Data Conversion – Služi za konverziju podataka u određeni tip.
- Derived Column – Služi za rad sa kolonama, dodavanje novih...
- Sort – Za sortiranje i uklanjanje redova duplikata
- Fuzzy Grouping – Ova komponenta je dosta važna za ovaj projekat, zbog oscijalcija u imenima trenera, stadiona i služi da spoji slične nazive, koji mogu biti greške u kucanju itd..
- Script Component – U ovom projektu je služila za normalizaciju vrednosti imena runde sezone, jer su vrednosti dosta varirale, za istu stvar bilo je više načina beleženja.

- Multicast – Za dupliranje podataka, kreiranje dve verzije od jedne, korišćeno kod dimenzije trenera.

Za svaku tabelu kreiran je surogat ključ a takođe u bazi su čuvani i prirodni ključevi kada je to bio slučaj.

Dimenzija Stadion

Za punjenje dimezije Stadion korišćen je fajl *games.csv* iz kog je uzeto obeležje *stadium* koje predstavlja ime stadiona na kome je odigrana utakmica. Bilo je potrebno transformisati vrednost obeležja da bi moglo da se izvrši grupisanje na osnovu sličnih imena i takođe je izvršena izmena nedostajućih vrednosti koje su sačuvane kao „Nepoznato“ što će biti praksa i u nastavku. Nakon toga je izvršeno grupisanje sličnih imena i tako obrađeni podaci su ubačeni u tabelu *DimStadium*.



7 - DimStadium

Dimenzija Sezona

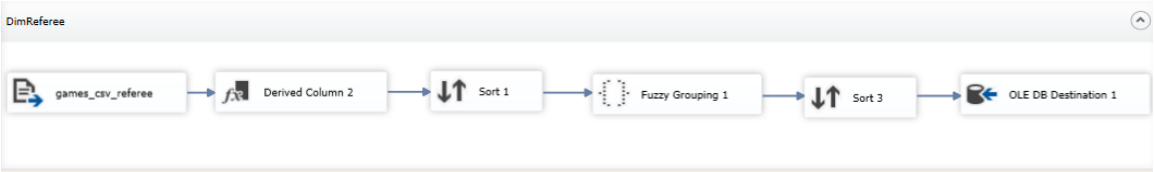
Što se tiče dimenzije sezona tu je takođe korišćen *games.csv* fajl koji sadrži kolone *season* i *round*. Podaci o rundi sezone morali su biti na neki način normalizovani jer su za istu rundu bili različiti nazivi. Za to je korščen *Script Component* koji normalizuje to kroz kod. Na kraju je izvršeno soritranje i podaci su ubačeni u *DimSeason*.

8 - DimSeason

Dimenzija Sudija

Punjenje dimenzije *Referee* je izvršeno uzimajući vrednosti kolone *referre* iz *games.csv*. Nakon toga je takođe izvršena zamena nedostajučih vrednosti sa „Nepoznato“ i

pripremanje podataka za grupisanje na osnovu imena. Nakon tog koraka izvršeno je sortiranje i ubacivanje podataka u tabelu *DimReferee.*



9 - DimReferee

Dimenzija Takmičenje

Prilikom punjenja ove dimenzione tabele kao izvor korišćen je *competitions.csv* fajl iz kog je uzet prirodni ključ i obeležja koja su nam od značaja. Tu je takođe bila izvršena izmena nedostajućih vrednosti sa „Nepoznato“. Nakon čega je usledilo sortiranje i ubacivanje podataka u *DimCompetition.*

10 - DimCompetition

Dimenzija Klub

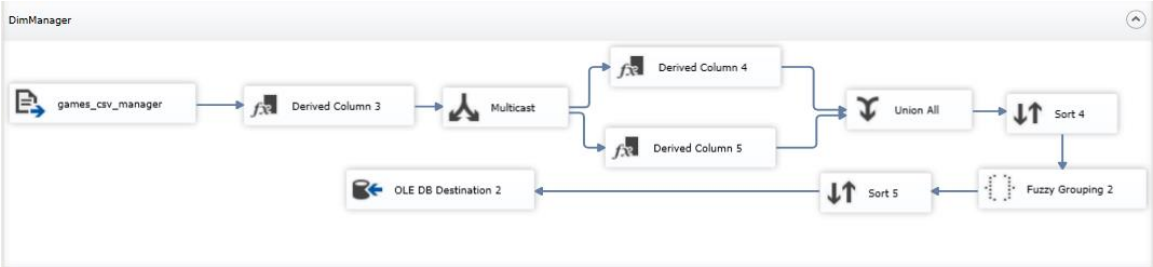
Na isti način kao i za tabelu takmičenja i ovde je čuvan prirodni ključ a podaci su učitani iz fajla *clubs.csv*. Nakon čega je izvršena zamena nedostajućih vrednosti sa „Nepoznato“ i učitavanje u *DimClub*.

11 - DimClub

Dimenzija Trener

Ovde je najveći posao urađen iz razloga što u izvornom fajlu *games.csv* imamo dve kolone koje sadrže ime trenera, *home_manager_name* i *away_manager_name*. Iz tog razloga prvo je izvršena zamena nedostajućih vrednosti i onda izvršeno kopiranje podatka da bi se mogle kolone zameniti istim imenom i na osnovu toga izvršiti spajnjei sortiranje preko kod takođe uklanjamo duplikate. Nakon toga je izvršeno grupisanje na osnovu slučnosti i ti podaci su

smešteni u bazu u tabelu *DimManager*.



12 - DimManager

Dimenzija vreme

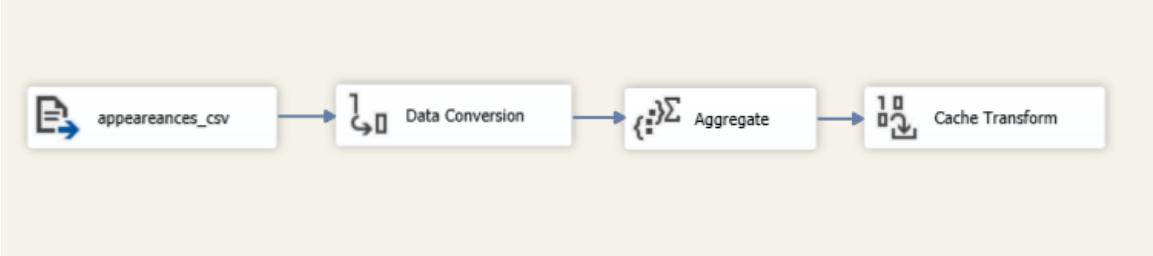
Za punjenje vremenske dimenzije kreiran je *Execute SQL task* kome je prosleđen kod kojim se puni table svim potrebnim podacima za interval od početka 2012. do kraja 2023. godine.

```
DECLARE @StartDate date = '2012-01-01';
DECLARE @EndDate   date = '2023-12-31';

SET DATEFIRST 1;

;WITH DateSpan AS (
    SELECT @StartDate AS d
    UNION ALL
    SELECT DATEADD(DAY, 1, d) FROM DateSpan WHERE d < @EndDate
)
INSERT INTO ProjectDW.DimDate
(
    date_full_date,
    date_day,
    date_month,
    date_year,
    date_quarter,
    date_day_name,
    date_month_name,
    date_week_of_year,
    date_is_weekend
)
SELECT
    d,
    DAY(d),
    MONTH(d),
    YEAR(d),
    DATEPART(QUARTER, d),
    DATENAME(WEEKDAY, d),
    DATENAME(MONTH, d),
    DATEDIFF(WEEK, DATEFROMPARTS(YEAR(d),1,1), d) + 1,
    CASE WHEN DATEPART(WEEKDAY, d) IN (6,7) THEN 1 ELSE 0 END
FROM DateSpan
WHERE NOT EXISTS (SELECT 1 FROM ProjectDW.DimDate x WHERE x.date_full_date = d)
OPTION (MAXRECURSION 0);
```

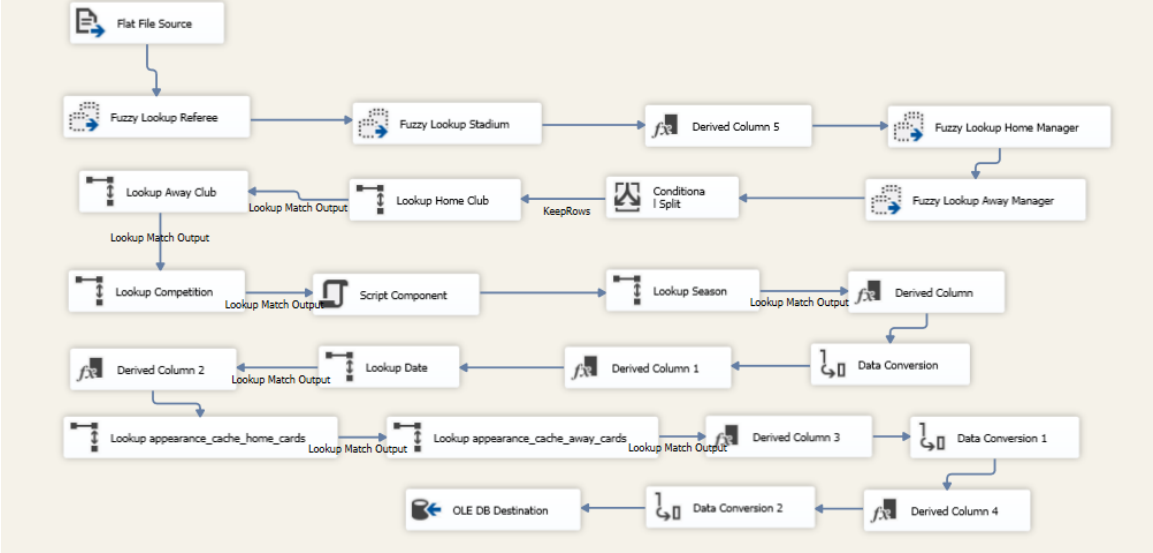
13 – DimDate

5.2 Punjenje činjenične tabele

Ovo je najkompleksniji deo ETL proces, jer je sada potrebno sve prethondo kreirane podatke i dimenzije povezati sa činjenciom, koja je centar dešavanja i ovog skladišta podatka. Da bi pre sveg bilo moguće ubaciti određene mere u činjeničnu tabelu potrebno je kreirati keš memoriju koja sadrži podatke o broju kartona a ti podaci potiču iz *appearances.csv* izvornog fajla. Uzeti su id kluba i utakmice tako i broj žutih i crrvenih kartona. Na osnovu toga kreirane su nove kolone *total_yellow_cards* i *total_red_cards* i ti podaci su grupisani po klubu i utakmici, služiće kasnije za potrebne analize i punjenje mera činjenične tabele.



14 – Appeareances cache

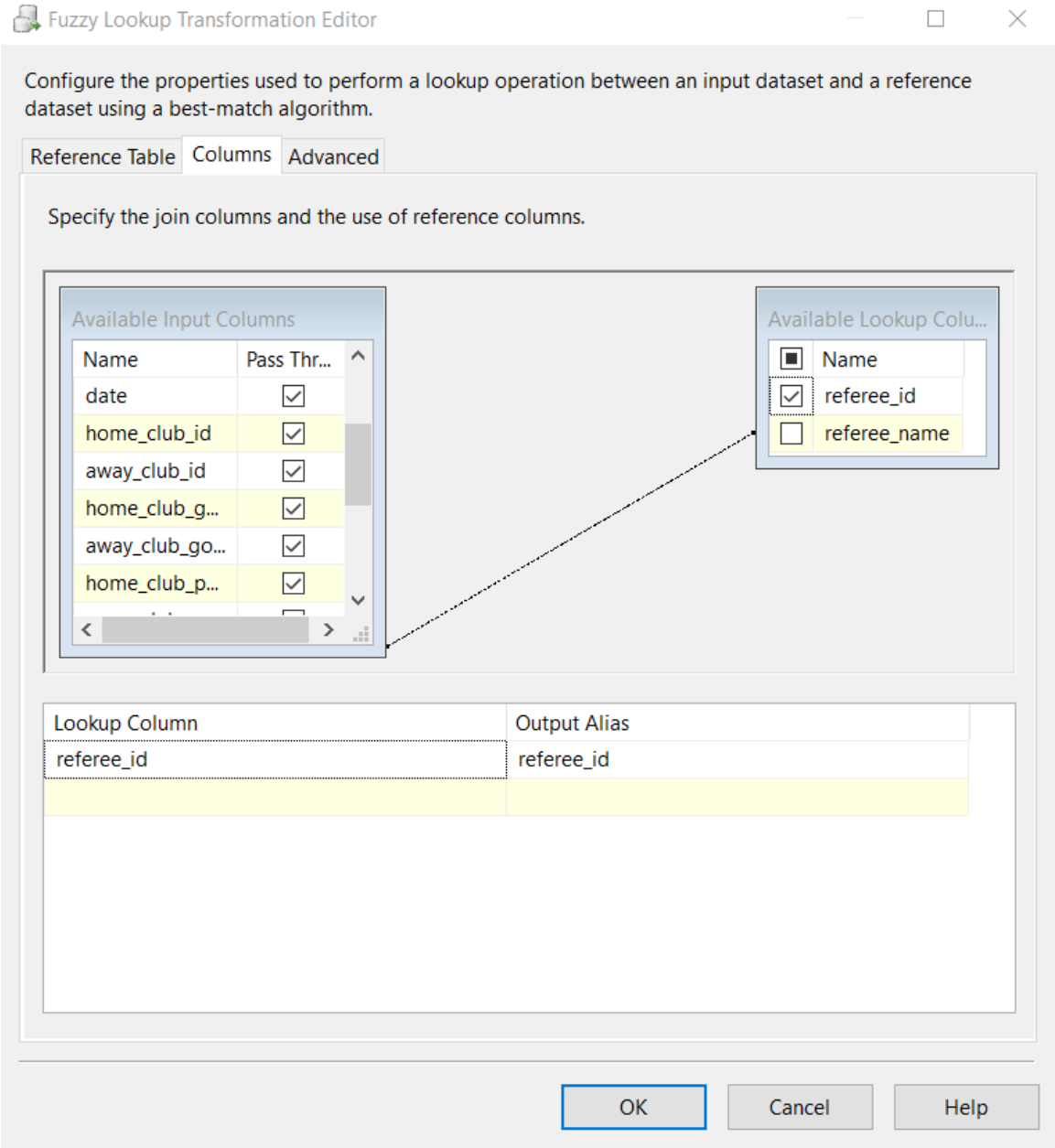


15 – Punjenje činjenične tabele

Neke od komponenata koje su korišćene u ovom procesu su:

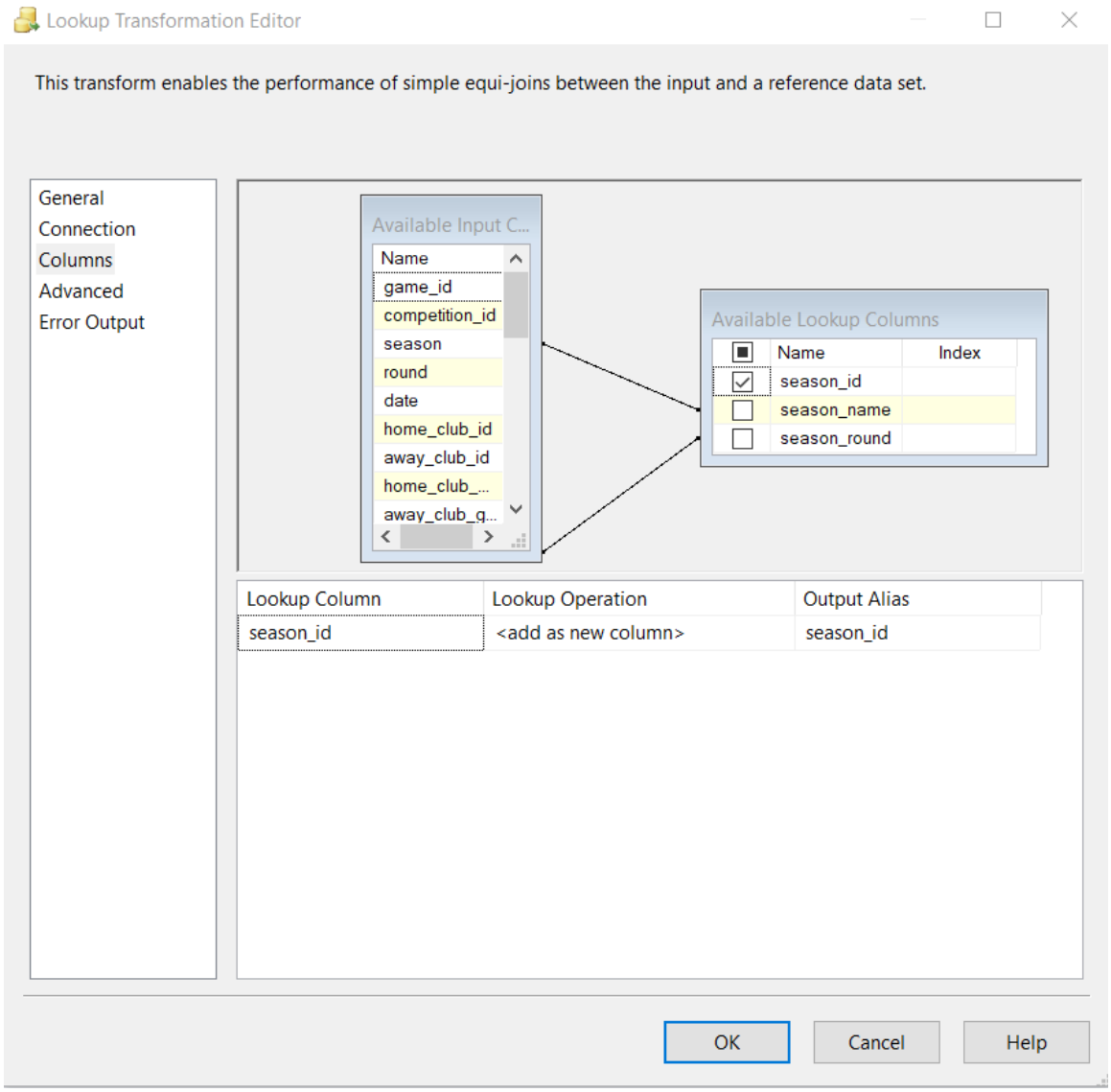
- Fuzzy Lookup – pronalazi najbliža poklapanja u referentnoj tabeli
- Lookup – traži tačno poklapanje u referentnoj tabeli

Proces punjenja činjenične table započinje sa učitavanjem fajla *games.csv* iz koga uzimamo obeležja preko kojih ćemo izvršiti spajanje sa dimenzionim tabelama. Tabele *Referee, Stadium* i *Manager* spojene su korišćenjem *fuzzy lookup*-a jer su podaci u *games.csv* u teksutalnom obliku, naravno da bi spajanje bilo moguće potrebno je takođe i ove podatke prilagoditi na isti način kao što je to rađeno u dimenzijama, nakon prilagođavanja izvršeno je spajanje po tekstualnim obeležjima a uzeti su id-jevi iz tih dimenzionh tabela.



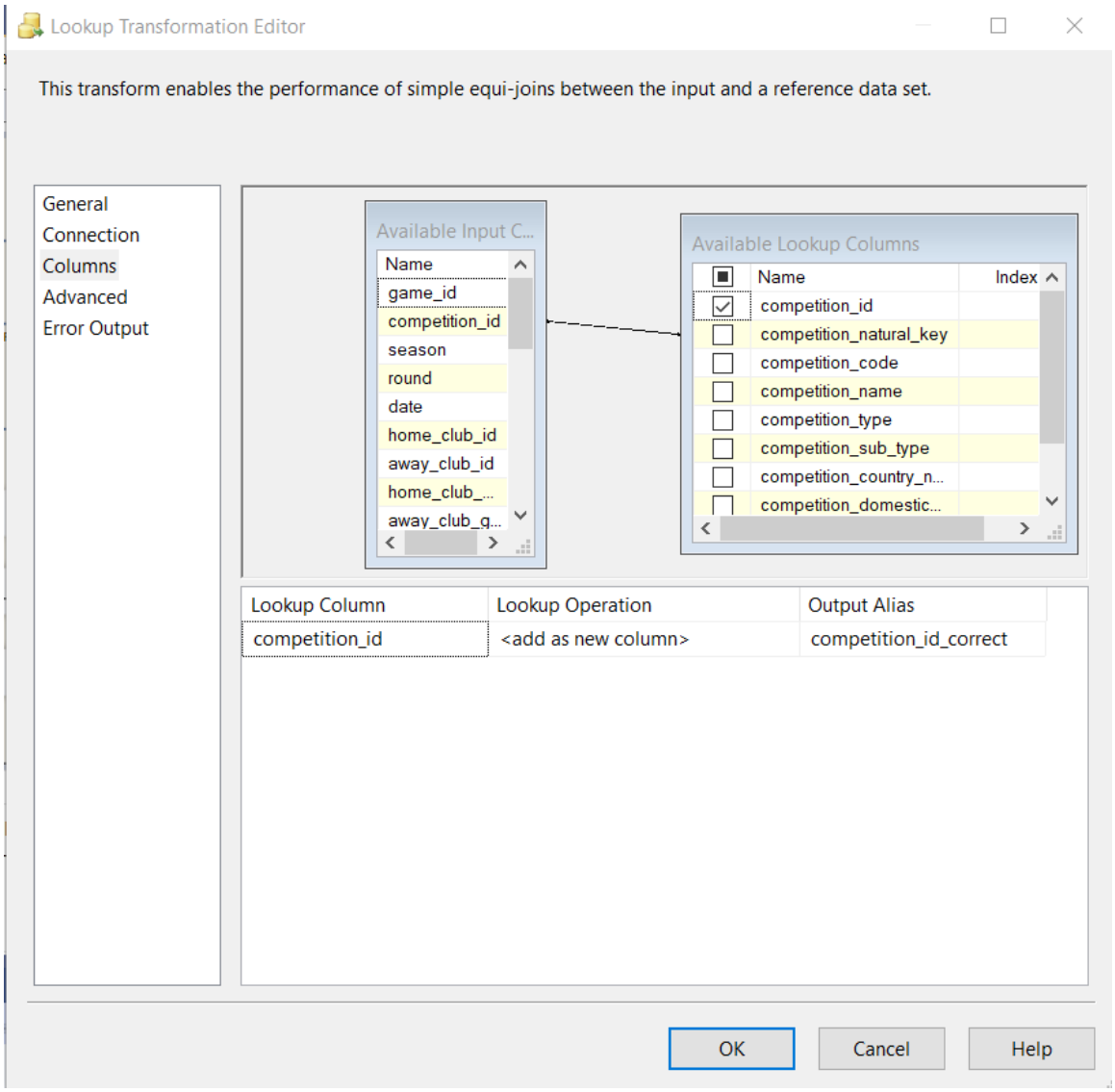
16 – Fuzzy Lookup Referee

Za ostale dimenzione tabele korišćen je običan *Lookup*. Jer u *games.csv* sadrže id-jeve putem kojih je moguće spojiti tabele lepo. Naravno kao što sam pomeuo i gore, ukoliko je vršena transformacija nekih obeležja po kojima se tabele spajaju potrebno je te iste transformacije sprovesti i ovde da bi spajanje bilo tačno.



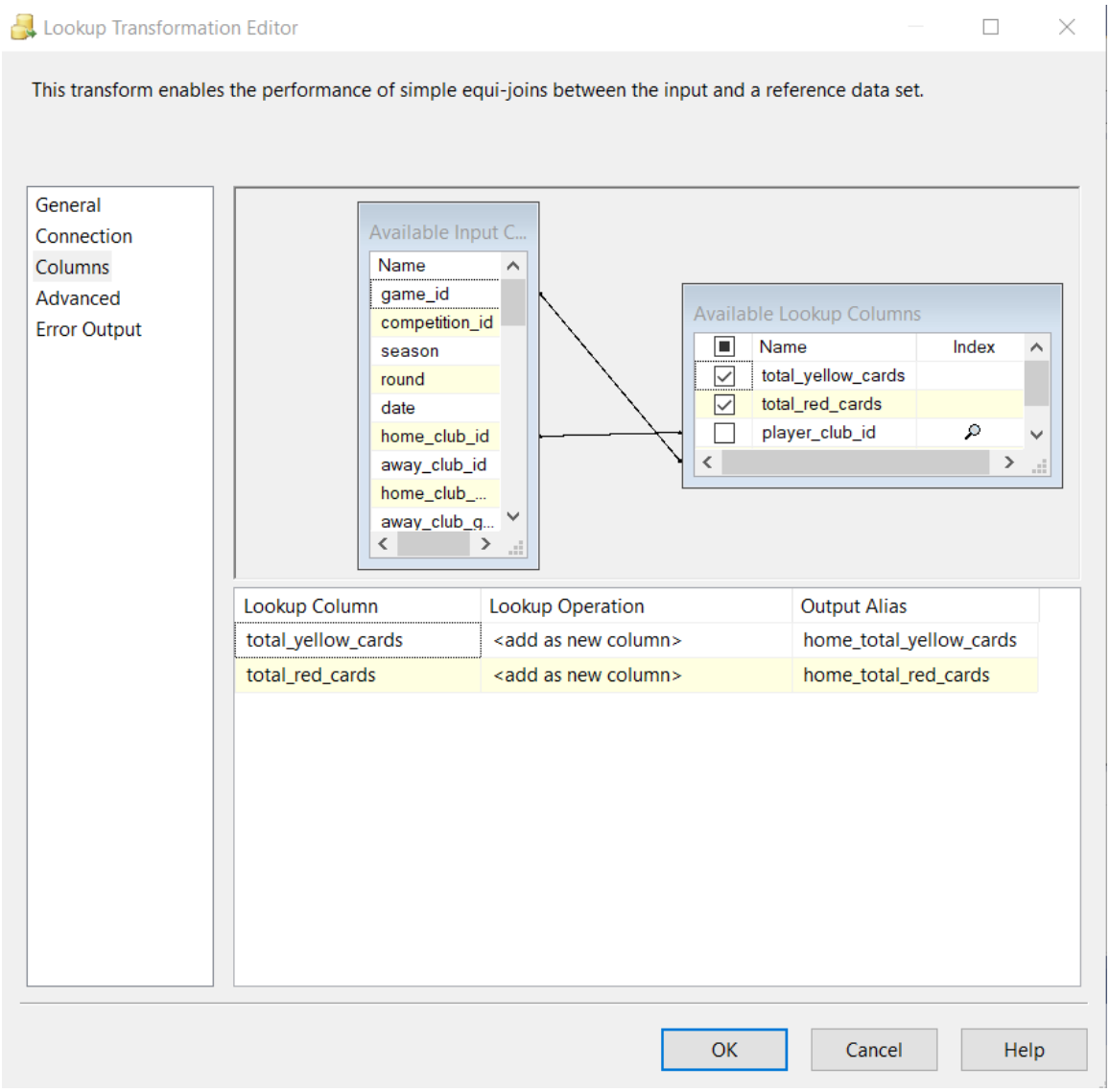
17 – Lookup Season

Izuzetak su tabele *club* i *competition* gde je spajanje sa dimenzionim tabelama vršeno preko prirodnog ključa.



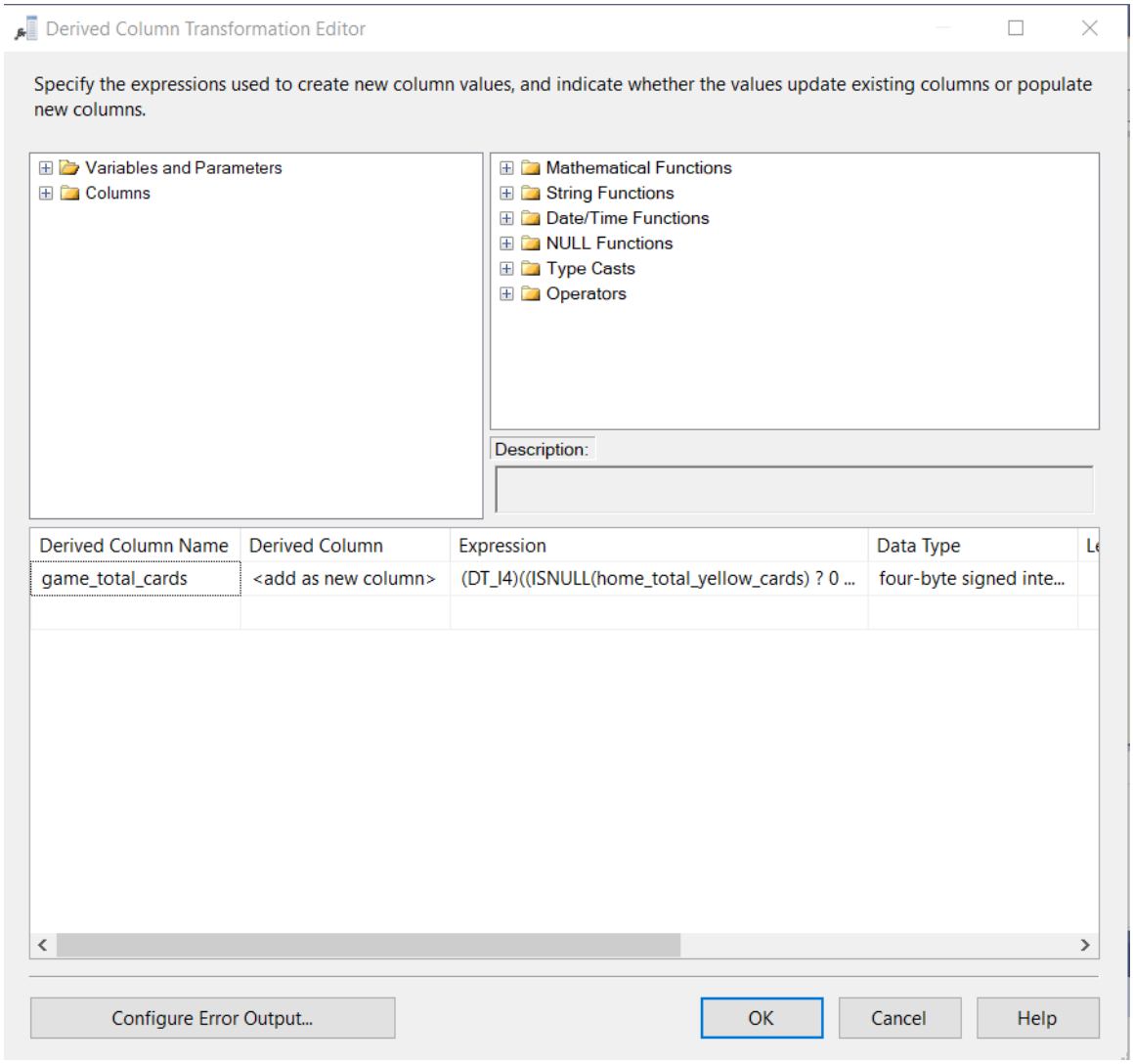
18 – Lookup Competition

Kada su tabele spojene na pravi način potrebno je i pokupiti podatke iz keš memorije o broju kartona po utakmici, domaćeg i gostujućeg tima.



19 – Lookup cache

Jedan od poslednjih koraka jeste računanje mera koje se dobijaju primenom računskih operacija, neke od tih mera su *game_total_cards, both_teams_score*....



20 – Dodavanje mera

I na samom kraju sledi ubacivajne tih podatka u tabelu *FactGame* u bazi podataka. Sa tim je završeno punjenje tablea a i sam ETL proces. U narednim koracima ovi podaci će biti korišćeni za dobijanje odgovora na prethodno postavljena pitanja.

6. Prikaz izveštaja

Nakon završenog ETL proces kreirani su izveštaji koji predstavljaju odgovor na definisana pitanja. Za početak je kreiran novi *SQL Server Reporting Services* projekat. Krairana je konekcija ka bazi podataka koja predstavlja naše skladište. I to je osnova iz koje ćemo kupiti podatke za izveštaje.

Prvi izveštaj – Najzanimljivija takmičenja po sezoni

Prvi izveštaj kriran je kao odgovor na prvo pitanje. Upit uzima mečeve iz *FactGame* spaja ih sa *DimCompetition* po id-ju i *DimSeason*. Agregira po takmičenju i sezoni i računa prosek golova po utakmici i broj mečeva. Sama tabela izveštaja sadrži kolone *Competiton, Country, Average goals, Total matches*. Podaci su grupisani po sezoni i svaka sezona može biti pregledana kada se klikne na nju. Takođe za svaku kolonu je urađeno interaktivno sortiranje. Najzanimljivje takmičenje je označeno belim redom u tabeli.

Most Entertaining Competitions by Season

Season: 2014

Competition	Country	Average goals	Total matches

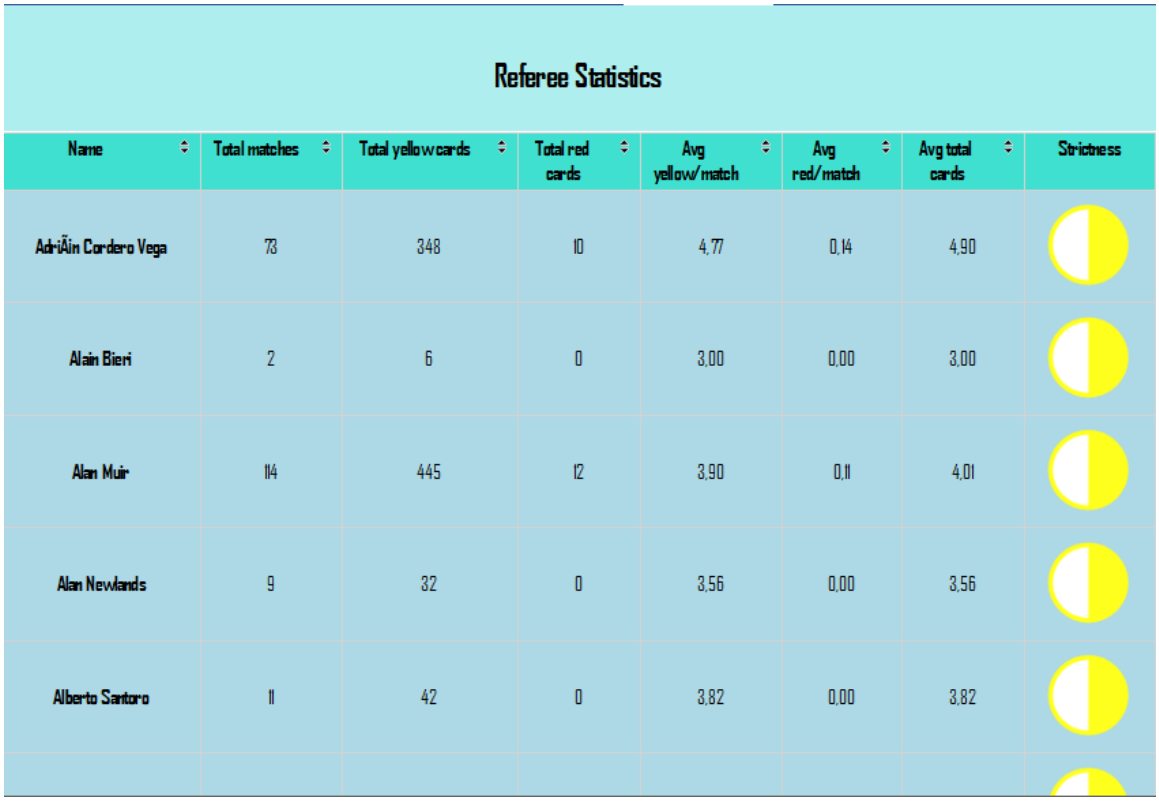
Season: 2015

Competition	Country	Average goals	Total matches
Allianz-cup	Portugal	3,533	15
Belgian-supercup	Belgium	1,000	1
Bundesliga	Germany	2,830	306
Community-shield	England	1,000	1
Copa-del-rey	Spain	3,147	34
Dfb-pokal	Germany	3,733	15
Dfl-supercup	Germany	11,000	1
Eredivisie	Netherlands	2,980	306
Europa-league	Europe Tournament	2,723	101
Europa-league-qualifikation	Europe Tournament	2,417	12
Fa-cup	England	2,345	29
Italy-cup	Italy	3,067	15
Johan-cruijf-schaal	Netherlands	3,000	1

21 – Izveštaj 1

Drugi izveštaj – Statisika sudija

Ovaj izveštaj je od velikog značaja za sudijsku komisiju koja je zadužena za delegiranje sudija za mečeve. Na osnovu ove statisike mogu videti temperament sudije i izabrati sudiju koji odgovara profilu utakmice. Upit spaja *FactGame* sa *DimReferee* i po sudiji sabira žute i crvene kartone i broj suđenih mečeva. Zatim računa proseke kartona po meču. Takođe ovde je moguće isto izvršiti selekciju strogoće sudija i na osnovu toga videti te sudije, tabela sadrži polja ime sudije, ukupno mečeva, broj žutih i crvenih kartona i njihov prosek po meču. Takođe poslednje polje je indikator koji boji krug u zavisnosti od stepena strogoće. Na dnu je *footer* koji sadrži podate o paginaciji.



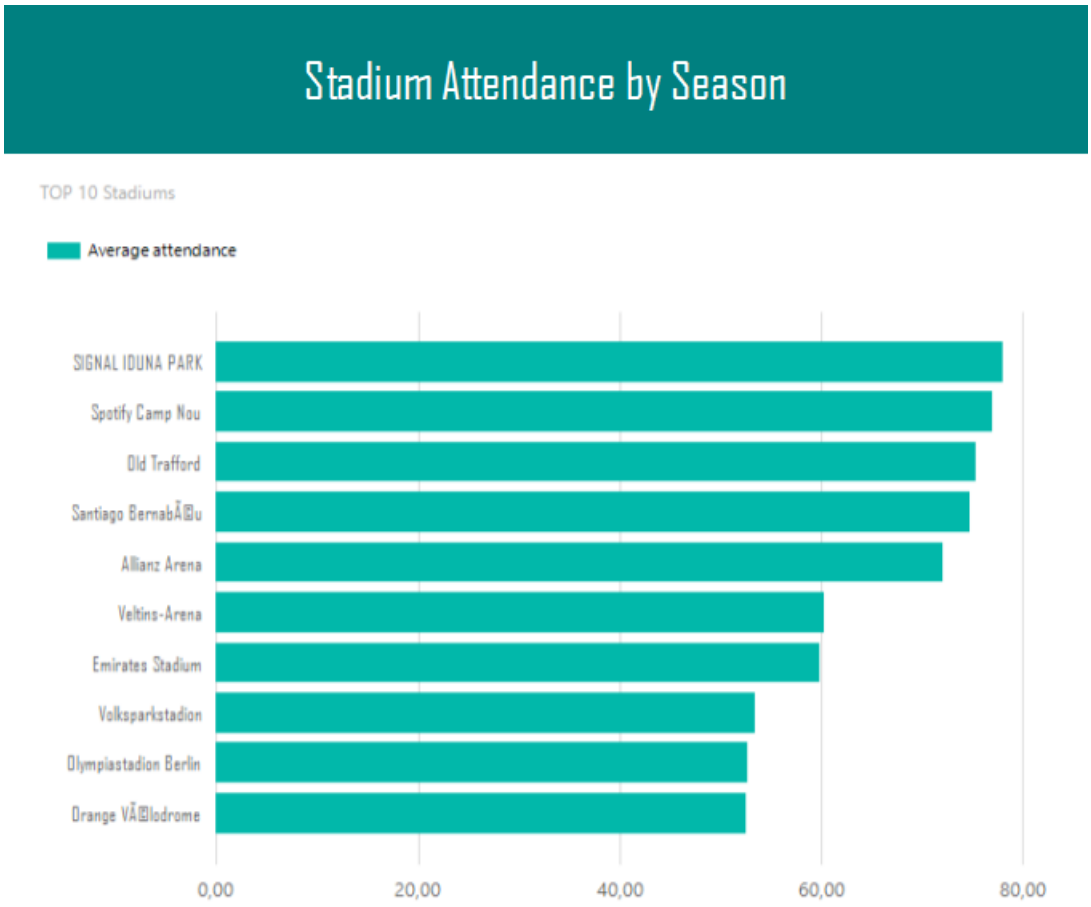
Referee Statistics

Name	Total matches	Total yellow cards	Total red cards	Avg yellow/match	Avg red/match	Avg total cards	Strictness
AdriÃin Cordero Vega	73	348	10	4,77	0,14	4,90	
Alain Bieri	2	6	0	3,00	0,00	3,00	
Alan Muir	114	445	12	3,90	0,11	4,01	
Alan Newlands	9	32	0	3,56	0,00	3,56	
Alberto Santoro	11	42	0	3,82	0,00	3,82	

22 – Izveštaj 2

Treći izveštaj – Prosečna posećenost stadiona tokom sezona

Ovaj izveštaj daje prikaz najposećenih stadiona tokom sezone. Sadrži deset najposećenijih stadiona po godini. Filtriranje je takođe moguće izvršiti po godini. Upit povezuje *FactGame* sa *DimStadium* i *DimSeason* i po stadionu i sezoni računa prosečnu posećenost i broj mečeva. Za prikaz je korišćen *chart* čija vertikalna osa prikazuje klubove, a horizontalna vrednost prosečne posete po stadionu. Ovo je vrlo važan faktor za neke kompanije koje žele da reklamiraju svoj brend, zatim za same turiske jer takvi stadioni privlače simpatije. Takođe ovaj izveštaj je važan i organima reda i mira jer to prikazuje koliko je potrebno angažovati pripadnika policije.



23 – Izvestaj 3

Četvrti izveštaj – Timska statistika

Ovaj izveštaj je izuzetno važan za klub i navijače. Vrh kluba, trener i navijači žele jasnu sliku o tome kako se timovi ponašaju u različitim uslovima, kolika im je stopa pobeda, koliko često igraju nerešeno i kakva je gol-razlika... Izveštaj je moguće filtrirati po učinku na domaćem terenu i gostujućem terenu. Podaci su prikupljeni iz *FactGame* i *DimClub* na osnovu podatak iz činjenične tabele. Tabela sadrži polja ime kluba, ukupno mečeva, ukupno pobeda, remija i poraza, zatim broj poena u proseku po meču, gol razliku i procenat pobeda. Poslednje tri kolone menjaju boju u zavisnosti od rezultata. Takođe svaku kolono moguće je interaktivno sortirati, a na dnu stranice je *footer* koji sadrži podatke o paginaciji.

Teams statistics Home / Away							
Name	Total games	Total wins	Total draws	Total losses	Points per game	Goal difference	Win rate
1-fc-koln	136	46	43	47	1,33	5	33,82%
1-fc-nurnberg	21	4	7	10	0,90	-7	19,05%
1-fc-union-berlin	70	38	22	10	1,94	45	54,29%
1-fsv-mainz-05	156	63	41	52	1,47	36	40,38%
aalborg-bk	130	56	31	43	1,53	42	43,08%
aarhus-gf	110	44	28	38	1,45	18	40,00%
aberdeen-fc	165	92	33	40	1,87	108	55,76%

24 – Izveštaj 4

Peti izveštaj – Statisika trenera

Ovaj izveštaj kreiran je sa ciljem da pruži opširnu statiskitku menadžera tokom njhovih karijera. On može poslužiti navijačima kao argument prilikom diskusija i takođe kao dodatni izvor znanja, al najbitniji je rukovodstvu kluba koji na ovaj način može da selektira trenera kojeg će „juriti“ na tržištu. Ovde se takođe uzimaju podaci iz *FactGame* koja se spaja sa *DimManager* preko id-a trenera. Izveštaj sadrži polja ime trnera, ukupno mečeva, pobeda, remija i poraza. Putem *bar charta* prikazuje procenat pobeda trenera, a kasnije polja proseka poena po meču i gol razlike koja su obojena u zavisnosti od kategorija vrednosti. I na kraju imamo prosek datih i primljenih golova po meču. Svaka kolona može biti interaktivno sortitrana i na dnu stranice imamo *footer* koji sadrži podatke o paginaciji.

Managers Statistics

Name	Total games	Wins	Draws	Losses	Win rate	Points per	Goal difference	GF (Average)	GA (Average)
Michael Beale	18	16	1	1	88,89%	2,72	30	2,67	1,00
Hansi Flick	79	64	8	7	81,01%	2,53	152	2,94	1,01
Jupp Heynckes	41	33	3	5	80,49%	2,49	78	2,88	0,98
Kenneth Andersen	25	19	3	3	76,00%	2,40	28	2,28	1,16
Luis Enrique	169	128	20	21	75,74%	2,39	333	2,82	0,85
Johnny Heitinga	15	11	2	2	73,33%	2,33	23	2,40	0,87
Ange Postecoglou	83	61	8	14	73,49%	2,30	130	2,52	0,95
Pep Guardiola	453	327	57	69	72,19%	2,29	722	2,45	0,85
Paulo Bento	30	21	5	4	70,00%	2,27	39	1,80	0,50
Oleksandr Kucher	20	14	3	3	70,00%	2,25	26	2,15	0,85

25 – Izveštaj 5

Šesti izveštaj – Ferplej rangiranje ekipa po sezonama

Izveštaj koji je od velikog značaja za nagradu koju dodeljuje svetska fudbalska ogranizacija za ekipu koja je najpoštenija, u prevodu ima najmanje kartona tokom sezone. Podaci su dobijeni spajanjem *FacGame i DimClub* za dobijanje imena kluba. Izveštaj je vizuelizovan preko *chart*-a. On prikazuje top 5 klubova sa najmanjim indexom, to govori o tome da ima ti kulobi imaju najmanje kartona u sezoni koja jje označena. Izveštaji mogu biti kreirani po sezonama. Takođe ekipa koja ima najbolji *fair-play index*, njen broj će biti označen zlatnom bojom.

26 – Izveštaj 6

7. Zaključak

Implementirano je celovito skladište podataka sa zvezdastom šemom (jedna činjenica, osam dimenzija), stabilnim ETL-om u SSIS-u i skupom SSRS izveštaja koji znatno ubrzavaju i olakšavaju analizu u odnosu na direktno izveštavanje iz operativne baze. Centralizovani, očišćeni podaci sada služe kao jedinstven izvor istine i omogućavaju korisnicima da donose brže i bolje odluke, uz jasne uvide o ligama, sezonama, menadžerima, stadionima i disciplini. Glavna ograničenja su kvalitet izvora (nedostajuće vrednosti, fuzzy mapiranja), a najveći izazovi nadalje su održavanje kvaliteta podataka i performansi kako obim raste. U narednoj fazi sistem može biti proširen u potpunu analitičku platformu tako što bi uveli AI modele koji, na osnovu događaja i podataka, prave predikcije po kolu i sezoni (plasman), verovatnoće ishoda utakmica, očekivani broj golova (*over/under*), projektovani broj kartona i kornera, xG/xGA metrike i „*over/underperformance*“, projekcije forme i umora, simulacije scenarija (promene postave/menadžera), preporuke optimalne postave i izmena, kao i rane alarme za anomalije u performansama i rizike, uz interaktivne dashboarde za praćenje u realnom vremenu.